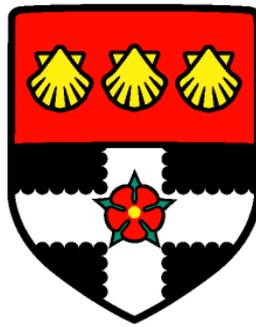


UNIVERSITY OF READING

Department of Geography and Environmental Science



The Art of Streamflow Forecasting over Europe

Louise Arnal

A thesis submitted for the degree of Doctor of Philosophy

September 2019

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Louise Arnal



For Aleix, who would have finished his PhD around the same time.

Acknowledgements

I would like to thank all the supportive and inspiring people who have kept me motivated during this grand voyage.

To Hannah Cloke and Liz Stephens for guiding, trusting and supporting me in every new avenue I aimed to explore (may it be a sensitivity analysis or art).

To Florian Pappenberger for being a great mentor too, inspiring some of the craziest research ideas I've had.

To the ECMWF EFAS and GloFAS teams, and the Water@Reading group for being a friendly, energetic and inspirational bunch.

To Rebecca Emerton and Jess Neumann, my "unicorn sisters", for their friendship and all the fun times we've had at conferences.

To Maria-Helena Ramos, for inspiring me to pursue a career in the field and for introducing me to the world of serious games.

To Andy Wood for the warm welcome and inspirational collaborative work at NCAR.

To Linus Magnusson and David Lavers, for their support and friendship throughout the IMPREX project and various IMPREX meetings across Europe.

To Katie Cooper for being a great mentor and friend, and for her encouragement to pursue my SciArt dream.

To my family and friends - Aleix, Greg, Katya, Alex, Alexandra, Esteban, Juan, Maria and Cécilia - for being there in the best and most challenging moments of this PhD.

To Julius Kreißig for his unconditional support, especially when I believed in myself the least. And for patiently listening while I talked on and on about flood forecasting.

To my IMPREX and HEPEX "families", for making me feel part of a wider enthusiastic community.

To the inspiring artists I met at the JRC SciArt summer school.

To my examiners and chair, Ed Hawkins, Pablo Suarez and Sally Lloyd-Evans, for an insightful discussion during my PhD viva.

Abstract

While floods are among the most damaging natural hazards, they have helped shape human developments over the last millennia, fostering scientific understanding and technological advances in an attempt for their mitigation. We now more skilfully predict floods at increasing lead times, through probabilistic hydro-meteorological forecasting. But we are now facing new challenges.

Have we reached the limits of predictability with seasonal streamflow forecasting? This thesis contributed to the implementation and design of operational seasonal streamflow outlooks, as part of the European and the Global Flood Awareness Systems. Openly available, they give users an overview of potential streamflow changes on sub-seasonal to seasonal timescales. The analysis of both systems highlighted current limits in seasonal predictability and the importance of initial hydrological conditions and the land surface memory. To tackle these limits of predictability, a sensitivity analysis was developed to guide developments for tangible future seasonal streamflow forecast improvements.

Are technical and scientific advances increasing faster than the rate at which forecasts are usable for decision-making? As shown by the application of serious games and research interviews at the Environment Agency (to guide a successful transition to probabilistic forecasts for flood early warning in England), science (e.g. forecast skill) is not necessarily a limiting factor for decision-making. Improved communication between scientists and decision-makers, aimed for instance at understanding the complex landscape in which decision-makers operate, is key to a successful adoption of the latest science in practice. Art can help bridge the communication gap, and this thesis culminated in an art exhibition, 'Gambling with floods?', at The Museum of English Rural Life (Reading, UK) from 1 to 15 November 2019, created to reach a wide audience.

Overall, this thesis has shown that a closer interaction between decision-makers, scientists and artists is urgently needed for a co-leadership on improving science for society.

Contents

Declaration	iii
Acknowledgements	vii
Abstract	ix
Contents	xi
1 Introduction	1
1.1 Wider context and motivation.....	1
1.2 Thesis objectives and structure	3
2 Insights from decision-making activities and serious games on the use of probabilistic hydrological forecasts for flood early warning	6
2.1 Background and aim.....	6
2.2 Willingness-to-pay for a probabilistic flood forecast: a risk-based decision-making game	7
2.2.1 Introduction.....	8
2.2.2 Set-up of the decision-making game.....	11
2.2.3 Results.....	20
2.2.4 Discussion.....	34
2.2.5 Conclusions	38
2.2.6 Resources.....	39
2.2.7 Appendix	40
2.2.8 Lessons learnt for improvements in future serious games.....	41
2.3 Stepping in the boots of a flood incident manager: an IMPREX serious game	42
2.3.1 Overview of the game storyline.....	42
2.3.2 What scientific lessons does the game communicate?.....	43
2.3.3 In-depth game design.....	43
2.3.4 Feedback and game release.....	46
2.4 Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity.....	47
3 Operational seasonal hydro-meteorological forecasting: from the global to the local scale	50
3.1 Background and aim.....	50
3.2 Skilful seasonal forecasts of streamflow over Europe?.....	52

3.2.1	Introduction.....	53
3.2.2	Data and methods.....	56
3.2.3	Hindcast evaluation strategy	58
3.2.4	Results.....	63
3.2.5	Discussion.....	69
3.2.6	Conclusions	76
3.3	The EFAS and GloFAS operational seasonal hydrological outlooks.....	78
3.3.1	EFAS-Seasonal.....	78
3.3.2	GloFAS-Seasonal.....	80
3.4	The 2013/14 Thames basin floods: do improved meteorological forecasts lead to more skilful hydrological forecasts at seasonal time scales?.....	81
4	Towards tangible seasonal streamflow forecast improvements	85
4.1	Background and aim	85
4.2	An efficient approach for estimating streamflow forecast skill elasticity.....	86
4.2.1	Introduction.....	87
4.2.2	Methods, data, and evaluation strategy.....	91
4.2.3	Results.....	97
4.2.4	Discussion.....	105
4.2.5	Conclusions	109
4.3	Scopes for improving EFAS-Seasonal	109
4.4	The forecasting paradox: do better computing resources make us worse forecasters?.....	114
4.4.1	Introduction.....	114
4.4.2	A novel concept for improved seasonal streamflow forecasting.....	115
4.4.3	Discussion.....	116
5	Using probabilistic hydrological forecasts for flood early warning: a real-life case study	119
5.1	Background and aim	119
5.2	“Are we talking just a bit of water out of bank? Or is it Armageddon?” Front line perspectives on transitioning to probabilistic fluvial flood forecasts in England....	119
5.2.1	Introduction.....	121
5.2.2	Context: the Environment Agency’s flood incident management strategy	123
5.2.3	Methods	127
5.2.4	Results.....	129
5.2.5	Discussion and recommendations.....	143
5.2.6	Conclusions	150

5.2.7	Appendix	151
6	The science and art of predicting floods.....	165
6.1	Background.....	165
6.1.1	Why (not) combine science and art?	165
6.1.2	Art as a scientific process	166
6.1.3	Science as an artistic inspiration.....	168
6.2	'Gambling with floods?' A science and art exhibition by L. Arnal	172
6.2.1	My science and art practice	172
6.2.2	Art exhibition.....	173
7	Conclusions.....	177
7.1	Lessons learnt.....	178
7.1.1	Objective 1: Decision-makers' requirements when using probabilistic hydrological forecasts for flood early warning	178
7.1.2	Objective 2: Current capabilities in seasonal hydrological forecasting on the global, continental and basin scales and operational implementation of a global and European seasonal streamflow outlook	179
7.1.3	Objective 3: A cost-efficient method for tangible seasonal streamflow forecast improvements.....	181
7.1.4	Objective 4: Facilitating decision-makers' use of probabilistic flood forecasts	183
7.1.5	Objective 5: Art as a tool to bridge the gap between science and society.....	184
7.2	Key contributions.....	185
7.3	Next steps	186
7.4	Closing remarks	188
	References.....	189
	Appendix.....	211
A1:	Willingness-to-pay for a probabilistic flood forecast: a risk-based decision-making game	212
A2:	Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity.....	233
A3:	Skilful seasonal forecasts of streamflow over Europe?.....	257
A4:	Developing a global operational seasonal hydro-meteorological forecasting system: GloFAS-Seasonal v1.0.....	274

A5: The 2013/14 Thames basin floods: do improved meteorological forecasts lead to more skilful hydrological forecasts at seasonal time scales?.....	295
A6: An efficient approach for estimating streamflow forecast skill elasticity.....	313
A7: The sensitivity of sub-seasonal to seasonal streamflow forecasts to meteorological forcing quality, modelled hydrology and the initial hydrological conditions.....	364
A8: Flexible operational seasonal river flow forecasting.....	510
A9: “Are we talking just a bit of water out of bank? Or is it Armageddon?” Front line perspectives on transitioning to probabilistic fluvial flood forecasts in England....	512

Chapter 1

Introduction

1.1 Wider context and motivation

Floods are among the most damaging earthly natural hazards. According to the UN Office for Disaster Risk Reduction, floods have accounted for 40% of the total economic damages in Europe between 1989 and 2008 (UNISDR, 2009). Not to mention their direct impacts on lives and livelihoods. Floods have also shaped human developments over the last millennia, fostering scientific understanding and technological advances in an attempt for their mitigation.

Around 4000 B.C., the Nile was dammed to manage flooding and increase agricultural yields. Towns of Mesopotamia were protected from flooding with high walls. And Chinese history is full of accounts of irrigation and flood control constructions. As time went by, the understanding of *hydrology* as a science developed.

During the first century B.C., Marcus Vitruvius put forward a philosophical theory of the water cycle, describing how precipitation falling in the mountains infiltrated the Earth's surface and led to streams in the lowlands (Acworth, 2009). While hydrological concepts were known, their science remained qualitative and empirical until the 1930s, with the advent of the first quantitative hydrological models (Sherman, 1932; Horton, 1940). The widely used 'bucket style' rainfall-runoff models nowadays remain largely the same as when they were first developed in the 1970s (Hartmann et al., 2002; Pagano et al., 2013).

The science and technology of weather forecasting, growing in parallel, led to the first mathematical computerised weather forecast in the 1950s, which could be produced for practical use soon after (Shuman, 1989). Combined with hydrological concepts and models, *hydrological forecasting* was born. Since the emergence of hydrological forecasting, and with an ever-growing scientific understanding and technological capabilities, one of the grand challenges in the field remains the production of better hydrological forecasts and at longer lead times. Extending hydrological predictability is a necessity for society for increased disaster preparedness and higher economic benefits (Emerton et al., 2016).

In the early 1900s, Church's work on forecasting streamflow a season ahead paved the way for extending hydrological predictability to seasonal timescales (Linsley, 1967). Through his work and snow surveys, Church understood the importance of snowpack measurements in winter to infer streamflow for the following spring. This proved vital for power generation: "The heavy year of 1910–11 came with menace and fear. The Sierra Pacific Power Company begged the use of our snow data to determine how much moisture was latent on the watershed." (Church, 1935). Whilst useful to an extent, drawbacks of more basic seasonal hydrological forecasting methods (such as the one used by Church) led to the incorporation of seasonal meteorological forecasts (being developed in parallel) to drive seasonal hydrological predictions (Pagano and Garen, 2006). Seasonal hydrological forecasts can indicate potential changes in streamflow in the following months, increasing the lead time at which managers can make decisions in a range of water sectors.

But progress does not come without a cost. The longer the forecast lead time, the larger the forecast uncertainty. In the early twentieth century, Poincaré put forward the idea that small perturbations of the initial conditions of a non-linear system (like the weather and the water cycle) could lead to widely diverging outcomes. This idea of the propagation of uncertainty was subsequently studied quantitatively in the context of weather forecasting by Thompson and Lorenz (Bauer et al., 2015). Now widely known as the 'chaos theory', coined by its founder, Lorenz, it has shaped the way we produce hydro-meteorological forecasts today (Lorenz, 1963). Recognising that this uncertainty limits predictive skill of future weather conditions, scientists adopted an ensemble approach to try and characterise weather forecast uncertainty. This ensemble approach was later adopted in hydrology and for flood forecasting (Cloke and Pappenberger, 2009).

This steady accumulation of knowledge and technological advances (coined the 'quiet revolution' by Bauer et al. (2015)) has led to improved ensemble hydro-meteorological forecasts at longer lead times (e.g. seasonal timescales). But as seasonal hydrological forecasts symbolise our technological and scientific progress, we are now faced with new challenges.

Methods to improve ensemble seasonal hydrological forecasts are plentiful and include: increasing spatial and temporal resolutions, calibrating hydrological models to better represent reality, using better post-processing methods and running an increasing number of ensemble members. Yet, operational streamflow forecast quality has not significantly improved in the last decade, despite the costly research and operational developments they

are receiving (Pagano et al. 2004a; Welles et al. 2007). Have we reached the limits of predictability in seasonal streamflow forecasting?

Societal needs have shaped science for the last millennia, but water managers and decision-makers have had to deal with far from perfect forecasts. To this day, the application of ensemble (or probabilistic) hydrological forecasts for flood early warning remains a challenge. This is mainly due to the complexity of transforming probabilistic information into a binary decision with high stakes. On seasonal timescales, challenges may seem even bigger given the increase of forecast uncertainty with lead time. Yet, the potential of operational seasonal hydro-meteorological forecasts to give earlier indications of possible future flood events and increase the lead times at which we can prepare is immense, and a topic that deserves further attention. Given the challenges, it may appear as though computer capabilities and the production of scientific knowledge are increasing faster than the rate at which forecasts are usable for decision-making (Pielke, 1997).

Due to climate change, the trend is for an increase in the risk of coastal (from sea level rise) and inland flooding (from an increase in the frequency and intensity of heavy precipitation events) in Europe in the future (UNISDR, 2009; IPCC, 2014). In this context, it is becoming vital to tackle these new challenges and reconnect science, practice and society.

1.2 Thesis objectives and structure

This thesis is part of the IMPREX (IMproving PRedictions and management of hydrological EXtremes) H2020 project and aligns itself with the project's wider aim: to improve the predictability of hydrological extremes in Europe.

This thesis guides the readers through the art of streamflow forecasting over Europe, presenting some of the latest challenges in operational hydro-meteorological forecasting. Namely, hydrological forecasting at longer lead times (i.e. seasonal timescales), the use of ensemble (or probabilistic) hydrological forecasts for flood early warning and bridging the gap between science and society. By targeting the full forecasting chain, from operational forecast development to forecast communication and use for decision-making, this thesis combines perspectives from science, practice and art to address the following objectives:

- 1 Explore decision-makers' requirements when using probabilistic hydrological forecasts for flood early warning.
- 2 Investigate the current capabilities in seasonal hydrological forecasting on the global, continental and basin scales, and implement a global and European operational seasonal streamflow outlook.

- 3 Develop a cost-efficient method for tangible seasonal streamflow forecast improvements and apply it over Europe.
- 4 Suggest ways to facilitate decision-makers' intake of probabilistic flood forecasts.
- 5 Explore art as a tool to bridge the gap between science and society.

This thesis explores the above five objectives in a variety of ways, including: the co-creation of two serious games and a decision-making activity, the publication of four first-author (of which one is in review) and three co-author peer-reviewed published articles and a first-author peer-reviewed published IMPREX deliverable, the co-development of two operational seasonal streamflow outlooks, research interviews and the creation of an art exhibition.

The structure of this thesis follows an 'end-to-end-to-end approach', an iterative feedback process between decision-making and research (Morss et al., 2005). User perspectives are first reported (Chapter 2), followed by state-of-the-art research (Chapters 3 and 4), and the thesis finally closes with more user perspectives (Chapter 5) and science and art (Chapter 6). All the elements that contribute to this thesis were reformatted as thesis chapters, including the published articles and deliverable (published versions of all first-author and co-author publications are provided in the thesis Appendix). Fig. 1 presents an overview of the structure of the results chapters of this thesis (Chapters 2 to 6).

Chapter 2 addresses the first objective of this research through two serious games and a decision-making activity. These activities were designed to investigate users' perceived challenges and opportunities associated with using short-term to seasonal probabilistic hydrological forecasts for flood early warning. One of the serious games was designed specifically to communicate these challenges and opportunities and introduce a wider audience to forecast-based decision-making for flood early warning.

In Chapter 3, the second objective of this research is addressed. It presents two state-of-the-art seasonal streamflow outlooks, co-designed operationally as part of this PhD. The European Flood Awareness System (EFAS)-Seasonal system is presented, and its performance is assessed over Europe as well as for the 2013/14 Thames river flood event. The Global Flood Awareness System (GloFAS)-Seasonal system is presented, and a quick overview of its performance over the global river network is given.

Chapter 4 investigates the third objective of this thesis. This chapter presents a cost-efficient sensitivity analysis method that enables the identification of the relative contributions of error sources in seasonal streamflow forecasts. This method is

subsequently applied to the EFAS-Seasonal forecasting system, for which results are presented. These findings inspired the idea of a novel seasonal hydrological forecasting system which reflects the system's predictability sources, finally presented in this chapter.

Chapter 5 contributes to the first and fourth research objectives, building on findings from Chapters 2, 3 and 4. Through research interviews, it explores the context in which UK Environment Agency (EA) decision-makers operate (objective 1) and provides recommendations to facilitate a smooth transition to probabilistic forecasts for flood early warning in England (objective 4).

Finally, Chapter 6 addresses the last objective of this thesis, exploring the use of art to bridge the gap between science and society. The main element of this chapter is a science and art exhibition, created as part of this PhD on this thesis' topic: seasonal streamflow forecasting and flood early warning in Europe.

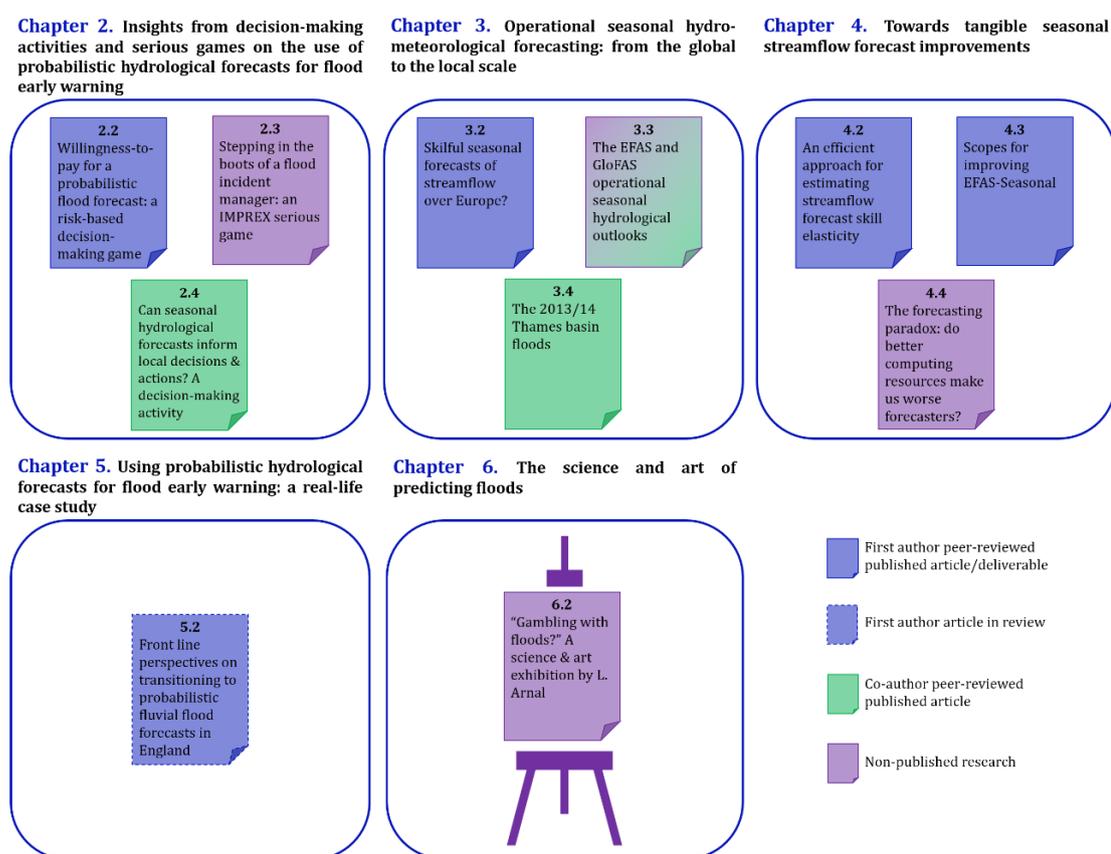


Figure 1. Overview of the result chapters of this thesis.

As has become evident from this introduction, streamflow forecasting is an art in that it is a skill acquired through centuries of practice, which has emerged from the fruitful combination of several disciplines (such as philosophy, environmental and social sciences), each with their own creative contributions to the topic at heart.

Chapter 2

Insights from decision-making activities and serious games on the use of probabilistic hydrological forecasts for flood early warning

2.1 Background and aim

We live in a world in which the disconnect between scientists, decision-makers and society is hindering progress. The uptake of new science by decision-makers has far from mirrored the extent of scientific advancements. Scientific developments are not always reflecting user needs. And scientific and decision-making progresses are both affected by the rise of a post-factual society. This disconnect is perhaps even more noticeable for science and decision-making associated with high societal stakes, such as the anticipation of floods.

In this disconnected world, tools and methods can be used to foster engagement and bridge the gap between scientists, decision-makers and society. Citizen science, the collection and analysis of data by members of the public as part of a collaborative project with scientists (e.g. the “weather rescue” project¹), can for example help enhance engagement between scientists and society by building awareness about a specific scientific topic and creating a sense of ownership. Serious games are another very popular tool which can be used to this end. They are “games developed for a purpose beyond entertainment alone.” Serious games “address societal challenges (e.g., ecological, social, economic, environmental, or a combination thereof)” and “tend to combine elements of entertainment (e.g., fun, suspense, mystery, inquiry) with elements of learning (e.g., developing knowledge, insights, skills).” (Aubert et al., 2019). In the last decade, the amount of environmental serious games has soared up (e.g. see the Red Cross/Red Crescent Climate Centre² and HEPEX³ online resources for examples). By providing a platform for sharing knowledge, serious games can build awareness and explain (e.g. the latest scientific developments), change views or offer an alternative perspective, enhance commitment and create “a sense of ownership of the

¹ www.zooniverse.org/projects/edh/weather-rescue/

² www.climatecentre.org/resources-games/games

³ hepex.irstea.fr/resources/hepex-games

decision”, giving players a chance to engage and give feedback on the serious topic at heart (Aubert et al., 2019).

In this chapter, we present two serious games and a decision-making activity, co-developed as part of this PhD. The first section of this chapter presents results from a serious game co-designed in collaboration with the HEPEX (ensemble hydrological forecasting) community. The second section introduces an online serious game co-designed with IMPREX project partners. The third section presents and summarises findings from a decision-making activity co-developed as part of IMPREX. While they were all designed with distinct storylines and for varied purposes, these serious games and activities all focus on flood early warning (from short to seasonal timescales). The aim of this chapter is to identify decision-makers’ requirements when using probabilistic hydrological forecasts for flood early warning.

2.2 Willingness-to-pay for a probabilistic flood forecast: a risk-based decision-making game

This section has been published in Hydrology and Earth System Sciences (HESS) with the following reference:

Arnal, L., M.-H. Ramos, E. Coughlan de Perez, H. L. Cloke, E. Stephens, F. Wetterhall, S. J. van Andel, and F. Pappenberger, 2016: Willingness-to-pay for a probabilistic flood forecast: a risk-based decision-making game, *Hydrol. Earth Syst. Sci.*, 20, 3109-3128, doi:10.5194/hess-20-3109-2016*

The contributions of the authors of this paper are as follows: M.-H. Ramos (collaborator: IRSTEA), E. Coughlan de Perez (collaborator: Red Cross/Red Crescent Climate Centre, Institute for Environmental Studies, VU and International Research Institute for Climate and Society), H. L. Cloke (supervisor: academic), E. Stephens (supervisor: academic), F. Wetterhall (collaborator: ECMWF), S. J. van Andel (collaborator: UNESCO-IHE) and F. Pappenberger (collaborator: ECMWF). L.A., M.-H.R., F.W., S.J.v.A. and F.P. designed the serious game. L.A. posed the research questions, designed and carried out the analysis, and wrote the paper. L.A., M.-H.R., H.L.C. and E.S. interpreted the results and commented on the

* ©2016. The Authors. Hydrology and Earth System Sciences, a journal of the European Geosciences Union published by Copernicus. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.

manuscript. Overall, 70% of the game design, 90% of the research and 90% of the writing was undertaken by L.A.

The published article can be found in the thesis Appendix A1 and the game can be played online (instructions and links on the HEPEX games resources page³, under “Pay-for-a-forecast game”).

Abstract. Probabilistic hydro-meteorological forecasts have over the last decades been used more frequently to communicate forecast uncertainty. This uncertainty is twofold, as it constitutes both an added value and a challenge for the forecaster and the user of the forecasts. Many authors have demonstrated the added (economic) value of probabilistic over deterministic forecasts across the water sector (e.g. flood protection, hydroelectric power management and navigation). However, the richness of the information is also a source of challenges for operational uses, due partially to the difficulty in transforming the probability of occurrence of an event into a binary decision. This paper presents the results of a risk-based decision-making game on the topic of flood protection mitigation, called “How much are you prepared to pay for a forecast?”. The game was played at several workshops in 2015, which were attended by operational forecasters and academics working in the field of hydrometeorology. The aim of this game was to better understand the role of probabilistic forecasts in decision-making processes and their perceived value by decision-makers. Based on the participants’ willingness-to-pay for a forecast, the results of the game show that the value (or the usefulness) of a forecast depends on several factors, including the way users perceive the quality of their forecasts and link it to the perception of their own performances as decision-makers.

2.2.1 Introduction

In a world where hydrological extreme events, such as droughts and floods, are likely to be increasing in intensity and frequency, vulnerabilities are also likely to increase (WMO, 2011; Wetherald and Manabe, 2002; Changnon et al., 2000). In this context, building resilience is a vital activity. One component of building resilience is establishing early warning systems, of which hydrological forecasts are key elements.

Hydrological forecasts suffer from inherent uncertainties, which can be from diverse sources, including the model structure, the observation errors, the initial conditions (e.g. snow cover, soil moisture, reservoir storages) and the meteorological forecasts of precipitation and temperature (Verkade and Werner, 2011; He et al., 2009). The latter variables are fundamental drivers of hydrological forecasts and are therefore major sources

of uncertainty. In order to capture some of this uncertainty, there has been a gradual adoption of probabilistic forecasting approaches, with the aim of providing forecasters and forecast users with additional information not contained in the deterministic forecasting approach. Whereas “a deterministic forecast specifies a point estimate of the predictand (the variate being forecasted)”, “a probabilistic forecast specifies a probability distribution function of the predictand” (Krzysztofowicz, 2001). For operational forecasting, this is usually achieved by using different scenarios of meteorological forecasts following the ensemble prediction approach (Buizza, 2008; Cloke and Pappenberger, 2009).

Many authors have shown that probabilistic forecasts provide an added (economic) value compared to deterministic forecasts (Buizza, 2008; Verkade and Werner, 2011; Pappenberger et al., 2015). This is due, for example, to the quantification of uncertainty by probabilistic forecasting systems, their ability to better predict the probability of occurrence of an extreme event and the fact that they issue more consistent successive forecasts (Dale et al., 2014; Cloke and Pappenberger, 2009). This probability of occurrence makes the probabilistic forecasts useful in the sense that they provide information applicable to different decision thresholds, essential since not all forecast users have the same risk tolerance (Michaels, 2015; Buizza, 2008; Cloke and Pappenberger, 2009). Probabilistic forecasts therefore enable the quantification of the potential risk of impacts (New et al., 2007) and, as a result, they can lead to more optimal decisions for many hydrological operational applications, with the potential to realise benefits from better predictions (Verkade and Werner, 2011; Ramos et al., 2013). These applications are, for example, flood protection (Stephens and Cloke, 2014; Verkade and Werner, 2011), hydroelectric power management (García-Morales and Dubus, 2007; Boucher et al., 2012) and navigation (Meissner and Klein, 2013). Moreover, the continuous increase in probabilistic forecast skill is very encouraging for the end-users of the probabilistic forecasts (Bauer et al., 2015; Magnusson and Källén, 2013; Simmons and Hollingsworth, 2002; Ferrell, 2009).

However, the communication of uncertainty through probabilistic forecasts and the use of uncertain forecasts in decision-making are also challenges for their operational use (Cloke and Pappenberger, 2009; Ramos et al., 2010; Michaels, 2015; Crochemore et al., 2015). One of the reasons why the transition from deterministic to probabilistic forecasts is not straightforward is the difficulty in transforming a probabilistic value into a binary decision (Dale et al., 2014; Demeritt et al., 2007; Pappenberger et al., 2015). Moreover, decision-makers do not always understand probabilistic forecasts the way forecasters intend them to (Handmer and Proudley, 2007). This is why it is essential to bridge the gap between

forecast production and hazard mitigation, and to foster communication between the forecasters and the end-users of the forecasts (Cloke and Pappenberger, 2009; Michaels, 2015).

As Michaels (2015) notes, “the extent to which forecasts shape decision making under uncertainty is the true measure of the worth of a forecast”. The potential added value of the forecast can furthermore only be entirely realised with full buy-in from the decision-makers. However, how much are users aware of this added value? How much are they ready to pay for a forecast? These are questions that motivated the work presented in this paper. In order to understand how users perceive the value of probabilistic forecasts in decision-making, we designed a risk-based decision-making game – called “How much are you prepared to pay for a forecast?” – focusing on the use of forecasts for flood protection. The game was played during the European Geophysical Union (EGU) General Assembly meeting 2015 (Vienna, Austria), at the Global Flood Partnership (GFP) workshop 2015 (Boulder, Colorado), as well as at Bristol University (BU) in 2015. Games are increasingly promoted and used to convey information of scientific relevance. They foster learning, dialogue and action through real-world decisions, which allow the study of the complexities hidden behind real-world decision-making in an entertaining and interactive set-up (Mendler de Suarez et al., 2012).

This paper presents the details of the game and the results obtained from its different applications. The participants’ perceived forecast value is analysed by investigating the way participants use the forecasts in their decisions and their willingness-to-pay (WTP) for a probabilistic forecast. The WTP is the amount an individual is inclined to disburse to acquire a good or a service, or to avoid something undesirable (Breidert et al., 2006; Leviäkangas, 2009). It is a widely and very commonly adopted method to make perceived value assessments and its use has been demonstrated in a meteorological context (Leviäkangas, 2009; Anaman et al., 1998; Rollins and Shaykewich, 2003; Breidert et al., 2006). Breidert et al. (2006) present a complete overview of the methods available, organised by data collection types. According to their classification, there exist two main WTP measuring approaches: the “revealed preference” and the “stated preference”. The former describes price-response methods (such as market data analysis, laboratory experiments and auctions, amongst others), while the latter refers to surveys in general. This experiment combines both “revealed preference” and “stated preference” methods. The design of the game is described in Sect. 2.2.2 and justified in terms of the purpose and contribution of the different components of the game to its main aim. The results and the

discussion promoted by the latter are subsequently presented in Sects. 2.2.3 and 2.2.4 respectively.

2.2.2 Set-up of the decision-making game

2.2.2.1 Experimental design

This game was inspired by the table game “Paying for Predictions”, designed by the Red Cross/Red Crescent Climate Centre (<http://www.climatecentre.org/resources-games/paying-for-predictions>). Its focus is however different. Here, our aim is to investigate the use of forecasts for flood protection and mitigation. Also, we strongly adapted the game to be played during conferences and with large audiences.

The set-up of the game (illustrated in Fig. 1a) was the following: participants were told that they were competing for the position of head of the flood protection team of a company. Their goal was to protect inhabitants of a fictitious town bordering a fictitious river against flood events, while spending as little money as possible during the game. The participant with the highest amount of money at the end of the game was chosen as head of the flood protection team. Each participant was randomly assigned a river (river yellow, river blue or river green) for the entire duration of the game. Each river had distinct initial river levels and rates of flood occurrences (see Table 1). Participants worked independently and had a worksheet to take notes (see Appendix A). An initial purse of 20000 tokens was given to each player to be used throughout the game.

Based on this storyline, the participants were presented the following sequence of events (illustrated in Fig. 1b): after being given their river’s initial level (ranging from 10 to 60 included), each participant was asked to make use of a probabilistic forecast (see Fig. 1b) of their river level increment after rainfall (ranging from 10 to 80 included) to decide whether they wanted to pay for flood protection or not. The cost of flood protection was 2000 tokens. They were informed, prior to the start of the game, that a flood occurred if the sum of the initial river level and the river level increment after rainfall (i.e. the actual river level after rainfall) reached a given threshold of 90. The probabilistic forecasts were visualised using boxplot distributions. They had a spread of about 10–20 and indicated the 5th and 95th percentiles as well as the median (i.e. 50th percentile) and the lower and upper quartiles (i.e. 25th and 75th percentiles respectively) of the predicted river level increment after rainfall. Forecasts were given to participants case by case (i.e. when playing the first case, they could only see the boxplot distribution of forecast river increment for case 1). Once the participants had made their decisions using both pieces of information (i.e. river level

before rainfall and forecast of river level increment), they were given the observed (actual) river level increment after rainfall for their rivers. If a flood occurred and the participant had not bought flood protection, a damage cost (i.e. price paid when no protection was bought against a flood that actually happened) of 4000 tokens had to be paid.

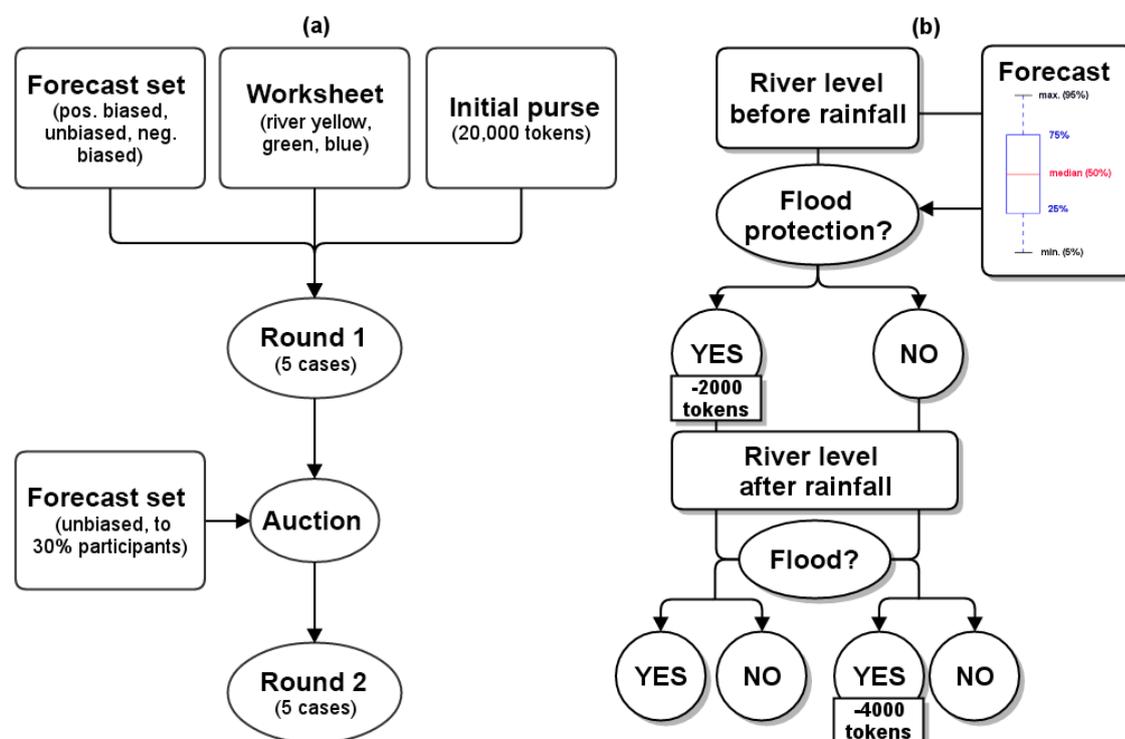


Figure 1. (a) Experiment set-up and (b) flow diagram of the game decision problem for one case.

The monetary values (initial purse, price of flood protection and damage cost) were deliberately chosen. The price of a protection was set to 2000 tokens such that if a participant decided to buy flood protection every time during the game (i.e. two rounds of five cases each, thus ten times) they would have no tokens left in their purse at the end of the game. This was done in order to discourage such a behaviour. The damage cost was set to twice the flood protection cost as this was estimated to be a realistic relation between the two prices based on Pappenberger et al. (2015). The latter states that the avoided damages due to early flood warning amount to a total of about 40%. Here, for simplicity, we used a percentage of 50%.

Once the context was explained, the participants were then told that they would first play one round of five independent cases, which would each be played exactly according to the sequence of events presented, and for which they would have to record their decisions on the worksheet they were provided (see Appendix A). The game had a total of two rounds of five cases each. This specific number of cases and rounds was chosen because of the time constraint to play the game during conferences (the game should last around 20–30 min

only). Table 1 presents the total number of flood events for each round and each river. The number of flood events was different for every river for each round as river level values were randomly generated for the purpose of the game. This allowed the exploration of the influence of different flood frequencies in round 1 on the participants' WTP for a second forecast set. The number of flood events was however sampled to some extent in order to obtain decreasing (increasing) numbers of flood events between the two rounds for the blue (yellow) river, or constant throughout the two rounds for the green river. This was done to investigate the effect of the change (or not) in flood frequency between rounds 1 and 2 on the participants' strategies throughout the game.

Table 1. Number of flood events for each round of the game and each river.

Round	River		
	Yellow	Green	Blue
1	1	2	3
2	3	2	1
Total	4	4	4

During the first round of the game, the participants had forecasts of river level increments to help their decisions. These forecasts were however not available for all participants in the second round, but were sold between the two rounds through an auction. The purpose and set-up of each round and the auction are explained in the following paragraphs.

2.2.2.1.1 Round 1

The objective of the first round was to familiarise the participants with the probabilistic forecasts they were given to help them in their decisions, and to create a diversity amongst the decision-makers in terms of

- their river behaviour: which is why different rivers, each with different flood frequencies and different initial levels, were assigned to the participants;
- the money they would spend during this round and have in hand for the ensuing auction (before round 2);
- the quality of their forecasts in the first round: to this end, different forecast sets were distributed to the players for round 1.

This diversity was triggered in round 1 in order to analyse whether or not the WTP for a second forecast set, measured in the auction performed before round 2, was dependent on

any of the factors inherent to the first round (i.e. river-specific flood frequency, money left in purse, or quality of the forecasts).

Before the start of the first round each participant was given a forecast set containing probabilistic forecasts of their river level increment after rainfall for the five cases of round 1. Participants were however not aware that three different forecast sets were produced for each of the rivers. One set had only forecasts with a positive bias (forecast sets 1), the second set had only unbiased forecasts (forecast sets 2) and the third set only forecasts with a negative bias (forecast sets 3). There were therefore nine different sets of forecasts which were distributed randomly amongst the audience prior to the start of the game. The three different forecast types were obtained by varying the position of the observation inside the forecast distribution. The unbiased forecasts had the observations fall between the lower and upper quartiles of their distributions, while the biased forecasts had the observations fall outside of the lower and upper quartiles of their distributions, leading to over- (positively biased forecast sets) or under-predictions (negatively biased forecast sets) of the observations.

The quality of each forecast set can be represented in terms of the number of correct forecast flood events (given a forecast percentile threshold) with respect to the number of observed flood events. For each forecast set type and each river, the number of forecast flood events during the first round was calculated by adding the median of the forecast river level increment to the initial river level for each case. A forecast is referred to as a false alarm if this sum forecasts a flood (i.e. it exceeds the flood threshold) but the flood is subsequently not observed. It is referred to as a hit if the sum forecasts the flood and the flood is subsequently observed. A miss is an observed flood that was not forecast. The numbers of hits, misses and false alarms are usually gathered in a contingency table as a matrix (e.g. Table 2): hits are placed on top, left, misses on bottom, left, and false alarms on top, right. The place on bottom, right is usually not considered in the evaluation of forecasts as it represents situations of low interest to a forecaster (i.e. when floods are neither forecast nor observed). Table 2 displays the nine contingency tables we obtain considering each forecast set type and each river. Each participant would find themselves in one of the contingency tables represented. We can see the higher number of total misses (false alarms) considering all rivers together in negatively (positively) biased forecast sets, and the absence of these in the unbiased forecast sets.

After all the five cases of round 1 were played, participants were asked to rate their performance as a decision-maker and the quality of their forecast set for round 1 on a scale

from “very bad” to “very good” (the option “I don’t know” was also available) (see Appendix A).

Table 2. Contingency table for each river and forecast set type for the first round (considering the 50th percentile, i.e. the median forecast). The numbers for a specific river-forecast set type represent, clockwise from the top left, hits (*italics*), false alarms (**bold**), correct negatives (–) and misses (regular).

Forecast set type	River					
	Yellow		Green		Blue	
Positively biased	<i>1</i>	1	<i>2</i>	1	<i>3</i>	2
	0	–	0	–	0	–
Unbiased	<i>1</i>	0	<i>2</i>	0	<i>3</i>	0
	0	–	0	–	0	–
Negatively biased	<i>0</i>	0	<i>2</i>	0	<i>2</i>	0
	1	–	0	–	1	–

2.2.2.1.2 Auction

The auction was carried out after round 1 in order to measure the participants’ WTP for a second forecast set and to evaluate its dependencies on any of the elements of the game in round 1. The auction was implemented as follows.

At the end of the first round participants were asked to transfer the remaining tokens from round 1 to the second round. They were then told that the forecasting centre distributing the probabilistic forecasts now wanted the decision-makers to pay for the forecast sets if they wanted to have access to them for the second round. Furthermore, they were informed that only 30% of them could get a second forecast set for this round. This percentage was chosen in order to restrict the number of participants that could buy a forecast set (and create a competitive auction), while keeping a high enough number of participants playing with a forecast set in round 2 for the analysis of the results.

Participants were then asked to make a sealed bid, writing down on their worksheets the number of tokens they were willing to disburse from their final purse of round 1 to obtain a set of probabilistic forecasts for all five cases of round 2. After the bids were made, a forecast set was distributed to the participants within the highest 30% of the bids. This was done through an auction. It was carried out by asking the participants whether any of them wrote down a bid superior or equal to 10000 tokens. If any participants did, they raised their hands, after which a forecast set – for the same river as the river assigned to them at the beginning of the game – was given to them. The auction continued by lowering the

number of tokens stated to the participants until all forecast sets for round 2 were distributed. Each participant having bought a forecast set for round 2 was then asked to disburse the number of tokens they paid for this forecast set from their remaining purse from round 1.

We note that participants were not told that the forecasts for the second round were all unbiased forecasts. Once again, the quality of the forecasts was kept secret in order for the participants to assign a value to the second forecast set that would strictly be related to the conditions under which they played the first round.

2.2.2.1.3 Round 2

The second round was played in order to measure the added value of an unbiased forecast set, compared to no forecast set at all, to the decisions of the participants on protecting or not against floods. Moreover, as the winner of the game was determined by the number of tokens left in their purse at the end of the game, this round would give a chance to participants who bought a second forecast set to make up for the money spent with the auction, during round 2.

The second round developed similarly to the first round, with five independent cases of decision-making, with the exception that only participants who bought a second forecast set could use it to make their decisions. Participants who did not buy a second forecast set did not have any forecasts on which to base their decisions.

After the five cases were played, the participants were asked to once again answer a set of questions (see Appendix A). They were asked to rate their performance as a decision-maker in the second round, on a scale from “very bad” to “very good” (the option “I don’t know” was also available). Participants without a second forecast set were invited to provide a justification for not purchasing a set of forecasts for this round. Participants who had bought a second forecast set were also asked to rate the quality of their forecast set for round 2 (on a scale from “very bad” to “very good”; the option “I don’t know” was also available) and whether those were worth the price they had paid for them. If not, they were asked to provide a new price that they would have rather paid.

The winner was finally determined by finding the player with the largest number of tokens in their purse at the end of the game.

2.2.2.2 Objectives and evaluation strategy

The main aim of this paper is to investigate the participants' WTP for a probabilistic forecast set in the context of flood protection, following the game experiment designed as presented in the previous paragraphs. It unfolds into two objectives that were pursued in the analysis of the results:

- 1 to analyse how participants used the information they were provided (probabilistic forecast sets) in this risk-based decision-making context, and
- 2 to characterise the participants' WTP for a probabilistic forecast set for flood protection.

We assess these objectives through six questions, which are presented below, together with the evaluation strategy implemented.

2.2.2.2.1 Did the participants use their forecasts and, in this case, follow the 50th percentile of their forecast during the decision-making process?

This first question was investigated using the results of the first round. We first wanted to know whether the players were actually using their forecasts to make their decisions. Moreover, we searched for clues indicating that the participants were following the 50th percentile (i.e. the median) of the probabilistic forecasts. This was done in order to see whether the 50th percentile was considered by the players as the optimal value to use for the decision-making process under this specific flood risk experiment. Additionally, this question relates to an intrinsic characteristic of the use of probabilistic forecasts for decision-making, which is the difficulty in transforming the probabilistic values into a binary decision (Dale et al., 2014; Demeritt et al., 2007; Pappenberger et al., 2015). The way in which probabilistic flood forecasts are used depends on attitudes of decision-makers towards risk, the uncertainty and the error in the information provided to them (Demeritt et al., 2007; Ramos et al., 2013), and decisions can vary from one participant to the next provided the same information (Crochemore et al., 2015).

Question one was explored by looking at the worksheets collected in order to infer from the decisions taken by the participants whether or not they most probably used the median of their forecasts to consider whether the river level would be above, at or under the flood threshold. In cases where the decisions did not coincide with what the median forecast indicated, other factors that could also influence the decisions were considered, such as (a) the flood frequency of each river and their initial river levels, (b) the forecast set type each participant had (i.e. biased – positively or negatively – or unbiased) and (c) the familiarity

of the participants with probabilistic forecasts and decision-making (given their occupation and years of experience).

2.2.2.2.2 Was there a correspondence between the way participants perceived the quality of their forecasts in round 1 and their “true” quality?

A well-known effect, called the “cry wolf”, was studied for weather-related decision-making by LeClerc and Joslyn (2015). It describes the reluctance of users to comply with future alarms when confronted in the past with false alarms. This leads to the second question which was explored in this paper: was there a correspondence between the way participants perceived the quality of their forecasts in round 1 and their “true” quality? Our aim here is to investigate whether the participants were more sensitive to false alarms or misses. The participants’ answers to the question on their forecast set quality for the first round (see Appendix A) were analysed against their “true” quality. The latter was measured in terms of forecast bias, calculated from the hits, false alarms and misses presented in Table 2. A bias value was computed for each forecast set type of each river (i.e. each contingency table; there were therefore nine different bias values in total) with the following equation:

$$\text{Bias} = \frac{\text{hits} + \text{false alarms}}{\text{hits} + \text{misses}} \quad (1)$$

A bias value equal to one is a perfect value (which corresponds to unbiased forecasts), and a value less than (superior to) one indicates under- (over-)prediction.

2.2.2.2.3 Did the participants’ perceptions of their own performance coincide with their “true” performance?

We also looked at the perception the participants had of their own performance. The answers to the question “How was your performance as a decision-maker” (see Appendix A) were assessed against the participants’ “true” performances (in rounds 1 and 2), which were calculated in terms of the money participants spent as a consequence of their decisions. The following general formula (n being the round number) was used:

$$\text{Performance} = \frac{\text{Money spent round n}}{\text{Optimal}} \quad (2)$$

The performance is expressed relatively to an optimal performance, which is the minimum amount a participant could have spent, given the river they were assigned, defined as

$$\text{Optimal} = \text{Protection cost} \times \text{Number of floods in round n} \quad (3)$$

A performance value of one indicates an optimal performance. Performance values greater than one indicate that participants spent more money than the minimum amount necessary to protect the city from the observed floods. The greater the value, the higher the amount of money unnecessarily spent.

2.2.2.2.4 What was the participants' willingness-to-pay for a probabilistic forecast set?

The auction was incorporated into the experiment in order to explore the WTP of participants for a probabilistic forecast set, considering the risk-based decision-making problem proposed by the game. To characterise this WTP, the bids were analysed and their relationships with several other aspects of the game were explored to explain the differences (if any) in the bids. These aspects were the following.

- The way participants used the forecasts. Here we try to learn about the effectiveness of the information on the user, which is an attribute of the value of information (Leviäkangas, 2009). It is assumed that a participant is not expected to be willing to disburse any money for information they are not using. The answers to question one (i.e. "Did the participants use their forecasts and, in this case, follow the 50th percentile of their forecast during the decision-making process?") are used here.
- The money available to participants after round 1 to make their bids. As participants were informed at the beginning of the game that the winner would be the player with the highest number of tokens in purse at the end of the game, the tokens they had in hand for the auction (after round 1) may have restricted them in their bids. The bids are thus also explored relative to the number of tokens in hand at the time of the auction.
- The forecast set type. The bias of the forecasts during round 1 could also have been a potential determinant of participants' WTP for a forecast set in round 2.
- The river flood frequency. This was different for all the rivers in the first round and could be an element of the relevance of the information, another attribute of the value of information (Leviäkangas, 2009). Indeed, one could ask: "If my river never floods, why should I pay for forecasts?".
- The years of experience and occupation. This might influence the familiarity participants may have with the use of probabilistic forecasts for decision-making.

2.2.2.2.5 Did participants with a forecast set perform better than those without?

Round 2 was led by a central question: did participants with a forecast set perform better than those without? It was investigated by looking at the performance of participants in

round 2, calculated from Eq. (2). While we expect players with more (unbiased) information to make better decisions, other factors could have influenced the trust participants had in the information during round 2, such as, for instance, the quality of the forecasts experienced by participants in round 1 or the flood events observed in the river in round 2, compared to the experience participants had previously had in round 1.

2.2.2.2.6 What were the winning and losing strategies (if any)?

Finally, from the final results of the game, a question arose: what were the winning and losing strategies (if any)? This question was explored by looking at the characteristics (e.g. river assigned, forecast set type in round 1, performances in both rounds, purchase of a second forecast set) and decisions of the participants during the game, in order to distinguish common attributes for the winning and losing strategies.

Furthermore, an “avoided cost” was calculated for each river based on the difference between the tokens spent by participants without a second forecast set and the tokens spent by participants with a second forecast set, during round 2. It represents the average number of tokens participants without a second forecast set lost by protecting when a flood did not occur or by not protecting when a flood did occur, compared to participants with a second forecast set. This “avoided cost” was measured and compared to the average bid of participants for each river in order to evaluate participants’ estimation of the value of the forecasts compared to their “true” value in terms of the money they enabled the participants with a second forecast set to save in the second round. An average “new bid” was also calculated by replacing the bids of participants who had said that their forecast set in the second round was not worth the price they had paid initially, with the new bids they would have rather paid (see Appendix A). This average “new bid” was compared to the “avoided cost” and the actual average bid obtained from the auction.

2.2.3 Results

The results are based on the analysis of 129 worksheets from the 145 worksheets collected. The remaining 16 worksheets were either incomplete or incorrectly completed and were thus not used. Table 3 shows the distribution of the 129 worksheets among the three forecast set types and the three rivers.

Table 3. Distribution of the 129 worksheets collected for the analysis per river (yellow, green and blue) and forecast set type (positively biased, unbiased and negatively biased).

Forecast set type	River			Total
	Yellow	Green	Blue	
Positively biased	15	11	18	44
Unbiased	13	21	9	43
Negatively biased	11	19	12	42
Total	39	51	39	129

The game was played at the different events mentioned in the introduction. The participants present at those events displayed a diversity in terms of their occupation and years of experience. This was surveyed at the beginning of the game and is presented in Fig. 2, for all the participants as well as for each river and forecast set type separately. Participants were mainly academics (postdoctoral researchers, PhDs, research scientists, lecturers, professors and students), followed by professionals (forecasters, operational hydrologists, scientists, engineers and consultants). The majority had less than 5 years of experience.

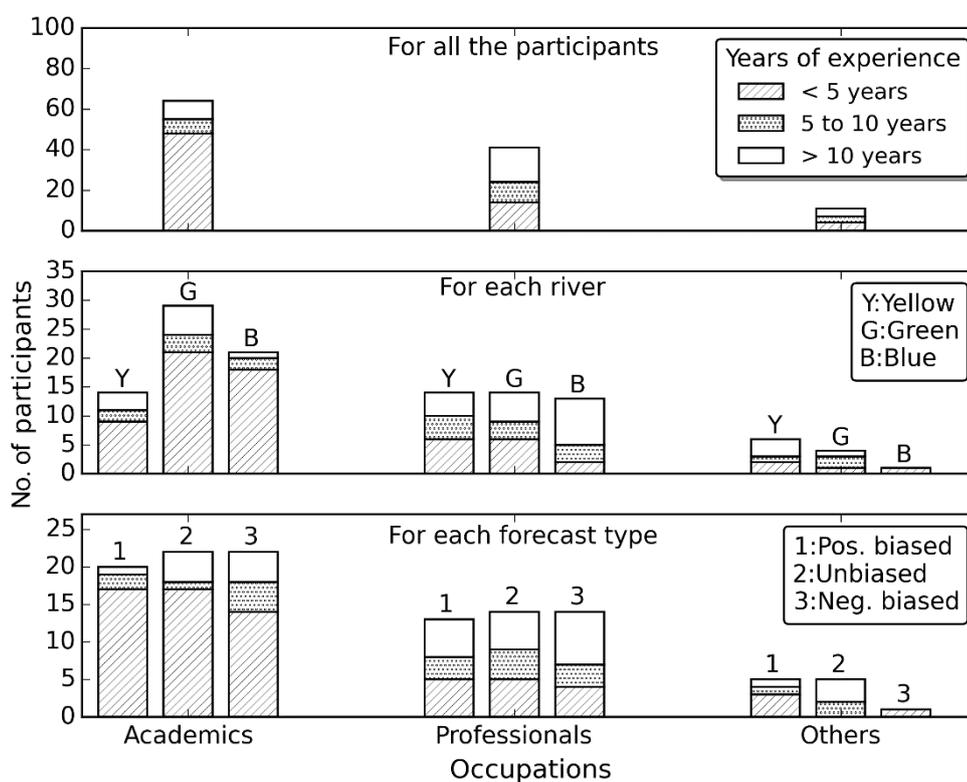


Figure 2. Number of participants according to occupation and years of experience. The categories of occupations are academics (postdoctoral researchers, PhDs, research scientists, lecturers, professors and students), professionals (forecasters, operational hydrologists, scientists, engineers and consultants) and others. Top: overall participant distribution;

middle: distribution according to their river; bottom: distribution according to the forecast quality types (1: positively biased, 2: unbiased and 3: negatively biased).

2.2.3.1 Participants were using the forecasts, but consistent patterns of use are difficult to detect

Figure 3 presents, on the one hand, the final purses of all the participants at the end of round 1, according to their river and forecast set type (columns and rows respectively), and, on the other hand, the final purses that participants would have had if they had made their decisions according to the median of their forecasts. Participants in charge of the yellow river (first column) ended the first round with, on average, more tokens than the others. Participants playing with the blue river (last column) are those who ended round 1 with less money in purse, on average. This is due to the higher number of flood events for the blue river in round 1 (see Table 1). There are also differences in terms of final purses for the participants assigned the same river but given a different forecast set type. Overall, participants who had unbiased forecasts (middle row) ended the first round with on average more money than the other players. These results are an indication that the participants were using their forecasts to make their decisions.

In order to see whether the participants were using the median values of the forecasts, a forecast final purse was computed considering the case where the participants followed the median of their forecasts for all the cases of the first round (red vertical lines shown in Fig. 3). If the participants had followed the median values of the forecasts during the entire first round, their final purses would have been equal to this value. Although this is almost the case for participants with unbiased forecast sets (for all rivers), for participants with the yellow river and positively biased forecast sets and the green river and negatively biased forecast sets, it is not an overall generally observed behaviour.

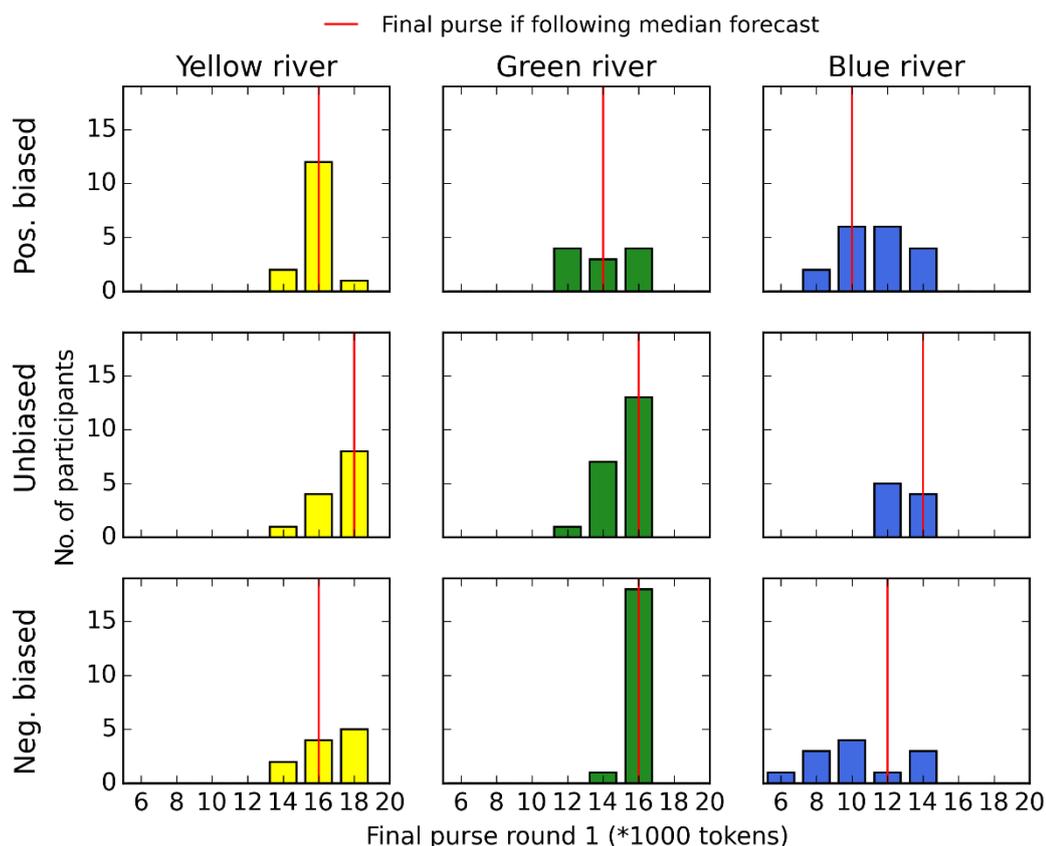


Figure 3. Participants' round 1 final purses for each river (from the leftmost to the rightmost column: the yellow, the green and the blue river) and for each forecast set type (from the top to the bottom row: positively biased, unbiased and negatively biased). The red lines show the final purses that the participants of a given river-forecast set type group would have gotten if they had followed the median of their forecasts for all five cases of the first round.

Could some participants have discovered the bias in their forecasts and adjusted them for their decisions? Although it is hard to answer this question from the worksheets only, some of the decisions taken seem to support this idea. Figure 4 presents in more detail the results for the blue river in the first round. The forecast final levels are shown as boxplots for each forecast set type and for each of the five cases of round 1. These are the levels the river would reach if the initial level is added to the percentiles of the forecasts for each case. The bars at the bottom of the figure show the percentages of participants whose decisions differed from what the median of their forecast final level indicated (i.e. participants who bought (or did not buy) protection while no flood (or a flood) was predicted by the median of their forecast).

When comparing cases 1 and 4, for which the initial river levels and the observed and forecast final river levels were the same, we would not expect any changes in the way participants were using their forecasts. This is however not true. Figure 4 shows that the percentages of participants not following their forecast median differs between the two cases. For instance, about 80% of the participants with negatively biased forecast sets

(under-predicting the increment of the river level) did not follow the median forecast in case 1, and did not protect against the predicted flood by their median forecast, while this percentage drops to about 20% in case 4. The fact that they were not consistently acting the same way may be an indication that they found out the bias in the forecasts and tried to compensate for it throughout round 1. We can also see that, in general, the lowest percentages of participants not following the median forecast are for the unbiased forecast set. This is especially observed in the cases where the forecast final levels given by the median forecast are well above or below the flood threshold (cases 1, 2, 4, 5). The fact that from case 1 to case 4, for unbiased forecast sets, we moved from about 10% of participants not following the median forecast to 0%, may also indicate that they built confidence in their forecasts (at least in the median value) along round 1, by perceiving that the median forecast could be a good indication of possible flooding or not in their river.

Figure 4 also shows that some participants with unbiased forecasts did not always follow the median of their forecasts (for instance, cases 1, 3 and 5). Additional factors may therefore have influenced the way participants used their forecasts. A number of worksheets indicated that the distance of the initial river level to the flood threshold could have been influential. In a few cases where the median forecast clearly indicated a flood, while the initial river level was low, some players did not purchase any flood protection. This can be observed in Fig. 4 for case 1, for example, for participants with positively biased or unbiased forecast sets. The inverse situation (i.e. the initial river level was high, but the river level forecast by the median was low, below the flood threshold) was also observed and is illustrated in Fig. 4 for case 2 and negatively biased forecast sets. Hence, in some cases, the initial river level seemed to also play a role in the decisions taken.

There are indications that the participants could also have used other percentiles of the forecast to make their decisions, especially in cases where the median of the forecast was marginally above or below the flood threshold. For example, in case 4, the entire unbiased forecast lies above the flood threshold and all the participants chose the same and correct action. In cases where the 5th or 95th percentiles of the forecast fell above or below the flood threshold, the participants showed less consistent decisions (e.g. case 3 for unbiased forecast sets).

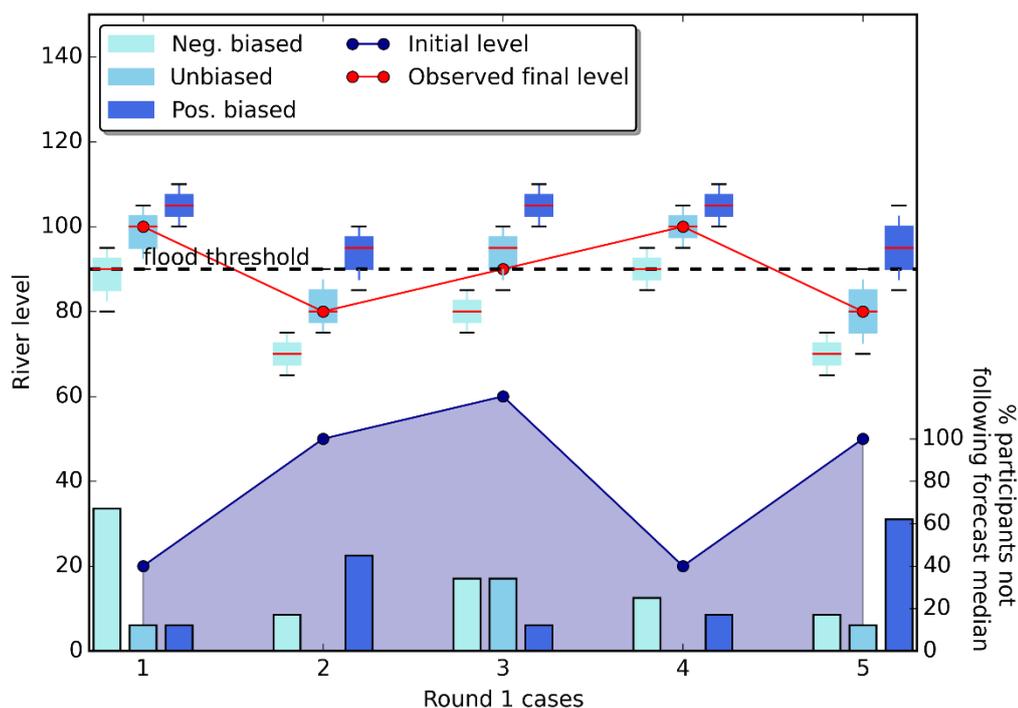


Figure 4. Observed initial and final river levels for the blue river for each case of the first round. The boxplots show the forecast final river levels by each forecast set type (negatively biased, unbiased and positively biased). The bars display the percentages of participants whose decisions did not correspond to what their forecast median indicated.

Other possible influencing factors, such as occupation and years of experience, were also investigated (not shown). No strong indications that these factors could have played a role in the participants' decision-making were however found.

2.2.3.2 Participants were overall less tolerant to misses than to false alarms in round 1

Figure 5 displays the cumulative percentages of participants having answered that the quality of their forecast set in round 1 (see Appendix A) was "very bad" to "very good", as a function of the "true" quality of the corresponding forecasts, measured by the forecast set bias (Eq. 1). While participants with forecast sets for which the bias equalled one (perfect value) mostly rated their forecasts "quite good" or "very good", the percentage of negative perceptions of the quality of the forecasts increases with increasing or decreasing forecast bias.

It is interesting to note that participants with forecasts biased towards over-prediction never rated their forecasts as "very bad". Also noteworthy is the very good rating given by participants with the most negatively biased forecasts (bias of 0). These participants belonged to the yellow river and had negatively biased forecasts in round 1. There was only one flood event for river yellow in the first round, which occurred at the end of the round and which was missed by the negatively biased forecasts. During the analysis of the results,

it was observed that only about 25% of the yellow river participants given the negatively biased forecasts did not purchase flood protection for this flood. An explanation for this low percentage could be that participants had time to learn about their forecasts' quality until the occurrence of the flood at the end of the first round. This low number of participants who actually suffered from their negative bias and the presence of only one miss out of the five cases of round 1 could therefore justify the good rating of their forecasts by those participants.

Overall, forecasts exhibiting under-prediction seem to be less appreciated by the participants. This could be an indication that participants were less tolerant to misses, while they accepted better forecasts leading to false alarms (over-predictions). This is contrary to the “cry wolf” effect, and could be explained by the particular game set-up for which the damage cost (4000 tokens) was twice the protection cost (2000 tokens).

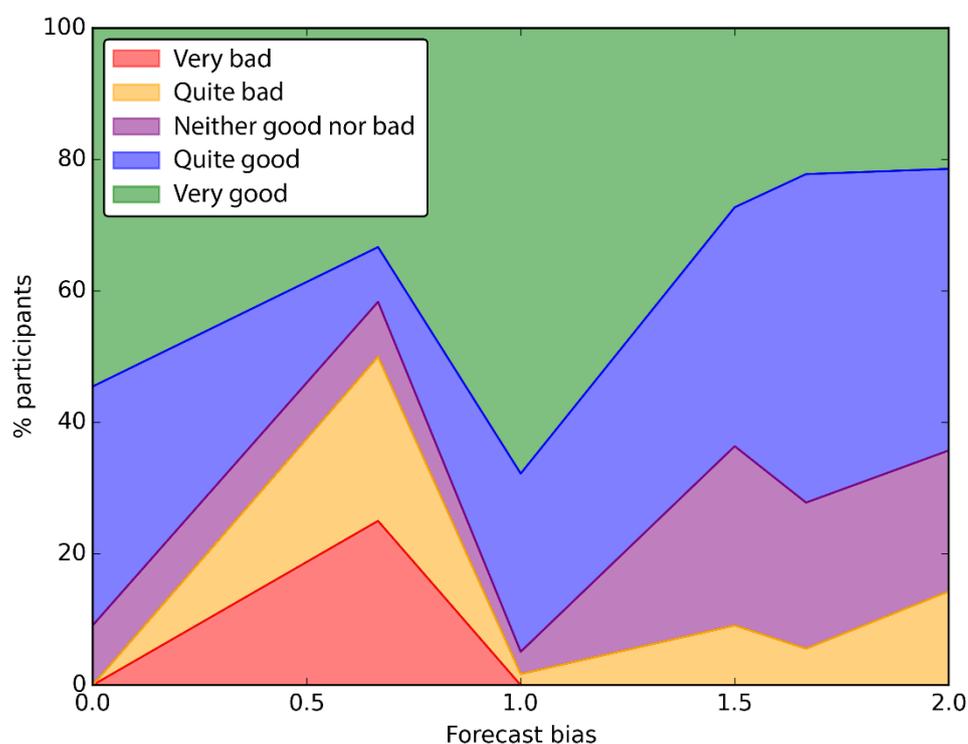


Figure 5. Cumulative percentages of participants who rated their forecast quality from “very bad” to “very good”, as a function of the forecast set bias (“true” forecast quality; Eq. 1) in round 1. A bias equal to one indicates perfect forecasts; a bias less than (superior to) one indicates under- (over-)prediction.

2.2.3.3 Participants had a good perception of their good (or bad) performance during the game and related it to the quality of their forecasts

Figure 6a illustrates the answers to the question “How was your performance as a decision-maker in round 1?” as a function of the participants’ “true” performance (calculated from Eq. (2), i.e. the ratio to an optimal performance). The figure shows the distribution of

participants across all perceived–actual performance combinations, for all rivers and forecast set types combined. The perceived decision-maker performance is presented on a scale from “very bad” to “very good”. An overall positive relationship between the participants’ perceived performance and their “true” performance is observed: the best performances (i.e. performance values of one or close to one) are indeed associated with a very good perception of the performance by the decision-makers and vice versa. The same analysis carried out for the answers concerning round 2 (not displayed) showed similar results: the ratings participants gave to their performance were similarly close to their “true” performance.

Figure 6b looks at the relationship between the perceived decision-maker performance and the rating the decision-makers gave to their forecast set quality in round 1. A positive relationship can also be seen: the majority rated their performance and the quality of their forecast set as “quite good” and “very good”, while those who rated their performance “very bad” also considered their forecast set “very bad”. The rating participants gave to their performance was therefore closely connected to the rating they gave to their forecast set quality. This also contributes to the evidence that participants were using their probabilistic forecast sets to make their decisions. It is furthermore an indication that participants linked good forecast quality to good performance in their decision-making and vice versa.

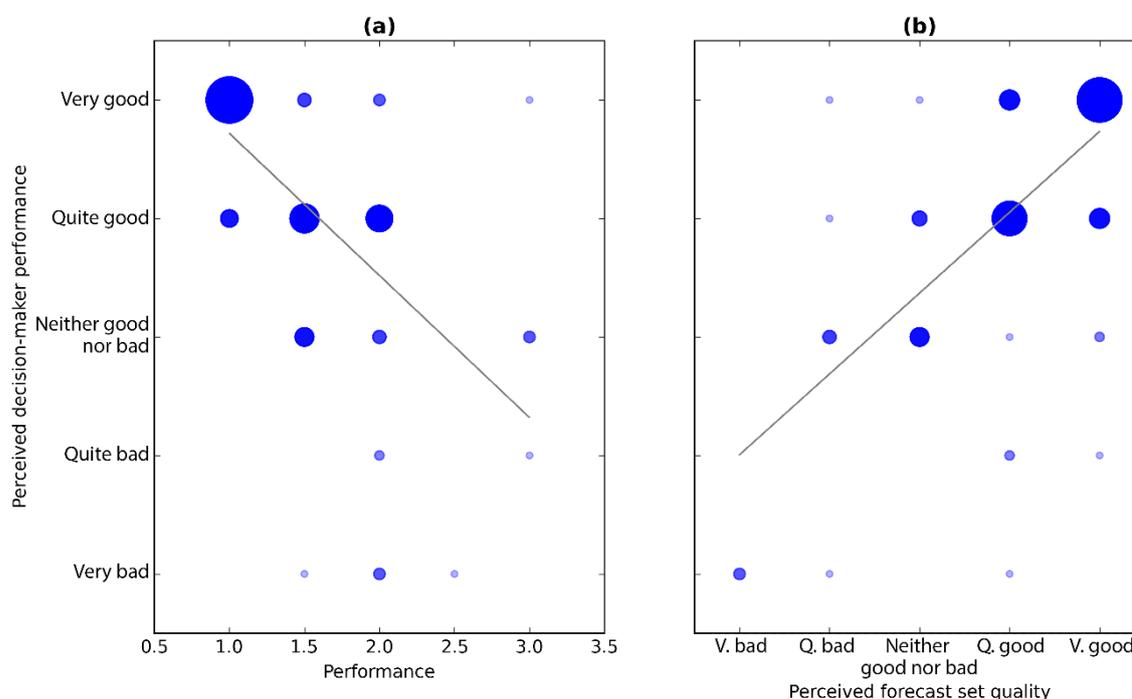


Figure 6. Number of participants having rated their performance as a decision-maker from “very bad” to “very good” in round 1, as a function of (a) their “true” performance (calculated from Eq. 2), and (b) their perceived forecast set quality. A performance value of one denotes a “true” performance equal to the optimal performance (Eq. 3). The larger the performance value, the more distant from optimal the decisions were during round 1. The size and the colour of the point indicate the

number of participants that fall into a specific perceived–actual performance combination or perceived performance–forecast set quality combination.

2.2.3.4 Several factors may influence the WTP for a forecast, including forecast quality and economic situation

Given the evidence that most participants were using their forecasts to make their decisions in round 1 (see Sect. 2.2.3.1), we now investigate their willingness-to-pay (WTP) for a new forecast set to be used in round 2.

Figure 7 shows the bids participants wrote on their worksheets prior to the auction, for a second forecast set, as a function of the number of tokens they had in their purses at the end of round 1. All bids are plotted and those from participants who succeeded in buying a second forecast set are displayed as red triangles in the figure. On average, participants were willing to pay 4566 tokens, which corresponds to 32% of the average number of tokens left in their purses. The minimum bid was zero tokens (i.e. no interest in buying forecasts for round 2), which was made by 10% of the players. Half of these players were participants who were assigned the blue river (the river for which players ended the first round with on average the lowest number of tokens in purse). The only three participants who never bought flood protection in the first round (i.e. who could be seen as “risk-seeking” players) made bids of zero, 3000 and 4000 tokens. The highest bid made was 14000 tokens, corresponding to 100% of the tokens left in that participant’s purse. However, this participant did not raise their hand during the auction to purchase a second forecast set. Nine participants (less than 10% of the total number of players) made a bid of 10000 tokens or above, corresponding to, on average, 77% of the tokens they had left in their purses. The total cost of protecting all the time for round 2 being 10000 tokens, as indicated in Fig. 7 by the dashed black line, bidding 10000 tokens or more for a second forecast set was clearly pointless. Half of these participants were players to which the yellow river was assigned (the river that experienced the least number of floods in round 1 and for which participants thus ended the first round with on average the highest number of tokens left in their purse) and eight out of these nine participants had a forecast set with a bias during the first round. These nine participants, who paid 10000 tokens or more for the second forecast set, were removed from the subsequent analyses of the auction results, as their bids suggest that they have not understood the stakes of the game.

From Fig. 7, there is a clear positive relationship between the maximum bids within each value of tokens left in purse and the tokens left in purse, as the participants did not disburse more tokens than they had left in their purse during the auction. When we look at the

evolution of the median of the bids with the number of tokens in purse, in general, the more tokens one had left in purse, the higher their WTP for a forecast set. Nonetheless, the WTP seems to have a limit. It can be seen that from a certain number of tokens left in purse, the median value of the bids remains almost constant (in our game case, at about a bid of 6000 tokens for participants with 12000 tokens or more in their purse). The number of tokens that the participants had in hand therefore only influenced to a certain extent their WTP for a second probabilistic forecast set.

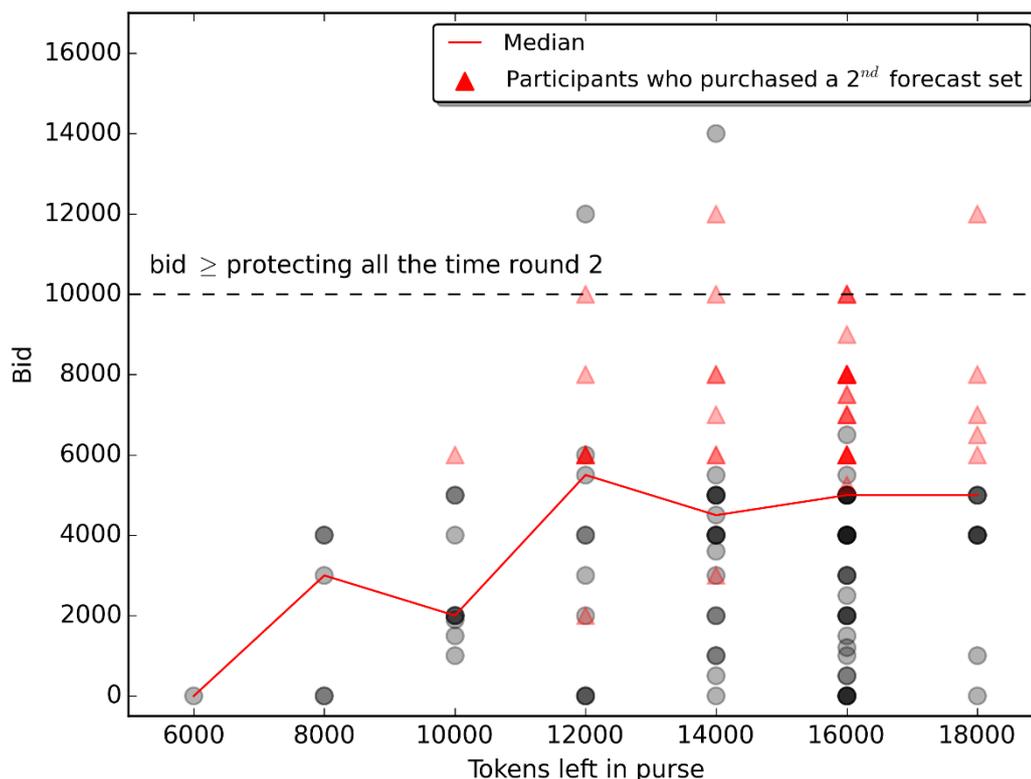


Figure 7. Bids declared by participants to purchase a forecast set for round 2, as a function of the number of tokens they had left in their purse at the end of round 1. The colour of the points indicates the number of participants that fall into a specific bid–tokens left in purse combination.

We also investigated whether the way participants perceived the quality of their forecast set in the first round was a plausible determinant of their WTP for another forecast set to be used in round 2. Figure 8 shows the % bids (i.e. bids expressed as a percentage of the tokens participants had left in their purse at the time of the auction) as a function of the rating participants gave to their forecast set quality in round 1 (from “very bad” to “very good”; see Appendix A). Firstly, it is interesting to observe that three participants judged their first forecast set to have been of “very bad” quality but were nonetheless willing to disburse on average 50% of the tokens they had left in purse. Those bids were however quite low, 4000 tokens on average. Moreover, players who rated their first forecast set from “quite good” to “very good” were on average willing to disburse a larger percentage of their

tokens than candidates who rated their previous forecast set from “quite bad” to “neither good nor bad”. Therefore, the way participants rated the quality of their first forecast set was to a certain degree influential on their WTP for a second forecast set.

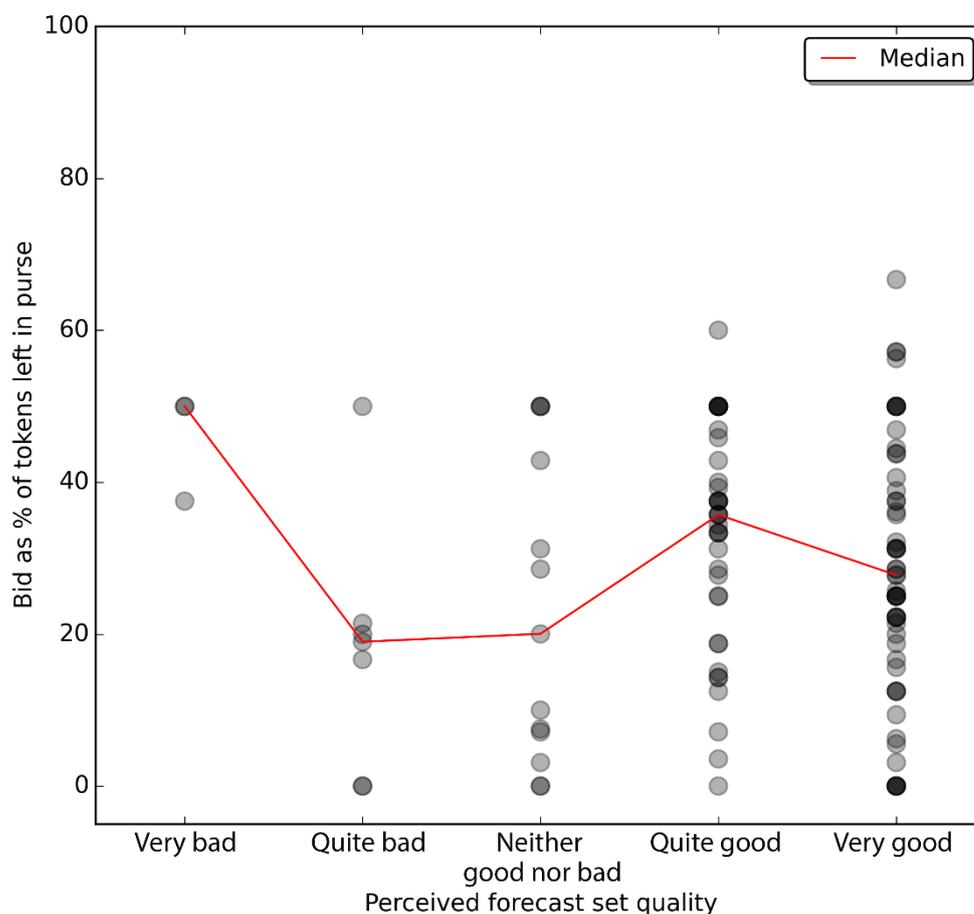


Figure 8. Participants’ % bids, bids expressed as a percentage of the tokens participants had left in their purse at the time of the auction, as a function of the rating they gave to their forecast set quality in round 1 (from “very bad” to “very good”). The colour of the points indicates the number of participants that fall into a specific bid-perceived forecast set quality combination.

During the auction following the closed bids, 44 forecast sets were distributed to the participants who made the highest bids, in order to be used in round 2. Table 4 shows that participants who purchased these second forecast sets were quite well distributed among the different forecast set types of round 1, with a slightly higher frequency of buyers among participants who had played round 1 with unbiased forecasts; 42% of all participants with unbiased forecasts purchased a second forecast set, while 30% (31%) of participants with positively biased (negatively biased) forecasts bought a second forecast set. Buyers also belonged more often to the group assigned river green (48%, or 41% of all green river participants), followed by rivers yellow (32%, or 36% of all yellow river participants) and blue (20%, or 23% of all blue river participants). The higher percentage of green river participants buying a second forecast set could have been due to a combination of the river

green flood frequency in round 1 (not as low as for the yellow river, making it more relevant for green river participants to buy a second forecast set) and of money left in purse (on average, not as low as for the blue river participants). The buyers of the second forecast sets are displayed as red triangles in Fig. 7. We note that these red triangles are not necessarily the highest bid values in the figure, since we plot results from several applications of the game (in one unique application, they would coincide with the highest bids, unless a participant had a high bid but had not raised their hand during the auction to buy a second forecast set). Differences in the highest bids among the applications of the game could be an indication that the size (or type) of the audience might have had an impact on the bids (i.e. the WTP for a probabilistic forecast). Our samples were however not large enough to analyse this aspect.

Table 4. Distribution of the 44 forecast sets sold during the auction, per river (yellow, green and blue) and forecast set type (positively biased, unbiased and negatively biased).

Forecast set type	River			Total
	Yellow	Green	Blue	
Positively biased	5	2	6	13
Unbiased	6	9	3	18
Negatively biased	3	10	0	13
Total	14	21	9	44

Participants who did not purchase a second probabilistic forecast set (85 players in total) stated their reason for doing so. The majority of them (66%, or 56 players) said that the price was too high (which means, in other words, that the bids made by the other participants were too high, preventing them from purchasing a second forecast set during the auction). Ten participants (12%) argued that the model did not seem reliable. Most of these participants were among those who had indeed received a forecast set with a bias in the first round. The rest of the candidates who did not purchase a second forecast set (22%, or 19 players) wrote down on their worksheet the following reasons.

- *Low flood frequency in the first round* – a participant assigned the yellow river wrote: “Climatology seemed probability of flood = 0.2”.
- *Assessment of the value of the forecasts difficult* – a participant wrote: “No information for the initial bidding line”; and another wrote: “Wrong estimation of the costs versus benefits”.
- *Preference for taking risks* – “Gambling” was a reason given by a player.

- *Enough money left in purse to protect all the time during round 2* – which can be an indication of risk-averse behaviour coupled with economic wealth and no worries of false alarms.
- *Not enough money left in purse to bid successfully* – a participant wrote: “The purse is empty due to a lot of floods”.

2.2.3.5 Decisions are better when they are made with the help of unbiased forecasts, compared to having no forecasts at all

The analysis of the results of round 2 allowed us to compare the performance of participants with and without a forecast set. Overall, participants without a second forecast set had an average “true” performance value of 3.1, computed as shown in Eq. (2) and over the five cases of round 2. The best performance was equal to the optimal performance (“true” performance value equal to 1) and the worst performance reached a value of 6. Comparatively, participants with a second forecast set had an average “true” performance of 1.2, thus much closer to the optimal performance than the average performance of participants without a second forecast set. The best performance in this group also equalled the optimal performance, while the worst performance value was 2.5, much lower (i.e. thus much closer to the optimal value) than the worst performance value of participants making their decisions without any forecasts. These numbers clearly indicate that the possession of a forecast set in the second round led to higher performances and to a lower spread in performances within the group of players with a second probabilistic forecast set (compared to players without forecasts in round 2).

Does this conclusion however depend on the participants’ performances in round 1? Do you need to be a good decision-maker to benefit from the forecasts in hand? Our results suggest otherwise. All the participants with a bad performance in the first round and a forecast set in round 2 had a good performance in the second round. This indicates that even if those participants had a bad performance in round 1, they took advantage of the forecasts and had a good performance in round 2. Additionally, 57 out of 59 participants with a good performance in round 1 and no forecasts in round 2 had a bad performance in the second round. This therefore indicates that no matter how well the participants performed in round 1, the possession of a forecast set led to better decisions in round 2.

All the participants without a second forecast set who were assigned the yellow river missed the first two floods in the second round. Some of these participants purchased flood protection for all or some of the subsequent cases, while the others never bought any protection. It could have been due to the low flood frequency of their river in the first round

(see Table 1). This behaviour was not observed for the green river participants without a second forecast set, for which a very diverse sequence of decisions was seen in the second round. As for the blue river participants without any second forecast set, most of them missed the first flood event that occurred in round 2 and, subsequently, purchased flood protection for a few cases where no flood actually occurred. These decision patterns were not observed for participants with a second forecast set within each river, who took more consistently right decisions.

The large majority of participants with a second forecast set in round 2 (41 out of 44) rated their forecasts as either “quite good” or “very good”, which was expected since all the forecasts were unbiased in round 2. The three remaining participants said that their second forecast set was “neither good nor bad” or “quite bad”. These participants all had biased forecasts in the first round and their behaviour during round 2 suggested that they might have been influenced by the bias in their forecasts for round 1.

2.2.3.6 Overall winning strategies would combine good performance with an accurate assessment of the value of the forecasts

The average final purse at the end of round 2 was 3149 tokens (3341 tokens for participants without a second forecast set and 2778 tokens for participants with a second forecast set), remaining from the 20000 tokens initially given to each participant. The minimum final purses observed were zero tokens or less. Twenty-five participants, out of the total of 129 players, finished the game with such amounts of tokens. Out of these 25 participants, 22 had received a biased forecast set in the first round. From the analysis of the game worksheets, we could detect three main losing strategies followed by these 25 participants who finished with zero tokens or less in purse.

- Eighteen participants, most of them blue river players, had an “acceptable to bad” performance in round 1 (performances ranging between 1.3 and 3), did not purchase a second forecast set, and performed badly in round 2 (performances ranging between 2.3 and 6).
- Four players, mostly in charge of the yellow river, had a “good to bad” performance in round 1 (performances ranging between 1 and 3), purchased a second forecast set for 10000 tokens or higher, and performed very well in round 2 (performances of 1).
- Three participants, all green river players, had a “good to acceptable” performance in round 1 (performances ranging between 1 and 1.5), bought a second forecast set

for 6000–8000 tokens, but performed badly in round 2 (performances ranging between 2 and 2.5).

The winners of the game, six players in total, finished round 2 with 8000 or 12000 tokens in their purse. Half of these participants were assigned the green river and the other half the blue river. Apart from one participant, all had received a biased forecast set in the first round. Most participants had a “good to acceptable” performance in the first round (performances ranging between 1 and 1.7), did not purchase any forecast set and had a “good to bad” performance in the second round (performances ranging between 1 and 3). Their performance in round 2 did not lead to large money losses, as it did for yellow river participants, which can be explained by the fact that they did not have so many flood events in this round (see Table 1).

The average “avoided cost”, the average bid for a second forecast set and the average “new bid” are presented in Table 5 for each river. By comparing the “avoided cost” with the average bid for each river, it is noticeable that the average bid was larger than the “avoided cost” of each river. On average participants paid 1000 tokens more for their second forecast set than the benefit, in terms of tokens spared in the second round, that they derived from having this forecast set. This could explain why none of the winners of the game had a forecast set in the second round. From the average “new bid”, it is evident that participants would have liked to pay less on average than what they originally paid for their second forecast set. For all the rivers, the average “new bid” is closer to the “avoided cost” than the average bid of participants during the auction.

Table 5. Average values of “avoided cost” for round 2, average bid for a second forecast set and average “new bid” if forecasts were considered not worth the price originally paid. Values are in tokens and for the three different rivers.

River	Average “avoided cost”	Average bid	Average “new bid”
Yellow	7251	7929	7083
Green	5829	7083	6224
Blue	5711	6889	5875

2.2.4 Discussion

2.2.4.1 Experiment results and implications

It was clear during the game that most participants had used the probabilistic forecasts they were given at the beginning of the game to help them in their decisions. This was an

important issue in our game since it was an essential condition to then be able to evaluate how the participants were using their forecasts and to understand the links between the way they perceived the quality of their forecasts and the way they rated their performance at the end of a round. There was evidence that participants were mostly using the 50th percentile of the forecast distributions, but, interestingly, the median alone could not explain all the decisions made. Other aspects of the game might have also shaped the participants' use of the information, such as the discovery, during the first round, of the forecast set bias (i.e. two out of three forecast sets were purposely biased for round 1). This was also mentioned by some participants at the end of some applications of the game, who said that the fact of noticing the presence of a bias (or suspecting it, since they were not told beforehand that the forecasts were biased) led them to adjust the way they were using the information. This could suggest that forecasts, even biased, can still be useful for decision-making, compared to no forecasts at all, if users are aware of the bias and know how to consider it before making a decision.

Interestingly, in the analysis of the worksheets, there was an indication that the players had, however, different tolerances to the different biases. Indeed, a lower tolerance for under-predictive forecasts than for over-predictive forecasts was identified. Biased forecasts were hence problematic for the users and indicative of the manner in which the information was used. This strongly indicates that there is an important need for probabilistic forecasts to be bias-corrected previously to decision-making, a crucial aspect for applications such as flood forecasting, for instance (Hashino et al., 2007; Pitt, 2008).

There was additionally evidence that, in a few cases, some participants with unbiased forecasts did not use their forecasts (when considering the 50th percentile as key forecast information). The analysis suggested that the players' risk perception, triggered by the initial river level or the proximity of the forecast median to the flood threshold, might have been a reason for this. This led to less consistent actions, where participants based their decisions on extremes of the forecast distribution (other percentiles of the forecast) or on no apparent information contained in the forecast distribution. A similar finding was reported by Kirchhoff et al. (2013) through a case study in America, where it was found that the perception of a risk was a motivational driver of a water manager's use of climate information. There is a constant effort from forecasters to produce and provide state-of-the-art probabilistic forecasts to their users. However, it was seen here that even participants with unbiased forecasts did not always use them. This is an indication that further work needs to be done on fostering communication between forecasters and users, to promote an enhanced use of the information contained in probabilistic forecasts.

From the results, it also appeared that the participants had an accurate perception of their decision-maker performance and related it to the quality of their forecasts. This implies that participants viewed their forecasts as key elements of their decision-making. This result is very encouraging for forecasters and also bears important implications for the real world. It could indeed suggest that decision-makers forget that their own interpretation of the forecasts is as important as the information held in the forecast itself, as there is a myriad of ways to interpret and use probabilistic forecasts for decision-making. The choice of the percentile on which the decisions are based is an example of such an interpretation. This could potentially mean that decision-makers will tend to blame (thank) the forecast providers for their own wrong (good) decisions.

Many papers have shown, through different approaches, the expected benefits of probabilistic forecasts vs. deterministic forecasts for flood warning (e.g. Buizza, 2008; Verkade and Werner, 2011; Pappenberger et al., 2015; Ramos et al., 2013). However, many challenges still exist in the operational use of probabilistic forecasting systems and the optimisation of decision-making. This paper is a contribution to improve our understanding of the way the benefits of probabilistic forecasts are perceived by the decision-makers. It proposes to investigate it from a different perspective, by allowing, through a game experiment, decision-makers to bid for a probabilistic forecast set during an auction. The auction was used in this paper as an attempt to characterise and understand the participants' WTP for a probabilistic forecast in the specific flood protection risk-based experiment designed for this purpose. Our results indicate that the WTP displays dependencies on various aspects.

The bids were to a certain extent influenced by the participants' economic situation. They were on average positively related to the money available to participants during the auction. Nonetheless, this was mainly a factor for participants who had little money left in their purses at the time of the auction. The participants' perceived forecast quality was also a factor influencing their WTP for another forecast set. Players who had played the first round with biased forecasts were less prone to disburse money for another forecast set for the second round. There was moreover an indication that the flood frequency of the river might have influenced the WTP for a forecast set. Some players in charge of a river with only one flood event in the first round (i.e. low flood risk) did not consider beneficial the purchase of a forecast set for the second round. The participants' risk perception was therefore an important element of their WTP for a probabilistic forecast. The more risk-averse participants did not buy a second forecast set as they had enough money to protect all the time; "gambling" was also stated as a reason for not buying a second forecast set.

Seifert et al. (2013) have similarly shown that “the demand for flood insurance is strongly positively related to individual risk perceptions”.

These results show that the perceived benefit of probabilistic forecasts as a support of decision-making in a risk-based context is multifaceted, and varies not only with the quality of the information and its understanding, but also with the relevance and the risk tolerance of the user. This further demonstrates that more work is needed not solely to provide guidance on the use of probabilistic information for decision-making, but also to develop efficient ways to communicate the actual relevance and evaluate the long-term economic benefits of probabilistic forecasts for improved decisions in various applications of probabilistic forecasting systems within the water sector. This could additionally provide insights into bridging the gap between the theoretical or expected benefit of probabilistic forecasts in a risk-based decision-making environment and the perceived benefits by key users.

2.2.4.2 Game limitations and further developments

This paper aimed to depict behaviours in the flood forecasting and protection decision-making context. Although game experiments offer a flexible assessment framework, compared to real operational configurations, it is however extremely complex to search for general explanatory behaviours in such a context. This is partially due to the uniqueness of individuals and the interrelated factors that might influence decisions, which are both aspects that are difficult to evaluate when playing a game with a large audience. A solution to overcome this, as proposed by Crochemore et al. (2015), could be to prolong the game by incorporating a discussion with the audience or with selected individuals, aiming at understanding the motivations hidden underneath their decisions during the game. Having more time available to apply the game would also allow one to play more cases in each round, bringing additional information to the analysis and clarifying key aspects of the game, such as the effect of the bias on the participants’ use of the forecasts and on their WTP for more forecasts. Co-designing such an experiment with social anthropologists could bring to light many more insights into participants’ decision-making behaviours.

Being set up as a game, this study also presents some limitations. As mentioned by Breidert et al. (2006), a source of bias in such studies is their artificial set-up. Indeed, under those circumstances, participants are not directly affected by their decisions, as they neither use their own money nor is the risk a real one. This might lead them to make decisions which they would normally not make in real life or in operational forecasting contexts.

Moreover, in our game, the costs given to both flood protection and flood damages were not chosen to represent the real costs that one encounters in real environments. First, real costs in integrated flood forecasting and protection systems are difficult to assess, given the complexity of flood protection and its consequences. Secondly, the external imposed conditions for playing our game (i.e. the fact that we wanted to play it during oral talks in conferences, workshops or teaching classes, with expected eclectic audiences of variable sizes, having a limited amount of time, and using paper worksheets to be collected at the end of the game for the analysis) were not ideal to handle any controversy on the realism (or absence of realism) of the game scenario.

It is however arguable whether the game results could be a reflection of the experiment set-up, and hence of the parameters of the game (the protection and damage costs, the number of flood events, etc.). For instance, the higher damage costs might have influenced the participants' tolerance to misses and false alarms. Further developments could include testing the influence of the parameters of this experiment on its results as a means of analysing the sensitivity of flood protection mitigation to a specific decision-making setting.

Additionally, the small sample size of this experiment limited the statistical significance of its results. Replicating it could ascertain some of the key points discussed, leading to more substantial conclusions, and improve our understanding of the effect of the professional background of the participants on their decisions.

Finally, the experiment's complex structure was its strength as well as its weakness. When analysing the game results, the chicken and egg situation arose. Several factors of the participants' use of the forecasts and of their WTP for a forecast set were identified, but it was not possible to measure causalities. It would therefore be interesting to carry out further work in this direction, together with behavioural psychologists, by, for instance, testing the established factors separately.

2.2.5 Conclusions

This paper presented the results of a risk-based decision-making game, called "How much are you prepared to pay for a forecast?", played at several workshops and conferences in 2015. It was designed to contribute to the understanding of the role of probabilistic forecasts in decision-making processes and their perceived value by decision-makers for flood protection mitigation.

There were hints that participants' decisions to protect (or not) against floods were made based on the probabilistic forecasts and that the forecast median alone did not account for

all the decisions made. Where participants were presented with biased forecasts, they adjusted the manner in which they were using the information, with an overall lower tolerance for misses than for false alarms. Participants with unbiased forecasts also showed inconsistent decisions, which appeared to be shaped by their risk perception; the initial river level and the proximity of the forecast median to the flood threshold both led the participants to base their decisions on extremes of the forecast distribution or on no apparent information contained in the forecast.

The participants' willingness-to-pay for a probabilistic forecast, in a second round of the game, was furthermore influenced by their economic situation, their perception of the forecasts' quality and the river flood frequency.

Overall, participants had an accurate perception of their decision-making performance, which they related to the quality of their forecasts. However, there appeared to be difficulties in the estimation of the added value of the probabilistic forecasts for decision-making, thus leading the participants who bought a second forecast set to end the game with a lower amount of money in hand.

The use and perceived benefit of probabilistic forecasts as a support of decision-making in a risk-based context is a complex topic. The paper has shown the factors that need to be considered when providing guidance on the use of probabilistic information for decision-making and developing efficient ways to communicate their actual relevance for improved decisions for various applications. Games such as this one are useful tools for better understanding and discussing decision-making among forecasters and stakeholders, as well as highlighting potential factors that influence decision-makers and that deserve further research.

2.2.6 Resources

This version of the game is licensed under CC BY-SA 4.0 (Creative Commons public license). It is part of the activities of HEPEX (Hydrologic Ensemble Prediction Experiment) and is freely available at www.hepex.org. This game was inspired by the Red Cross/Red Crescent Climate Centre game "Paying for Predictions" (<http://www.climatecentre.org/resources-games/paying-for-predictions>).

Acknowledgements. The authors gratefully acknowledge financial support from the Horizon 2020 IMPREX project (grant agreement no. 641811) (project IMPREX: www.imprex.eu). The authors would like to thank the participants of the game who very enthusiastically took part in this experiment. Furthermore, we would like to acknowledge

L. Crochemore, A. Ficchi, C. Poncelet and P. Brigode for their valuable help with the game preparation and worksheet distribution at EGU 2015. Finally, we would like to thank C. Bachofen and everyone who tested and gave suggestions to improve the game during its development.

2.2.7 Appendix

Appendix A. Example of a worksheet distributed to the game participants (here for river blue and the set 1 of positively biased forecasts: BLUE-1).

BLUE-1



How much are you prepared to PAY for a forecast?

Occupation (student, PhD candidate, scientist, operational hydrologist, forecaster, professor, lecturer, other):

.....

How many years of experience do you have? < 5 years 5 to 10 years > 10 years

Flood protection = -2,000 tokens; flood without protection = -4,000 tokens

Flood occurs at 90 or above

Round	Case	River level before rainfall (10-60)	Flood protection?	River level increment (10-80)	River level after increment	Flood? (≥ 90)	Tokens spent	Purse (20,000)
1	1		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	2		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	3		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	4		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	5		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		

- How was your forecast set in Round 1?
 - very bad quite bad neither good nor bad quite good very good I don't know
- How was your performance as a decision-maker in Round 1?
 - very bad quite bad neither good nor bad quite good very good I don't know

Round 2

Do not forget to transfer your final Round 1 purse to Round 2 (in the brackets under 'Purse')

Round	Case	River level before rainfall (10-60)	Flood protection?	River level increment (10-80)	River level after increment	Flood? (≥ 90)	Tokens spent	Purse (.....)
<i>Bid: tokens. Did you buy a probabilistic forecast set? YES / NO</i>								
<i>If yes, deduct the money you paid for it here:</i>								
2	1		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	2		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	3		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	4		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	5		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		

- How was your performance as a decision-maker in Round 2?
 very bad quite bad neither good nor bad quite good very good I don't know
- For the people who DID NOT buy a forecast set:
 - Why didn't you buy a forecast set?
 The model did not seem reliable
 The price was too high
 Other reason (explain):
- For the people who DID buy a forecast set:
 - How was your forecast set in Round 2?
 very bad quite bad neither good nor bad quite good very good I don't know
 - Were the forecasts worth what you paid for them? Yes No
 - If not, how many tokens would you now pay for them? tokens

Please return this worksheet into the envelope and give it to one of the assistants before you leave.
 Thank you for your participation!
 We hope you enjoyed it!

2.2.8 Lessons learnt for improvements in future serious games

The HEPEX serious game described above provided an overview of participants' willingness-to-pay for a probabilistic flood forecast and its underlying factors. However, the game's design and application could benefit from further improvements. The game was designed with many interrelated factors, which effects on the participants' strategies during the game could not be fully disentangled. Future game designs should aim to explore one or two factors at most for more significant findings. The participants' strategies were inferred from the results of this game but could not be verified. Designing and running this serious game with a social scientist or an anthropologist could have shed light on participants' reasoning throughout the game. This was to some extent captured using "empathy maps"

during the IMPREX decision-making activity (Sect. 2.4). Because this game was designed for exploratory purposes, its design was not memorable (unlike the design of the IMPREX online serious game; Sect. 2.3), and it contained no clear take-away message (i.e. the take-away message could have been the fact that probabilistic flood forecasts are valuable for decision-making. However, here, the winner of the game was not always a participant who had purchased a second probabilistic forecast set). Allowing for ample discussion time at the end is a vital aspect of any exploratory serious game, as it allows to explore participants' viewpoints and crystallise the game's key messages. Finally, this serious game was a simplification of reality (e.g. the costs of taking action vs. damage costs). This should be designed with more care in the future as it could have influenced the participants' strategies.

2.3 Stepping in the boots of a flood incident manager: an IMPREX serious game

In this section, we present an online serious game co-created as part of the IMPREX H2020 project. This serious game was developed to engage with a broad non-expert audience (from high school students onwards) to let the player experience first-hand the opportunities and challenges associated with using probabilistic hydrological information for flood early warning. Featuring on the final webpage of the IMPREX project, this serious game serves as a hook, raising awareness about the real-world relevance of research and innovation projects such as IMPREX, and inviting players to find out more about the project after they have played the game.

2.3.1 Overview of the game storyline

In this game, the player is the protagonist of an interactive story, taking on the role of a flood incident commander who has to use 5-day probabilistic flood forecasts to make a series of decisions. The player's goal is to manage a flood forecasting institute and its two teams (a team of forecasters and a flood response team) to ultimately protect a fictitious city against floods. The actions and decisions that the player can make are varied and include: looking at the latest flood forecast, asking forecasters questions about the forecast production, asking the flood response team what the situation looks like on the ground and ultimately, deciding whether to: do nothing, alert the public of an imminent flood and put up flood defences or evacuate the population.

2.3.2 What scientific lessons does the game communicate?

By introducing and guiding the player through the operational world of forecast-based decision-making for flood preparedness, this online serious game shows how state-of-the-art forecast products (available in and outside of IMPREX) can be used for local decision-making. It highlights the opportunities and challenges associated with this by giving an overview of the range of actions and decisions that can be made (based on probabilistic information) ahead of a potential flood event.

While this serious game was developed for the flood risk management sector, the concepts it communicates are relevant to decision-making processes in other water sectors addressed by the IMPREX project (e.g. navigation, reservoir management and agriculture).

2.3.3 In-depth game design

This game was designed collaboratively by IMPREX partners, with the aim to communicate these scientific lessons and guide the player through the operational forecasting chain (from forecast production to decision-making; see Fig. 9 for a flowchart of the game concept). In the game, the player has access to probabilistic flood forecasts, which they can use to make a decision (see Fig. 10, top row). However, the forecasts are not trivial to interpret, and the player is prompted to seek the help of a team of forecasters (forecast room) and flood responders (incident room) to understand, interpret and use the forecasts (see Fig. 10, second and third rows). In the incident room, the player can also find out about the current state of the river and ask for recommendations on what decision they should make (doing nothing, alerting the population, putting up flood defences or evacuating the population; see Fig. 10, fourth row). However, the flood response team is only here to guide the player and no clear answer is given to them as to what decision they should make. Using all the sources of information available to them in the game, the player has to make a decision alone.

In the game, the player's score is a measure of their popularity, which varies with the decisions they make. For example, evacuating the population while no flood occurs will cause a decrease in popularity, reflecting the costs of the evacuation, as well as the loss in public trust given the false alarm. Once in the game, the player can boost their popularity through public engagement (see Fig. 10, bottom row). This was designed to reflect the opportunities and challenges (monetary gains/losses and public trust/distrust) associated with flood forecast-based decision-making and public response.

This serious game was professionally designed into an online game by Arctik (IMPRES partner), to make it aesthetically pleasing and easily accessible by the target audience.

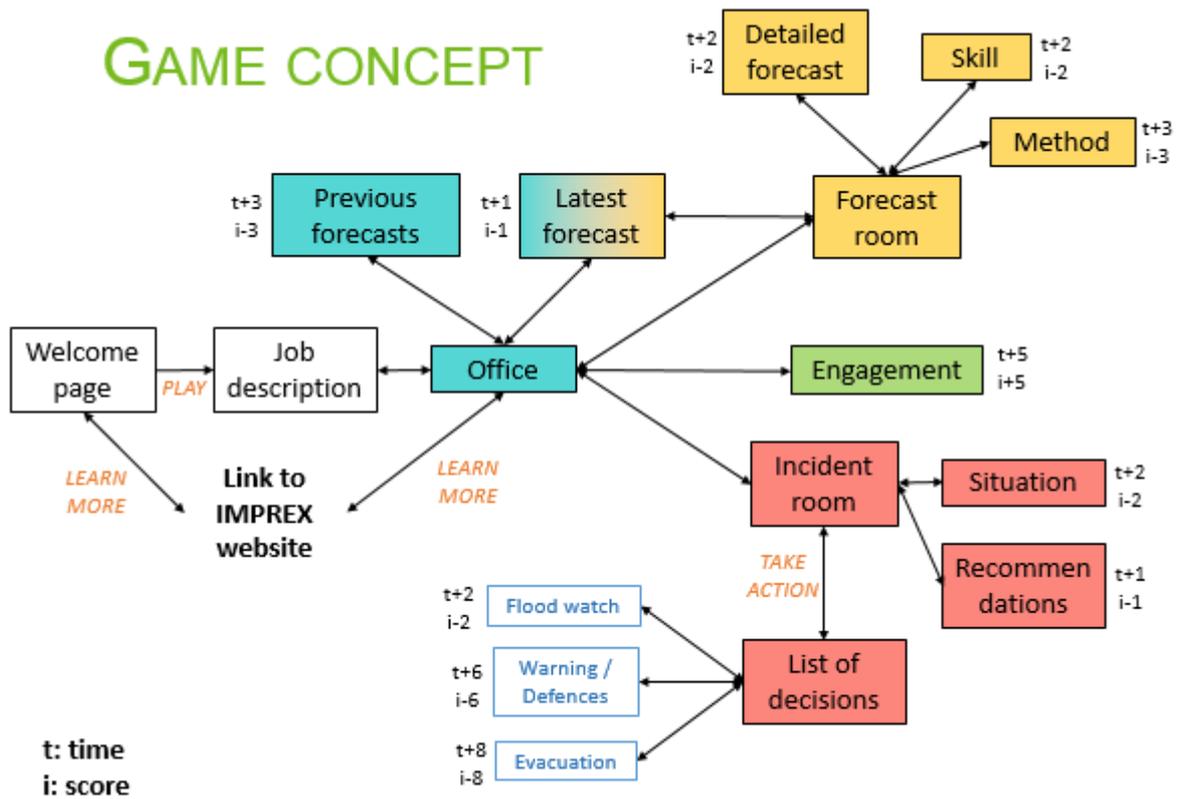
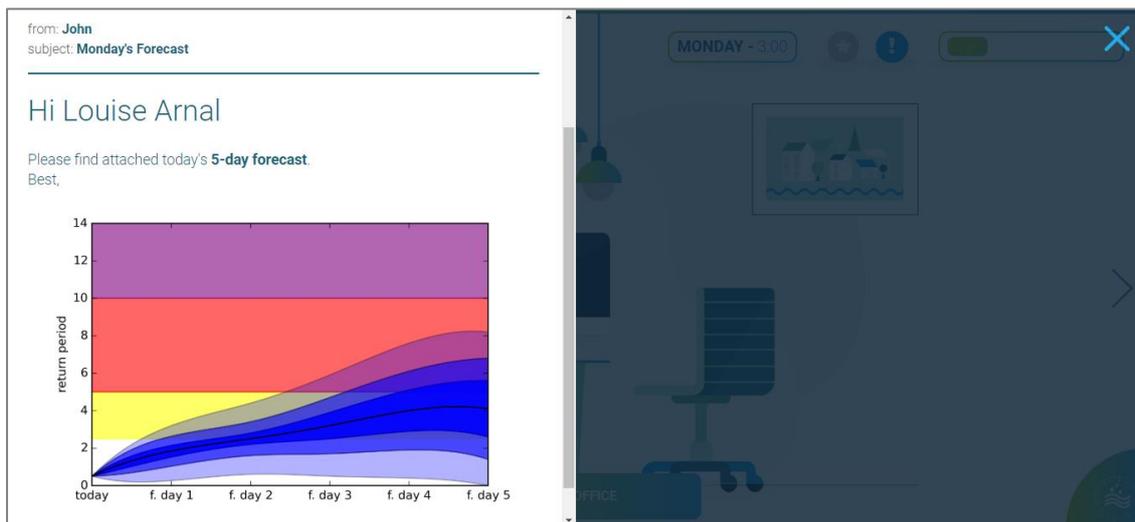
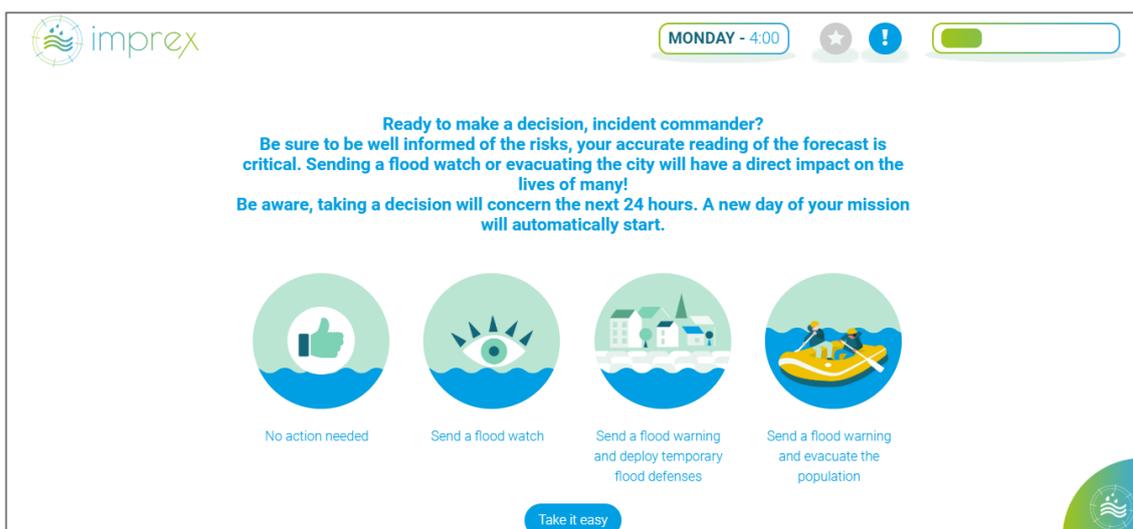
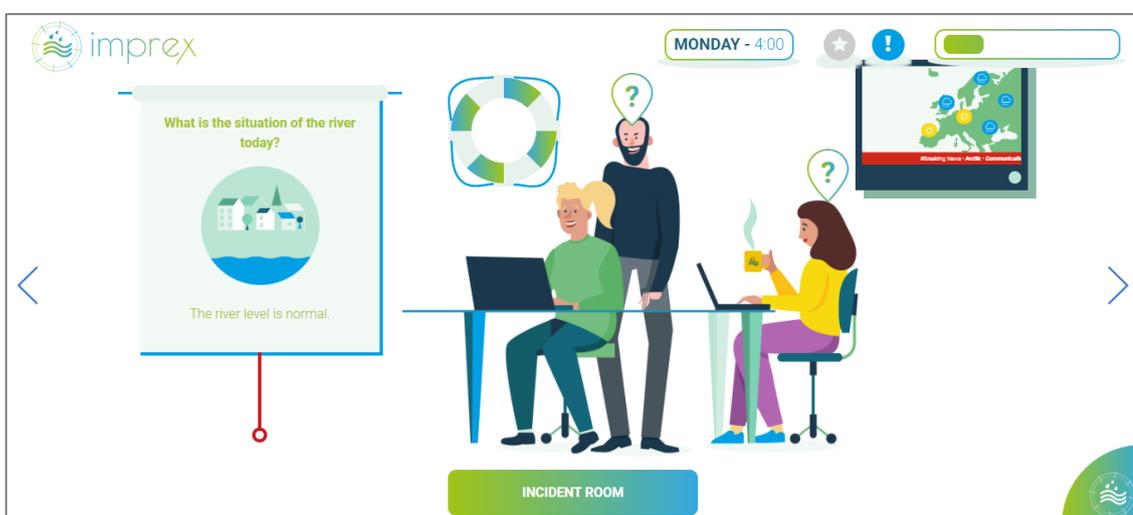
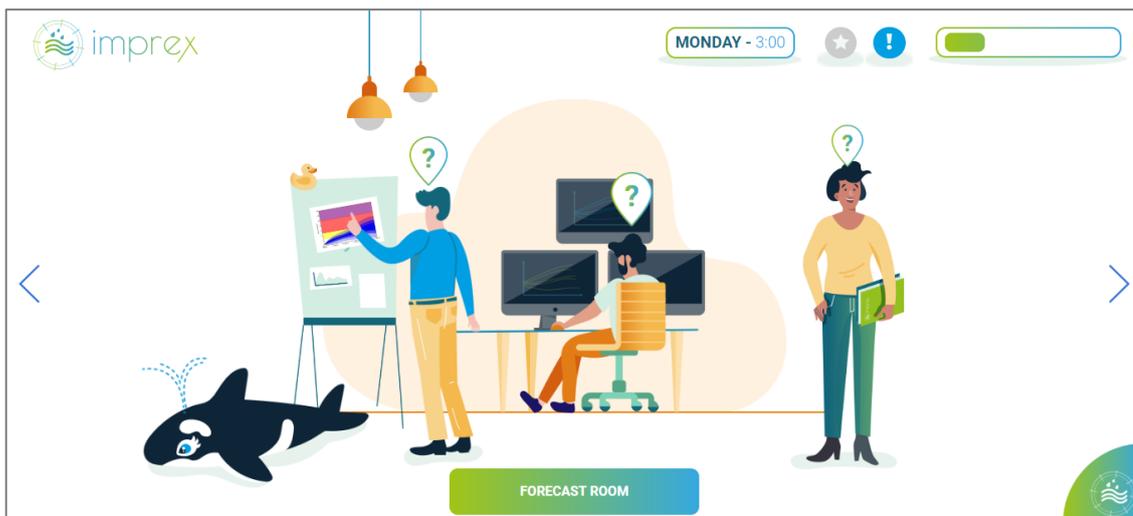


Figure 9. Flowchart of the IMPRES online serious game concept.





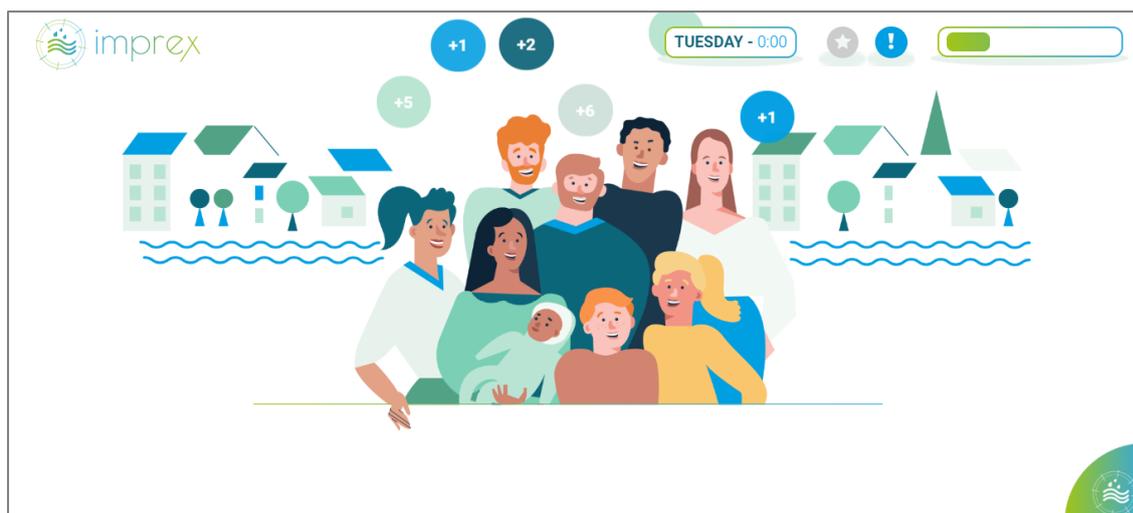


Figure 10. Screenshots of the online IMPREX online serious game. From top to bottom: Example of probabilistic flood forecast available to player, forecast room, incident room, decision-making page and popularity boost page.

2.3.4 Feedback and game release

The beta version of this online serious game was released to a selected audience (a mix of forecasters, decision-makers and non-experts) to obtain some feedback regarding its appearance and the lessons it conveys. This feedback was then incorporated to design the final game version.

This online serious game was released during the European Geoscience Union (EGU) General Assembly 2019. It was furthermore presented at the European Climate Change Adaptation (ECCA) conference 2019 and during the UK Flood Forecasting Centre (FFC) 10th year anniversary event. It is available to play online⁴.

Acknowledgements. The original idea of the game was co-developed by Louise Arnal (ECMWF/University of Reading), Florian Pappenberger (ECMWF), Maria-Helena Ramos (IRSTEA) and Hannah Cloke (University of Reading). It was further enhanced by the contributions from Louise Crochemore (SMHI), Matteo Giuliani (Politecnico di Milano) and Emma Aalbers (KNMI) with website design by Arctik for the IMPREX Risk Outlook. The game is part of a series of games that have been developed within the framework of the international community initiative HEPEX.

⁴ www.imprex.arctik.tech

2.4 Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity

This section introduces a decision-making activity, created for an IMPREX stakeholder focus group for the West Thames (UK) on 6th November 2017. This focus group was co-organised by the University of Reading and the Environment Agency, and supported by the European Centre for Medium-Range Weather Forecasts (ECMWF). The motivation behind the design of this activity was to create a platform to share new scientific research (on the topic of seasonal hydrological forecasting for flood early warning) with decision-makers of the water sector in the West Thames. Additionally, this provided space for discussions with local experts, highlighting potential factors that might influence decision-makers' uptake of probabilistic seasonal hydrological forecasts and that deserve further attention.

This is the summary of a paper, presenting a co-author contribution arising through collaboration during this PhD, and has the following reference:

Neumann, J. L., L. Arnal, R. E. Emerton, H. Griffith, S. Hyslop, S. Theofanidi and H. L. Cloke, 2018b: Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity, *Geosci. Commun.*, 1, 35-57, doi:10.5194/gc-1-35-2018*

L.A. co-designed the decision-making activity, co-organised the set-up of the focus group and took part in delivering it. Additionally, L.A. commented on the manuscript before and during publication.

The published article can be found in the thesis Appendix A2.

Seasonal hydrological forecasts indicate how the river and groundwater levels are expected to change in the coming months. While this is valuable for informing future flood or drought risk and water availability, studies investigating how seasonal hydrological forecasts can be used for decision-making are still limited. A decision-making activity was designed to capture how different water sector users (flood and drought forecasters, water resource managers and groundwater hydrologists in the West Thames, UK) (might) interpret and use probabilistic seasonal hydrological forecasts. It is important to note that while all users present at this focus group were familiar with seasonal hydrological forecasts, they do not all currently use them operationally.

* ©2018. The Authors. Geoscience Communication, a journal of the European Geosciences Union published by Copernicus. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.

For this activity, participants were given three sets of progressively confident and locally tailored seasonal hydrological hindcasts (a combination of operational – e.g. EFAS-Seasonal, see Chapter 3 for more information - and hypothetical products) for a flood event in three months' time (i.e. the actual 2013/14 Thames river floods). The participants were asked to play with their “day-job” hat on and were not told whether the hindcasts represented a flood, drought, or business-as-usual scenario. We observed that participants increased their decision/action choice with more confident and locally tailored information, with clear differences in the choices made across the various water sector users. Forecasters and groundwater hydrologists were most likely to request further information about the situation, inform other organisations and implement actions for preparedness. On the other hand, water resource managers more consistently adopted a “watch and wait” approach. Local knowledge, risk appetite, and experience of previous floods were identified as key factors of the decision-making process.

Further discussions highlighted that forecast uncertainty is not necessarily a barrier to use, if information about forecast quality is shared, but that seasonal hydrological forecasts need to be disseminated at a finer spatial scale to aid local decision-making. Additionally, seasonal hydrological forecast information communicated in a combination of different formats (e.g. maps, texts, hydrographs and tables), tailored to different user groups, was identified as being beneficial for interpretation and use. In summary, “engagement is user-specific and seasonal hydrological forecasts have the potential to be more useful if they could be presented at a scale which matches that employed in decision-making” (Neumann et al., 2018b). A wider conclusion of this focus group was that an improved communication between scientists and users is necessary to ensure that users are kept up to date with scientific developments.

In Chapter 2, insights on factors that shape decision-makers' use of probabilistic forecasts for flood early warning were captured through a HEPEX serious game and an IMPREX decision-making activity.

Results from both activities suggested that probabilistic flood forecasts, even biased or uncertain, can still be useful for decision-making, if users are aware of the forecast quality. This hints that, as scientists and forecast providers, we have a responsibility to understand and adequately communicate the qualities and limitations of the forecasts we provide to users. As stated by Neumann et al. (2018b), "we do not want to be in the position whereby seasonal hydrological forecast skill has improved but the credibility and reliability of the information is questioned by decision-makers who have not been kept up to date with developments."

Chapter 3 addresses our current capabilities in terms of seasonal hydrological forecasting and flood early warning, from the global to the local scale.

Chapter 3

Operational seasonal hydro-meteorological forecasting: from the global to the local scale

3.1 Background and aim

Seasonal hydrological forecasts provide insights into the likely evolution of hydrological variables (e.g. streamflow, soil moisture, groundwater levels, etc.) over the next few months. They have the potential to inform a range of water sectors, from flood and drought early warning to reservoir management, navigation and agriculture. Yet, these forecasts are relatively recent in the world of operational hydro-meteorological forecasting and, as a result, their application still lags behind.

The predictability of seasonal hydrological forecasts originates from two main sources. Notably, the knowledge i) of the initial hydrological states (e.g. initial streamflow, groundwater levels and snowpack, etc.) and ii) of the future meteorological conditions (i.e. temperature, precipitation and evaporation). The seasonal predictability of future meteorological conditions is driven by large-scale climate patterns, such as the El Niño–Southern Oscillation (ENSO), the North Atlantic Oscillation (NAO) and the Pacific–North American (PNA) pattern (Yuan et al., 2015b).

While there exist a number of seasonal hydrological forecasting studies (see Sect. 3.2 for references), operational systems are found in limited numbers. These include: the European Service for Water Indicators in Climate Change Adaptation (SWICCA; Copernicus, 2019), the Australian Government Bureau of Meteorology Seasonal Streamflow Forecasts (Bennett et al., 2016), the United States’ National Hydrologic Ensemble Forecast Service (HEFS; Demargne et al., 2014) and the Hydrological Outlook UK (Prudhomme et al., 2017).

Seasonal hydrological forecasting systems are different in their configurations, reflecting the wide range of climates and landscapes (and hence hydro-meteorological processes and predictability sources) for which they were designed. Their configuration additionally echoes the amount of resources (in terms of data, time and money) available for their development and (operational) maintenance. There exist two main branches of seasonal hydrological forecasting methods: statistical and physically-based (or model-based). Statistical seasonal hydrological forecasts are produced by regressing the hydrological

variable of interest on a range of predictors (e.g. historical meteorological observations, large-scale climate indices and teleconnection patterns, and initial hydrological conditions) (Robertson and Wang, 2012; Mendoza et al., 2017). Model-based seasonal hydrological forecasts are produced by forcing a hydrological model with meteorological data. These meteorological data can for example be historical meteorological observations (this forecasting method is referred to as ESP, which nowadays stands for Ensemble Streamflow Prediction) (Twedt et al., 1977; Day, 1985). Or they can be seasonal meteorological forecasts (interchangeably referred to as “seasonal climate forecasts” in this thesis), output from a general circulation model (Mo and Lettenmaier, 2014; Bennett et al., 2016; Greuell et al., 2018; Wanders et al., 2019). Hybrid statistical and model-based seasonal hydrological forecasting methods are a more recent endeavour (Slater et al., 2018).

Statistical seasonal hydrological forecasting methods are relatively fast to run. However, they are not always reliable as the stationary nature of their statistical relationships can misrepresent changing temporal patterns in the seasonality of the hydrological variable of interest and shifts in the relationship between that variable and its predictors over time (Slater et al., 2018; for example, as can be expected from climate change). While model-based seasonal hydrological forecasts can capture these shifting temporal patterns and physically represent coupled atmosphere-ocean-land interactions, they are more resource-intensive. Additionally, until recently, their development and application has been hindered by limits in the seasonal meteorological forecast predictability, especially over Europe (Arribas et al., 2010; Doblas-Reyes et al., 2013).

Yet, recent years have seen unprecedented improvements in seasonal meteorological forecasting (Scaife et al., 2014; MacLachlan, 2015; ECMWF, 2017b), encouraging the development of model-based seasonal hydro-meteorological forecasting systems. Given their design, these systems reflect the latest scientific findings in the field of seasonal hydro-meteorological predictability, as well as technical advances in hydrological modelling and numerical weather prediction (including data assimilation and post-processing methods). As such, evaluating the performance of model-based seasonal hydro-meteorological forecasts can bring us closer to a complete understanding of state-of-the-art scientific and technical capabilities in seasonal hydro-meteorological forecasting.

In the context of a changing climate, hydro-meteorological patterns and the occurrence of extreme hydro-meteorological events are changing (IPCC, 2014), rendering seasonal hydro-meteorological forecasting vital for decision-making. Indeed, understanding and

improving the predictability of extreme hydro-meteorological events on the seasonal to interannual timescale is a stepping stone to climate change adaptation.

In this chapter, we present two recently operational model-based seasonal hydro-meteorological forecasting systems: EFAS-Seasonal and GloFAS-Seasonal. These forecasting systems were developed as part of the Copernicus Emergency Management Service (EMS) European and Global Flood Awareness Systems (EFAS and GloFAS, respectively). In the first section of this chapter, EFAS-Seasonal is assessed in terms of its overall performance, as well as its potential usefulness for predicting lower and higher than normal streamflows months in advance in Europe. The second section introduces the EFAS-Seasonal and GloFAS-Seasonal operational outlooks and gives a brief overview of the performance of GloFAS-Seasonal on the global scale. The third section demonstrates the potential usefulness of EFAS-Seasonal for predicting the 2013/14 Thames basin floods. The aim of this chapter is to assess the current capabilities in terms of seasonal streamflow forecasting on the global, continental and basin scales, with a particular focus on flood early warning.

3.2 Skilful seasonal forecasts of streamflow over Europe?

This section has been published in Hydrology and Earth System Sciences (HESS) with the following reference:

Arnal, L., H. L. Cloke, E. Stephens, F. Wetterhall, C. Prudhomme, J. Neumann, B. Krzeminski and F. Pappenberger, 2018: Skilful seasonal forecasts of streamflow over Europe?, *Hydrol. Earth Syst. Sci.*, 22, 2057-2072, doi:10.5194/hess-22-2057-2018*

The contributions of the authors of this paper are as follows: H. L. Cloke (supervisor: academic), E. Stephens (supervisor: academic), F. Wetterhall (collaborator: ECMWF), C. Prudhomme (collaborator: ECMWF), J. Neumann (collaborator: academic), B. Krzeminski (collaborator: ECMWF) and F. Pappenberger (collaborator: ECMWF). F.W., B.K. and F.P. produced the seasonal streamflow hindcast dataset. L.A. conceived and posed the research questions, designed and carried out the analysis, and wrote the paper. L.A., H.L.C., E.S., F.W., C.P. and F.P. interpreted the results. H.L.C., E.S., F.W., C.P., J.N. and F.P. commented on the manuscript. Overall, 100% of the research and 85% of the writing was undertaken by L.A.

* ©2018. The Authors. Hydrology and Earth System Sciences, a journal of the European Geosciences Union published by Copernicus. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.

The published article can be found in the thesis Appendix A3. An overview of EFAS-Seasonal was additionally written by L.A. and published as part of a book chapter written by Hirpa et al. (2018).

Abstract. This paper considers whether there is any added value in using seasonal climate forecasts instead of historical meteorological observations for forecasting streamflow on seasonal timescales over Europe. A Europe-wide analysis of the skill of the newly operational EFAS (European Flood Awareness System) seasonal streamflow forecasts (produced by forcing the Lisflood model with the ECMWF System 4 seasonal climate forecasts), benchmarked against the ensemble streamflow prediction (ESP) forecasting approach (produced by forcing the Lisflood model with historical meteorological observations), is undertaken. The results suggest that, on average, the System 4 seasonal climate forecasts improve the streamflow predictability over historical meteorological observations for the first month of lead time only (in terms of hindcast accuracy, sharpness and overall performance). However, the predictability varies in space and time and is greater in winter and autumn. Parts of Europe additionally exhibit a longer predictability, up to 7 months of lead time, for certain months within a season. In terms of hindcast reliability, the EFAS seasonal streamflow hindcasts are on average less skilful than the ESP for all lead times. The results also highlight the potential usefulness of the EFAS seasonal streamflow forecasts for decision-making (measured in terms of the hindcast discrimination for the lower and upper terciles of the simulated streamflow). Although the ESP is the most potentially useful forecasting approach in Europe, the EFAS seasonal streamflow forecasts appear more potentially useful than the ESP in some regions and for certain seasons, especially in winter for almost 40% of Europe. Patterns in the EFAS seasonal streamflow hindcast skill are however not mirrored in the System 4 seasonal climate hindcasts, hinting at the need for a better understanding of the link between hydrological and meteorological variables on seasonal timescales, with the aim of improving climate-model-based seasonal streamflow forecasting.

3.2.1 Introduction

Seasonal streamflow forecasts predict the likelihood of a difference from normal conditions in the following months. Unlike forecasts at shorter timescales, which aim to predict individual events, seasonal streamflow forecasts aim at predicting long-term (i.e. weekly to seasonal) averages. The predictability in seasonal streamflow forecasts is driven by two components of the Earth system, the initial hydrological conditions (IHC; i.e. of snowpack, soil moisture, streamflow and reservoir levels, etc.) and large-scale climate patterns, such

as the El Niño–Southern Oscillation (ENSO), the North Atlantic Oscillation (NAO), the Pacific–North American (PNA) pattern and the Indian Ocean Dipole (IOD) (Yuan et al., 2015b).

The first seasonal streamflow forecasting method, based on a regression technique developed around 1910–1911 in the United States, harnessed the predictability from accurate IHC of snowpacks to derive streamflow for the following summer (Church, 1935). This statistical method recognized antecedent hydrological conditions and land surface memory as key drivers of streamflow generation for the following months.

Alongside the physical understanding of streamflow generation processes came technical developments, such as the creation of the first hydrological models and the acquisition of longer observed meteorological time series, which led to the creation of the first operational model-based seasonal streamflow forecasting system. This system, called extended streamflow prediction (ESP; i.e. note that ESP nowadays stands for ensemble streamflow prediction, although it refers to the same forecasting method), was developed by the United States National Weather Service (NWS) in the 1970s (Twedt et al., 1977; Day, 1985). The ESP forecasts are produced by forcing a hydrological model, initialized with the current IHC, with the observed historical meteorological time series available. The output is an ensemble streamflow forecast (where each year of historical data is a streamflow trace) for the following season(s) (Twedt et al., 1977; Day, 1985). The quality of the ESP forecasts can be high in basins where the IHC dominate the surface hydrological cycle for several months (the exact forecast quality depending on the time of year and the basin’s physiographic characteristics; Wood and Lettenmaier, 2008).

In basins where the meteorological forcings drive the predictability, however, the lack of information on the future climate is a limitation of the ESP forecasting method and might result in unskilful ESP forecasts. This drawback led to the investigation of the use of seasonal climate forecasts, in place of the historical meteorological inputs, to feed hydrological models and extend the predictability of hydrological variables on seasonal timescales (Pagano and Garen, 2006). This investigation was made possible by technical and scientific advances. Scientifically, seasonal climate forecasts were improved greatly by the understanding of ocean–atmosphere–land interactions and the identification of large-scale climate patterns as drivers of the hydro-meteorological predictability (Goddard et al., 2001; Troccoli, 2010). This was technically implementable with the increase in computing resources, making it possible to run dynamical coupled ocean–atmosphere–land general circulation models on the global scale at high spatial and temporal resolutions (Doblas-

Reyes et al., 2013). An additional technical challenge, the coarse spatial resolution of seasonal climate forecasts compared to the finer resolution of hydrological models, had to be addressed. To tackle this issue, many authors have explored different ways of downscaling climate variables for hydrological applications (Maraun et al., 2010, and references therein).

While climate-model-based seasonal streamflow forecasting experiments are more common outside of Europe, for example for the United States (Wood et al., 2002, 2005; Mo and Lettenmaier, 2014), Australia (Bennett et al., 2016), or Africa (Yuan et al., 2013a), they remain limited in Europe, with a few examples in France (Céron et al., 2010; Singla et al., 2012; Crochemore et al., 2016), in central Europe (Demirel et al., 2015; Meißner et al., 2017), in the United Kingdom (Bell et al., 2017; Prudhomme et al., 2017) and at the global scale (Yuan et al., 2015a; Candogan Yossef et al., 2017). This is because, although the quality of seasonal climate forecasts has increased over the past decades, there remains limited skill in seasonal climate forecasts for the extra-tropics, particularly for the variables of interest for hydrology, notably precipitation and temperature (Arribas et al., 2010; Doblas-Reyes et al., 2013).

In Europe, the NAO is one of the strongest predictability sources of seasonal climate forecasts; it is associated with changes in the surface westerlies over the North Atlantic and Europe, and hence with changes in temperature and precipitation patterns over Europe (Hurrell, 1995; Hurrell and Van Loon, 1997). It was shown to affect streamflow predictability, especially during winter (Dettinger and Diaz, 2000; Bierkens and van Beek, 2009; Steirou et al., 2017), in addition to the IHC and the land surface memory. It was furthermore shown to be an indicator of flood damage and occurrence in parts of Europe (Guimarães Nobre et al., 2017).

As the quality and usefulness of seasonal streamflow forecasts increase, their usability for decision-making has lagged behind. Translating the quality of a forecast into an added value for decision-making and incorporating new forecasting products into established decision-making chains are not easy tasks. This has been explored for many water-related applications, such as navigation (Meißner et al., 2017), reservoir management (Viel et al., 2016; Turner et al., 2017), drought-risk management (Sheffield et al., 2013; Yuan et al., 2013a; Crochemore et al., 2017), irrigation (Chiew et al., 2003; Li et al., 2017), water resource management (Schepen et al., 2016) and hydropower (Hamlet et al., 2002), but seasonal streamflow forecasts have yet to be adopted by the flood preparedness community.

The European Flood Awareness System (EFAS) is at the forefront of seasonal streamflow forecasting, with one of the first operational pan-European seasonal hydrological forecasting systems. The aim of this paper is to bridge the current gap in pan-European climate-model-based seasonal streamflow forecasting studies. Firstly, the setup of the newly operational EFAS climate-based seasonal streamflow forecasting system is presented. A Europe-wide analysis of the skill of this forecasting system compared to the ESP forecasting approach is then presented, in order to identify whether there is any added value in using seasonal climate forecasts instead of historical meteorological observations for forecasting streamflow on seasonal timescales over Europe. Subsequently, the potential usefulness of the EFAS seasonal streamflow forecasts for decision-making is assessed.

3.2.2 Data and methods

3.2.2.1 EFAS hydrological simulation and seasonal hindcasts

The data used in this paper include a streamflow simulation and two seasonal streamflow hindcasts (Fig. 1). Further information on these datasets is given below.

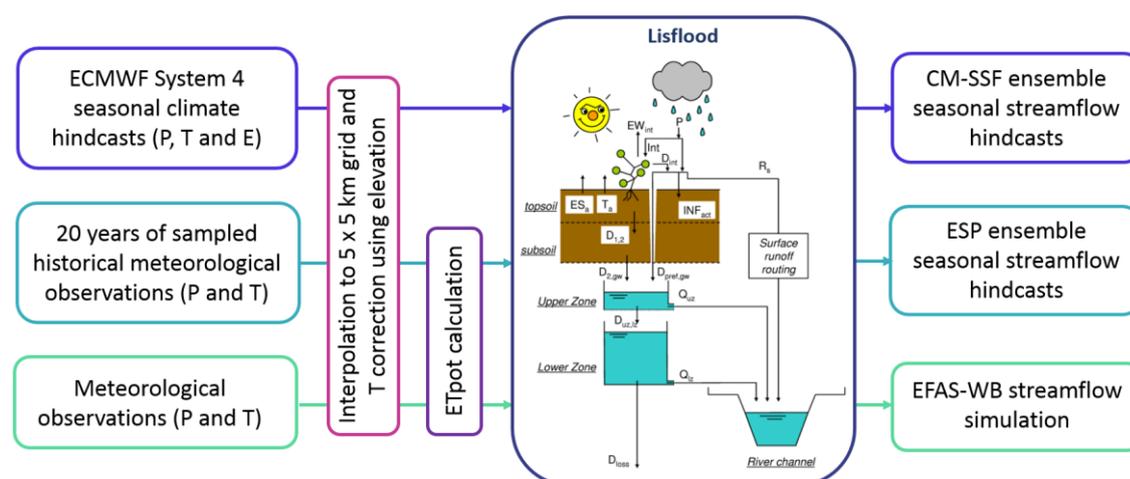


Figure 1. Schematic of the EFAS-WB streamflow simulation and of the CM-SSF and ESP seasonal streamflow hindcast generation, where P is precipitation, T is temperature, E is evaporation and ETpot is potential evapotranspiration. The Lisflood model diagram was taken from Burek et al. (2013).

3.2.2.2 Hydrological modelling and streamflow simulation

The Lisflood model was used to produce all the simulations and hindcasts used in this paper. Lisflood is a GIS-based hydrological rainfall–runoff–routing distributed model written in the PCRaster Dynamic Modelling Language, which enables it to use spatially distributed maps (i.e. both static and dynamic) as input (De Roo et al., 2000; Van Der Knijff et al., 2010). The Lisflood model was calibrated to produce pan-European parameter maps. The calibration was performed for 693 basins from 1994 to 2002 using the Standard

Particle Swarm Optimisation 2011 (SPSO-2011) algorithm. The calibration was carried out for parameters controlling snowmelt, infiltration, preferential bypass flow through the soil matrix, percolation to the lower groundwater zone, percolation to deeper groundwater zones, residence times in the soil and subsurface reservoirs, river routing and reservoir operations for a few basins. The results were validated with the Nash–Sutcliffe efficiency (NSE) for the validation period 2003–2012. In validation (calibration), Lisflood obtained a median NSE of 0.57 (0.62). Basins with large discrepancies between the observed and simulated flow statistics were situated mainly on the Iberian Peninsula and on the Baltic coasts (see Zajac et al., 2013, and Smith et al., 2016, for further details).

The Lisflood model is run operationally in EFAS, with the simulation domain covering Europe at a 5×5 km resolution. A reference simulation, called the EFAS water balance (EFAS-WB), is available on a daily time step starting from February 1990. Lisflood simulates the hydrological processes within a basin (most of which are mentioned above), starting from the previous day's IHC (e.g. snow cover, storage in the upper and lower zones, soil moisture, initial streamflow, reservoir filling) and forced with the most recent observed meteorological fields (i.e. of precipitation, potential evapotranspiration and temperature; provided by the EFAS meteorological data collection centres). The observed meteorological fields are daily maps of spatially interpolated point measurements of precipitation (from more than 6000 stations) and temperature (from more than 4000 stations) at the surface level. These same data are used to produce interpolated potential evapotranspiration maps from the Penman–Monteith method (Alfieri et al., 2014). All meteorological variables are interpolated on a 5×5 km grid using an inverse distance weighting scheme and the temperature is first corrected using the elevation (Smith et al., 2016).

The EFAS-WB is the best estimate of the hydrological state at a given time and for a given grid point in EFAS and is thus used as initial conditions from which the seasonal hydrological forecasts are started.

3.2.2.3 Ensemble seasonal streamflow hindcasts

In this paper, two types of ensemble seasonal streamflow hindcasts are used: the ensemble streamflow prediction (ESP) hindcast (hereafter referred to as ESP) and the System 4-driven seasonal streamflow hindcast (hereafter referred to as CM-SSF (climate-model-based seasonal streamflow forecast)), following the notation from Yuan et al. (2015b)).

They are both initialized from the EFAS-WB, on the first day of each month, to produce a new ensemble streamflow forecast up to a lead time of 7 months (215 days), with a daily

time step. Both hindcasts are generated from February 1990 for the same European domain as the EFAS-WB, at the same 5×5 km resolution. The unique difference between the ESP and the CM-SSF is the meteorological forcing used to drive the hydrological model, described below.

The ESP is produced by driving the Lisflood model with 20 (the number of years of data available at the time the hindcast was produced) randomly sampled years of historical meteorological observations (i.e. the same as the meteorological observations used to produce the EFAS-WB, excluding the year of meteorological observations corresponding to the year that is being forecasted). A new 20-member ESP is thus generated at the beginning of each month and for the next 7 months.

The CM-SSF is produced by driving the Lisflood model with the ECMWF System 4 seasonal climate hindcast (Sys4, i.e. of precipitation, evaporation and temperature). Sys4 has a spatial horizontal resolution of about 0.7° (approximately 70 km). It is re-gridded to the Lisflood spatial resolution using an inverse distance weighting scheme and the temperature is first corrected using the elevation. Sys4 is made up of 15 ensemble members, extended to 51 every 3 months (Molteni et al., 2011). From 2011 onwards the Sys4 forecasts were run in real time and all contained 51 ensemble members. A new 15- to 51-member CM-SSF is hence produced at the beginning of each month and for the next 7 months. Operationally, the CM-SSF forecasts are currently used in EFAS to generate a seasonal streamflow outlook for Europe at the beginning of every month.

3.2.3 Hindcast evaluation strategy

For this study, monthly region specific discharge averages of the hindcasts (CM-SSF and ESP) and EFAS-WB were used. The specific discharge is the discharge per unit area of an upstream basin. For this paper, the gridded daily specific discharge was calculated by dividing the gridded daily discharge output maps (of the hindcasts and the EFAS-WB) by the Lisflood gridded upstream area static map. Subsequently, the gridded daily specific discharge maps were used to calculate daily region averaged specific discharges (for each region in Fig. 2) by summing up the daily specific discharge values of each grid cell within a region, divided by the number of grid cells in that region. Finally, monthly specific discharge region averages were calculated for each calendar month.

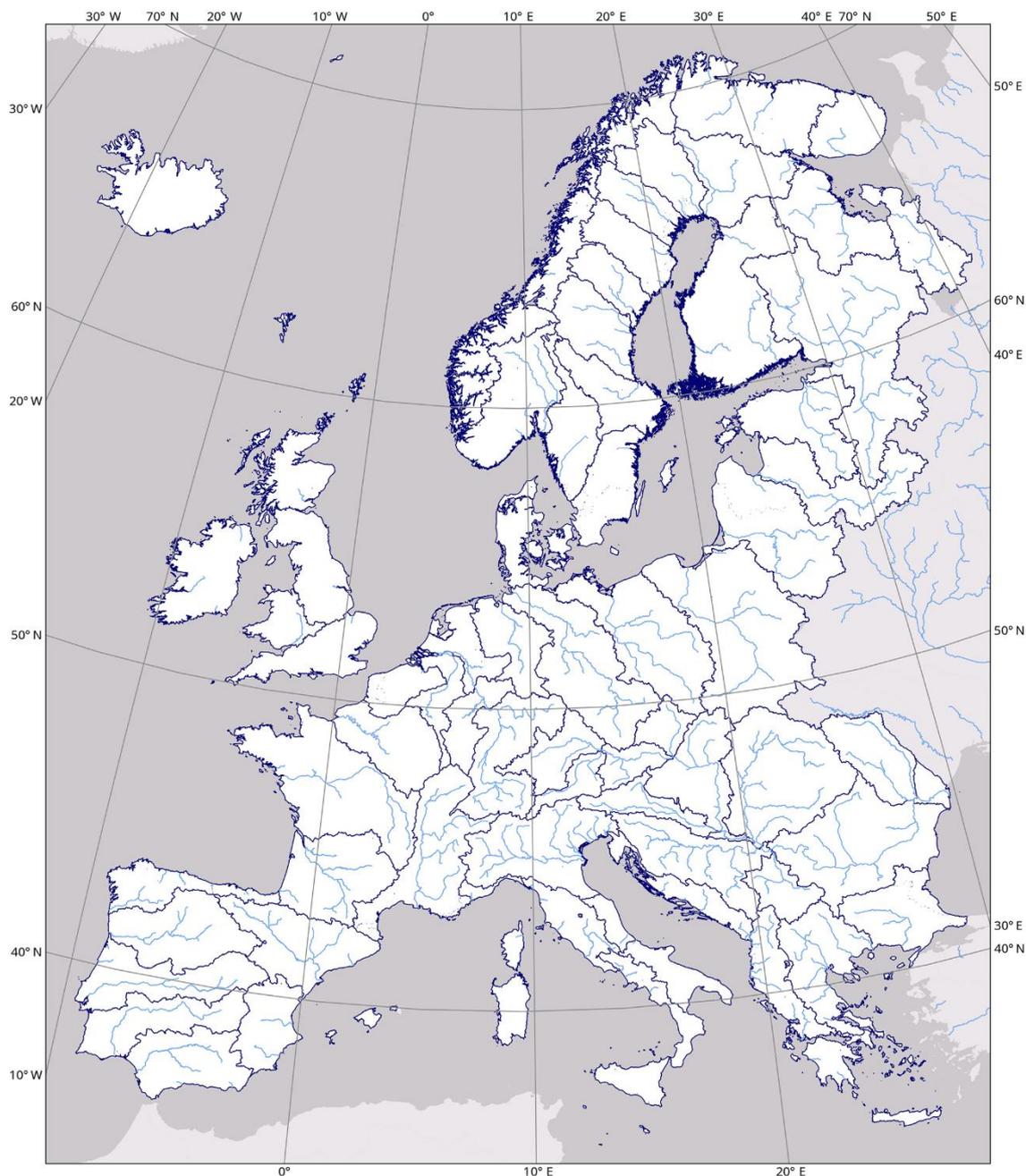


Figure 2. Map of the 74 European regions (dark blue outlines) selected for the analysis of the CM-SSF and the ESP.

The regions displayed in Fig. 2 were created by merging several basins together (basins used operationally in EFAS for the shorter timescale forecasts), while respecting hydroclimatic boundaries. They were chosen for the analysis presented in this paper for two main reasons. Firstly, they are the regions used operationally to display the EFAS seasonal streamflow outlook. Secondly, they were created in order to capture large-scale variability in the weather.

The analysis of the hindcasts was performed on monthly specific discharge (hereafter referred to as streamflow) region averages for hindcast starting dates spanning February

1990 to November 2016 (included; approximately 27 years of data), with 1 to 7 months of lead time. In this paper, 1 month of lead time refers to the first month of the forecast (e.g. the January 2017 streamflow for a forecast made on 1 January 2017). Two months of lead time is the second month of the forecast (e.g. the February 2017 streamflow for a forecast made on 1 January 2017), etc. Monthly averages were selected for the analysis presented in this paper as it is a valuable aggregation time step for decision-makers for many water-related applications (as shown in the literature for applications such as, for example, navigation (Meißner et al., 2017), reservoir management (Viel et al., 2016; Turner et al., 2017), drought-risk management (Yuan et al., 2013a), irrigation (Chiew et al., 2003; Li et al., 2017) and hydropower (Hamlet et al., 2002)).

Several verification scores were selected in order to assess the hindcasts' quality. These verification scores were chosen to cover a wide range of hindcast attributes (i.e. accuracy, sharpness, reliability, overall performance and discrimination). All of these verification scores, except for the verification score selected to look at hindcast discrimination, are the same as chosen in Crochemore et al. (2016), and are described below. The EFAS-WB streamflow simulations were used as a proxy for observation against which the seasonal streamflow hindcasts were evaluated, hence minimizing the impact of model errors on the hindcasts' quality.

3.2.3.1 Hindcast accuracy

Both hindcasts (CM-SSF and ESP) were assessed in terms of their accuracy, the magnitude of the errors between the hindcast ensemble mean and the "truth" (i.e. the EFAS-WB). For this purpose, the mean absolute error (MAE) was calculated for each region, target month (i.e. the month that is being forecast) and lead time (i.e. 1 to 7 months). The lower the MAE, the more accurate the hindcast.

3.2.3.2 Hindcast sharpness

Both hindcasts were also assessed in terms of their sharpness, an attribute of the hindcast only, which is a measure of the spread of the ensemble members of a hindcast. In this paper, the 90% interquartile range (IQR; i.e. the difference between the 95th and 5th percentiles of the hindcast distribution) was calculated for each region, target month and lead time. The lower the IQR, the sharper the hindcast.

3.2.3.3 Hindcast reliability

Both hindcasts were additionally assessed in terms of their reliability, the statistical consistency between the hindcast probabilities and the observed frequencies. For this purpose, the probability integral transform (PIT) diagram was calculated for each region, target month and lead time (Gneiting et al., 2007). The PIT diagram is the cumulative distribution of the PIT values as a function of the PIT values. The PIT values measure where the “truth” (i.e. EFAS-WB) falls relative to the percentiles of the hindcast distribution. For a perfectly reliable hindcast, the “truth” should fall uniformly in each percentile of the hindcast distribution, giving a PIT diagram that falls exactly on the 1-to-1 diagonal. A hindcast that systematically under- (over-) predicts the “truth” will have a PIT diagram below (above) the diagonal. A hindcast that is too narrow (i.e. underdispersive; hindcast distribution smaller than the distribution of the observations) (large (i.e. overdispersive; hindcast distribution greater than the distribution of the observations)) will have a transposed S-shaped (S-shaped) PIT diagram (Laio and Tamea, 2007).

In order to compare the reliability across all regions, target months and lead times, the area between the PIT diagram and the 1-to-1 diagonal was computed for all PIT diagrams (Renard et al., 2010). The smaller this area, the more reliable the hindcast.

Furthermore, to disentangle the causes of poor reliability, the spread and bias of the hindcasts were calculated for all PIT diagrams, using two measures first introduced by Keller and Hense (2011): β -score and β -bias, respectively. By definition, a perfectly reliable hindcast (with regards to its spread) will have a β -score of zero (to which a tolerance interval of ± 0.09 was added), whereas a hindcast that is too narrow (large) will have a negative (positive) β -score (outside of the tolerance interval). A perfectly reliable hindcast (with regards to its bias) will have a β -bias of zero (to which a tolerance interval of ± 0.09 was added), whereas a hindcast that systematically under- (over-) predicts the “truth” will have a negative (positive) β -bias (outside of the tolerance interval).

3.2.3.4 Hindcast overall performance

The hindcasts were furthermore assessed in terms of their overall performance from the continuous rank probability score (CRPS), calculated for each region, target month and lead time (Hersbach, 2000). The CRPS is a measure of the difference between the hindcast and the observed (i.e. EFAS-WB) cumulative distribution functions. The lower the CRPS, the better the overall performance of the hindcast.

In this paper, the skill of the CM-SSF is benchmarked with respect to the ESP in order to identify whether there is any added value in using Sys4 instead of historical meteorological

observations for forecasting the streamflow on seasonal timescales over Europe. To this end, skill scores were calculated for the MAE, IQR, PIT diagram area and CRPS, using the following equation:

$$\text{Skill score} = 1 - \frac{\text{SCORE}_{\text{CM-SSF}}}{\text{SCORE}_{\text{ESP}}} \quad (1)$$

Skill scores were calculated for each region, target month and lead time and will be referred to as MAESS, IQRSS, PITSS and CRPSS, respectively. Skill scores larger (smaller) than zero indicate more (less) skill in the CM-SSF compared to the ESP. A skill score of zero means that the CM-SSF is as skilful as the ESP. Note that as the ESP is not a “naive” forecast, using it as a benchmark might lead to lower skill than benchmarking the CM-SSF against, for example, climatology.

3.2.3.5 Hindcast potential usefulness

For decision-making, the ability of a seasonal forecasting system to predict the right category of an event (e.g. above or below normal conditions) months ahead is of great importance (Gobena and Gan, 2010). In this paper, the potential usefulness of the CM-SSF and the ESP to forecast lower and higher than normal streamflow conditions within their hindcasts is assessed.

To do so, the relative operating characteristic (ROC) score, a measure of hindcast discrimination (Mason and Graham, 1999), was calculated. The thresholds selected to calculate the ROC are the lower and upper terciles of the EFAS-WB climatology for each season. They were calculated for the simulation period (February 1990 to May 2017), by grouping together EFAS-WB monthly streamflows for each month falling in a season (SON: September–October–November, DJF: December–January–February, MAM: March–April–May and JJA: June–July–August). For each season and each region a lower and upper tercile streamflow value was obtained, subsequently used as thresholds against which to calculate the probability of detection (POD) and the false alarm rate (FAR; with 0.1 probability bins) for both hindcasts, and for each region, season and lead time. Finally, the area under the ROC curve, i.e. the ROC score, was calculated for both hindcasts, for each region, season and lead time. The ROC score ranges from 0 to 1, with a perfect score of 1. A hindcast with a ROC score ≤ 0.5 is unskilful, i.e. less good than the long-term average climatology which has a ROC of 0.5, and is therefore not useful.

Because the ROC score was calculated from a low number of events (i.e. approximately 27 years \times 3 months in each season \times 1/3 (lower or upper tercile) = 27 simulated events), the

hindcasts were judged skilful and useful when their ROC score ≥ 0.6 instead of 0.5. Moreover, the CM-SSF was categorized as more useful than the ESP when the CM-SSF's ROC score was at least 10% larger than the ESP's ROC score.

3.2.4 Results

3.2.4.1 Overall skill of the CM-SSF

In the first part of the results, the skill of the CM-SSF (benchmarked with respect to the ESP) is presented, in terms of the accuracy (MAESS), sharpness (IQRSS), reliability (PITSS) and overall performance (CRPSS) in the hindcast datasets. This will benchmark the added value of using Sys4 against the use of historical meteorological observations for forecasting the streamflow on seasonal timescales over Europe.

As shown by the MAESS boxplots (Fig. 3), the CM-SSF appears on average more accurate than the ESP for the first month of lead time only, for all seasons excluding spring (MAM). Beyond 1 month of lead time, the CM-SSF becomes on average as or less accurate than the ESP. There are however noticeable differences between the different seasons. The CM-SSF shows the largest improvements in the average accuracy compared to the ESP in winter (DJF) and for the first month of lead time. For longer lead times (i.e. 2 to 7 months), the accuracy of the CM-SSF is on average quite similar to that of the ESP in autumn (SON) and winter, and on average lower in spring and summer (JJA). The boxplots for the CRPSS look very similar to the MAESS boxplots, the main difference being the lower average scores for 2 to 7 months of lead time in autumn and winter (Fig. 3).

The boxplots of the IQRSS show that the CM-SSF predictions are on average as sharp as those of the ESP for the first month of lead time (slightly sharper in autumn; Fig. 3). For 2 to 7 months of lead time, in autumn and winter, the CM-SSF predictions are on average sharper than those of the ESP, whereas in spring and summer, the CM-SSF predictions are on average slightly less sharp than the ESP predictions.

As shown by the boxplots of the PITSS (Fig. 3), the CM-SSF predictions are less reliable than the ESP predictions for all seasons and months of lead time. For the first month of lead time and all seasons, 10–20% of the ESP hindcasts and less than 5% of the CM-SSF hindcasts are reliable (Fig. 4; n.b. the percentages represent the portion of hindcasts meeting a certain criteria, with the total number of hindcasts in a given season represented by the hindcasts for all target months falling in this season and for all 74 European regions). About 40–60% of the ESP hindcasts are not reliable for the first month of lead time and all seasons due to the ensemble spread. Approximately half of these hindcasts are too large, while the other

half (slightly more in autumn and winter) are too narrow. Furthermore, 50–80% of the ESP hindcasts under-predict the simulated streamflow for the first month of lead time and all seasons. The percentage of reliable (unreliable) ESP hindcasts increases (decreases) with lead time, as the effect of the IHC fades away. About 70–90% of the CM-SSF hindcasts are too narrow for the first month of lead time and all seasons. With increasing lead time, the percentage of CM-SSF hindcasts that are too narrow (large) decreases (increases), especially in spring. Approximately 40–50% of the CM-SSF hindcasts over-predict the simulated streamflow in spring and summer for the first month of lead time (and increasingly over-predict with longer lead times). In autumn and winter, about 70% of the CM-SSF hindcasts under-predict the simulated streamflow for the first month of lead time (and increasingly under-predict with longer lead times).

For all verification scores, the boxplots for autumn and winter are slightly smaller than for spring and summer, hinting at a smaller variability in the verification scores amongst regions and target months in autumn and winter than in spring and summer. Furthermore, the presence of the boxplots above the zero line (i.e. no skill line) for all lead times suggests that the CM-SSF is more skilful than the ESP for some regions and target months, beyond the first month of lead time.

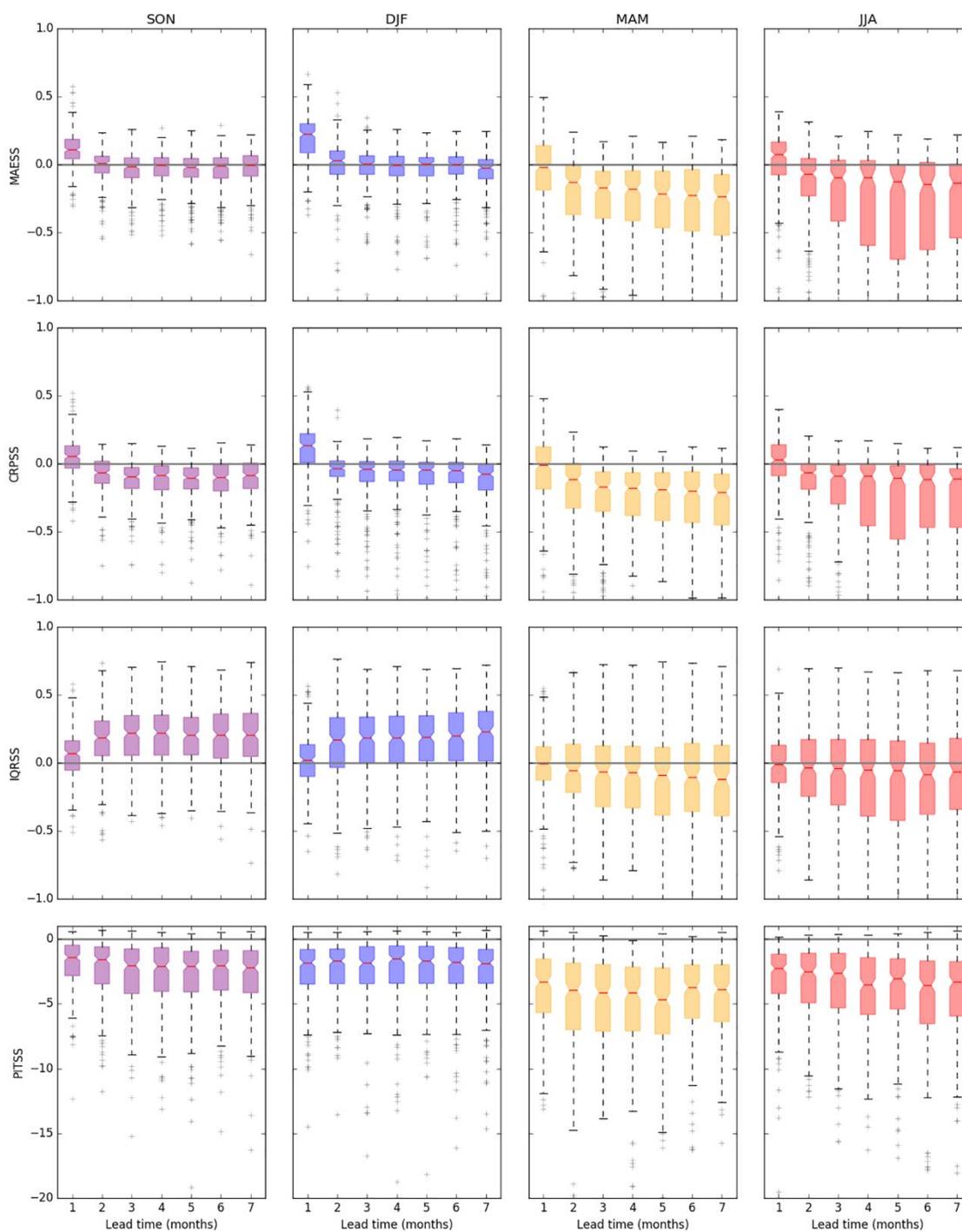


Figure 3. Boxplots of the MAESS, CRPSS, IQRSS and PITSS (from the top to bottom rows) for all four seasons (SON, DJF, MAM and JJA from the left-most to right-most columns) as a function of lead time (i.e. 1 to 7 months). The boxplots contain the scores for all target months falling in a given season and all 74 European regions. For all scores, values larger (smaller) than zero indicate that the CM-SSF is more (less) skilful than the ESP (benchmark). Where the skill is zero, the CM-SSF is as skilful as the ESP for the hindcast period. Note that the PITSS plots have a different y-axis scale.

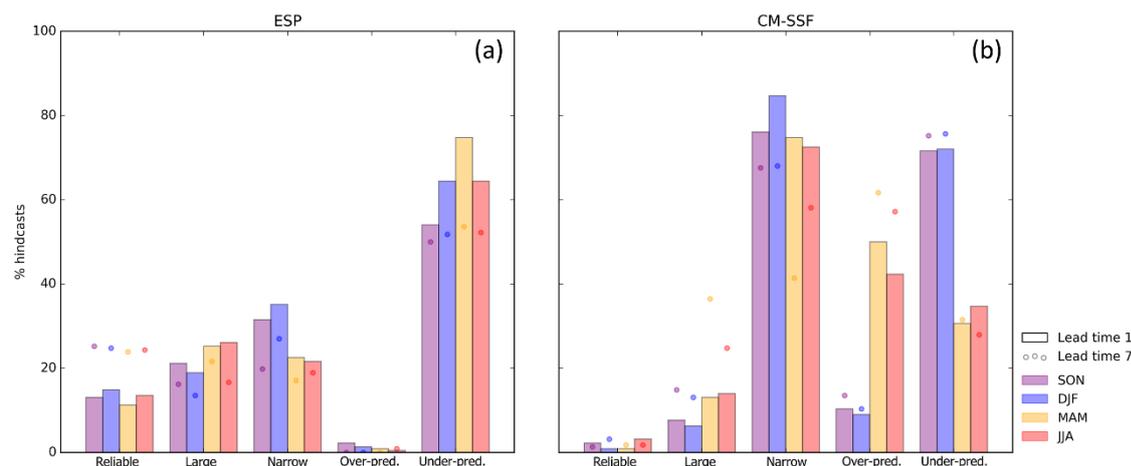


Figure 4. Plots of the percentage of the ESP (a) and the CM-SSF (b) hindcasts falling into each reliability category (reliable – in terms of both spread and bias, too large, too narrow, over-predicting and under-predicting) for all four seasons (SON, DJF, MAM and JJA from the left-most to right-most bars in each reliability category). The results are shown as bar charts for the first month of lead time and as circles for the seventh month of lead time. These lead times were selected for display to highlight the evolution of reliability between the first and last months of the hindcast. The percentages were calculated from hindcasts for all target months falling in a given season and all 74 European regions.

3.2.4.2 Potential usefulness of the CM-SSF

In the second part of the results, the potential usefulness of the CM-SSF compared to the ESP is described for decision-making. Here, potential usefulness is defined as the ability of the forecasting systems to predict lower or higher streamflows than normal, as measured with the ROC score.

Generally, either of the two forecasting systems (CM-SSF or ESP) is capable of predicting skilfully whether the streamflow will be anomalously low or high in the coming months (Fig. 5). However, for a few seasons and regions, none of the two forecasting systems is skilful at predicting lower and/or higher streamflows than normal. This is especially noticeable in winter.

For most seasons and regions, the ESP is more skilful than the CM-SSF at predicting lower and higher streamflows than normal. However, in winter for most regions and during other seasons for several regions, the CM-SSF appears more skilful than the ESP. Regions where the CM-SSF best predicts lower and higher streamflows than normal at most lead times are summarized in Table 1 for all four seasons and the lower and upper terciles of the simulated streamflow.

Table 1. Regions where the CM-SSF is more skilful than the ESP at predicting anomalously low (lower tercile; first column) or high (upper tercile; second column) streamflows for all four seasons (SON, DJF, MAM and JJA from the top to bottom rows). This is a summary of the information displayed in Fig. 5.

	Lower tercile	Upper tercile
SON	<ul style="list-style-type: none"> – Few regions in Fennoscandia – Po River basin (northern Italy) – Elbe River basin (south of Denmark) – Upstream of the Rhine River basin – Upstream of the Danube River basin – Duero River basin (Iberian Peninsula) 	<ul style="list-style-type: none"> – Few regions in Fennoscandia – Iceland – Parts of the Danube River basin – Segura River basin (Iberian Peninsula)
DJF	<ul style="list-style-type: none"> Many regions except – in most of Fennoscandia north of the Baltic Sea, – parts of central Europe. 	Same as lower tercile.
MAM	<ul style="list-style-type: none"> – Few regions on the Iberian Peninsula – Few regions in the western part of central Europe 	Same as lower tercile.
JJA	<ul style="list-style-type: none"> – Few regions in the United Kingdom (UK) – Ireland – North-western edge of the Iberian Peninsula – Regions in Fennoscandia around the Baltic Sea – Regions south of the North Sea 	<ul style="list-style-type: none"> – Northern part of the UK – Ireland – North-western edge of the Iberian Peninsula – Regions in Fennoscandia around the Baltic Sea – Around the Elbe River basin – Upstream of the Danube River basin – Along the Adriatic Sea in Italy

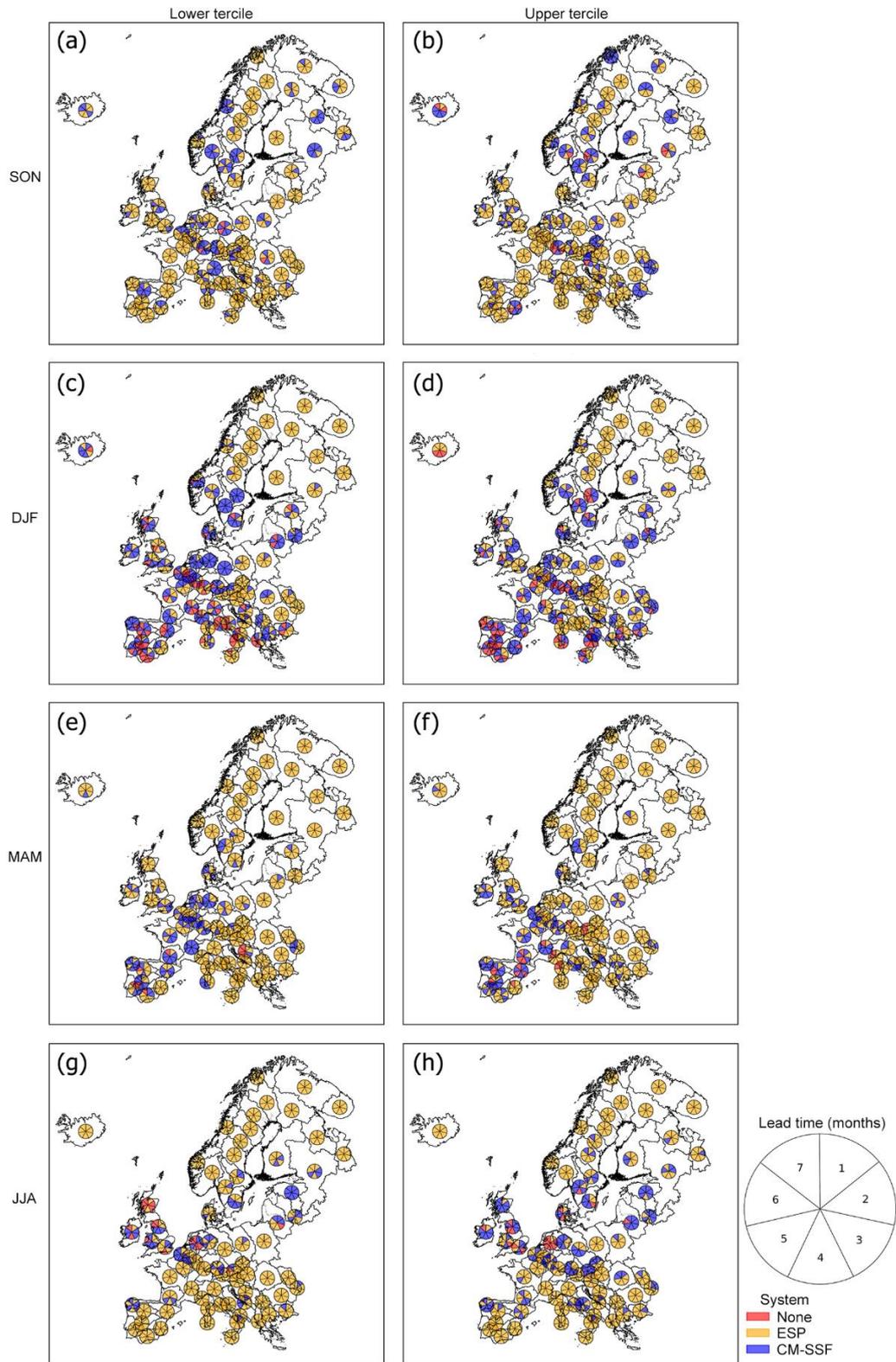


Figure 5. Maps of the best system (as measured with the ROC score) for all four seasons (SON, DJF, MAM and JJA) and the lower and upper simulated streamflow seasonal terciles (left-most and right-most columns, respectively) in each region from (a) to (h). The pie charts display the best system for each lead time (i.e. 1 to 7 months), as shown in the example pie chart on the bottom right of this figure. There are three possible cases: (1) neither the ESP nor the CM-SSF is skilful (red colours), (2) the ESP is skilful and better than the CM-SSF (yellow colours), and (3) the CM-SSF is skilful and better than the ESP (blue colours).

3.2.5 Discussion

3.2.5.1 Does seasonal climate information improve the predictability of seasonal streamflow forecasts over Europe?

On average over Europe and across all seasons, the CM-SSF is skilful (in terms of hindcast accuracy, sharpness and overall performance, using the ESP as a benchmark) for the first month of lead time only. This means that, on average, Sys4 improves the predictability over historical meteorological information for pan-European seasonal streamflow forecasting for the first month of lead time only. At longer lead times, historical meteorological information becomes as good as or better than Sys4 for seasonal streamflow forecasting over Europe. Crochemore et al. (2016) and Meißner et al. (2017) similarly found positive skill in the seasonal streamflow forecast (Sys4 forced hydrological model compared to an ESP) for the first month of lead time, after which the skill faded away for basins in France and central Europe, respectively. Additionally, on average over Europe and across all seasons, the CM-SSF is less reliable than the ESP for all lead times. This is due to a combination of too narrow and biased CM-SSF hindcasts, where the bias depends on the season that is being forecasted. As mentioned in Sect. 3.2.2, the ESP is not a “naive” benchmark, which might partially explain the limited predictability gained from Sys4.

The predictability varies per season and the CM-SSF predictions are on average sharper than and as accurate as the ESP predictions in autumn and winter beyond the first month of lead time (and increasingly sharper with longer lead times). The CM-SSF however tends to systematically under-predict the autumn and winter simulated streamflow (and increasingly under-predicts with longer lead times). In spring and summer, the CM-SSF predictions are on average less sharp and less accurate than the ESP predictions, and they tend to systematically over-predict the simulated streamflow (and increasingly over-predict with longer lead times).

The added predictability gained from Sys4 was shown to lead to skilful CM-SSF predictions of lower and higher streamflows than normal for specific seasons and regions. The CM-SSF is more skilful at predicting anomalously low and high streamflows than the ESP in certain seasons and regions, and noticeably in winter in almost 40% of the European regions, mostly clustered in rainfall-dominated areas of western and central Europe. Several authors have discussed the higher winter predictability over (parts of) Europe, with examples in basins in France (Crochemore et al., 2016), central Europe (Steirou et al., 2017), the UK (Bell et al., 2017) and the Iberian Peninsula (Lorenzo-Lacruz et al., 2011). Bierkens and van Beek (2009) additionally showed that there was a higher winter predictability in

Scandinavia, the Iberian Peninsula and around the Black Sea. Our results are mostly consistent with these findings, except for Scandinavia, where the ESP is more skilful than the CM-SSF in winter. Bierkens and van Beek (2009) produced the seasonal streamflow forecast analysed in their paper by forcing a hydrological model with resampled years of historical meteorological information based on their winter NAO index. However, Sys4 has difficulties in forecasting the NAO over Europe (Kim et al., 2012), which could have led to these inconsistent results with the ones presented by Bierkens and van Beek (2009).

In spring, the CM-SSF is more skilful than the ESP at predicting lower and higher streamflows than normal beyond 1 month of lead time in approximately 15% of the European regions, and mostly in regions of western Europe. This could be due to a persistence of the skill from the previous winter through the land surface memory (i.e. groundwater-driven streamflow or snowmelt-driven streamflow), as highlighted by Bierkens and van Beek (2009) for Europe, Singla et al. (2012) for parts of France, Lorenzo-Lacruz et al. (2011) for the Iberian Peninsula and Meißner et al. (2017) for the Rhine. Moreover, it could be that most of the gained predictability occurs in March, a transition month between the more predictable winter (as mentioned above) and spring, as discussed by Steirou et al. (2017). The ESP is overall more skilful than the CM-SSF at predicting the spring streamflow in snow-dominated regions (e.g. most of Fennoscandia and parts of central and eastern Europe). This hints at the importance of the IHC (i.e. of snowpack) and the land surface memory for forecasting the spring streamflow in snow-dominated regions in Europe.

The added predictability from Sys4 for forecasting lower and higher streamflows than normal is limited in summer and autumn for most regions. The CM-SSF is more skilful at predicting anomalously low and high streamflows than the ESP in about 10–20% of the European regions during those seasons. Other studies have found similar patterns for (parts of) Europe; these include less skill in summer than in winter overall for basins in France (Crochemore et al., 2016), less skill for the low flow season (July to October) for basins in central Europe (Meißner et al., 2017), negative correlations in summer and autumn seasonal streamflow forecasts in central Europe as the influence of the winter NAO fades away (Steirou et al., 2017), and less skill overall in summer than in winter in Europe (Bierkens and van Beek, 2009). The lower CM-SSF skill for predicting lower and higher streamflows than normal in summer could additionally be due to the convective storms in summer over Europe, which are hard to predict, and to the fact that it is the dry season in most of Europe, where rivers are groundwater fed. Therefore, in this season, the quality of the IHC controls the streamflow predictability.

While the CM-SSF is most skilful (in terms of hindcast accuracy, sharpness and overall performance, using the ESP as a benchmark) in autumn and winter and most potentially useful in winter, this does not appear to correlate with high performance in the Sys4 precipitation and temperature hindcasts (as seen on the maps of correlation for Sys4 precipitation and temperature for all four seasons (SON, DJF, MAM and JJA) and with 2 months of lead time (as identified in this paper); available at https://meteoswiss.shinyapps.io/skill_metrics/, Forecast skill metrics, 2017). Over Europe, the Sys4 precipitation and temperature hindcasts are the most skilful in summer and the least skilful in autumn and winter. Moreover, the regions of high CM-SSF skill for predicting lower and upper streamflows than normal do not clearly correspond to regions of high performance in the Sys4 precipitation and temperature hindcasts. Weisheimer and Palmer (2014) looked at the Sys4 hindcasts reliability, and categorised temperature hindcasts as “marginally useful” in winter, with improved forecasts over Europe in summer. They categorised precipitation hindcasts as mostly “marginally useful”, with “not useful” forecasts for dry winters over Northern Europe and “dangerously useless” to “not useful” forecasts for dry summers. These reliability patterns do not appear to clearly correlate with patterns observed in the CM-SSF usefulness. Overall, differences between the streamflow and the meteorological predictability could be partially induced by the different benchmark used to evaluate the skill of the CM-SSF (i.e. the ESP) compared to the one used to look at the performance of the Sys4 precipitation and temperature hindcasts (i.e. ERA-Interim for both sources). However, these results clearly indicate that looking at the performance of the Sys4 precipitation and temperature hindcasts only does not give a good indication of the skill and potential usefulness of the seasonal streamflow hindcasts over Europe, and that marginal performance in seasonal climate forecasts can translate through to more predictable seasonal streamflow forecasts, and vice versa. The added predictability in the CM-SSF could be due to the combined predictability in the precipitation and temperature hindcasts, as well as a lag in the predictability from the land surface memory.

In most regions and for most seasons, at least one of the two forecasting systems (CM-SSF or ESP) is able to predict lower or higher streamflows than normal. However, in winter, the number of regions and lead times for which none of the forecasting systems are skilful increases. This could be because in winter, many regions experience weather-driven high streamflows and the performance of Sys4 is limited at this time of year (as mentioned above). In those regions, the seasonal streamflow forecasts could be improved either by improving the IHC, through for example data assimilation, or by improving the seasonal climate forecasts.

Overall, the ESP appears very skilful at forecasting lower or higher streamflows than normal, showing the importance of IHC and the land surface memory for seasonal streamflow forecasting (Wood and Lettenmaier, 2008; Bierkens and van Beek, 2009; Yuan et al., 2015b).

3.2.5.2 What is the potential usefulness and usability of the EFAS seasonal streamflow forecasts for flood preparedness?

What appears like little added skill does not necessarily mean no skill for the forecast users and can in fact be a large added value for decision-making (Viel et al., 2016). The ability of a seasonal streamflow forecasting system to predict the right category of an event months ahead is valuable for many water-related applications (e.g. navigation, reservoir management, drought-risk management, irrigation, water resource management, hydropower and flood preparedness). From the results presented in this paper, it appears that either of the two forecasting systems (CM-SSF or ESP) is capable of predicting lower or higher streamflows than normal months in advance, thanks to the predictability gained from the IHC, the land surface memory and the seasonal climate hindcast in some regions and for certain seasons.

However, as highlighted by White et al. (2017), there is currently a gap between usefulness and usability of seasonal information. What is a useful scientific finding does not automatically translate into usable information which will fit into any user's decision-making chain (Soares and Dessai, 2016). While several authors have already investigated the usability of seasonal streamflow forecasts for applications such as navigation (Meißner et al., 2017), reservoir management (Viel et al., 2016; Turner et al., 2017), drought-risk management (Sheffield et al., 2013; Yuan et al., 2013a; Crochemore et al., 2017), irrigation (Chiew et al., 2003; Li et al., 2017), water resource management (Schepen et al., 2016) and hydropower (Hamlet et al., 2002), its application to flood preparedness is still left mostly unexplored. One exception being Neumann et al. (2018a), who look at the use of the CM-SSF to predict the 2013/14 Thames basin floods. This is partially due to the complex nature of flood generating mechanisms, still poorly studied on seasonal timescales beyond snowmelt-driven spring floods, as well as the fact that seasonal forecasts reflect the likelihood of abnormal seasonal streamflow totals, but without much skilful information on the exact timing, location and severity of the impact of individual flood events within that season. Coughlan de Perez et al. (2017) looked at the usefulness of seasonal rainfall forecasts for flood preparedness in Africa and highlighted the complexities behind using these forecasts as a proxy for floodiness (for a discussion on floodiness, see Stephens et al.,

2015). Furthermore, decision-makers in the navigation, reservoir management, drought-risk management, irrigation, water resource management and hydropower sectors are familiar with working on long timescales (i.e. several weeks to months ahead). In contrast, the flood preparedness community is currently mostly used to working on timescales of hours to a couple of days.

The Red Cross Red Crescent Climate Centre has recently designed a new approach that harnesses the usefulness of seasonal climate information for decision-making for disaster management. This approach, called “Ready-Set-Go!”, is made up of three stages. The “Ready” stage is based on seasonal forecasts, where they are used as monitoring information to drive contingency planning (e.g. volunteer training). The “Set” stage is triggered by sub-seasonal forecasts, used as early-warning information to alert volunteers. Finally, the “Go!” stage is based on short-range forecasts and consists in the evacuation of people and the distribution of aid (White et al., 2017). Using a similar approach, seasonal streamflow forecasts could complement existing forecasts at shorter timescales and provide monitoring and early-warning information for flood preparedness. Such an approach however requires the use of consistent forecasts from short to seasonal timescales. In this context, moving to seamless forecasting is becoming vital (Wetterhall and Di Giuseppe, 2018).

Soares and Dessai (2016) also identified the accessibility to the information, enhanced by collaborations and ongoing relationships between users and producers, as a key enabler of the usability of seasonal information. International projects, such as the Horizon 2020 IMPREX (IMproving PRedictions and management of hydrological EXtremes) project (van den Hurk et al., 2016), alongside promoting scientific progress on hydrological extremes forecasting from short to seasonal timescales over Europe, gather together forecasters and decision-makers and can effectively demonstrate the added value of the integration of seasonal information in decision-making chains. The Hydrologic Ensemble Prediction EXperiment (HEPEX) is another international initiative that brings together researchers and practitioners in the field of ensemble prediction for water-related applications. It is an ideal environment for collaboration and fosters communication and outreach on topics such as the usefulness and usability of seasonal information for decision-making.

3.2.5.3 Aspects for future work

In this paper, terciles of the simulated streamflow are used. However, and because the application of the EFAS seasonal streamflow forecasts is of particular relevance for flood preparedness, the evaluation of the hindcasts for lower and higher streamflow extremes

(for example the 5th and 95th percentiles, respectively) would be more relevant and might give very different results. This was not done in this paper as the time period covered by the seasonal streamflow hindcasts (i.e. approximately 27 years) was not long enough for statistically reliable results for lower and higher streamflow extremes. The limited hindcast length is a common problem in seasonal predictability studies. Increasing the hindcast length back in time could lead to more stable Sys4 hindcasts and hence to more stable and potentially skilful seasonal streamflow hindcasts (Shi et al., 2015).

Furthermore, in this paper, the hindcasts were analysed against simulated streamflow, used as a proxy for observed streamflow. This is necessary because it enables an analysis of the quality of the hindcasts over the entire computation domain, rather than at non-evenly spaced stations over the same domain (Alfieri et al., 2014). Further work could however include carrying out a similar analysis for selected river stations in Europe, in order to account for model errors in the hindcast evaluation.

The calculation of the verification scores (excluding the ROC) was made by randomly selecting 15 ensemble members from the 51 ensemble members of the CM-SSF hindcasts, for starting dates for which the ensemble varies between 15 and 51 members (i.e. hindcasts made on 1 January, March, April, June, July, September, October and December; this is due to the split between 15 and 51 ensemble members in the Sys4 hindcasts, as described in Sect. 3.2.2.3 of this paper). In order to investigate the potential impact of this evaluation strategy on the results presented in this paper, the CRPSS was calculated for 15 and 51 ensemble members of the CM-SSF hindcasts for starting dates for which 51 ensemble members are available for the full hindcast period (i.e. hindcasts made on 1 February, May, August and November). This is displayed in Fig. 6 for all hindcast starting dates, lead times (i.e. 1 to 7 months) and regions combined. Overall, it is apparent that while the effect of this evaluation strategy is small, it could have had significant impacts in terms of the skill differences shown in Fig. 3. Most points on Fig. 6 are situated marginally above the 1-to-1 diagonal, signifying that the skill of the system might have been under-predicted to some extent in this analysis.

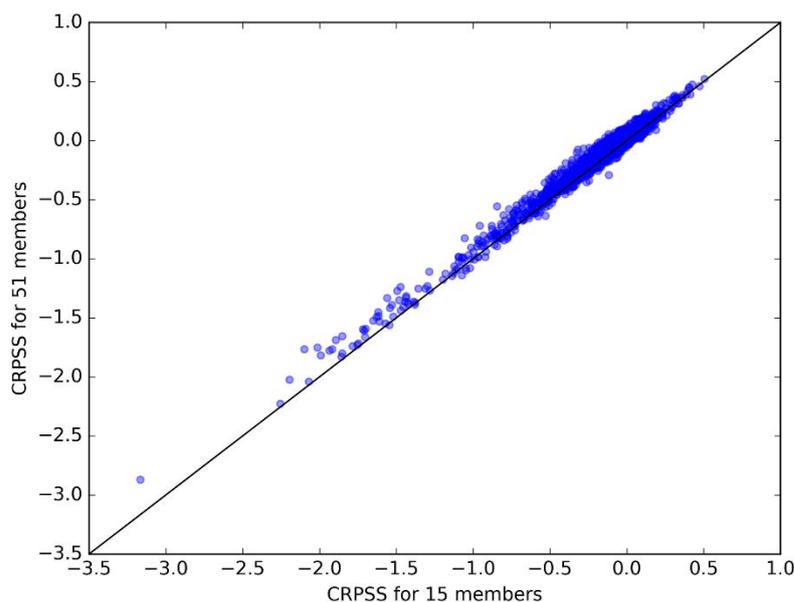


Figure 6. CRPSS calculated for the CM-SSF against the ESP (benchmark) for hindcasts made on 1 February, May, August and November, all lead times (i.e. 1 to 7 months) and all 74 European regions. The x-axis (y-axis) contains the CRPSS calculated from 15 (all 51) ensemble members of the CM-SSF.

The next version of the ECMWF seasonal climate forecast, SEAS5, was released in November 2017. Future work could include forcing the Lisflood model with SEAS5 and comparing the obtained seasonal streamflow hindcasts to the CM-SSF presented in this paper. This should indicate whether developments to the seasonal climate forecast translate through to better pan-European seasonal streamflow forecasts, which is of particular interest for regions and seasons when neither the ESP nor the CM-SSF is currently skilful. Additionally, alternative seasonal climate forecasting systems exist that may yield larger seasonal streamflow forecast improvements. For example, Scaife et al. (2014) reported that high levels of predictability were obtained for the European winter climate and the surface NAO using the Met Office Global Seasonal forecast System 5 (GloSea5). Forcing the Lisflood model with these improved GloSea5 forecasts would form valuable future research.

The operational EFAS medium-range streamflow forecasts are currently post-processed as a means to improve their reliability (Smith et al., 2016, and references therein). Results from this paper have shown that the CM-SSF is mostly unreliable (with regards to the EFAS-WB) and could hence benefit from post-processing of the seasonal climate forecast. However, post-processing techniques used for the EFAS medium-range streamflow forecasts might not be suitable for the CM-SSF, as the seasonal climate forecast used for the latter should be post-processed in terms of its seasonal anomalies rather than for errors in the timing, volume and magnitude of specific events. This is currently being considered for

operational implementation within EFAS and is an active area of discussion within the EFAS user community.

For the analysis presented in this paper, the CM-SSF was benchmarked against the ESP. Several other techniques exist for seasonal streamflow forecasting, such as statistical methods using predictors ranging from climate indices to antecedent observed precipitation and crop production metrics, to mention a few (e.g. Mendoza et al., 2017; Slater et al., 2017). Further analysis could include benchmarking the CM-SSF against one or multiple statistical methods, to assess the relative benefits of various seasonal streamflow forecasting techniques.

In this paper, the ability of both systems (CM-SSF and ESP) to forecast lower and higher streamflows than normal was explored, with several hypotheses made to link the streamflow predictability to regions' hydro-climatic processes. This includes the higher potential usefulness of the ESP in forecasting the spring streamflow in snow-dominated regions and the summer streamflow in regions where rivers are groundwater fed. In these regions and for these seasons, the IHC and the land surface memory drive the predictability. The CM-SSF provides an added potential usefulness in winter in the rainfall-dominated regions of central and western Europe, where the skill appears to persist through to spring due to the land surface memory (i.e. groundwater-driven streamflow and snowmelt-drive streamflow). While further exploration of these hypotheses is outside of the scope of this paper, future work is required to disentangle the links between the added predictability from Sys4 and the basins' hydro-climatic characteristics, for example, understanding the predictability in snow-dominated basins, arid regions and temperate groundwater-fed basins.

In this context, additional work to further disentangle and quantify the contribution of both predictability sources (seasonal climate forecasts versus IHC) to seasonal streamflow forecasting quality over Europe could be carried out by using the EPB (end point blending) method (Arnal et al., 2017b).

3.2.6 Conclusions

In this paper, the newly operational EFAS seasonal streamflow forecasting system (producing the CM-SSF forecasts by forcing the Lisflood model with the ECMWF System 4 seasonal climate forecasts (Sys4)) was presented and benchmarked against the ESP forecasting approach (ESP forecasts produced by forcing the Lisflood model with historical meteorological observations) for the hindcast period 1990 to 2017. On average, Sys4

improves the predictability over historical meteorological information for pan-European seasonal streamflow forecasting for the first month of lead time only (in terms of hindcast accuracy, sharpness and overall performance). However, the predictability varies per season and the CM-SSF is more skilful on average at predicting autumn and winter streamflows than spring and summer streamflows. Additionally, parts of Europe exhibit a longer predictability, up to 7 months of lead time, for certain months within a season. In terms of hindcast reliability, the CM-SSF is on average less skilful than the ESP for all lead times, due to a combination of too narrow and biased CM-SSF hindcasts, where the bias depends on the season that is being forecasted.

Subsequently, the potential usefulness of the two forecasting systems (CM-SSF and ESP) was assessed by analysing their skill in predicting lower and higher streamflows than normal. Overall, at least one of the two forecasting systems is capable of predicting those events months in advance. The ESP appears the most skilful on average, showing the importance of IHC and the land surface memory for seasonal streamflow forecasting. Nevertheless, for certain regions and seasons the CM-SSF is the most skilful at predicting anomalously low or high streamflows beyond 1 month of lead time, noticeably in winter for almost 40% of the European regions. This potential usefulness could be harnessed by using seasonal streamflow forecasts as complementary information to existing forecasts at shorter timescales, to provide monitoring and early-warning information for flood preparedness.

Overall, patterns in skill in the CM-SSF are however not mirrored in the Sys4 precipitation and temperature hindcasts. This suggests that using seasonal climate forecast performance as a proxy for seasonal streamflow forecasting skill is not adequate and that more work is needed to understand the link between meteorological and hydrological variables on seasonal timescales over Europe.

Data availability. The data from the European Flood Awareness System are available to researchers upon request (subject to licensing conditions). Please visit www.efas.eu for more details.

Competing interests. The authors declare that they have no conflict of interest.

Special issue statement. This article is part of the special issue “Sub-seasonal to seasonal hydrological forecasting”. It is not associated with a conference.

Acknowledgements. Louise Arnal, Hannah L. Cloke and Jessica Neumann gratefully acknowledge financial support from the Horizon 2020 IMPREX project (grant agreement

641811) (project IMPREX: www.imprex.eu). Louise Arnal's time was additionally partly funded by a University of Reading PhD scholarship. Fredrik Wetterhall, Christel Prudhomme and Blazej Krzeminski's work was supported by the EFAS computational centre in support to the Copernicus Emergency Management Service/Early Warning Systems (Flood) (contract no. 198702 from JRC-IES). Elisabeth Stephens is thankful for support from the Natural Environment Research Council and Department for International Development (grant number NE/P000525/1) under the Science for Humanitarian Emergencies and Resilience (SHEAR) research programme.

3.3 The EFAS and GloFAS operational seasonal hydrological outlooks

The EFAS and GloFAS seasonal hydrological outlooks are operational products of the Copernicus EMS. Louise Arnal led the design of EFAS-Seasonal and helped with the design of GloFAS-Seasonal. Both outlooks were created to extend the lead time of EFAS and GloFAS services to the sub-seasonal to seasonal timescale. They show the likelihood of the European or global river network to be unusually dry or wet within the forecast horizon. EFAS-Seasonal and GloFAS-Seasonal have the potential to give earlier warnings of floods and droughts, for increased preparedness and disaster risk reduction.

3.3.1 EFAS-Seasonal

EFAS-Seasonal (introduced in Sect. 3.2; ECMWF, 2017a) was made operational in December 2016, extending the lead time of EFAS forecasting products from about two weeks to eight weeks. It is one of the first operational pan-European sub-seasonal to seasonal hydrological outlooks.

EFAS-Seasonal shows a map (Fig. 7) of predicted streamflow anomaly and its probability of occurrence in the next eight weeks for 94 European river basins (note that the EFAS domain, and hence number of river basins, was extended since Sect. 3.2, and the corresponding paper, was written). Each river basin can be clicked on to call up a hydrograph (Fig. 8), showing the ensemble streamflow forecast river basin weekly averages, relevant climatological thresholds (10th and 90th yearly percentiles) and the current simulated streamflow (referred to as EFAS-WB in Sect. 3.2), once available. A new EFAS-Seasonal outlook is available at the start of each month.

The latest EFAS-Seasonal outlook is openly available via the EFAS web interface¹. More information about EFAS-Seasonal can be found on the EFAS webpage².

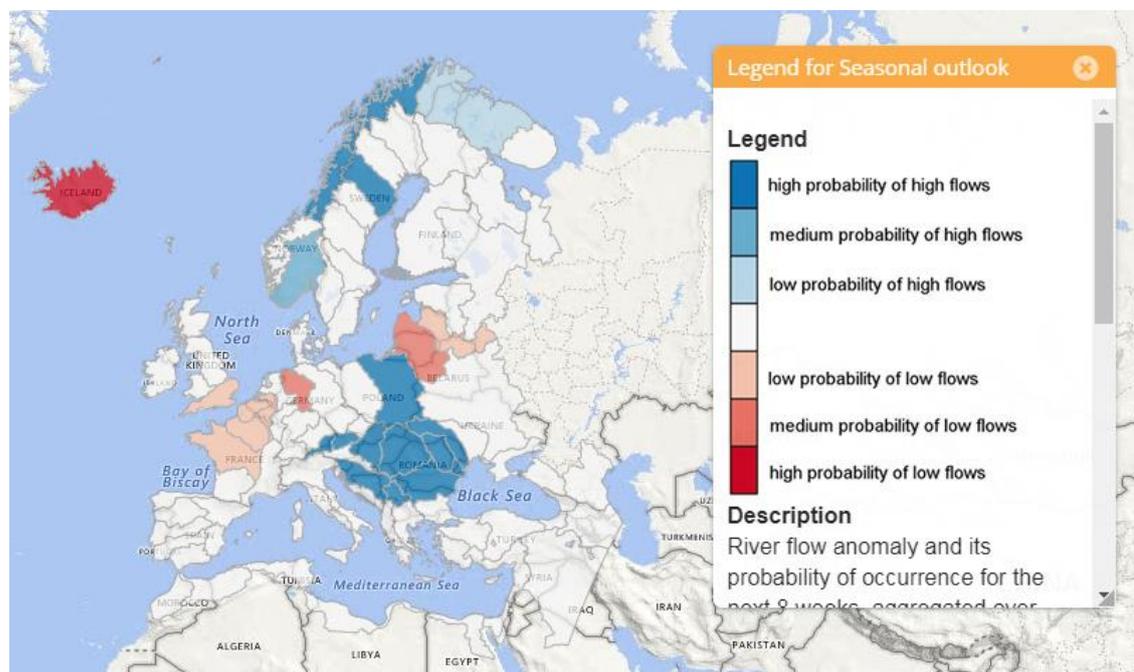


Figure 7. EFAS-Seasonal example map.

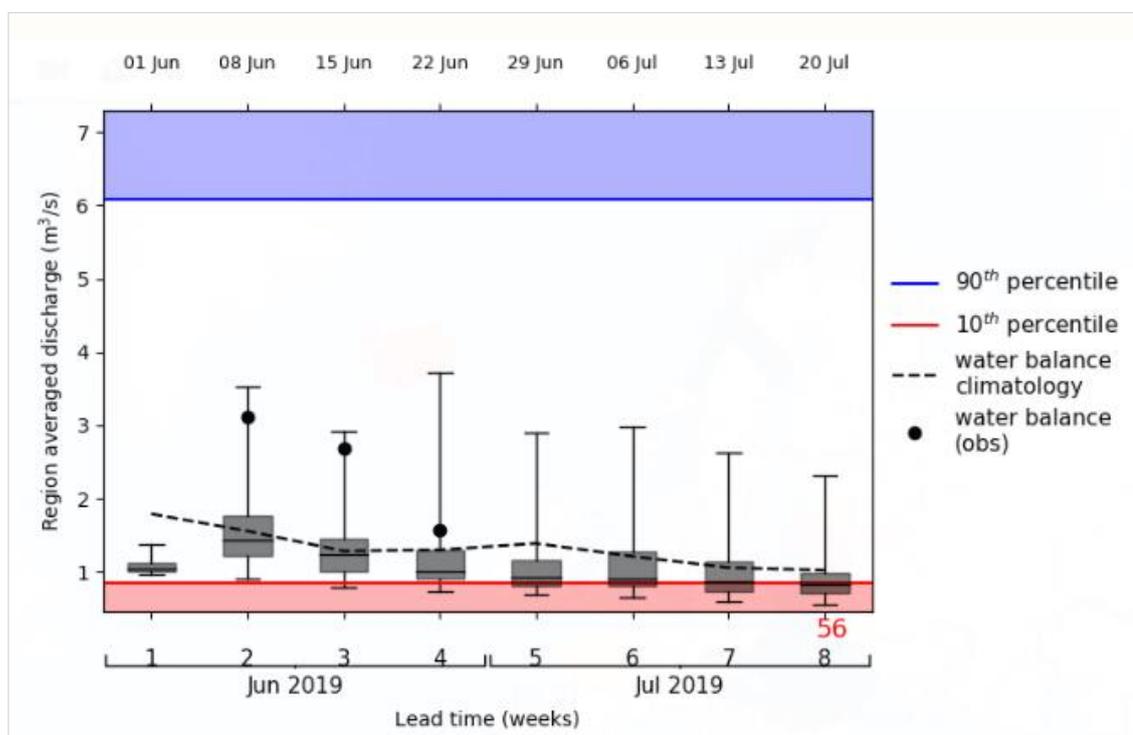


Figure 8. EFAS-Seasonal example hydrograph.

¹ www.efas.eu

² www.efas.eu/en/seasonal-outlook

The EFAS-Seasonal streamflow forecasts are produced following the approach described in Sect. 3.2. However, ECMWF's latest seasonal meteorological forecast version (SEAS5; ECMWF, 2017b) are now used for their production.

EFAS-Seasonal was designed in close co-operation between UoR and EFAS scientists (from ECMWF and the JRC). It was presented at the EFAS annual meeting 2017, during which feedback was gathered from the users. This feedback helped shape GloFAS-Seasonal.

3.3.2 GloFAS-Seasonal

The design of GloFAS-Seasonal (along with a performance assessment of the system) is documented in a paper, presenting a co-author contribution arising through collaboration during this PhD, and has the following reference:

Emerton, R., E. Zsoter, L. Arnal, H. L. Cloke, D. Muraro, C. Prudhomme, E. M. Stephens, P. Salamon and F. Pappenberger, 2018: Developing a global operational seasonal hydro-meteorological forecasting system: GloFAS-Seasonal v1.0, *Geosci. Model Dev.*, 11, 3327-3346, doi:10.5194/gmd-11-3327-2018*

L.A. co-designed the GloFAS-Seasonal operational product (led by Rebecca Emerton) and commented on the manuscript before and during publication.

The following is a summary of the paper, fully available in the thesis Appendix A4.

GloFAS-Seasonal (ECMWF, 2017a) became operational in Autumn 2017. Its design is very similar to EFAS-Seasonal, with the exception of a few additional elements, developed based on the feedback received about EFAS-Seasonal.

GloFAS-Seasonal shows a map of the forecast for 305 major world river basins, highlighting the predicted streamflow anomaly and its probability of occurrence in the next 18 weeks (approximately four months), averaged over these river basins and displayed for each pixel of the global river network. When a station is clicked on, two products are shown:

- A hydrograph, showing the ensemble streamflow forecast weekly averages and relevant climatological thresholds (20th and 80th weekly percentiles).

* ©2018. The Authors. Geoscientific Model Development, a journal of the European Geosciences Union published by Copernicus. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.

- A consistency diagram, displaying the probability of threshold exceedance for the latest four consecutive forecasts.

A new GloFAS-Seasonal outlook is available at the start of each month.

The latest GloFAS-Seasonal outlook is openly available via the GloFAS web interface³. More information about GloFAS-Seasonal can be found on the GloFAS webpage⁴.

The GloFAS-Seasonal streamflow forecasts are produced by forcing the HTESSEL land surface model with ECMWF's SEAS5 seasonal meteorological forecasts. The obtained runoff from HTESSEL is subsequently routed through the global river network using Lisflood.

An initial evaluation of the performance of the system (presented in Emerton et al. (2018)) indicates that GloFAS-Seasonal streamflow forecasts are potentially useful (i.e. more skilful at predicting low and high streamflow events than a long-term average climatology) out to several months ahead over the global river network. Additionally, GloFAS-Seasonal forecasts are on average more reliable than climatology. The performance of the system however varies by region and season, and the forecasts were found to over-predict streamflow in general.

3.4 The 2013/14 Thames basin floods: do improved meteorological forecasts lead to more skilful hydrological forecasts at seasonal time scales?

In Sect. 3.2, we have seen that streamflow forecasts have limited skill on seasonal timescales over Europe. However, results indicated that seasonal meteorological forecasts can provide some added predictability for certain European regions, especially in winter, leading to potentially useful streamflow forecasts of higher than normal streamflows up to seven months ahead. Furthermore, the IMPREX decision-making activity presented in Chapter 2, Sect. 2.4, revealed that skill does not necessarily pose a barrier to use, but that hydrological forecasts need to be provided at a higher spatial resolution to support local decision-making. This section presents an evaluation of the performance of EFAS-Seasonal for predicting the 2013/14 Thames basin floods.

This is the summary of a paper, presenting a co-author contribution arising through collaboration during this PhD, and has the following reference:

³ www.globalfloods.eu

⁴ www.globalfloods.eu/general-information/forecast-viewer-info

Neumann, J. L., L. Arnal, L. Magnusson and H. Cloke, 2018a: The 2013/14 Thames basin floods: Do improved meteorological forecasts lead to more skillful hydrological forecasts at seasonal time scales?, *J. Hydrometeorol.*, 19, 6, 1059-1075, doi:10.1175/JHM-D-17-0182.1*

L.A. co-designed the analysis presented in this paper, provided some technical support for the analysis, including running forecasting experiments to produce seasonal hydrological hindcasts, analysing the quality of the hindcasts produced and providing guidance on plotting the data. Additionally, L.A. commented on the manuscript before and during publication.

The published article can be found in the thesis Appendix A5.

During the winter of 2013/14, the Thames river basin experienced 12 major Atlantic depressions, which led to extensive and prolonged fluvial and groundwater flooding. From atmospheric relaxation experiments (i.e. conditions within specified regions are relaxed toward a reanalysis; see method in Magnusson (2017)), this exceptional weather was found to coincide with highly anomalous meteorological conditions across the globe, through teleconnection patterns.

Building on these findings, this study looked at whether improved seasonal meteorological forecasts translate to more skilful seasonal hydrological forecasts for this flood event. To this end, atmospheric relaxation experiments produced at ECMWF were applied through the Lisflood hydrological model to reforecast the 2013/14 Thames basin floods on a seasonal timescale. Results were investigated for three Thames sub-basins of distinct hydrogeological characteristics and compared to the default EFAS-Seasonal hydrological forecasts.

Results highlighted that the tropics played an important role in the development of extreme conditions over the Thames river basin. While both the tropical Atlantic and tropical Pacific relaxation experiments captured seasonal meteorological flow anomalies, the greatest seasonal hydrological forecasting skill was found for the tropical Atlantic. The seasonal hydrological forecasts associated with the tropical Atlantic relaxation experiments could have indicated a potential flood event up to two months ahead. Surprisingly, the seasonal hydrological forecasts associated with the north-eastern Atlantic relaxation (closest relaxation region to the UK) were confident but largely under-predicted the hydrological extremes. All relaxation experiments produced more skilful seasonal hydrological forecasts

* ©2018. The Authors. Journal of Hydrometeorology, a journal of the American Meteorological Society. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.

than EFAS-Seasonal. The basins' hydrogeological characteristics were additionally found to be an important factor of seasonal streamflow forecast skill. The streamflow was most skilfully forecasted for the Thames sub-basin with a large drainage area and a high peak flow. Permeable lithology and the basin's antecedent conditions were both found to influence the skill of seasonal groundwater level forecasts.

Atmospheric relaxation experiments can improve our understanding of global teleconnection patterns and the potential predictability of extreme events such as the 2013/14 basin floods. While these experiments were conducted in retrospect, they indicate potentially achievable future improvements in operational hydro-meteorological forecasting. Results showed that seasonal hydrological forecasting skill differed from considering the meteorology alone, highlighting the need to consider both components jointly when investigating such high-impact hydro-meteorological events.

Chapter 3 explored the performance of the EFAS-Seasonal and the GloFAS-Seasonal operational hydro-meteorological forecasting systems over Europe and the globe, respectively. Findings showed that both systems could be potentially useful for predicting low and high streamflow events out to several months ahead. However, although seasonal hydrological predictability has improved over the last decades, the performance of EFAS-Seasonal and GloFAS-Seasonal (and similar systems) remains limited.

Chapter 4 presents a sensitivity analysis method which disentangles the relative contributions of predictability sources to seasonal streamflow forecasting skill. This method can easily be applied to any seasonal hydro-meteorological forecasting system to assess how to efficiently improve the forecasts it produces.

Chapter 4

Towards tangible seasonal streamflow forecast improvements

4.1 Background and aim

Hydro-meteorological forecasts currently offer a limited predictability of events on the seasonal timescale (see Chapter 3). This barrier in predictability can arise from various parts of the forecasting chain, such as errors in the: i) hydrological model, ii) forecasts' initial hydrological conditions, iii) seasonal meteorological forcings (i.e. forecasts, historical observations, etc.), and iv) due to the inherent chaos of nature and the growing uncertainties with time.

Given the variety of these error sources, seasonal hydro-meteorological forecast developments are equally varied and include:

- i Hydrological modelling developments: through model calibration, improvements of the model physics, increase in the model's spatial and temporal resolutions, or the combination of several hydrological models to better capture model uncertainties (Wanders et al., 2019).
- ii Initial hydrological conditions' improvements: through data assimilation (Liu et al., 2012), initialising the hydrological model with multiple sets of initial hydrological conditions or more frequent forecast initialisations to better characterise the initial conditions' uncertainty (Wetterhall and Di Giuseppe, 2018).
- iii Reduction of seasonal meteorological forecasts' errors: by improving the model physics, increasing the general circulation model's spatial and temporal resolutions, increasing the number of meteorological ensemble members or combining multiple meteorological model outputs to better characterise uncertainty (Mo and Lettenmaier, 2014), or post-processing the meteorological forecasts (Crochemore et al., 2016).
- iv Improvements in the science of seasonal predictability through research experiments and case studies (Neumann et al., 2018a; see Chapter 3, Sect. 3.4).

Many of the methods mentioned above are applicable to hydro-meteorological forecasting on a range of timescales; only a few are unique to seasonal forecasting. While implementing these methods can be costly (in terms of time invested, computer-power and money), they might not lead to proportionate seasonal hydro-meteorological forecast improvements for predicting events of societal relevance. In fact, operational streamflow forecast quality has not significantly improved in the last decade, despite the costly research and developments they are receiving (Pagano et al., 2004a; Welles et al., 2007). Adding to this is the threat from the currently observed and future expected consequences of climate change (i.e. in Europe these are an increase in the risk of coastal and inland flooding, from sea level rise and from an increase in the frequency and intensity of heavy precipitation events, respectively; IPCC, 2014). In this context, it is vital to prioritise seasonal hydro-meteorological forecast developments to maximise improvements in seasonal predictability of hydrological extremes for forecast users and society.

In this chapter, we present a novel efficient sensitivity analysis method which can disentangle the dominant error sources (i.e. targeting sources ii) and iii) mentioned above) in seasonal streamflow forecasts. The first section of this chapter introduces the method, tested over 18 US basins (or catchments). In the second section, we present results of the method applied to EFAS-Seasonal. The third section presents the novel concept of a 'flexible seasonal hydro-meteorological forecasting system', inspired by the sensitivity analysis' results. The aim of this chapter is to propose novel methods for tangible seasonal streamflow forecast improvements.

4.2 An efficient approach for estimating streamflow forecast skill elasticity

This section has been published in Journal of Hydrometeorology (JoHM) with the following reference:

Arnal, L., A. W. Wood, E. Stephens, H. L. Cloke and F. Pappenberger, 2017b: An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity, *J. Hydrometeorol.*, 18, 1715–1729, doi:10.1175/JHM-D-16-0259.1*

The contributions of the authors of this paper are as follows: A. W. Wood (collaborator: NCAR), E. Stephens (supervisor: academic), H. L. Cloke (supervisor: academic) and F.

* ©2017. The Authors. Journal of Hydrometeorology, a journal of the American Meteorological Society. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.

Pappenberger (collaborator: ECMWF). A.W.W. provided the dataset. L.A. and F.P. conceived the experiment. L.A. posed the research questions, designed and carried out the analysis, and wrote the paper. L.A., A.W.W., E.S., H.L.C. and F.P. interpreted the results and commented on the manuscript. Overall, 95% of the research and 90% of the writing was undertaken by L.A.

The published article can be found in the thesis Appendix A6.

Abstract. Seasonal streamflow prediction skill can derive from catchment initial hydrological conditions (IHCs) and from the future seasonal climate forecasts (SCFs) used to produce the hydrological forecasts. Although much effort has gone into producing state-of-the-art seasonal streamflow forecasts from improving IHCs and SCFs, these developments are expensive and time consuming and the forecasting skill is still limited in most parts of the world. Hence, sensitivity analyses are crucial to funnel the resources into useful modeling and forecasting developments. It is in this context that a sensitivity analysis technique, the variational ensemble streamflow prediction assessment (VESPA) approach, was recently introduced. VESPA can be used to quantify the expected improvements in seasonal streamflow forecast skill as a result of realistic improvements in its predictability sources (i.e. the IHCs and the SCFs) – termed “skill elasticity” – and to indicate where efforts should be targeted. The VESPA approach is, however, computationally expensive, relying on multiple hindcasts having varying levels of skill in IHCs and SCFs. This paper presents two approximations of the approach that are computationally inexpensive alternatives. These new methods were tested against the original VESPA results using 30 years of ensemble hindcasts for 18 catchments of the contiguous United States. The results suggest that one of the methods, end point blending, is an effective alternative for estimating the forecast skill elasticities yielded by the VESPA approach. The results also highlight the importance of the choice of verification score for a goal-oriented sensitivity analysis.

4.2.1 Introduction

Unprecedented increases in computer capabilities have shaped the last several decades' advances in numerical weather prediction (NWP), and with them, the development of environmental forecasting and modeling systems. This has led to a shift in the strategy of operational forecasting centers toward more integrated modeling and forecasting approaches, such as coupled systems and Earth system models (ESMs), with the final aim to extend the limits of predictability (i.e. from sub-seasonal to seasonal forecasting). These developments are supported by the assimilation of more and better-quality observation

data as well as the increase in model resolutions and complexity. However, such advances can be very expensive and data hungry and may not yield proportional improvements.

Seasonal hydrological forecasts are predictions of the future states of the land surface hydrology (e.g. streamflow), up to a few months ahead. They are valuable for applications such as reservoir management for hydropower, agriculture and urban water supply, spring flood and drought prediction, and navigation, among others (Clark et al. 2001; Hamlet et al. 2002; Chiew et al. 2003; Wood and Lettenmaier 2006; Regonda et al. 2006; Luo and Wood 2007; Kwon et al. 2009; Cherry et al. 2005; Viel et al. 2016). They have the potential to provide early warning for increased preparedness (Yuan et al. 2015b). Traditionally, seasonal streamflow forecasts have relied upon land surface memory, the persistence in the land surface (e.g. catchment) initial hydrological conditions (IHCs; of soil moisture, groundwater, snowpack, and the current streamflow). IHCs are one of the most important predictability sources of seasonal streamflow forecasts and were thus the starting point for the development of the ensemble streamflow prediction (ESP) approach in the 1970s (Wood et al. 2016b). The ESP was first developed and used for reservoir management purposes. It is produced by running a hydrological model with observed meteorological inputs to produce current observed IHCs, from which the forecast is started, and the forcing over the forecast period is undertaken using an ensemble of historical meteorological observations (Day 1985). The ESP method assumes that the model states to initialize a forecast are perfectly estimated, while the future climate is completely unknown. However, the skill of the ESP decreases significantly after one to a few months of lead time over most parts of the world because of a decrease in the land surface memory with time. The achievable predictability from the ESP thus depends on the persistence of the IHCs, which can vary as a function of the season (i.e. the transition between dry and wet seasons can, for example, be hard to forecast) and the location and size of the catchment (i.e. the streamflow in a large catchment with a slow response time and/or situated in a region with negligible precipitation inputs during the forecast period will for example be easier to forecast; Wood and Lettenmaier 2008; Shukla et al. 2013; van Dijk et al. 2013; Yuan et al. 2015b).

More recently, seasonal climate predictability derived from large-scale climate precursors [e.g. El Niño–Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO)] has been used to enhance seasonal streamflow forecasting (e.g. Wood et al. 2002; Yuan et al. 2013b; Demargne et al. 2014; Mendoza et al. 2017). Such systems produce streamflow forecasts by initializing a hydrological model to estimate IHCs and forcing the model with inputs based on seasonal climate forecasts (SCFs; of temperature and precipitation) instead of historical observations. Their skill is also still limited because of the rapid decrease in

precipitation forecasting skill beyond two weeks of lead time, and the skill is variable in both space and time (Yuan et al. 2011; van Dijk et al. 2013; Slater et al. 2016). In Europe, for instance, the skill is higher in winter in regions where the winter precipitation is highly correlated with the NAO. Regions with high skill include the Iberian Peninsula, Scandinavia, and regions around the Black Sea (Bierkens and van Beek 2009). In the contiguous United States (CONUS), the skill is on average higher over (semi)arid western catchments, due to the persistence of the IHCs influence up to three months of lead time. The skill can be higher in some regions of the western CONUS (i.e. California, the Pacific Northwest, and Great Basin) in the winter and fall due to higher precipitation forecasting skill in strong ENSO phases (Wood et al. 2005).

Increasing the seasonal streamflow forecast skill remains a challenge: one that is being tackled by improving IHCs and SCFs using a variety of techniques. Techniques include model developments and data assimilation and can vary in computational expense. However, over the past several decades, it has been shown that operational streamflow forecast quality has not significantly improved (Pagano et al. 2004a; Welles et al. 2007). This is the motivation for the use of sensitivity analysis techniques to guide future forecasting developments for seasonal streamflow forecasting and is the basis for this paper.

It is in this context that the attribution of seasonal streamflow forecast uncertainty to the IHC and SCF errors has been researched extensively. Wood and Lettenmaier (2008) introduced a method based on two hindcasting end points: the ESP and the reverse-ESP. In contrast to the ESP, which only represents the uncertainty in the future climate, the reverse-ESP only represents the uncertainty in IHCs by using an ensemble of initial model states taken from historical simulations to initialize a prediction forced by a single set of observed meteorological inputs. Typically, the input uncertainty attenuates over a period of months as the influence of the perfect future climate input increasingly determines model states.

Comparing the skill of the ESP versus reverse-ESP seasonal streamflow forecasts allows one to identify the dominant predictability source (and conversely uncertainty source) of seasonal streamflow forecasting (i.e. the IHCs or the SCFs), and its evolution in both space and time. It was successfully used to disentangle the relative importance of initial conditions and boundary forcing errors on seasonal streamflow forecast uncertainties by several authors: for example, for catchments in the United States (Wood and Lettenmaier 2008; Li et al. 2009; Shukla and Lettenmaier 2011), in France (Singla et al. 2012), in Switzerland (Staudinger and Seibert 2014), in China (Yuan et al. 2016; Yuan 2016), and in

the Amazon (Paiva et al. 2012), as well as for the entire globe (Shukla et al. 2013; Yossef et al. 2013; MacLeod et al. 2016). This work is instructive as it enables the dominant predictability source to be identified (i.e. where efforts and resources should be targeted) to focus improvement, which could potentially lead to more skillful seasonal streamflow predictions.

This method was extended by Wood et al. (2016a, hereafter W16) via a method called variational ensemble streamflow prediction assessment (VESPA), which involves assessing intermediate IHC and SCF uncertainty points between the perfect and climatological points applied in ESP and reverse-ESP. The approach allows the calculation of a metric called “skill elasticity”, that is, the sensitivity of streamflow forecast skill to IHC and SCF skill changes. A key drawback of the VESPA approach, however, is that it is computationally intensive. For each catchment and initialization month of a forecast, the response surface was defined through the use of dozens of multidecadal variable-skill ensemble hindcasts, ultimately amounting to millions of simulations. In contrast, the ESP and reverse-ESP skill can be estimated from a single set of ensemble hindcasts spanning a historical period. The IHC and SCF skill variation method was also highly specific to the particular model state configuration and involved a relatively simplistic linear blending procedure. The elasticity calculations were furthermore based only on a single verification score of forecast skill (i.e. coefficient of determination R^2) for the analysis. An ensemble forecast has many attributes, for example, the skill, the reliability, the resolution, and the uncertainty of the forecast, among others. To obtain a complete picture of the forecast quality, the scores should encompass many of these attributes, as each verification score will give us different information about the forecast quality.

The drawbacks of VESPA motivate us to assess two computationally inexpensive methods of estimating the forecast skill elasticities, using only the original ESP and reverse-ESP results that depend on the single hindcast series as mentioned above. The two methods are termed end point interpolation (EPI) and end point blending (EPB). In the first part of this paper, we compare results from the two methods tested on 18 catchments of the CONUS to the original results from the VESPA, using a single verification score. The objective of this part is to investigate whether the new methods can discriminate the influence of IHC and SCF errors on seasonal streamflow forecasting uncertainties and to assess the ability of those new methods to correctly estimate the forecast skill elasticities. In the second part, additional verification scores are applied for streamflow forecast verification, supporting the second objective of the paper, which is to explore the sensitivity of the results obtained from the two new methods and the VESPA approach to the choice of the verification score.

4.2.2 Methods, data, and evaluation strategy

4.2.2.1 The VESPA approach

In this work, as in W16, the term “perfect” refers to current observed meteorological data and the term climatological refers to the whole distribution of historical observed data. Figure 1 presents the ESP (Fig. 1a), the reverse-ESP (Fig. 1b), the climatology (Fig. 1c), and the VESPA forecast (Fig. 1d), as generated in W16. The ESP, the reverse-ESP, the perfect forecast, and the climatology are all end points of the uncertainty in the sense that the uncertainty in those forecasts is either perfect or climatological. They are the end points of the VESPA approach.

VESPA aims to produce streamflow forecasts from IHCs and SCFs with an uncertainty situated between the perfect and the climatological uncertainty (Fig. 1d). Forecasts were generated by linearly blending the climatological and perfect IHCs (i.e. model moisture states) and the climatological and perfect SCFs (i.e. meteorological forcings of precipitation, evapotranspiration, and temperature), subsequently used to run the hydrological model. The weights used for blending the data were ($w = 0, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 1.0$), applied so that a weight of zero is the perfect knowledge and unity is the climatological knowledge, with w_{IHC} and w_{SCF} denoting the weights used to blend the IHCs and the SCFs, respectively (W16). An ESP forecast results from the weights $w_{\text{IHC}} = 0$ and $w_{\text{SCF}} = 1$, the reverse-ESP from $w_{\text{IHC}} = 1$ and $w_{\text{SCF}} = 0$, the perfect forecast from $w_{\text{IHC}} = 0$ and $w_{\text{SCF}} = 0$, and the climatology from $w_{\text{IHC}} = 1$ and $w_{\text{SCF}} = 1$.

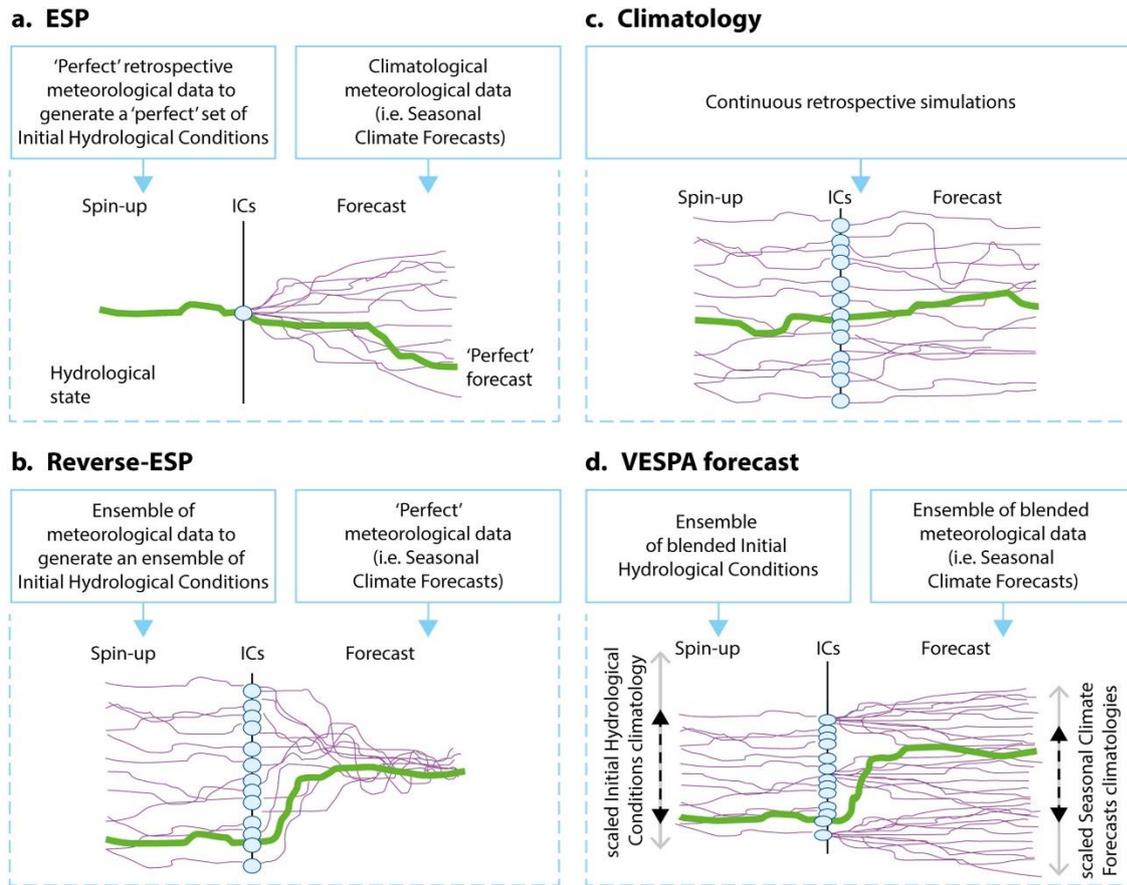


Figure 1. Schematic of (a) the ESP, (b) the reverse-ESP, (c) the climatology, and (d) the VESPA (this figure is adapted from Fig. 3 in W16).

To plot the skill of the VESPA forecasts as a function of the IHC and SCF skill, W16 used skill surface plots (Fig. 2), interpolating forecast skill results from different IHC and SCF weighting combinations. The axes represent the SCF and IHC skill, derived respectively from the blending weights w_{SCF} and w_{IHC} using the following equations (W16):

$$SCF \text{ skill} = 100 \times (1 - w_{SCF}^2) \quad \text{and} \quad (1)$$

$$IHC \text{ skill} = 100 \times (1 - w_{IHC}^2) \quad (2)$$

The SCF and the IHC skill values obtained from these equations are the percentage of climatological variance explained in the respective predictability source (i.e. SCF and IHC; W16). Each SCF skill–IHC skill combination corresponds to a specific VESPA forecast, the skill of which can be plotted on the skill surface plot (black plus signs in Fig. 2). The blue circles are the end points of the VESPA forecasts: the reverse-ESP (revESP in Fig. 2), the perfect forecasts, the ESP, and the climatology (climo in Fig. 2). The skill surface plots are hence a graphical representation of the response surface obtained from the VESPA sensitivity analysis.

The VESPA seasonal streamflow forecasts were generated by W16 using lumped Sacramento Soil Moisture Accounting (SAC-SMA) and SNOW-17 catchment models for unimpaired catchments. The models were forced with daily inputs in precipitation, temperature, and potential evapotranspiration and were calibrated and validated against observed daily streamflow from the U.S. Geological Survey (USGS). Eighty-one skill variations of a 30-yr hindcast (from 1981 to 2010) were produced for 424 catchments in the CONUS, starting at the beginning of each month (i.e. forecast initialization dates), with lead times up to 6 months.

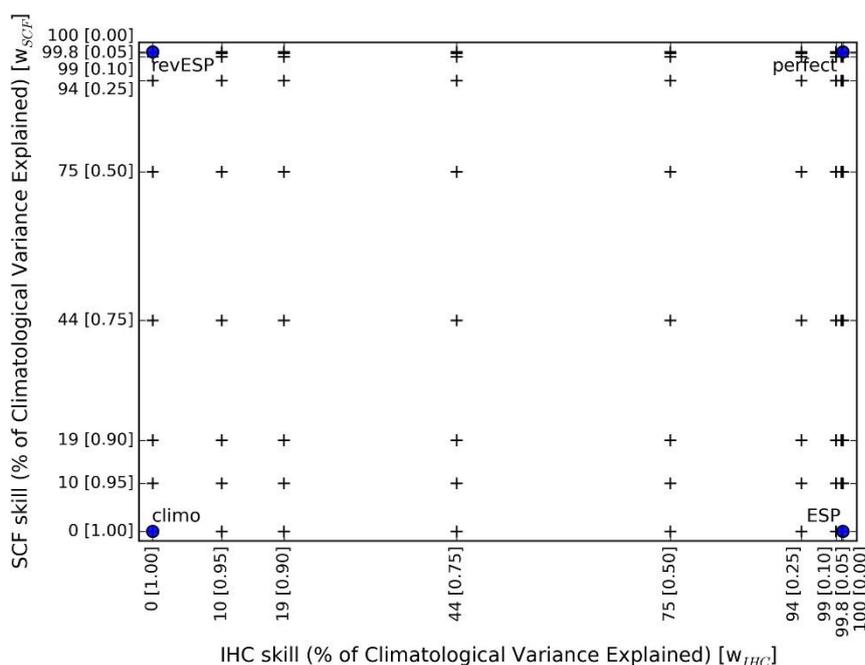


Figure 2. Schematic of a skill surface plot. The y and x axes display the SCF and the IHC skill, respectively. They are expressed as a percentage of the climatological variance explained in the respective predictability source. The blending weights, w_{SCF} and w_{IHC} , from which the skill values are derived are shown in square brackets in the figure.

4.2.2.2 Alternative methods to the VESPA approach

In this paper we present two alternative methods of the VESPA approach, the EPI and the EPB. These methods aim to reproduce the response surface obtained from the VESPA approach by using the same 30-yr hindcast ensembles produced by W16, aggregated over the first three months with zero lead time for each initialization date (referred to as 3-month streamflow forecast hereafter) and corresponding exclusively to the end points (i.e. the ESP, the reverse-ESP, the perfect forecast, and the climatology).

The two new methods were tested for a subset of the CONUS-wide catchment dataset presented in W16 (Fig. 3), comprising 18 catchments from the large USGS Hydro-Climatic Data Network (HCDN; Lins 2012). The 18 selected catchments cover a large range of

hydrometeorological conditions, including the maritime climate regime of the U.S. West Coast catchments; the humid regime of the eastern United States (south of the Great Lakes) with rainfall-driven runoff and variable winter snow in the most northern catchments; and the Intermountain West and northern Great Plains regions, where streamflow is greatly influenced by the snow cycle.

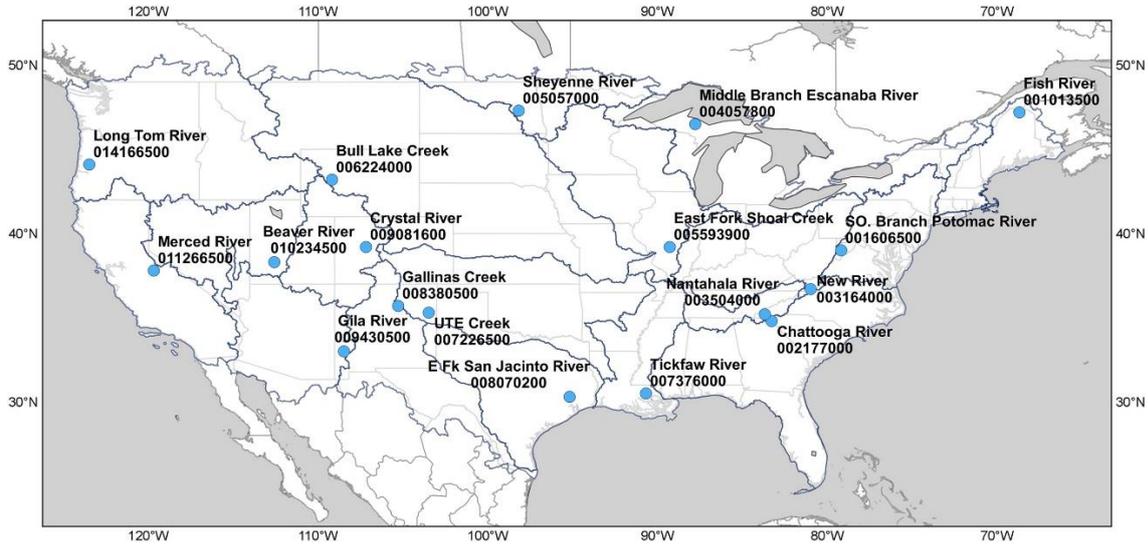


Figure 3. Map of the 18 catchments of the CONUS selected for the analysis and the HCDN regions (dark blue outlines).

4.2.2.2.1 End point interpolation

The EPI produces a response surface by interpolating the forecast skill of the end points throughout the skill surface plot. Both linear (i.e. linear barycentric interpolation) and cubic interpolation techniques were tested. However, results will be shown for the linear interpolation only as the cubic interpolation did not provide noticeable improvements to the linear interpolation given that the interpolation is based on only four points situated at the corners of the response surface. The linear EPI was performed for each forecast initialization date and for each catchment.

4.2.2.2.2 End point blending

The EPB generates hindcasts for each w_{SCF} - w_{IHC} combination (i.e. each plus sign in Fig. 2; w_{SCF} and w_{IHC} are selected to be the same blending weights used by W16, for the purpose of comparison). For each combination point, a new ensemble of 100 members was generated by blending the four end points, given a specific weighted average. The percentage of each end point used [EP(%); i.e. the number of members randomly selected from each end point], was calculated for each combination point using the following equation:

$$EP(\%) = (1 - |x_{EP} - w_{IHC}|) \times (1 - |y_{EP} - w_{SCF}|) \quad (3)$$

where x_{EP} and y_{EP} are the w_{IHC} and w_{SCF} values of the end point for which the percentage is calculated, respectively. For example, if the w_{IHC} and w_{SCF} match the end point values, 100% of the EPB hindcast members are resampled from that end point (i.e. the end point skill is reproduced). This was done for each forecast initialization date and for each catchment.

To produce the skill surface plots for the EPB method, the SCF and IHC skill was calculated using the same equations as in W16 [i.e. Eqs. (1) and (2), respectively].

4.2.2.3 The evaluation strategy

The aim of this paper is to compare two computationally inexpensive alternative methods to the VESPA approach, the EPI and the EPB. To this end, the paper unfolds into two distinct objectives. First, we want to investigate whether the EPI and/or the EPB can discriminate the influence of IHC and SCF errors on seasonal streamflow forecasting uncertainties and reproduce VESPA skill elasticity estimates. This will validate the use of one or both methods as alternative to the VESPA approach. Second, we want to explore the sensitivity of the results obtained from the EPI, the EPB, and the VESPA methods to the choice of the verification score. This will be an attempt to demonstrate the importance of the choice of the verification score for forecast verification and communication.

4.2.2.3.1 Can EPI and EPB discriminate the influence of IHC and SCF errors on seasonal streamflow forecast uncertainties?

To explore the first objective of this paper, skill surface plots were produced for the EPI, the EPB, and the VESPA methods. As in W16, the seasonal streamflow forecast skill depicted in the skill surface plots was calculated from the R^2 of forecast ensemble means with the observations, where perfect forecasts (model simulations driven by the observed meteorology) were treated as observations to calculate the R^2 . As discussed at length in W16, this choice deliberately excludes the model errors as a source of forecast uncertainty.

The skill surface plots obtained from the EPI and the EPB methods were subsequently compared qualitatively and quantitatively to the skill surface plots obtained for the VESPA approach. The qualitative analysis consisted in visually inspecting the patterns contained in the skill surface plots, giving an indication of the dominant predictability source on the streamflow forecast skill. The quantitative analysis of the results was based on the calculation of the skill elasticities for the IHCs and the SCFs (E_{IHC} and E_{SCF} , respectively), for the EPI, the EPB, and the VESPA methods, averaged across three transects of a quadrant situated in the center of the response surface, according to the following equations:

$$E_{IHC} = 100 \times \left\{ \frac{S(F[75,19]) - S(F[19,19])}{75\% - 19\%} + \frac{S(F[75,44]) - S(F[19,44])}{75\% - 19\%} + \frac{S(F[75,75]) - S(F[19,75])}{75\% - 19\%} \right\} / 3 \quad (4)$$

$$E_{SCF} = 100 \times \left\{ \frac{S(F[19,75]) - S(F[19,19])}{75\% - 19\%} + \frac{S(F[44,75]) - S(F[44,19])}{75\% - 19\%} + \frac{S(F[75,75]) - S(F[75,19])}{75\% - 19\%} \right\} / 3 \quad (5)$$

The numerators, expressed as $S(F[\cdot]) - S(F[\cdot])$, contain the gradients in the streamflow forecast skill between IHC skill (or SCF skill) values of 75% and 19% (the denominator). The values in between the square brackets of the numerator are the IHC skill followed by the SCF skill values, which indicates a certain $w_{SCF} - w_{IHC}$ combination point in the example skill surface plot in Fig. 2. In the denominator, the IHC and SCF skill gradients are gradients in the percentage of the climatological variance explained in the respective predictability source. The skill elasticities (E_{IHC} and E_{SCF}) are positively oriented, where a skill elasticity of zero is obtained when the predictability source has no influence on the skill of the streamflow forecast, while positive (negative) elasticities mean that an improvement in the predictability source will lead to higher (lower) streamflow forecast skill. The skill elasticities were calculated for all three methods and for the 3-month streamflow forecasts produced for each catchment and forecast initialization date.

The only difference between Eqs. (4) and (5) and the skill elasticities calculated in W16 is that they chose to calculate skill elasticities around the ESP point in the skill surface plots. Here, we choose to calculate skill elasticities across a quadrant within the skill surface plot (between 75% and 19% of the climatological variance explained in the IHC and the SCF) in order for the skill elasticity values calculated in this paper to reflect the forecast skill gradients within the response surface. This is done differently to W16 because the aim of this paper is to compare (qualitatively and quantitatively) the skill surface plots obtained from the EPI and the EPB methods to the VESPA approach.

4.2.2.3.2 What is the sensitivity of the response surface to the choice of the verification score?

To investigate the second objective of this paper, several verification scores were calculated for each method (i.e. the EPI, the EPB, and the VESPA approach). These scores were selected in order to cover key attributes of the forecasts verified, and they include

- the mean absolute error (MAE) of forecast ensemble means, relative to the perfect forecasts and
- the continuous rank probability score (CRPS) and its decomposition:

- the potential CRPS (CRPS_{spot}), where CRPS_{spot} = resolution - uncertainty, and
- the reliability part of the CRPS (CRPS_{reli}).

The potential CRPS is the CRPS value that a forecast with perfect reliability would have. The uncertainty is the variability of the observations and the resolution is the ability of the forecast to distinguish situations with distinctly different frequencies of occurrence. The CRPS reliability is a measure of the bias and the spread of the system.

The CRPS was chosen as it is a widely used score to assess the overall quality of an ensemble hydrometeorological forecast. The CRPS moreover has the advantage that it can be decomposed into different scores in order to look at the many different attributes of an ensemble forecast. The CRPS for a single forecast is equivalent to the MAE, which is why the latter was chosen.

For all of the above verification scores, the corresponding skill scores were calculated for each point of the skill surface plots from

$$\text{skill score}_{\text{forecast}} = 1 - \frac{\text{score}_{\text{forecast}}}{\text{score}_{\text{reference}}} \quad (6)$$

where the $\text{score}_{\text{reference}}$ is the score of the climatology point, for each method, each initialization date, and each catchment. A perfect forecast results in a forecast skill score of unity and a forecast with equal quality as the reference forecast corresponds to a skill score of zero. Any forecasts of lower quality than the reference forecast produce negative skill score values. Skill scores were calculated in order to have a similar score range as the R^2 (i.e. a climatological score of zero and a perfect score of one), for comparative purposes.

Skill elasticities were subsequently calculated for all the skill scores, using Eqs. (4) and (5), for all three methods and for the 3-month streamflow forecasts produced for each catchment and forecast initialization date. From these skill elasticity values, the influence of improvements in the IHCs and SCFs on the seasonal streamflow forecast skill can be assessed, in terms of the forecasts' overall performance (considering the mean of the ensemble or the full ensemble spread, from the MAE and the CRPS, respectively), their resolution and uncertainty (CRPS_{spot}), and their reliability (CRPS_{reli}).

4.2.3 Results

4.2.3.1 Can EPI and EPB discriminate the influence of IHC and SCF errors on seasonal streamflow forecast uncertainties?

For the first part of this study, the Crystal River (Colorado; USGS gauge 009081600), a snowmelt-driven catchment, will be used as a test case to illustrate the skill surface plots obtained from the EPI and the EPB methods, compared to the VESPA approach. Precipitation is the highest in winter and spring in this catchment and falls as snow between November and April. In April, the snow starts melting and consequently the soil moisture and streamflow both increase.

Figure 4 displays the skill surface plots obtained for the VESPA (Fig. 4a), the linear EPI (Fig. 4b), and the EPB methods (Fig. 4c), from R^2 for the 3-month streamflow forecast for the Crystal River, for initializations on the first of each month (each row in Fig. 4).

Figures 4d and 4e show the differences between the skill surface plots obtained for the VESPA and EPI methods and the VESPA and EPB methods, respectively. A first visual comparison of the skill surface plots obtained from the linear EPI method (Fig. 4b) and the EPB method (Fig. 4c) with those obtained from the VESPA approach (Fig. 4a) for the Crystal River tells us that the skill surface plots obtained from all three methods are very similar. For each initialization date, the orientation of the gradients in streamflow forecast skill appears identical. The EPI and the EPB methods seem to correctly indicate the dominant predictability source on the 3-month streamflow forecast skill, for each initialization date for this catchment. Similar results were obtained for the other 17 catchments (see Figs. S1–S17 in the supplemental material in thesis Appendix A6). Forecasts made on the first of February, March, and September show a sensitivity to the SCF skill (i.e. horizontal or near to horizontal orientation of the streamflow forecast skill gradients), while all other forecasts are dominantly sensitive to the IHC skill (i.e. vertical or near to vertical orientation of the streamflow forecast skill gradients).

The gradients in streamflow forecast skill contained in the EPI skill surface plots (Fig. 4b) differ moderately from the gradients obtained from the VESPA approach (Fig. 4a). This can be observed in Fig. 4d, showing the differences between the skill surface plots obtained for both methods. The VESPA approach gives very strong gradients, causing a rapid decrease in streamflow forecast skill with a decrease in one of the predictability sources' skill, depending on the initialization date. In comparison, the EPI method indicates a gradual decrease in streamflow forecast skill with a decrease in one of the two predictability sources, depending on the initialization date. The streamflow forecast skill gradients produced by the EPI method are a reflection of the interpolation method used (i.e. here linear), and because the corner points lack information about describing curvature of the surface at interior points, they cannot fully capture nonlinearities in the skill gradients

across the skill surface. For some interior points, this limitation of the EPI method could estimate very different skill elasticities than those obtained from the VESPA approach.

The skill surface plots produced by the EPB method (Fig. 4c) show minor differences in the streamflow forecast skill gradients when compared to the skill surface plots generated by the VESPA approach (Fig. 4a). This can be seen in Fig. 4e, which shows the differences between the skill surface plots obtained for both methods. To further inspect those differences, they will be explored quantitatively (i.e. by comparing the skill elasticities) below.

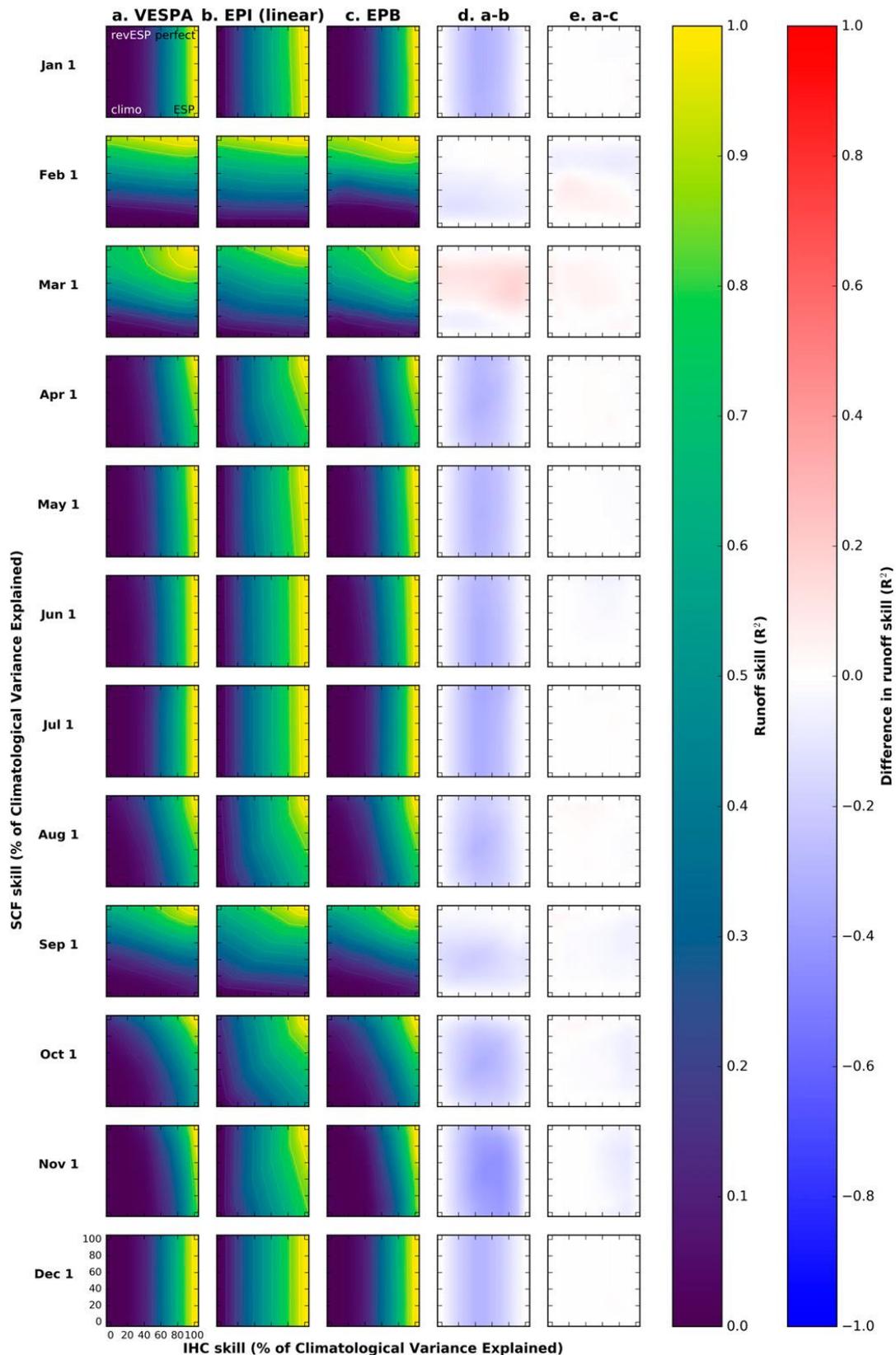


Figure 4. Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R^2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for the Crystal River (USGS gauge 009081600), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.

To quantify the accuracy of the patterns contained in the EPI and the EPB skill surface plots compared to the patterns of the VESPA skill surface plots, IHC and SCF skill elasticities (i.e. E_{IHC} and E_{SCF} , respectively) were calculated across a quadrant situated within the response surface for all three methods, for the 18 catchments and each forecast initialization date, from Eqs. (4) and (5), respectively. Figure 5 presents the skill elasticities for nine of the 18 catchments (the plots for the other nine catchments are shown in Fig. S18, in the supplemental material in thesis Appendix A6). Each plot corresponds to a catchment and shows the skill elasticities obtained from the VESPA, the linear EPI, and the EPB methods as a function of the forecast initialization date. From the nine different plots, the skill elasticities given by the EPB method appear almost identical to the VESPA approach, whereas the skill elasticities obtained from the EPI method differ in some places. This confirms that the patterns of the EPB method are very similar to the patterns of the VESPA approach, with it being the closest out of the two tested methods.

The value of the SCF skill elasticity (i.e. E_{SCF}) in relation to the value of the IHC skill elasticity (i.e. E_{IHC}), for a given method, indicates the dominant predictability source on the 3-month streamflow forecast skill (here calculated from the R^2). For a selected method, equal SCF and IHC skill elasticity values signifies that equal improvements in both the SCFs and the IHCs will lead to equal improvements in the streamflow forecast skill. If E_{SCF} is superior (inferior) to E_{IHC} , it reflects a larger potential increase in streamflow forecast skill by improving the SCFs (IHCs). Although the EPI method almost always indicates the same dominant predictability source as the two other methods, the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill (i.e. the exact values of the skill elasticities) often differs. For many catchments and forecast initialization dates, the EPI appears to underestimate the skill elasticities produced by the VESPA method.

The nine different catchments for which the skill elasticities are presented in Fig. 5 display three different types of behavior, best captured by the VESPA approach and the EPB method. For the three catchments in Fig. 5 (left), improvements in the IHCs would yield the highest improvements in the 3-month streamflow forecast skill for spring to summer initializations (April–August for the Crystal River, March–July for the Fish River, and March–June for the Middle Branch Escanaba River) and in the winter (October–January for the Crystal River, November–December for the Fish River, and in December for the Middle Branch Escanaba River). SCF improvements would lead to better 3-month streamflow forecast skill for forecasts initialized in the late winter and summer to fall (February–March and September for the Crystal River, February and August–October for the Fish River, and January–February and July–September for the Middle Branch Escanaba River). For the

three catchments in Fig. 5 (middle), a notable feature is that the 3-month streamflow forecast skill would benefit from SCF improvements for summer initializations (June–September for the Chattooga and the Nantahala Rivers and July–September for the New River). Finally, for the three catchments in Fig. 5 (right), the 3-month streamflow forecast skill would benefit from improvements in the SCFs for all initialization dates. This is true with the exception of forecasts initialized in December for East Fork Shoal Creek. It is important to note that there is uncertainty around these estimates. However, this is a good first indication of the sensitivity of 3-month streamflow forecast skill (measured from the R^2) to IHC and SCF errors for each forecast initialization date and each catchment.

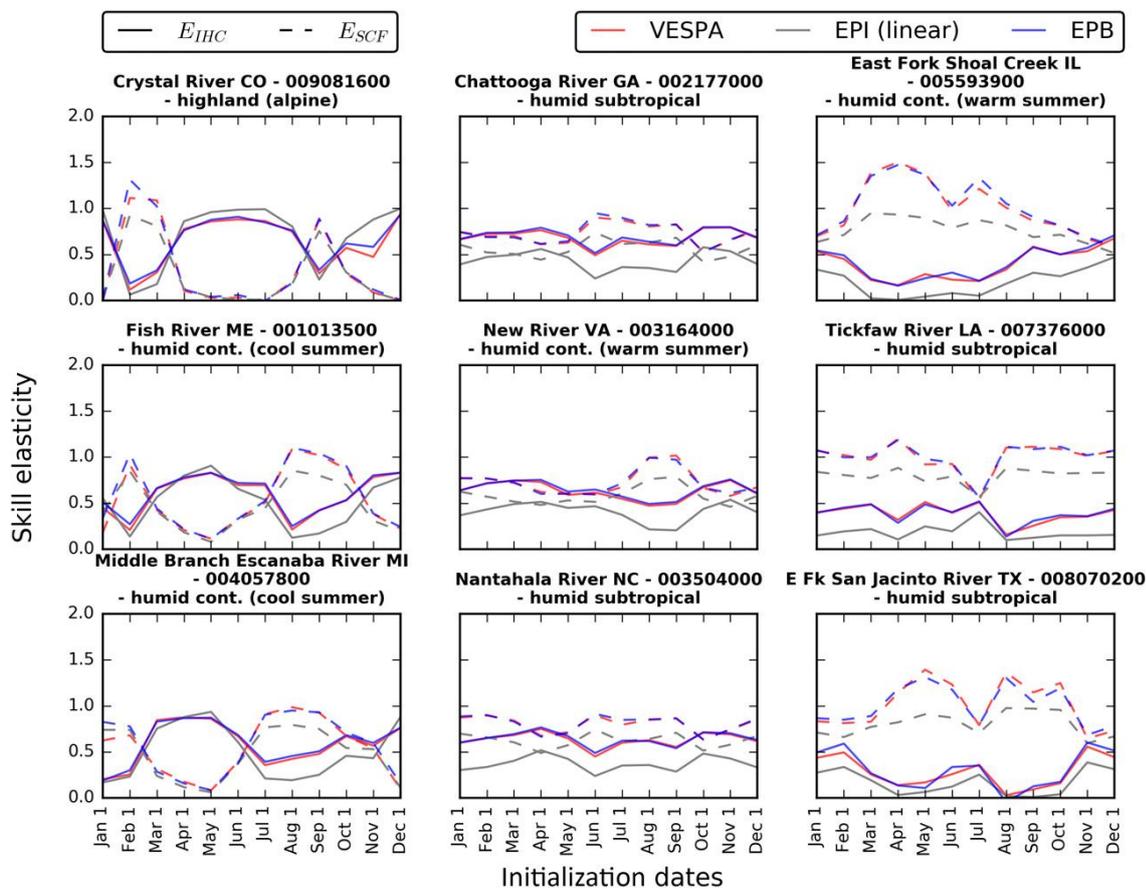


Figure 5. Streamflow forecast skill elasticities for the IHCs (i.e. E_{IHC} , solid line) and the SCFs (i.e. E_{SCF} , dashed line), calculated across a quadrant situated within the 3-month streamflow forecast skill surface plots for the VESPA (red), the linear EPI method (gray), and the EPB method [blue; using Eqs. (4) and (5)]. Each plot shows the evolution of the IHC and SCF skill elasticities with the initialization date for a given catchment. The climatological regions of the catchments are indicated in the plots' headings. The skill surface plots from which these skill elasticities were calculated are presented in Fig. 4 and Figs. S1–S17.

The skill elasticities produced by the EPB method appear to be almost identical to the skill elasticities obtained from the VESPA approach, with occasional marginal differences. This suggests that the EPB method captures nearly exactly the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill, obtained from the VESPA approach. Both

methods additionally indicate the same dominant predictability source: the predictability source which, once improved, could lead to the largest increase in 3-month streamflow forecast skill. The EPB method will therefore be used as an alternative to the VESPA approach to investigate the second objective of this paper.

4.2.3.2 What is the sensitivity of the response surface to the choice of the verification score?

To investigate the sensitivity of the response surface to the choice of the verification score, and therefore to the attribute of the forecast, several scores were computed to evaluate the streamflow forecast quality. The R^2 , the mean absolute error skill score (MAESS), and the continuous rank probability skill score (CRPSS) were calculated to evaluate the forecasts' overall performance in terms of the ensemble mean and the entire ensemble. The potential CRPSS (CRPSSpot) was computed to look at the forecasts' resolution and uncertainty, and the CRPSS reliability (CRPSSreli) was computed to look at the forecasts' reliability. The Crystal River (USGS gauge 009081600) will here again be used as a test case to illustrate this part of the results.

Figure 6 presents the IHC and SCF skill elasticities [i.e. E_{IHC} and E_{SCF} ; in Fig. 6 (top) and Fig. 6 (bottom), respectively] as a function of forecast initialization date for the Crystal River catchment. These are calculated from Eqs. (4) and (5), for all the mentioned verification scores, for the VESPA approach (Fig. 6a) and the EPB method (Fig. 6b). If we compare the skill elasticities obtained from the VESPA approach with the skill elasticities obtained from the EPB method, it appears that both methods produce very similar elasticities for the R^2 , the MAESS, and the CRPSS. This further confirms the results of the first part of the analysis, which highlighted the similarity of the EPB results to the VESPA results and extends it to multiple attributes of the seasonal streamflow forecasts. However, slight differences between the skill elasticities produced by the two methods can be observed for the CRPSSpot, and significant differences exist for the CRPSSreli. These dissimilarities are discussed further below.

If we now compare the skill elasticities obtained for the various verification scores for both methods, it is clear that the R^2 , the MAESS, the CRPSS, and the CRPSSpot give very similar skill elasticities. This hints that those verification scores overall agree on the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill. However, a few dissimilarities can be observed for some of the forecast initialization dates. This is, for example, the case for forecasts made in the spring and in summer, where the E_{IHC} appears lower for the MAESS and the CRPSS (and the CRPSSpot for the VESPA approach) compared to the E_{IHC} obtained for the R^2 for both methods. It is also apparent for forecasts made on

the first of February, March, and September, where the E_{SCF} calculated for the MAESS and the CRPSS (and the CRPSSpot for the VESPA approach) is lower than the E_{SCF} obtained for the R^2 for both methods. For both examples, it infers that improvements in the IHC and the SCF skill could lead to larger improvements in the streamflow forecast skill in terms of the R^2 rather than in terms of the MAESS and the CRPSS (and the CRPSSpot for the VESPA approach). Overall, this indicates that the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill differs relative to the choice of the verification score.

While the R^2 , the MAESS, the CRPSS, and the CRPSSpot give a very similar picture, the skill elasticities obtained for the CRPSSreli appear very different, occasionally reaching negative values. These negative values indicate a loss in streamflow forecast skill (in terms of the forecast reliability) as a result of improvements in one of the two predictability sources, while all the other verification scores suggest a gain in streamflow forecast skill (in terms of the forecast ensemble mean and the ensemble overall performance, its resolution, and uncertainty) with improvements in one of the two predictability sources.

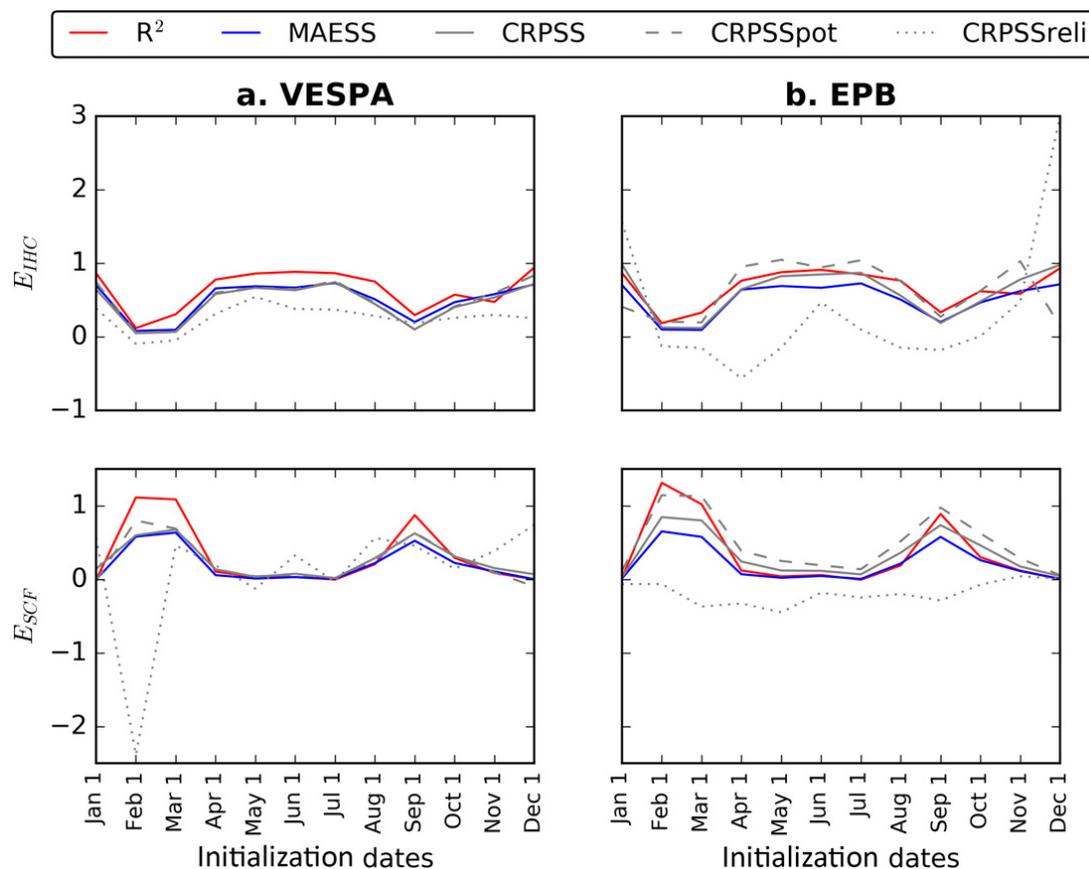


Figure 6. Streamflow forecast skill elasticities for the (top) IHCs (i.e. E_{IHC}) and (bottom) SCFs (i.e. E_{SCF}) as a function of forecast initialization dates, for hindcasts produced from 1981 to 2010 for the Crystal River (USGS gauge 009081600). These skill elasticities were calculated across a quadrant situated within the 3-month streamflow forecast skill surface plots [from Eqs. (4) and (5)] for several verification scores (R^2 in red, MAESS in blue, CRPSS in gray solid line, CRPSSpot in gray dashed line, and CRPSSreli in gray dotted line). The results are shown for (a) the VESPA approach and (b) the EPB method.

The substantial differences in skill elasticities obtained for the CRPSSreli from the VESPA versus EPB method suggest that there are limitations to the ability of EPB to reconstruct the full ensemble information present in VESPA, and of VESPA (applied with relatively small ensembles at the end points) to estimate sensitivities for complex verification scores such as reliability. The reliability verification score is influenced by the combination of bias, spread, and other ensemble properties and exhibits more noisy outcomes here than were obtained for other verification scores. A negative elasticity may occur because the ensemble spread has narrowed without sufficient improvements in bias, for instance. The behavior of the elasticity of reliabilities is even more difficult to diagnose, but we suspect that the presence of noise (erroneous local minima or maxima) or curvature in the associated VESPA skill surface greatly undermines the linear blending techniques.

Overall, these results suggest that improvements in the skill of either of the two predictability sources will impact streamflow forecast skill differently depending on the attribute (i.e. verification score) of the forecast skill that is considered and whether the ensemble mean or the full ensemble is used.

4.2.4 Discussion

4.2.4.1 Implications and limitations of the results

W16 introduced the VESPA approach, a sensitivity analysis technique used to pinpoint the dominant predictability source of seasonal streamflow forecasting (i.e. the IHCs and the SCFs), as well as quantifying improvements that can be expected in seasonal streamflow forecast skill as a result of realistic improvements in those key predictability sources. Despite being a powerful sensitivity analysis approach, VESPA presents two key limitations.

- 1) It is computationally intensive, requiring multiple ensemble hindcasts to define the skill response surface (81 were used in the VESPA paper vs one for the EPB and the EPI techniques).
- 2) It requires a complex state and forcing blending procedure that may introduce additional uncertainties, biases, or interactions between the predictability sources (Saltelli et al. 2004; Baroni and Tarantola 2014) that are not accounted for or difficult to quantify. This is not necessary in any of the end points required in the two approaches presented here, which rely instead on analyzing a single conventional hindcast dataset that is more likely to be feasible for forecasting centers.

The central aim of this paper was to address the first limitation of the VESPA approach by presenting two computationally inexpensive alternative methods: the EPI and the EPB methods. Both methods successfully identified the dominant predictability source of 3-month streamflow forecasts for a given catchment and forecast initialization date (i.e. given by the orientation of the streamflow forecast skill gradients in the skill surface plots). However, the EPB was more successful in reproducing the VESPA skill elasticities—the exact streamflow forecast skill gradients situated within the skill surface plots (for skill and accuracy verification scores including the R^2 , the MAESS, the CRPSS, and the potential CRPSS to a certain extent). These skill elasticities indicate the influence of changes in IHC and SCF skill on streamflow forecast skill.

The new methods, by differing in their setup from the VESPA approach, do not inherit the drawbacks specific to this approach and mentioned above. The EPI and the EPB methods nevertheless have their own limitations.

The EPI (both for the linear and cubic interpolation methods; the latter was not shown) did not fully capture the VESPA skill elasticities because of the nature of the method that produces predefined gradients within the skill surface plots (i.e. defined by the interpolation method used). Additionally, curvature or local minima or maxima (if any) of the response surface cannot be represented by the EPI method. The EPB, on the other hand, performs better at reflecting curvature in the skill response surface, hence local elasticities between the end points. The EPB method aimed at reproducing VESPA elasticities only by manipulating the output of a single hindcast dataset (interpreted as ESP, reverse-ESP, the perfect forecast, and climatology). The EPB method cannot match exactly the forecasts created by the VESPA approach, as it does not account for the idiosyncrasies in model forecast behavior, such as interactions between the predictability sources. Furthermore, it is likely that the more the model investigated is nonlinear or exhibits skill response thresholds, the more the results obtained from the EPB method will differ from the ones obtained from the VESPA approach. These results overall allow that the EPB method can be used as an inexpensive alternative method to the VESPA approach, yet with the potential limitations of the method stated above.

For the first part of the analysis, the streamflow forecast quality was evaluated in terms of the forecasts' skill from the R^2 . The use of multiple verification scores is, however, essential to obtain a more complete perspective of forecast quality. Thus, we explored the performance of the two new methods and the VESPA approach for a range of additional verification scores. The results, presented for the EPB method and the VESPA approach,

showed differences in the response surfaces obtained for the various verification scores (i.e. the R^2 , the MAESS, the CRPSS, and its decomposition). This suggests distinct sensitivities of the seasonal streamflow forecast attributes (i.e. overall performance of the forecast ensemble mean and its full ensemble, forecast resolution, uncertainty, and reliability) to changes in the IHC and SCF skill. Ideally, a sensitivity analysis should be goal oriented, that is, it should be performed with prior knowledge of the intended use of the results (Saltelli et al. 2004; Pappenberger et al. 2010; Baroni and Tarantola 2014), which may favor using one verification score over another.

This paper covered selected limitations of the work presented by W16. However, many areas were left unexplored and could be interesting topics in which to focus future research. First, a major area inherent to model-based sensitivity analyses is that their results are model dependent (Saltelli et al. 2000); thus, the extent to which they can be transferred to reality depends on the model fidelity. The results presented in this paper are specific to the forecasting system and similar systems on which this analysis was based and should be used as an indicator of catchment sensitivities. As noted in W16, an extension of the elasticity analysis to include observations and a model error component would provide valuable insights. Another possible approach could be to use the results from various forecasting systems as input to the sensitivity analysis, in order to achieve a multimodel consensus view of the skill. As shown in Cloke et al. (2017), a multimodel forcing framework can be highly beneficial for streamflow forecasting compared to a single model forecasting approach, provided the models are chosen judiciously so as to provide a rational characterization of forecasting uncertainty. Second, the dependence of blending technique performance versus VESPA on the characteristics of the skill surface (e.g. linear or nonlinear) bears further investigation. Finally, this sensitivity analysis leaves generic the concept of improvements in either of the predictability sources. Breaking down the IHC and the SCF into the different variables they represent is a point which merits further research. Additionally, the space-time nature of improvements may be consequential. This work could therefore be extended by studying the effect of degradations in the temporal and spatial accuracy of the input data, thereby indicating the relative value of improvements in the spatial or temporal predictability of a given variable, for a specific catchment and a specific time of the year.

4.2.4.2 The wider context

The new strategy of operational forecasting centers is to move toward more integrated operational modeling and forecasting approaches, such as land surface-atmosphere

coupled systems, and beyond that, Earth system models. These advances are enabled by the continuous growth of computing capabilities, a better understanding of physical processes and their interactions throughout all compartments of the Earth system, and the availability and use of more and better observation data (i.e. satellite data). Despite all these advances, most forecasts still reflect substantial uncertainty that grows with time and limits the predictability of observed events beyond a few weeks of lead time. The rapid progress has led our systems to be ever more data hungry as increases in model complexity and resolution are sought. These computationally expensive developments are not always feasible; hence, model developers must be creative and constantly weigh the costs and benefits of improving one aspect over another, such as increasing the resolution or complexity of the models (Flato 2011).

In this context, sensitivity analyses appear more than ever as a natural tool to establish priorities in improving predictions based on Earth system modeling. Such analyses are a powerful and valuable tool to support the examination of uncertainty and predictability across spatial and temporal scales and for various applications. They can be used for a large range of activities, including examining model structure, identifying minimum data standards, establishing priorities for updating forecasting systems, designing field campaigns, and providing realistic insights into the potential benefits of efforts to improve a forecasting system to managers with prior knowledge of their costs (Cloke et al. 2008; Lilburne and Tarantola 2009; W16).

However, sensitivity analyses must be easily reproducible to be effective in supporting each new model or forecast system update, and the results should easily be applied in order to constitute a “continuous learning process” (Baroni and Tarantola 2014).

Furthermore, climate change is likely to generate changes in the drivers of seasonal predictability with time (e.g. recent research has shown that climate change is likely to increase the frequency of extreme El Niño events; Cai et al., 2015). This further strengthens the need for easily reproducible sensitivity analyses to guide seasonal forecast developments in a changing climate.

As we balance the costs and benefits of forecast development, the wider context of the societal use of these forecasts must not be ignored. Indeed, an infinitely better hydro-meteorological forecast for a phenomenon with no societal impacts is not necessarily worthy of the development costs they incurred. With this in mind, sensitivity analyses must be user-oriented for guiding forecast developments of benefit to society.

In conclusion, a sensitivity analysis should be a simple, tractable tool for addressing a multifaceted challenge.

4.2.5 Conclusions

This paper presents two computationally inexpensive alternative methods to the VESPA approach for estimating forecast skill sensitivities and elasticities. Of these, the end point blending (EPB) method provides a useful substitute to the VESPA approach. Despite the existence of some differences between the EPB and the VESPA outcomes, the EPB successfully identifies the dominant predictability source (i.e. IHCs and SCFs) of seasonal streamflow forecast skill, for a given catchment and forecast initialization date. The EPB method can additionally reproduce the VESPA forecast skill elasticities, indicating the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill. The paper also draws attention to how the choice of verification score impacts the forecast's sensitivity to improvements made to the predictability sources. With a good understanding of the limitations of the methods, such a sensitivity analysis approach can be a valuable tool to guide future forecasting and modeling developments.

Acknowledgments. L. Arnal, A. W. Wood, H. L. Cloke, and F. Pappenberger gratefully acknowledge financial support from the Horizon 2020 IMPREX project (Grant Agreement 641811) (project IMPREX: www.imprex.eu). E. Stephens' time was funded by the Leverhulme Early Career Fellowship ECF-2013-492. We also acknowledge high-performance computing support from Yellowstone ([ark:/85065/d7wd3xhc](https://doi.org/10.7554/ark:/85065/d7wd3xhc)) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. Last, A. W. Wood is thankful for support from the U.S. Bureau of Reclamation under Cooperative Agreement R11AC80816 and from the U.S. Army Corps of Engineers (USACE) Climate Preparedness and Resilience Program (Award Number 1254557).

4.3 Scopes for improving EFAS-Seasonal

The EPB sensitivity analysis method presented in Sect. 4.2 was subsequently applied to EFAS-Seasonal (which performance over Europe was benchmarked in Chapter 3), as suggested in the future work section (Chapter 3, Sect. 3.2). The results were discussed in an IMPREX deliverable, of which this section presents a summary.

This IMPREX deliverable was led by L.A. and arose through collaboration during this PhD. It has the following reference:

Arnal, L. et al.: “IMPRES D4.2 - The sensitivity of sub-seasonal to seasonal streamflow forecasts to meteorological forcing quality, modelled hydrology and the initial hydrological conditions”. Deliverable of EU H2020 project “IMPRES – Improving predictions and management of hydrological extremes” (contract n° 641811), 2017a*

L.A. collaborated on laying out the structure of the deliverable, gathered the information and results from IMPRES partners, contributed parts of the results, analysed and interpreted most of the results and wrote about 75% of the document.

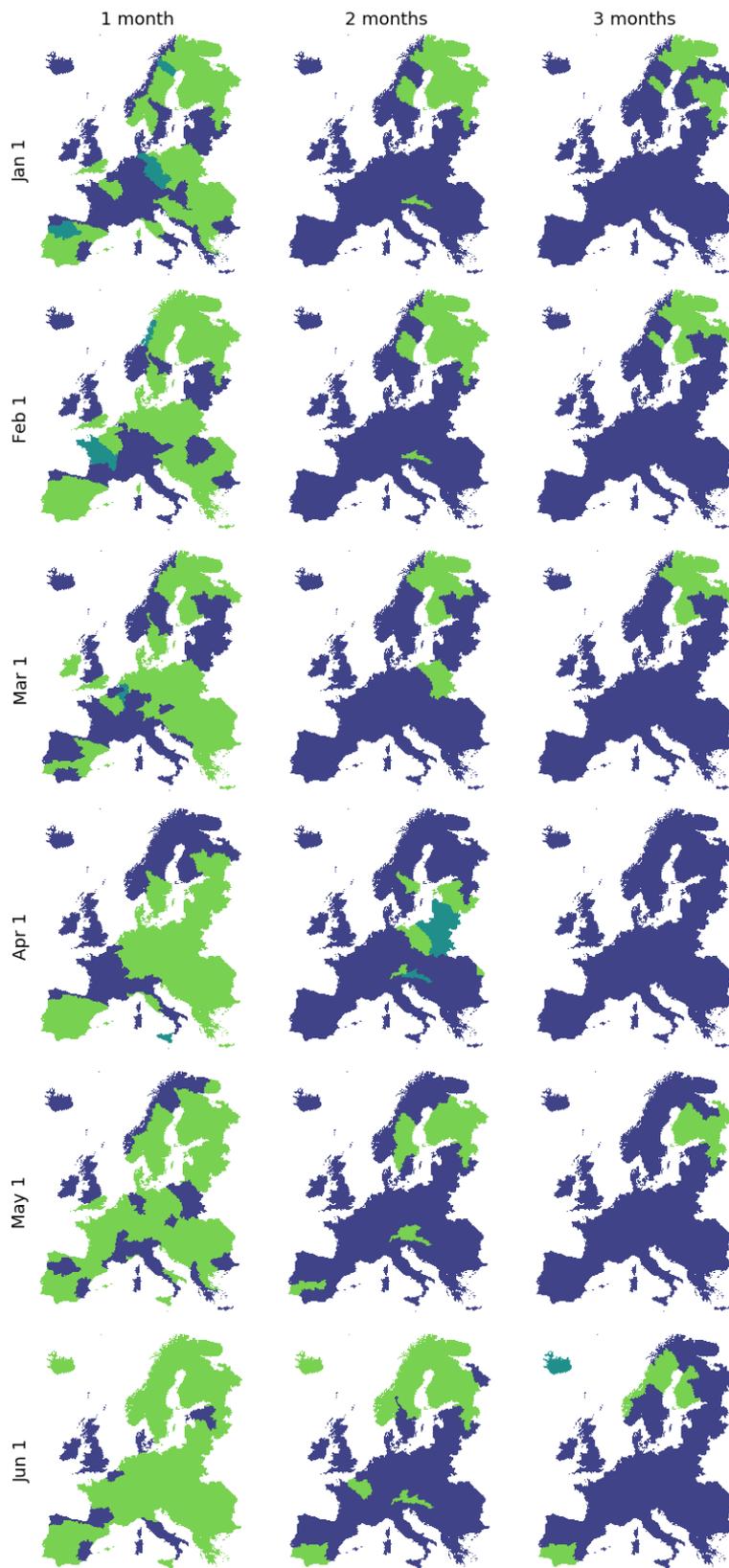
The full deliverable can be found in the thesis Appendix A7.

This deliverable is structured in three parts. The first part is an intercomparison of the performance of five seasonal hydro-meteorological forecasting systems (from IMPRES partners) over Europe. All streamflow forecasts were produced using the same seasonal meteorological forecasts (ECMWF’s System 4, with or without post-processing), but a different hydrological model. This intercomparison revealed that differences in the forecasting systems’ set up (in terms of the presence of post-processing and the hydrological model used) leads to large differences in their seasonal streamflow forecast performance. This in turn influences the systems’ strengths and weaknesses for several applications of the water sector (such as flood early warning, navigation and agriculture).

The second part of this deliverable is the application of the EPB sensitivity analysis method to EFAS-Seasonal and the German Federal Institute of Hydrology (BfG)’s seasonal streamflow forecasts, based on the forecasts’ CRPSS. The two forecasting systems displayed very similar results to one another; the rest of this summary focuses on the EFAS-Seasonal sensitivity analysis results (see Fig. 7). Overall, results suggested that, for the first month of lead time, the EFAS seasonal streamflow forecasts would more largely benefit from improvements to the initial hydrological conditions. After the first month of lead time, improving the seasonal meteorological forecasts would yield a higher seasonal streamflow forecast improvement. Results however varied by European region and forecast starting date, highlighting spatio-temporal differences in the seasonal streamflow generating mechanisms. Below is an overview of some of the key findings. For forecasts made in the summer (May to July), a larger number of European regions exhibited a streamflow sensitive to improving the initial hydrological conditions for the first month of lead time (compared to forecasts made in the winter). This could be due to the lower rainfall over Europe during the summer and groundwater-driven streamflow during those months. For

* This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 641811.

leeward Scandinavia, the initial hydrological conditions were found to dominate seasonal streamflow forecast quality for forecasts made in the winter through to the spring with up to three months of lead time. This is likely due to precipitation falling as snow during the winter, leading to groundwater-driven streamflow in the winter and snowmelt-driven streamflow in the following spring. For most of the Iberian Peninsula, a higher sensitivity to initial hydrological conditions was found for forecasts made in the summer (June to September) with up to three months of lead time. This might be caused by very low precipitation and groundwater-driven streamflow during those months.



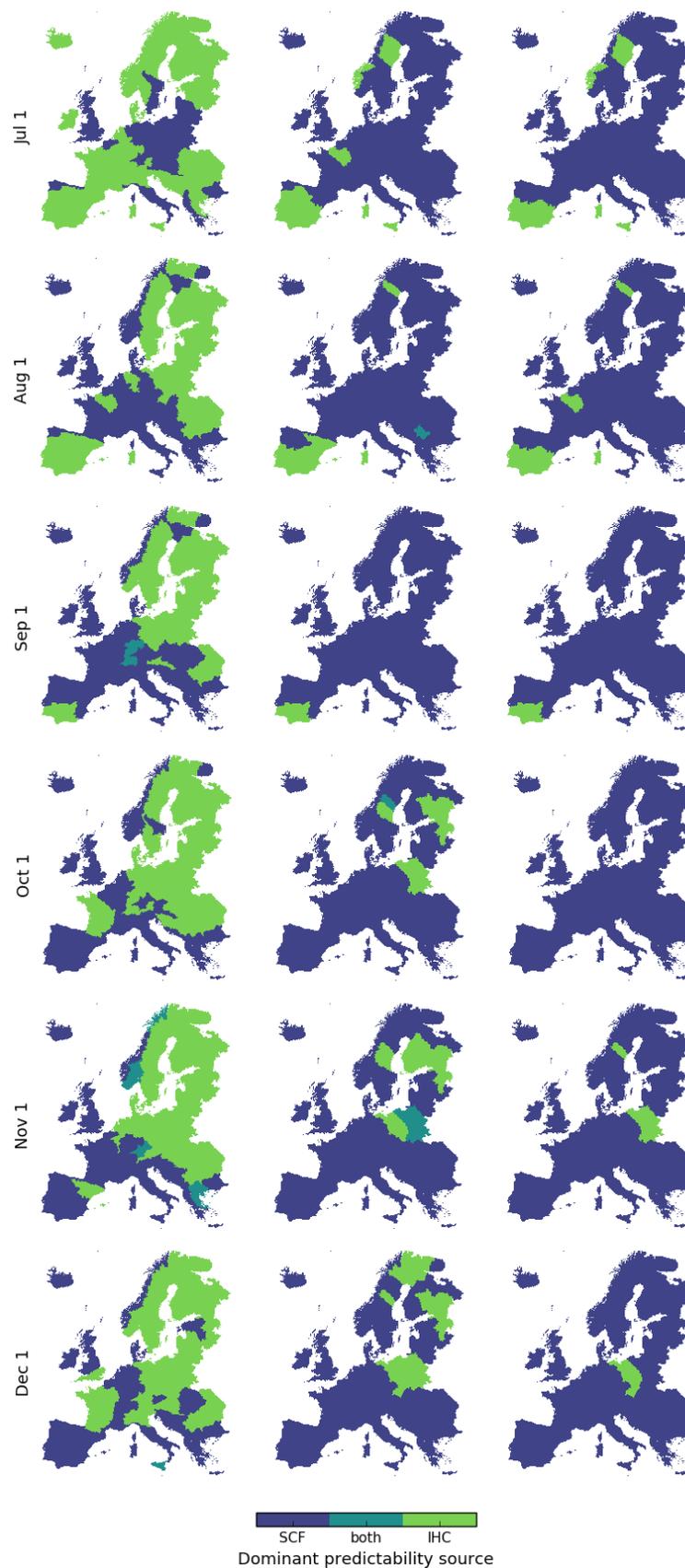


Figure 7. Maps of the EFAS-Seasonal streamflow forecasts' sensitivity to the seasonal meteorological forecast (SCF; in blue), the initial hydrological conditions (IHC; in green), or both (in blueish-green). The results are shown for each forecast starting date (rows) and for the first three months of lead time (columns). (Figure taken from IMPREX D4.2).

The third part of this deliverable aimed to identify key drivers (i.e. basins' physiographic-hydroclimatic characteristics) of seasonal hydrological forecasting skill over Europe (based on the Swedish Hydro-Meteorological Institute (SMHI)'s E-HYPE forecasting system). Results showed that seasonal hydrological forecast skill mainly depends on the basin's hydrological regime and is limited for flashy river basins (i.e. basins with strong flow dynamics over the year and less land surface memory). Basins' elevation and temperature biases were also identified to be important factors.

This deliverable provided a benchmark of state-of-the-art seasonal hydro-meteorological forecasting systems' performance over Europe. Additionally, it identified drivers of seasonal hydrological predictability and their impacts for a range of water sectors.

4.4 The forecasting paradox: do better computing resources make us worse forecasters?

This section introduces a novel concept for seasonal streamflow forecasting, presented at the European Geoscience Union (EGU) General Assembly 2017. The original poster can be found in the thesis Appendix A8.

4.4.1 Introduction

Our scientific understanding of hydro-meteorological predictability has improved considerably over the last decades. This has led to more diverse and improved operational hydrological forecasting methods. The coincidental growth in computer power over the years has rendered these improved methods technically implementable. Taking seasonal streamflow forecasting as an example, there now exist a wide range of operational forecasting methods, from statistical to dynamical, and a blend of both methods (see Chapter 3, Sect. 3.1).

However, as operational hydrological forecasting has improved and diversified, a given system will usually run a unique forecasting method in space and time. The choice of the forecasting method is often made on the basis that it is: i) scientifically valid and overall more skilful than a benchmark (for seasonal streamflow forecasting, this is generally the streamflow climatology or an ensemble streamflow prediction (ESP)) and ii) technically and financially implementable and sustainable (contributing to a reliable system, able to generate forecasts operationally with no delays). These criteria reflect the resources available at the time the forecasting method is implemented. Often, a forecasting method, given that it meets the above criteria, will be selected for operational implementation

because it is already running in research mode in-house and/or easy to implement as all or most parts of the forecasting chain are easily or readily available.

In addition, given the constant pressure faced by operational centres to demonstrate that they are advancing science, operational development might be driven by the implementation of 'low hanging fruits'. These are plentiful and might include: increased representation of uncertainty with more ensemble members, higher resolution forecasts, longer lead times and the incorporation of an increasing amount of observation data. While these relatively easy gains might provide slight forecast improvements, they do not necessarily translate the complexity of our scientific findings into operations.

As a result, the selected operational forecasting method might not be the most suitable (i.e. scientifically valid, useful for the community, implementable, sustainable and reliable) in all space and time dimensions, given the spatio-temporal variability of hydro-meteorological forecast predictability. This might in turn lead to a low forecast performance to generation cost ratio for certain river basins, temporal aggregations and forecast starting dates, where alternative (sometimes cheaper) forecasting methods could be more suitable. We argue that the variability of predictability sources in space and time should be reflected in the way we produce hydro-meteorological forecasts operationally (not only on seasonal timescales).

4.4.2 A novel concept for improved seasonal streamflow forecasting

Here, we introduce a novel concept for operational seasonal streamflow forecasting. A 'flexible seasonal hydro-meteorological forecasting method' with the potential to produce forecasts that are more skilful, cheaper and faster to run operationally, compared to a unique forecasting method. This concept builds on the work presented in Chapters 3 and 4.

Following this novel concept, the operational method adopted to produce each new seasonal streamflow forecast (for a given forecast starting date, river basin and output aggregation period and lead time) is chosen based on the decision tree illustrated in Fig. 8. Results of the EPB sensitivity analysis (performed beforehand; see Sect. 4.2) are first analysed. If they indicate that improving the initial hydrological conditions would lead to higher seasonal streamflow forecast improvements, the green branch of Fig. 8 is subsequently followed. In a second step, results of the comparison between the performance of seasonal streamflow forecasts produced 1) using seasonal meteorological forecasts as forcings (e.g. EFAS-Seasonal, 'EFAS-Seas.' on Fig. 8) and 2) using historical meteorological observations as forcings (e.g. ESP) are investigated (performed beforehand;

see Chapter 3, Sect. 3.2). The final operational method used is the one presenting the highest overall seasonal streamflow forecast performance, to which a data assimilation (DA) method is applied to improve the initial hydrological conditions. Going back a few steps, if improving the seasonal meteorological forcings would yield higher seasonal streamflow forecast improvements, the blue branch of Fig. 8 is subsequently followed. Looking at the comparison between EFAS-Seasonal and the ESP (in this example), the forecast with the highest overall performance is used operationally, with post-processing of the meteorological forcing to improve their quality.

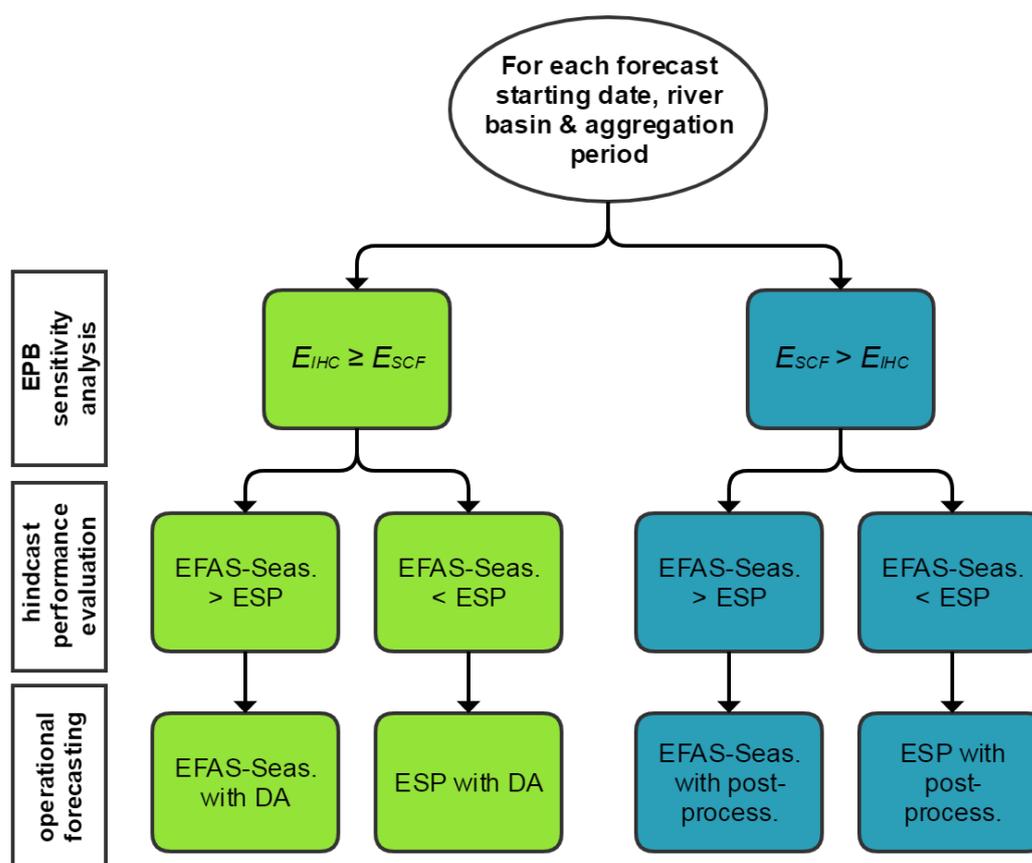


Figure 8. Decision tree of the ‘flexible seasonal hydro-meteorological forecasting’ novel concept, applied to EFAS-Seasonal.

4.4.3 Discussion

This flexible system balances the costs of operational forecast production (e.g. buying seasonal meteorological forecasts, running a data assimilation and a meteorological forecast post-processing) with forecast performance (knowledge gained from running an EPB sensitivity analysis and a hindcasts’ performance analysis). It is based on forecasts usually readily available at forecasting centres running seasonal hydrological forecasts and

based on methods which are easily implementable (i.e. EPB sensitivity analysis and hindcast performance analysis).

This novel concept builds on the flexible modelling approach proposed by Clark et al. (2015a; 2015b), a Structure for Unifying Multiple Modeling Alternatives (SUMMA), extending it to seasonal streamflow forecasting. It is different from a multi-modelling forecasting approach (whereby multiple hydrological models and/or seasonal meteorological forecasts are used), in that it first aims to understand the drivers of seasonal hydro-meteorological forecast predictability, based on which it subsequently proposes a cost-effective operational forecasting method. Contrastingly, multi-model forecasting aims to run all the models available operationally, to produce a relatively costly forecast that combines the various systems' strengths.

The main limitation of this novel concept is the fact that the flexible system might have to be set up differently for each streamflow temporal aggregation and lead time – i.e. it cannot run seamlessly in time. This limitation could be partly tackled if the decision tree is driven by decision-makers' requirements. Let us consider a decision-maker who needs a three-monthly streamflow average for reservoir operations in their river basin. They want an operational system which can provide reliable forecasts, especially for high streamflow events. Given these decision-making criteria, the flexible system could be tuned to produce three-monthly streamflow averages, based on the EPB sensitivity analysis and hindcasts evaluation performed for the ROC score for the streamflow upper tercile (see Chapter 3, Sect. 3.2.3.5; carrying out the EPB sensitivity analysis with the ROC score however first requires further research, as per the limitations mentioned in Chapter 4, Sect. 4.2). In other words, only one flexible system configuration would need to be run for this decision-maker for their specific application in the river basin. This novel concept could therefore be improved further by integrating decision-makers in the decision tree, for a coupled human-natural system approach (Li et al., 2017).

Seasonal streamflow forecasting and the EFAS-Seasonal operational system were used for illustrative purposes. However, this concept could be extended to a range of operational hydro-meteorological forecasting systems.

So far, we have touched on several important elements of a probabilistic hydro-meteorological forecasting chain. We discussed potential barriers and enablers of the use of probabilistic hydrological forecasts for flood early warning (Chapter 2). We then looked at the performance of state-of-the-art operational seasonal hydro-meteorological forecasting systems, with a particular focus on extending flood predictability on the seasonal timescale (Chapter 3). This chapter finally discussed methods to improve state-of-the-art seasonal forecasting systems.

However, as shown in Chapter 2, probabilistic forecast quality (and its improvement) is only one of the factors that influences their use in practice. To understand real-world opportunities and challenges associated with the uptake of state-of-the-art probabilistic flood forecasts in practice, the complex landscape in which decision-makers operate must be considered. Chapter 5 presents results from interviews at the UK Environment Agency (EA), carried out to understand the opportunities and challenges faced by the EA decision-makers in view of a transition to probabilistic forecasts for flood early warning in England.

Chapter 5

Using probabilistic hydrological forecasts for flood early warning: a real-life case study

5.1 Background and aim

In Chapter 2, insights were gained on the factors that shape decision-makers' use of probabilistic forecasts for flood early warning, through a serious game and decision-making activity. While serious games and activities can foster the uptake of serious information and promote discussions through their engaging formats, they are a simplification of reality (Aubert et al., 2019). Thus, the conclusions that can be drawn from these serious activities are to some extent partial and should be supported by further research.

Interviews can be an effective method to capture the full and complex landscape in which decision-makers operate in reality (Schoenberger, 1991; Pagano et al., 2004b). By providing interviewees with an understanding of the world from the perspective of the informants (i.e. decision-makers), they shed light on their unique perceptions (Sivle et al., 2014).

The Environment Agency (EA), responsible for managing risks of flooding in England, is in the process of a transition to operational probabilistic fluvial flood forecasts. With the aim to get a complete understanding of the real-life challenges and opportunities associated with flood early warning based on probabilistic forecasts, interviews were carried out with EA decision-makers. Based on these interviews and literature findings, recommendations were made to the EA to help a successful transition to probabilistic forecasts in operation. This chapter presents the following study.

5.2 “Are we talking just a bit of water out of bank? Or is it Armageddon?” Front line perspectives on transitioning to probabilistic fluvial flood forecasts in England

This section is in open discussion in Geoscience Communication (GC) with the following reference:

Arnal, L., L. Anspoks, S. Manson, J. Neumann, T. Norton, E. Stephens, L. Wolfenden, and H. L. Cloke, 2019: Are we talking just a bit of water out of bank? Or is it Armageddon? Front line

perspectives on transitioning to probabilistic fluvial flood forecasts in England, *Geosci. Commun. Discuss.*, doi:10.5194/gc-2019-18, in review*

The contributions of the authors of this paper are as follows: L. Anspoks (collaborator: EA), S. Manson (collaborator: EA), J. Neumann (collaborator: academic), T. Norton (collaborator: EA), E. Stephens (supervisor: academic), L. Wolfenden (collaborator: EA) and H. L. Cloke (supervisor: academic). H.L.C., L.An. and S.M. posed the original question. L.An., S.M., T.N. and L.W. brought L.Ar. up to speed about the EA and their decision-making practices. T.N. identified the interviewees. L.Ar., H.L.C., T.N. and E.S. designed the interviews. L.Ar. carried out the interviews and analysed the interview transcripts. L.Ar., J.N. and H.L.C. wrote the paper. H.L.C., S.M., J.N. and T.N. commented on the manuscript. Overall, 95% of the research and 85% of the writing was undertaken by L.Ar.

The published article (currently under discussion) can be found in the thesis Appendix A9.

Abstract. The inclusion of uncertainty in flood forecasts is a recent, important yet challenging endeavour. In the chaotic and far from certain world we live in, probabilistic estimates of potential future floods are vital. By showing the uncertainty surrounding a prediction, probabilistic forecasts can give an earlier indication of potential future floods, increasing the amount of time we have to prepare. In practice, making a binary decision based on probabilistic information is challenging. The Environment Agency (EA), responsible for managing risks of flooding in England, is in the process of a transition to probabilistic fluvial flood forecasts. A series of interviews were carried out with EA decision-makers (i.e. duty officers) to understand how this transition might affect their decision-making activities. The interviews highlight the complex and evolving landscape (made of alternative 'hard scientific facts' and 'soft values') in which EA duty officers operate, where forecasts play an integral role in decision-making. While EA duty officers already account for uncertainty and communicate their confidence in the system they use, they view the transition to probabilistic flood forecasts as both an opportunity and a challenge in practice. Based on the interview results, recommendations are made to the EA to ensure a successful transition to probabilistic forecasts for flood early warning in England.

* ©2019. The Authors. *Geoscience Communication*, a journal of the European Geosciences Union published by Copernicus. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.

We believe that this paper is of wide interest for a range of sectors at the intersection between geoscience and society. A glossary of technical terms is highlighted by asterisks in the text and included in Appendix A.

5.2.1 Introduction

One of the most recent and significant challenges in hydrology has been the inclusion of uncertainty information in flood forecasts. We live in a world where it is currently impossible to say with 100% certainty how the weather will evolve in the following days to months, or by how much exactly a river level is expected to change. This is due to the inaccurate measurement of hydro-meteorological observations*, errors in the mathematical models used to produce these forecasts (due to scientific and technical limitations) and, most importantly, nature's intrinsic chaos* (Lorenz, 1969; Buizza, 2008). In this world, probabilistic estimates of potential future floods are vital. Probabilistic forecasts* give a range of likely possible future outcomes, contrary to deterministic forecasts*, which indicate a single future possibility (Buizza, 2008). Probabilistic flood forecasts are generally produced by forcing* a hydrological model* with an ensemble* of future meteorological scenarios (Cloke and Pappenberger, 2009). By giving an idea of the uncertainty surrounding a prediction, probabilistic forecasts can give an earlier indication of potential future extreme events, such as floods, increasing the amount of time decision-makers have to prepare (Buizza, 2008; Stephens and Cloke, 2014).

In practice however, probabilistic forecasts can be challenging to use for operational decision-making*, given their uncertain nature (Nicholls, 1999; Cloke and Pappenberger, 2009; Demeritt et al., 2010; Nobert et al., 2010; Ramos et al., 2010; Stephens et al., 2019). Having to translate a range of possible outcomes into a binary decision (such as sending out a flood warning) is intricate and requires careful interpretation, an understanding of probabilities, risk*, uncertainty* (Dessai and Hulme, 2004) and of the systems modelled. Furthermore, probabilistic forecasts are designed to capture scenarios that may not always realise, which in turn could lead to false alarms*. Decision-making can be based on a set of rules, such as threshold exceedance (Dale et al., 2013). It is, for example, possible to take decisions (e.g. send a flood warning) when a pre-defined threshold is reached with a minimum forecast probability (Thielen et al., 2009). However, the decision-making process is generally based on, and influenced by, several additional factors. These include the type of event considered (e.g. a localised small flood event vs. a large scale extreme flood event), the costs of taking action vs. not taking action, experience of past events, the decision-maker's trust in the forecast (which can be built up over time), their risk aversion, and the

cultural context in which decisions are made (Cloke et al., 2009; Arnal et al., 2016; Neumann et al., 2018b).

The Environment Agency (EA)* is responsible for managing risks of flooding in England and their flood incident management strategy* is often shaped by major flood events (Werner et al., 2009; Stephens and Cloke, 2014; Pilling et al., 2016). In the 1990s and early 2000s, the UK policy shifted from a 'flood defence' to a 'flood risk management' strategy, on the back of the 1998 and 2000 floods (McEwen et al., 2012), which has led to more forecast-based decision-making. The summer 2007 UK floods boosted the development of the National Flood Forecasting System and the Flood Forecasting Centre (FFC*; a UK Met Office and EA partnership), with the aim to improve national flood warning services (Pitt, 2008; Stephens and Cloke, 2014). The winter 2013/14 UK floods further demonstrated the value of the FFC and the use of ensemble surge forecasts* for flood preparedness* (Stephens and Cloke, 2014). It was also during the 2013/14 floods that the EA started using two fluvial (or river) flood scenarios* (a reasonable worst case* and a best estimate*, instead of a single prediction) for flood incident management. Following this, Defra (the UK government Department for Environment, Food & Rural Affairs)* published a National Flood Resilience Review (NFRR) in 2016 (HM Government, 2016; House of Commons - Environment, Food and Rural Affairs Committee, 2016). This review aimed at understanding and increasing the UK's resilience to river and coastal flooding from extreme weather over the next ten years. The NFRR recommends a better integration of probabilistic weather forecasts into flood forecast products, for an improved characterisation of uncertainty and an enhanced communication of flood risk and likelihood to inform a range of flood management measures*.

While catastrophic events can foster the uptake of state-of-the-art science (e.g. probabilistic forecasts) for decision-making, achieving a complete and successful transition relies on many elements. For example, the use of ensemble surge forecasts in 2013/14 might not have been possible without the prior shift to a flood risk management mindset and the creation of the FFC. Moreover, we do not want to be in a situation where we require a catastrophic event in order to begin implementing the best science into risk management practice; it is vital to understand a country's and institution's cultural landscape to ensure that science is not being under- or misused (Golding et al., 2017). In the case of probabilistic forecasts, making sure that they add value rather than uncertainty to operational decision-making is key (Nobert et al., 2010). Interviews can be an effective method to capture an institution's complex cultural landscape (Schoenberger, 1991; Pagano et al., 2004b). They can provide interviewers with an understanding of the world (in this case the institution

world) from the perspective of the informants, shedding light on their unique perceptions and information only known to them (Sivle et al., 2014).

As outlined by the NFRR, the EA is in the process of a transition to probabilistic fluvial flood forecasts, from the two flood scenarios they currently use operationally (Orr and Twigger-Ross, 2009; Sene et al., 2009). To capture the EA's forecasting practice landscape and understand how this transition might affect their flood decision-making activities, a series of interviews were carried out in the summer 2018 with EA 'Monitoring and Forecasting Duty Officers' (MFDOs) and 'Flood Warning Duty Officers' (FWDOs). These two roles are at the heart of the EA's flood risk management decision-making chain. The outcomes of these interviews were used as a basis for this paper, with the aim to highlight the potential opportunities and challenges that this transition might translate to for the duty officers, ahead of it happening.

5.2.2 Context: the Environment Agency's flood incident management strategy

The Environment Agency (EA) is an executive non-departmental public body, sponsored by Defra. The EA has an operational responsibility to manage risks of flooding from rivers and the sea in England, by warning and informing the public and businesses about impending floods. Flood warnings are sent with a 2-hour minimum lead time*; however, different lead times have recently been introduced to take into account the type of flooding and catchment characteristics*; i.e. flash flooding vs. slow responding catchment. Under the Flood and Water Management Act 2010 (DEFRA, 2010), the EA takes a lead role on river and coastal flooding, whilst lead local flood authorities take a lead role on local flood risk (which covers flooding from other sources, including surface water, groundwater and minor watercourses). The EA also has a strategic overview role for all sources of flooding and works with lead local flood authorities by providing guidance, knowledge and support in responding to surface water flooding. The following schematic (Fig. 1) displays the EA's institutional landscape, with a particular focus on the flood incident management (FIM) information flow to and from MFDOs and FWDOs.

Historically, the EA was structured as a national body, delivering its work across England in six operational regional boundaries (i.e. regional boundaries were political delineations and were roughly aligned with the regional development fund boundaries). On 1st April 2014, the EA changed its operating structure to adopt area boundaries (i.e. broadly based on catchment delineations, but some catchments span different areas, especially at the borders with Wales and Scotland). These were aligned in 2016 with the Natural England (non-departmental public body, sponsored by Defra, and responsible for ensuring the protection and improvement of England's natural environment) boundaries. The EA is now operating over 14 areas with 7 forecasting centres (hereafter referred to as 'centres'; see Fig. 2).

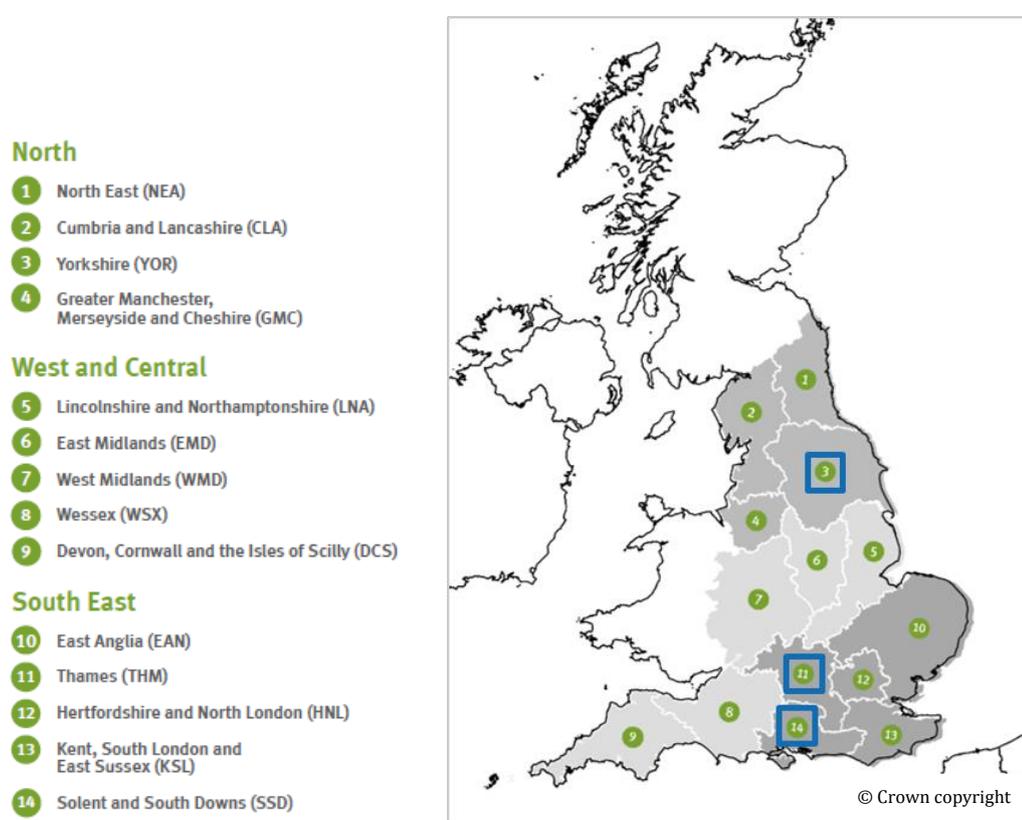


Figure 2. Map showing the geographical areas of the EA's operations (green numbered areas), highlighting the three areas which the centres where interviews were carried out are responsible for (blue boxes) (source: EA). The works published in this journal are distributed under the Creative Commons Attribution 4.0 License. This licence does not affect the Crown copyright work, which is re-usable under the Open Government Licence (OGL). The Creative Commons Attribution 4.0 License and the OGL are interoperable and do not conflict with, reduce or limit each other.

To help manage flood risk, the EA receive hydro-meteorological forecasts* produced by the Flood Forecasting Centre (FFC; see Fig. 1) on a daily basis (more or less frequently depending on the forecasting product* – see Sect. 5.2.4.1.1). The FFC is a partnership between the EA and the UK Met Office. It combines the hydrological and meteorological expertise from both institutes to provide hydro-meteorological forecasting products (for all

natural forms of flooding, including river, surface water, coastal and groundwater flooding) to emergency responders: category 1 (e.g. police services, fire and rescue authorities, including the EA for England), category 2 (e.g. utilities, telecommunications, transport providers, Highways Agency), Natural Resources Wales (for Wales) and the Met Office (for England and Wales).

The EA's FIM is based on the principle: 'think big, act early, be visible' (EA, 2018). This is part of a wider move from incident response to risk anticipation, with the aim to ensure that resources are put in place early and that the EA is prepared to scale-up or -down (i.e. preparations for measures implemented or not closer to the potential incident; e.g. expanded incident rotas with duty officers on standby, instigating requests for mutual aid to a different area, requests for equipment to support preventative and/or repair work, such as temporary barriers and pumps). As part of this strategy, the FFC forecasts are currently (and since the UK winter floods of 2013/14) used to produce two deterministic fluvial flood scenarios with a five-day lead time at the EA, a 'Best Estimate' and a 'Reasonable Worst Case'.

Several internal documents have been written to give guidance on how to use these scenarios to support decision-making for FIM activities, in line with the EA's principle. In summary, the Reasonable Worst Case gives an indication of what 'could' happen and should be used for preparation, information and response to flooding. The Best Estimate gives an indication of what 'should' happen and should be used as the basis for planning for warning. Together, the two scenarios provide the scale and size of the incident for planning and response preparations (FFC, 2017).

According to research done in the Thames river basin (UK), New et al. (2007) showed that probabilistic forecasts provide more informative results (enabling the potential risks of impacts to be quantified) than a scenario-based approach. The transition to the two scenarios can be seen as a stepping stone towards probabilistic fluvial flood forecasts. Ultimately, the EA would like to: 1) quantify uncertainty and communicate flood risk in a clear manner internally and externally, and 2) make decisions around incident preparation and escalation, operational activities and flood warnings effectively, intelligently and accurately. While the EA acknowledges that a potential benefit of probabilistic flood forecasts is the possibility to give earlier warnings, they question the extent to which probabilistic forecasts would reduce scientific and decision uncertainties in a FIM context (Orr and Twigger-Ross, 2009).

While work has already been done by the EA to investigate the technical feasibility of a transition to probabilistic fluvial flood forecasts (Orr and Twigger-Ross, 2009; Sene et al., 2009, Dale et al., 2013), this paper focuses on exploring the perceptions of the EA duty officers on the subject matter. This work is important as it will ensure the appropriate use of fluvial flood probabilistic forecasts for FIM decision-making activities, once operational. It should be noted that the EA already uses coastal flood probabilistic forecasts (Flowerdew et al., 2009); this work focuses on fluvial flooding. To this end, a series of interviews were carried out with EA 'Monitoring and Forecasting Duty Officers' (MFDOs) and 'Flood Warning Duty Officers' (FWDOs), as they are the two roles at the heart of the EA's internal forecast-led decision-making, building on the exchange between the MFDOs and the FFC (see Fig. 1; more information about their respective roles in Sect. 5.2.3.1 and 5.2.4.1).

5.2.3 Methods

5.2.3.1 Participants

The EA has several MFDO and FWDO roles, fulfilled by a number of different people. These are voluntary roles, added to the staff's day-to-day job, for which they follow relevant training. MFDOs receive, process and communicate forecast information to FWDO's, who are responsible for interpreting the information and working out the potential impacts on the ground. The duty officers' schedules are predetermined by a rota, and duty officers are on call for a period of one week at a time. During times of increased flood risk, when more forecasting or warning activities are required, additional rostering takes place. Duty officers receive a range of forecasts (nowcasting* products to monthly outlooks*) and are aware of potential situations from a month out. Five days ahead is when the activity really starts to build and is the focus of these interviews.

A total of six EA MFDOs and FWDOs from three different EA centres (one pair per centre) were interviewed to capture a range of perspectives in relation to this topic, following best practice (Sivle et al., 2014; participant information sheet provided as supplementary material in the thesis Appendix A9). Forecasting and decision-making varies between EA centres due to different management approaches and different types of geography and catchment response. To protect anonymity, the three centres where interviews were carried out are shown in terms of the wider area they are responsible for: 1) the Yorkshire area (YOR) in the North (area 3), 2) the Thames area (THM) in the South East (area 11), and 3) the Solent and South Downs area (SSD) in the South East (area 14) (Fig. 2).

MFDOs and FWDOs were interviewed in pairs as they are used to working together and the information they use sits between these two roles. The thought was that by talking to the MFDOs alone we would lose the element of “and so what?”, while talking to the FWDOs alone we would lose all the expertise about forecasting. All MFDOs and FWDOs interviewed had several years of experience and so were able to describe how current practice would change with a different type of forecast.

Participants were selected by EA study co-developer I1 to meet the above criteria. For the purpose of anonymity, the interviewees will thereafter be reported using codes. The three MFDOs interviewed will be referred to as MFDO1, MFDO2 and MFDO3, and the three FWDOs interviewed as FWDO1, FWDO2 and FWDO3 (interviewed pairs are however represented by the same number). As well as those from the MFDOs and FWDOs, quotes from two EA study co-developers are reported in this paper, I1 and I2, who helped the interviewer (Louise Arnal) by providing some context about the EA’s organisational landscape, forecasting systems and MFDO and FWDO roles prior to the three interviews.

5.2.3.2 Interviews

By design, qualitative, semi-structured interviews are used to understand interviewees’ perspectives, allowing the exploration of a research question that does not necessitate quantifying information and creating generalisations from the interview transcripts. The strength of such studies (compared to other survey methods) is that they are more sensitive to historical and institutional complexity and can capture the influence of local context (Schoenberger, 1991; Pagano et al., 2004b). Moreover, they are flexible, allowing the interviewer to remodel questions throughout an interview and from one interview to the next, to follow up on new information discovered (Sivle et al., 2014).

A fixed set of open-ended questions were prepared in advance to guide the discussion and allow for comparability across all three interviews. To prompt discussion, all three MFDO and FWDO pairs were asked the same opening question: “Could you please walk me through what you would do ahead of a potential flood event?” The following questions were also prepared in advance, but their order was changed, or they were skipped depending on whether the interviewees had already answered them:

- “Could you tell me about the uncertainties in the information you said you used in this context?”
- “How do you deal with these uncertainties?”
- “Could you tell me about how you communicate these uncertainties to each other?”

- “How would your job be influenced by a transition to probabilistic forecasts?”

Each interview lasted between 30 minutes and 1 hour 30 minutes. All interviews were conducted and digitally recorded by the first author (Louise Arnal) in meeting rooms at the corresponding EA centres.

5.2.3.3 Data analysis

All interviews were transcribed verbatim and transcripts were analysed qualitatively with respect to three main research questions. These research questions provide the structure for the results' section of this paper (Sect. 5.2.4).

- 1) What are the MFDOs' and FWDOs' roles and how do they interact with one another?
- 2) Where are the forecasts currently situated amidst their decision-making process?
- 3) Considering how the duty officers communicate confidence with one another at present, what might be the potential impacts of a transition to probabilistic forecasts on their roles?

Although interpretations might have been communicated by many interviewees, no frequencies are provided as quantitative generalisations cannot be inferred from this small and purposive sample. Following best practice, the results contain a mix of interviewees' perspectives, supported by quotes, and further interpretation of the interview transcripts by the authors, identifiable throughout the text (Davies et al., 2014).

5.2.4 Results

5.2.4.1 Roles and interactions between EA duty officers

Below, we summarise the MFDOs' and FWDOs' roles in an incident response context, using the interviewees' responses to the question: “Could you please walk me through what you would do ahead of a potential flood event?” It is worth noting that all interviewed pairs suggested the MFDO answer that question before the FWDO, indicating that the decision-making process starts with the MFDO.

“My role's an MFDO so generally if there's a flood event coming I should know before the FWDO, in theory” [MFDO2]

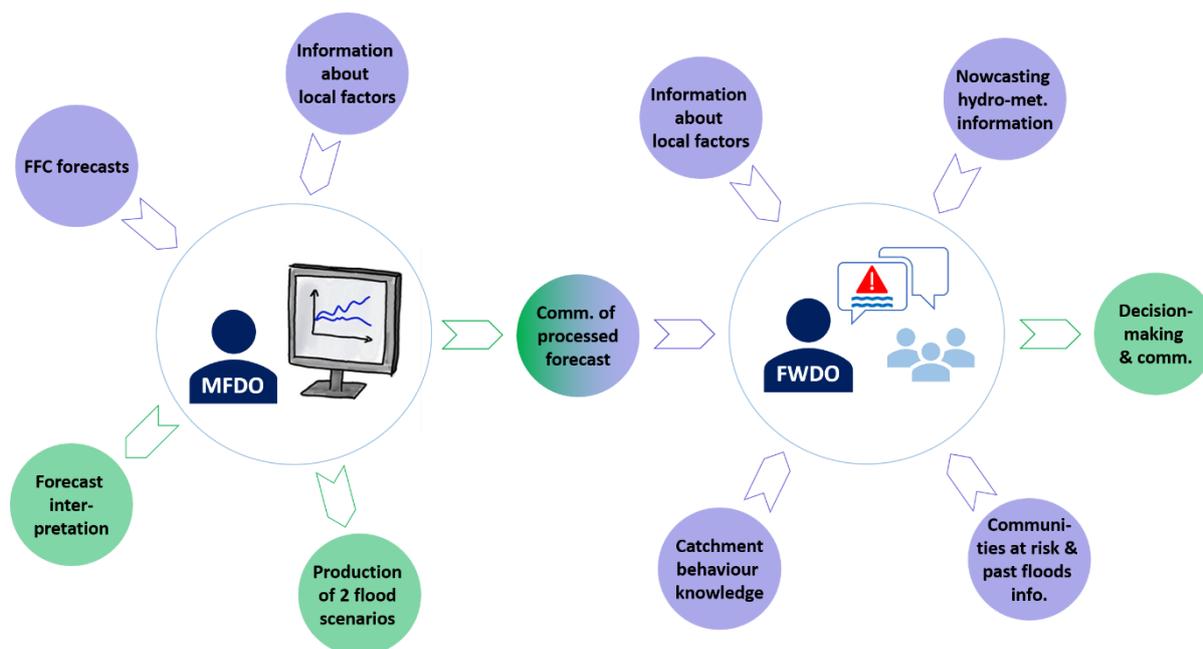


Figure 3. Roles and interactions between EA duty officers. Blue arrows and circles are for incoming information and green arrows and circles relate to outputs from either of the duty officers.

5.2.4.1.1 The role of Monitoring and Forecasting Duty Officers

“Ramping up to a flood event, the MFDO gathers that information, processes it and filters it, and passes that along to the area staff [FWDO].” [MFDO2]

What information do they use?

The MFDOs regularly receive FFC (Flood Forecasting Centre) national and county scale (i.e. area sub-divisions) flood risk forecasts and produce catchment/local scale flood forecasts, which they communicate with the FWDOs (see Fig. 3). The FFC generates three types of products:

- **Outlook products** – annual, seasonal and monthly assessments of flood risk;
- **Flood Guidance Statement (FGS)*** – a five-day forecast of flood risk for all sources of flooding, for England and Wales, at a county scale (see Appendix B, Fig. (a) for an example);
- **Hydro-Meteorological Services*** – detailed products communicating flood forecast data, currently comprising a Hydro-Meteorological Guidance, Forecast Meteorological Data and Heavy Rainfall Alerts (see Appendix B, Fig. (b), (c) and (d), respectively, for examples).

How do they use this information?

Based on this suite of information, the MFDOs decide whether they want to run the hydrological forecasting model, which sits in a separate system called the National Flood Forecasting System (NFFS; see Appendix B, Fig. (e) for an example). The decision can be triggered by the colours shown on the FGS (which communicates flood risk as a combination of likelihood and impact; e.g. high flood risk values on the FGS are more likely to lead to the MFDOs running the hydrological model). The NFFS allows users to explore the observed data (i.e. river levels and rainfall) and run hydrological and hydraulic models*. These models, forced with the FFC's deterministic weather forecast, provide a single trace of past and future (i.e. for the next five days) river level for specific areas. This initial forecast scenario is usually referred to as the 'Best Estimate' scenario, showing what 'should' happen. What 'could' happen (i.e. the 'Reasonable Worst Case' scenario) may not always be run.

"If there's uncertainty in the forecast like if there's showers [...] especially when they're thundery and they can give you really high totals in a very short space of time that's when you start to run 'What If' scenarios" [MFDO1]

'What If' scenarios (i.e. 'Reasonable Worst Case' scenarios) are additional forecasts run by the MFDOs by manually modifying the FFC's deterministic weather forecast (usually through the use of predefined factors applied over an entire catchment; e.g. 200% of catchment rainfall totals in the next 6 hours). They then run this 'modified weather forecast' through the hydrological/hydraulic models to obtain a new river level forecast scenario, often referred to as the 'Reasonable Worst Case' scenario. The MFDOs choose which What If scenario to run based on the FFC Hydro-meteorological Guidance and their own expert knowledge, to estimate the likelihood of both scenarios (the 'Best Estimate' and the 'Reasonable Worst Case').

"[The FFC] might give us a number of different scenarios and we tend to pick the worst one and then see what that does" [MFDO1]

A critical part of the MFDOs' role is to interpret the different forecasting products, which might sometimes be inconsistent (e.g. differences between the national and local scale pictures). The MFDOs usually do this by applying expert judgement based on knowledge of model performance and catchment response* to make a coherent story and put the information into context for the FWDOs.

The MFDOs decide when to pass the information on to the FWDOs, generally waiting for the forecast to be confident* before flagging a situation. The exact content of the

communication depends on each pair, but usually contains information about the scale of the event and their confidence in the forecast.

“Which scenario is going through which threshold [and] how likely that is to happen” [MFDO1]. “Approximate [...] scale of the event [...] are we talking just a bit of water out of bank? Or is it Armageddon?” [MFDO2]

The conversation can sometimes be bilateral, and the MFDOs might ask questions to the FWDOs.

“Can they provide information [...] in terms of local sensitivity [...] and are works going on in that catchment? Is there a gauge out of play?” [MFDO2]

5.2.4.1.2 The role of Flood Warning Duty Officers

“The role of the FWDO is to make sense of all that forecasting information and try and work out potentially what the impacts could be of that on the ground and then make decisions as to whether or not [they] issue flood alerts, flood warnings or severe flood warnings.” [FWDO1]

What information do they use?

The FWDOs' role is to combine several different types of information to decide whether to issue a flood alert or warning (see Fig. 3). The information available to them includes:

- **The processed hydro-meteorological forecast and interpretation from the MFDOs;**
- **Factors within the catchment** – factors that could influence river levels (e.g. blockage from a tree fallen down). This is ad-hoc information and comes from a variety of sources, including: information gathered from community contacts (flood wardens*, flood action groups*, etc.), from EA staff and duty officers, hydrometric data/CCTV images, details of consented works (i.e. work going on in a channel);
- **The situation on nowcasting meteorological products** – e.g. rainfall radar;
- **Information about the communities that might be affected** – e.g. have they been affected by many floods in the past;
- **Expert knowledge about catchment response.**

How do they use this information?

The FWDOs assess these various sources of information (e.g. in terms of their accuracy) to make a decision, knowing that they do not necessarily have all the information to make a judgement call.

“I look at the river level forecasts and then what I want to know from the MFDO is, does this account for the rain we’ve had? So, do you think this is likely to change? Is the forecast I’m seeing on my screen a good river level forecast? Or do we think it’s not picked something up properly?” [FWDO2]

According to an internal document on using the two flood scenarios in practice, the Best Estimate should be used as a basis to issue flood alerts or warnings. However, both scenarios are currently used for incident planning activities (e.g. resources needed for response) and communication with responders and communities, while flood alerts and warnings are mostly issued based on nowcasting products. This discrepancy could be due to the challenges associated with forecast accuracy* and lead time, specifically for surface water flooding* and rapid-response catchments*. This document does however encourage the use of the two scenarios for planning and flood warning activities whenever possible, in combination with expert judgement.

“The scenarios are planning scenarios and at some point [...] we move into operational now type forecasting. So normally we’d issue a flood warning with anywhere between 30 minutes to [...] six hours lead time, whereas these scenarios are generally two to five days ahead. So you wouldn’t normally [...] come up with a simple statement that will issue flood warnings based on the best estimate [...] and at some point we transition into something that’s more now that we use for operational decision making” [I1]

5.2.4.1.3 Communication between MFDOs and FWDOs

“The FWDO shouldn’t even really be thinking about anything until they’ve had a phone call from the MFDO [...]. Some FWDOs do go a bit more proactive than that, I think particularly the ones with the forecasting backgrounds almost can’t help themselves looking into it. And it depends on personality as well, some people hate the idea of being surprised by anything. But it does also depend on the MFDO.” [FWDO2]

There is usually a constant exchange of information between MFDOs and FWDOs, even when no major event is on the horizon. However, more recently, the level of activity in preparation for a potential event has increased. Since 2007 (this corresponds vaguely with the summer 2007 floods), the lead time for which forecasts are shown and on which MFDOs

and FWDOs can take action has increased from a few days to a few months ahead (based on the FFC's outlook products mentioned in Sect. 5.2.4.1.1). This is consistent with findings from Neumann et al. (2018b), who report that the EA currently uses long-range* (i.e. seasonal) hydrological forecasts mainly as supporting information, while relying on the shorter-range forecasts* for action.

“So even from a month out now we’re starting to become aware of potential situations [...], but [...] because [...] most of our products [...] are [...] based on that five-day forecast [...] that’s when the activity really starts to build” [MFDO1]

The communication between MFDOs and FWDOs varies across people and EA centres. Factors that might influence communication – in terms of its trigger, frequency and content – include the duty officers' personality, day-to-day job and level of experience. Some FWDOs are more proactive than others in obtaining the information needed to make a decision; some might wait to be contacted by the MFDOs with a processed forecast, and others monitor the situation on a daily basis (see quote from FWDO2 above). In some cases, the FWDO might contact the MFDO first to get more details about an area of concern to them.

“[...] and [...] then it’s [...] liaising with regional forecasting [the MFDOs] so they can give us any more detail or certainty or if we’re concerned about an area they can watch it a bit more for us [the FWDOs]” [FWDO3]

Duty officers' level of experience can also influence the content and interpretation of the conversation. Knowing each other helps interpret and gauge the confidence from each other's language, which MFDO2 refers to as 'nuanced communication'. Working with new duty officers can lead to misinterpretations and you might have to justify your position further and prompt them to obtain the information you need.

“I’ve known [FWDO1] for quite a while so when I’m on duty with [them] [...] I can sense [...] what sort of questions [they] want to ask, where [they’re] coming from. I think with less experienced duty officers it’s often more tricky to do that. So [...] the verbal communication that you go into with [FWDO1] for example might be a bit brief probably because I know that [they’ve] understood the message and interpreted the message well, whereas a new duty officer you might be spelling out [...] your position more, spending more time explaining why the uncertainty is such and how that may impact on the ground” [MFDO1]

“Knowing each other is really important because if I know it’s [MFDO2] on duty [they’ve] probably put that interpretation on already. If I get someone who’s reading off the

screen, I put the interpretation on and if we misjudge that and we both put it on we could end up getting it too low” [FWDO2]

Other factors that influence communication include the context of the event, duty officers’ geographical proximities and a centre’s practice. In some areas, the FWDOs will make the final call of warning the public or not, while in other areas, the MFDOs will tell the FWDOs when they need to issue a warning. In addition, MFDOs and FWDOs do not always sit in the same building or town. MFDOs work from forecasting centres, while FWDOs are based in Area offices or Area incident rooms, which influences their (mode of) communication (in person vs. via phone or emails).

“If these people [the FWDOs] were sitting geographically with these other people [the MFDOs], I think you’d get a better service” [I1]

5.2.4.2 The forecast, a small cog in a much bigger wheel

“Forecasting’s really important. It is, it should be really central to what we do [...] but actually it’s a small cog in the middle of a much bigger wheel.” [I1]

Forecasting supports incident response by providing a critical piece of information. However, duty officers have to consider a range of other sources of information and factors when making risk-based decisions.

“We always implore people to try and look at different sources of information” [I2]

These additional sources of information include river level correlations*, model performance*, local knowledge (i.e. knowledge of how a certain catchment behaves), personal experience, and internal and external considerations (see Fig. 4). This section gives a more detailed overview of these factors and their relevance for decision-making.

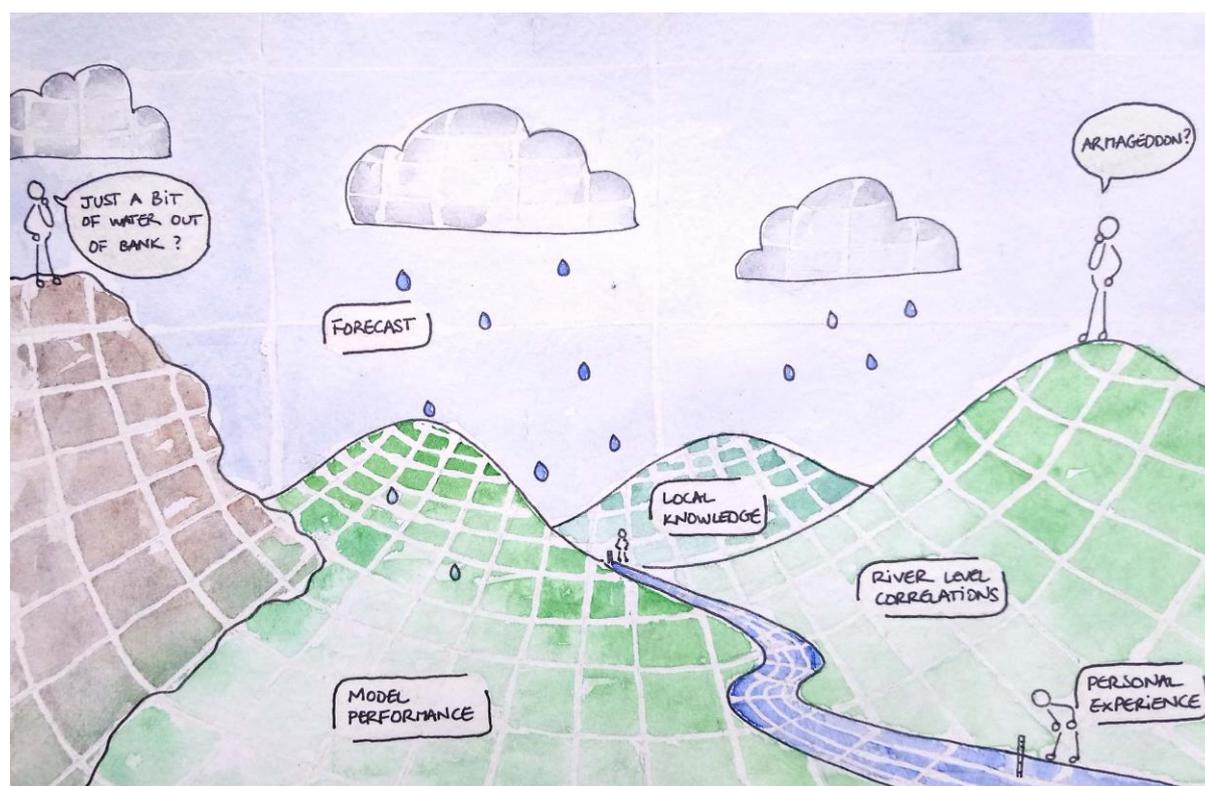


Figure 4. Complex decision-making landscape in which EA duty officers operate.

5.2.4.2.1 River level correlations and model performance

“The MFDO will be looking at how much rain is falling compared to what was forecast. You can check the river levels on the telemetry sites*, so you can see how fast they’re responding compared to the model and you can start to gauge how that catchment’s responding compared to what you thought it would do”

[MFDO1]

MFDOs might use several products to gain an understanding of model and forecast performance while the event unravels. More basic forecasting methods, like river level correlation tables, complement forecast information and aid the decision-making process. These correlations are based on a linear regression between peak levels upstream and downstream of a station. However, discrepancies between the forecasts and correlations can call into question the forecast accuracy.

“If the model says you’re going to get flooding, the correlation says we’re going to get flooding, we’ve had more rainfall than any previous event, you know that that decision’s [...] a clear one. If the model says flooding, the correlation says no you’re fine, and we’ve had somewhere in the middle in terms of rainfall, that’s when it gets difficult, because those borderline calls are really tricky to make” [I2]

The MFDOs' knowledge of the hydraulic/hydrological model performances, for certain types of events and catchments, is also key in interpreting the forecast. This is based on performance measures*, local feedback from real time river gauges*, experience and target lead times (i.e. the theoretical maximum lead time you have to send out a flood warning for a catchment before it floods, based on catchment size, gauge location and flood risk in that catchment). For certain types of events, such as convective rainfall events*, for which the duty officers know models are still limited, they might decide to issue a warning based on the 'Reasonable Worst Case', although it is *"technically against procedure"* [MFDO2].

The FFC meteorological products also communicate some sort of confidence, which the MFDOs can use to complement the hydrological models' performance information.

5.2.4.2.2 Local knowledge and personal experience

"Whilst we are very data reliant on the information coming through, there's also that experience that you know that certain watercourses are very slow responding and [...] no matter how much money we spend on your forecast, it's always not very good, you always delay it by a day and drop the peak by a bit. [...] Data is very important but that local experience is as important if not more so in certain circumstances" [MFDO2]

Local knowledge and personal experience are key ingredients for judgement, an important component of the decision-making process. This means duty officers can react appropriately to an event and add confidence to the forecast. As MFDO2 put it, *"experience is the unwritten part of the value that each role has"*.

Local knowledge is so important to decision-making that the interviewees believe it cannot be replaced by training, written material or fully automated systems.

"Some areas have very set triggers for a severe flood warning whereas other areas may just take it on a feel. [...] And each area has done it for a good reason, it's the local reasons for doing that but it isn't nationally consistent" [MFDO2]

"We have in the past looked at automated warnings [...], we can't automate them [...], there's a lot of personal interpretation and judgement [that] goes into it, and if a computer just hits a level and issues a warning, it's going to go wrong" [FWD02]

This also manifests itself in perceptions about how successfully duty officers can transfer to other centres or areas to help during an important flood event.

“One of the things we’re trying to do at the moment is to get mutual aid sorted out so that if a flood event happens in [some of the Northern areas] and their MFDOs [...] or the FWDOs are very [...] stretched [...] we can go [...] there, use their tools, their systems and do the same job. But whenever we’ve tried it the local knowledge is the key thing. Like knowing that this river responds particularly quickly and that we need to deal with it first before we move on to other ones that’s the sort of thing that even if you’re picking it up whilst you’re working in a different centre it’s affecting your ability to deliver the role at the time” [MFDO1]

Duty officers have access to tangible information about past flood events that can be useful for placing model information into context. The ‘Flood Intelligence Files’ compile information (e.g. highest events on record, what rainfall led to them, what the catchment state was at the time and any known impacts) for every gauge the EA is providing forecasts for.

How information is interpreted, risk appetite and past experience, can all affect decisions taken. There is the danger of following instincts too much and becoming biased towards issuing too many (i.e. risk-averse) or not enough warnings (i.e. risk-hungry), while in some cases decisions might never be forecast-led.

“Since the Boxing Day floods I think the next level of flooding after that there was some discrepancies amongst the area responses [...] they were a bit [...] jumpy [...] to not be caught out again which is understandable” [MFDO2]

“these kind of decisions about do we need to draw up a roster, do we need to be in the office overnight, a lot of that has probably been done on gut feel, probably this FWDO being the advisor. [...] Do we need to do whatever based on judgement, experience, feel for it. [...] I wouldn’t expect these people to actually be looking at any forecast and saying, based on this I will do” [11]

5.2.4.2.3 Internal and external considerations

“There are lots of external pressures as well, particularly as FWDO you can come under pressure from all different types of sources to make decisions and perhaps not based on the evidence that you’ve got for political reasons, [...] reputational reasons, organisation, in terms of being seen to be active, seen to [...] act early”

[FWDO1]

Decisions are not only dictated by the science, local knowledge or personal experience and differences, but might have to respond to internal and external considerations, especially during major events.

At an internal level, some areas and duty officers might be more forecast-led while others are more reliant on a nowcasting type approach. Discrepancies amongst the area responses are partially due to historical differences across the different areas and EA centres.

“There are definite differences between areas and [...] between individual staff, so [town X] are far more likely to issue flood alerts [...] purely on rainfall than [town Y] is, [town Y] will generally wait for a river level to rise and that develops I suppose out of slight historical differences and personalities involved” [FWDO2]

“Some other areas will issue messages based on forecast whereas, we were always told to base it on what’s happening, so we kind of wait to see if the rain comes in and then if anything happens issue. And we get marked on messages that we send out, so one of the things is the timeliness and if you’ve issued one, did it actually flood afterwards? So if you’re obviously issuing on a forecast, then you’re probably going to get scored low because it doesn’t always happen, so it’s difficult” [FWDO3]

There are exceptions to these procedures and FWDO3 mentions the possibility of issuing flood alerts based on the forecast when the impact is expected to occur overnight or if the forecast displays *“rarely high confidence”* of rainfall and *“if it’s a more prolonged event”* and *“you know the catchment’s already wet”*.

The EA’s principle, “think big, act early, be visible”, is an example of an internal consideration, which might influence the duty officers’ decision-making (EA, 2018). In what ways does the EA’s statutory warning responsibilities and principle influence decision-making? Does “act early” put the forecast in first place while “think big” and “be visible” move it to a secondary position?

“Our mantra to incident response is think big, act early so sometimes [...] there is a danger that you’re over responding. Somewhere you’re issuing alerts and warnings when actually the risk is low. So I think the role of the FWDO is to assimilate all that information, forecasting information and using it to help inform the instant response but also manage expectations” [FWDO1]

There is usually a political element (external consideration) to the response immediately following a very major flood, as the EA puts a greater focus on demonstrating to

communities and the government that they are being proactive in warning, informing, etc. There is also the need for the EA to align its message with actions of lead local flood authorities and responders and to think about public response.

“It’s managing expectations internally in terms of operational response and how this is going to potentially play out which [...] can still be quite hard to do but it’s even harder to do it externally with [the] mood of the public or even some of our professional partners, so local authorities are also obviously geared up to respond to flooding” [FWDO1]

To conclude this section, it is evident that the duty officers have to take different sources of information, besides the forecast, into consideration to make a decision. However, the forecast helps determine the timing of warning and response activities. Because the forecast plays a seemingly small part in a much bigger system, could that mean that the transition to a different type of forecast will have very minor impacts on the duty officers? Or on the contrary, could it unsettle this very complex machine?

5.2.4.3 What could a transition to probabilistic forecasting mean in practice?

5.2.4.3.1 Current practice: communicating confidence for decision-making at the EA

“Uncertainty is present in everything that we do and every bit of communication, [...] I don’t think I’ve ever been able to say something with 100% confidence, ever.” [MFDO2]

We have previously touched on the factors and uncertainties duty officers have to work with, including uncertainties in: the weather (and how it cascades down to the hydrological response), model performance, the different spatial scales of response (local vs. national), the situation on the ground (e.g. soil conditions prior to an event and river blockages), EA staff decisions and actions, and the public’s reaction to warnings.

Duty officers currently adapt the language they use to communicate these uncertainties internally and externally, based on their confidence level. According to internal EA guidelines, the language used should change according to the scenario used so that duty officers *“get used to the [...] way they’re working around scenarios and probabilistic forecasting” [I1]*.

“If messages around a ‘Reasonable Worst Case’ use, could or [...] is possible; if it’s a ‘Best Estimate’ use, we expect, it’s probable” [I1]

Between the MFDOs and the FWDOs, confidence and uncertainty appears to always be (based on these interviews) communicated, usually using the two flood forecast scenarios.

“I don’t think we can withhold uncertainty. One, the key role for MFDO is providing the forecast. So it’s getting the forecast as accurate as you can and then communicating it in the clearest way possible. So that’s often about interpreting the uncertainty and communicating it. So we often use the ‘Reasonable Worst Case’ and the ‘Best Estimate’ to do that” [MFDO1]

Messages to the public are also worded with care to communicate the appropriate level of risk and prompt appropriate response and also contain some information about confidence and uncertainty. These messages are usually free-text messages and will therefore vary across FWDOs.

“The message starts off with this flood warning has been issued for this place then it runs on after a while into detail which is where you can communicate those shades of grey” [FWDO2]

However, not all uncertainties are critical, and local knowledge and experience are key for the *“interpretation of the uncertainties”* [FWDO2] and their impact on the ground.

“Uncertainty from the forecasting point of view is always prevalent but understanding how it will impact the [...] area’s reaction is kind of the key thing” [MFDO2]

There is currently space for the communication of confidence at the EA and externally. This is a step towards probabilistic forecasting. But how big of a step is it? And how big of a step is still needed to reach that full transition to probabilistic flood forecasts?

5.2.4.3.2 The duty officers’ perceived opportunities and challenges

“Whether it creates as many problems as it solves, maybe” [I2]

The transition to probabilistic forecasts is a significant one, which appears to generate mixed feelings amongst the duty officers. It is undeniable that this transition will bring changes at the EA; as *FWDO2* put it, *“probabilistic forecasting is kind of a fresh start for everyone”*. This section presents the interviewees’ perspectives on the changes that will ensue from this transition, in terms of perceived opportunities (left wordcloud on Fig. 5), challenges (right wordcloud on Fig. 5) and neutral changes. Table 1 outlines these perspectives, split into six main topics and supported by quotes reported in Appendix C.

	things will need to change slowly to give duty officers time to build confidence in the new system (quote C4).
Decision-making	A few interviewees mentioned the fact that probabilistic forecasts will not solve the fundamental need of decision-making to be binary and saw this as a challenge (quotes C5 and C6). Others saw this as an opportunity for early warning and long-term planning (quotes O4 and O5)
Duty officers' roles	This transition was seen neither as an opportunity nor as a challenge by and for the MFDs. They simply stated how things might change for them (quotes N2 and N3). A few of the FWDs however thought that this might push more of the interpretation on to them (quotes C7 and C8). It is worth noting that none of the interviewees mentioned worries concerning potential impacts of this future transition on the communication and interaction between duty officers. The worries seem to mostly lie outside of their interaction (quote O6).
New staff training	An interviewee mentioned that probabilistic forecasts could help with new staff training, by increasing their understanding of catchment response (quote O7).

Some of the quotes reported in Appendix C might sound very extreme, which could be partly due to the way the questions that prompted them were phrased. However, it could also reflect personal resistance and should be explored further.

5.2.5 Discussion and recommendations

5.2.5.1 Considerations for a successful transition to probabilistic forecasts

Probabilistic forecasts have a great potential to capture extreme events (Stephens and Cloke, 2014), and their benefits (compared to deterministic forecasts) for flood warning are evident (Verkade and Werner, 2011; Pappenberger et al., 2015). However, despite the increasing lead times at which we can confidently predict floods, the uncertainty inherent in the chaotic natural system being modelled grows with increasing lead times, posing new problems. As science and decision-making are both individually progressing, adapting to their respective internal and external changes, there still lacks an ideal framework for the incorporation of new and 'uncertain' science in decision-making practices, and,

respectively, the uptake of decision-makers' perspectives in the design of scientific practice. Here, results from this study and relevant literature are joined to put forward elements that should be considered for a successful transition to probabilistic forecasts for flood warning in England.

From these interviews and previous EA studies, it is apparent that forecasts are one element in the complex decision-making landscape within which EA duty officers operate (Orr and Twigger-Ross, 2009; Dale et al., 2014). This landscape includes alternative 'hard scientific facts' (e.g. correlations, model performance and local knowledge to an extent), and 'soft values' (dependent on culture and context, personal experience and internal and external considerations) (Morss et al., 2005; Cloke et al., 2009; Arnal et al., 2016; Neumann et al., 2018b). Morss et al. (2005) found that "although flood management practitioners might appreciate more certain hydro-meteorological information, scientific uncertainty is often swamped by other factors [e.g. community perception, time, money and resource constraints] and thus is not a high priority." When uncertainties are evident and decision stakes are high, as is the case for the uncertainty communicated by probabilistic forecasts for flood incident management, traditional decision-making pathways could become ineffective and soft values might become more important than hard scientific facts (Funtowicz and Ravetz, 1993). In this specific study for example, an uncertain probabilistic forecast could lead to some duty officers reverting to the 'Best Estimate' and the river level correlations to make a decision, ignoring low probabilities of extreme events which could have ultimately led to an earlier flood warning.

Facing constantly evolving soft values, some decision-makers may find familiarity with the scientific methods they use reassuring, reducing their personal willingness to adopt new scientific methods (Morss et al., 2005; Ishikawa et al., 2011). This personal willingness was captured in the range of responses (perceived challenges and opportunities) obtained during the interviews. An institute's operating practice should reflect the complex landscape in which decision-makers operate, where the forecast plays an integral role in decision-making. To this end, the co-design of forecasting systems by both forecasters and users is necessary.

To do that, clear communication between forecasters and users is needed. However, language is perhaps one of the biggest barriers between scientists and decision-makers. It has been observed that "the way scientists referred to and discussed uncertainty sometimes confused practitioners" (Morss et al., 2005). Similarly, there is a lot of research done on the impacts of graphical representation of uncertainty in hazard forecasts on decision-making.

These have shown that great care has to be taken when designing and communicating uncertain information, as it can impact the nature of the actions taken (Bruen et al., 2010; Joslyn and Savelli, 2010; Stephens et al., 2012; Pappenberger et al., 2013; Sivle et al., 2014).

There is the common misconception amongst the scientific community that decision-makers want 100% certain information (Demeritt et al., 2013; Michaels, 2014). In reality, as shown in this paper, decision-makers appreciate that scientific information is uncertain, not unlike other types of information they use. Decision-makers want to see that uncertainty, which they do not necessarily perceive as a barrier to use (Morss et al., 2005; Bruen et al., 2010; Neumann et al., 2018b). One reason for this misconception might be the different ways scientists and decision-makers approach forecast uncertainty. Scientists see (the reduction of) forecast uncertainty as an end goal and “often deal with uncertainty by attempting to reduce, quantify, analyze, and/or assess it”. Decision-makers “view uncertainty as an unavoidable factor [...] all information about the future is uncertain [and] they must make decisions under uncertainty every day, in a complex, evolving social, institutional, and political environment” (Morss et al. 2005).

In this complex evolving landscape, decision-makers deal with forecast uncertainty similarly to other uncertainties they might face: under time and resources constraints. They assess the total uncertainty there is (the forecast uncertainty might sometimes be negligible compared to all the other factors at stake) in terms of its potential effect on the decision-making process and outcome (Morss et al., 2005). As mentioned by a few EA duty officers, uncertainty is prevalent in everything that they do, and the key is understanding what the impact of these uncertainties will be on the ground. It is crucial to develop a methodology for decision-makers to be able to use (forecast) uncertainty information optimally. A solution that does not require any additional time- and resource-consuming complex analyses, given the high stakes and strict deadlines decision-makers have to work with. Smith et al. (2018) argue that if there was a “greater involvement of decision-makers in the design and execution of uncertainty analyses”, “more purposeful evaluation and communication of uncertainty would certainly result”. This remains an open challenge to be tackled.

By design, probabilistic forecasts might contain some realisations that capture scenarios which do not always realise. This may lead to false alarms. Institutions can have specific risk perceptions and flood management priorities: seeking to avoid false alarms, or on the contrary, seeking to avoid missed flood events*, and the minimum/maximum lead time at which they (have to) issue flood warnings. This cultural landscape within which decision-

makers operate may have an impact on the decision-making outcome (as discussed in Sect. 5.2.4.2.3) and an institution's uptake of probabilistic flood forecasts in practice (Nobert et al., 2010; Ishikawa et al., 2011; McEwen et al., 2012; Demeritt et al., 2013; Michaels, 2014). A transition to probabilistic flood forecasts should be reflected in an institution's wider flood management priorities. This could be done, for example, by changing their internal communication pathways or their warning procedures (e.g. lead times at which they operate).

Very often however, the ability of an institution to pick up new information and methods is not only down to them, but could be influenced by the wider socio-political context and other key actors in the decision-making web (e.g. the government, local authorities, regulations and guidelines), additionally to the populations at risk and the way they respond to flood warnings (Dessai and Hulme, 2004; Morss et al., 2005; Parker et al., 2009). This is reflected in the interviewed EA duty officers' perceived challenges regarding 'Language & communication' and 'Binary decision-making' (Sect. 5.2.4.3.2). In the face of a socio-political context that is demanding ever more precise information and with the rise of a post-factual society, the general trust in science might be a limiting factor to the uptake of new science and institutions should trust their capacity to use uncertain probabilistic information (Soares and Dessai, 2015; Golding et al., 2017; Knudsen and de Bolsée, 2019).

It is also important to note that "moving to probabilistic forecasting from deterministic forecasting may trigger an institutional shift in who is responsible for decision making under uncertainty" (Michaels, 2014). Because making a decision based on probabilistic information is more nuanced than using deterministic information, the outcome will determine who will be 'blamed' and this ownership of the uncertainty judgment might have implications on the forecasters-users relationships (Michaels, 2014). This relates to some of the interviewed duty officers' fears of a transition to probabilistic forecasts at the EA, as it might move "the burden of making a decision further down the tree" (Sect. 5.2.4.3.2). In this context, a framework to engage with all key actors of the decision-making web ahead of and during a transition to probabilistic forecasts appears crucial. Ramos et al. (2010) advocated the use of integrated platforms to allow a continuous exchange between scientists and decision-makers in real-time. Similar studies on the provision of climate services have identified the lack of user engagement as a great limiting factor of the uptake of climate information in practice (Golding et al. 2017). It is evident that a transition to probabilistic forecasts is not only a scientific endeavour and feasibility studies should include other disciplines, such as social-science.

This paper and the supporting interviews intentionally excluded the notion of coastal flood forecasting as not all interviewees were familiar with and use these forecasts operationally (given the non-proximity of some of the EA centres where interviews were carried out with the coast). However, as mentioned in Sect. 5.2.2, the EA already uses coastal flood probabilistic forecasts and learning from this previous transition could provide valuable information for the current transition to probabilistic fluvial flood forecasts. This should be further investigated.

5.2.5.2 Recommendations to the EA

In light of the findings of this study, and other relevant studies, we make a list of recommendations to support the uptake of probabilistic forecasts at the EA. These ten recommendations are high priority actions for the EA as an institution. The service, role owners and those responsible for ensuring a quality service delivery should ensure that these recommendations are pursued, alongside technical work around the transition. Please note that these recommendations are not ranked in priority order for the EA, as some of these will be quicker and easier to implement and to demonstrate progress on.

- 1) Communicate (via engagement campaigns, videos, email newsletters, social media updates and webinars, etc.) with all key players in the decision-making chain (as well as external players such as the emergency responders and the public) to ensure that they are all aware that the transition to probabilistic forecasts will become operational practice.
- 2) Give appropriate and custom designed internal training to all key players (Nobert et al., 2010). Duty officers must receive training on how to make decisions based on probabilistic forecasts (for example in the form of decision-making activities and serious games - see the HEPEX[†] and the Red Cross Climate Centre[‡] resources for inspiration).
- 3) Expand existing EA communication structures to allow the co-design of the new products between forecast producers and users (Morss et al., 2005; Smith et al., 2018). Everyone using the forecasting products and systems at the EA should have the chance to have a say in how the system will look and function through a mutual design strategy. If the new system does not reflect the complex landscape in which

[†] hepex.irstea.fr/resources/hepex-games

[‡] www.climatecentre.org/resources-games/games

duty officers operate (a mix of 'hard scientific facts' and 'soft values'), probabilistic forecasts might end up being under- or misused.

- 4) Reach out to the community of practice in hydrological probabilistic forecasting, such as HEPEx[§] (community of international experts in the field of probabilistic hydrological forecasting and decision-making) and connect with institutes which have already gone through such a transition to gain insights and share best practice, as some elements might be transferrable (Nobert et al., 2010; Dale et al., 2014). This could be done through organised workshops, webinars and the establishment of an advisory group.
- 5) The way probabilistic information will be translated into meaningful content and communicated to the emergency responders and the public requires careful thought and design (Bruen et al., 2010; Joslyn and Savelli, 2010; Stephens et al., 2012; Pappenberger et al., 2013; Sivle et al., 2014). To this end, an interdisciplinary approach between forecasters and social-scientists would be greatly valuable as social-science can offer insights into the human response to warning messages. A tailored and inter-disciplinary study of the forecasting products using probabilistic information and used in the decision-making process is urgently required.
- 6) a) The EA's heterogeneity at the national level should be accounted for and addressed. Given the heterogeneity of the EA at a national level and the areas' diversity in terms of history and catchment response, we do not expect probabilistic forecasts to be welcomed similarly in all the EA centres. Efforts will therefore have to be made by the EA to achieve a simultaneous and homogeneous transition in all its centres.

b) Furthermore, the design of the new forecasting system should be homogenised at the national level (to allow for staff movement during major flood events), while accounting for the heterogeneity of local conditions, existing dynamics and institutional practices. This could be achieved through the co-design of the forecasting system with local duty officers (see recommendation 3).

[§] hepex.irstea.fr

- 7) Be prepared to move towards lead times that reflect the probabilistic forecast predictability. The optimal lead time to trigger action depends on both the probabilistic flood forecast quality and the actions' operational implementation time (Bischiniotis et al., 2019). While the EA operates with pre-defined lead times for each specific activity (e.g. it takes x hours/days to move equipment from A to B, or to deploy temporary defences), probabilistic forecasts could in theory provide earlier indications of potential future floods, giving the EA more time to prepare ahead of a flood event. To utilise probabilistic forecasts to their full potential, tailored studies should be performed during the EA system's co-design to adjust lead times (for planning and warning) on the probabilistic products and event types, with ample time for testing by the EA duty officers.
- 8) Under no circumstances should the old system be switched off as soon as the probabilistic system is operational. There should be a reasonable period of overlap between the two systems in order to give everyone some time to gradually adapt (Funtowicz and Ravetz, 1993). During that time of overlap, end-user feedback should be collected (Thielen et al., 2006). To avoid situations where the probabilistic forecast and the two scenarios show contrasting results, the new operating procedures need to specify that the probabilistic forecasts should be looked at first.
- 9) Update the duty officers' operating procedures. Clear guidelines should be provided to the duty officers on how to make a decision based on the new probabilistic products. These guidelines should include information such as: the various sources of information available to them for making a decision, how to interpret a probabilistic forecast, the forecast confidence at which certain decisions and actions should be made and the language that should be used.
- 10) Document this transition (in writing or through documentary-style interviews, etc.) to help other institutes and future transitions at the EA (Pielke, 1997). While this paper investigates how things might change, post-transition evaluation should seek to answer the question: "How did we do?"

Many of these recommendations are however general and could be applicable to other institutes and types of information.

5.2.6 Conclusions

The Environment Agency (EA) is in the process of a transition to probabilistic fluvial flood forecasts, from the two flood scenarios they currently use operationally for flood warning and incident management activities in England. State-of-the-art probabilistic forecasts can give an earlier indication of potential future extreme events, such as floods, increasing the amount of time decision-makers have to prepare. A series of interviews were carried out with EA 'Monitoring and Forecasting Duty Officers' (MFDOs) and 'Flood Warning Duty Officers' (FWDOs), two roles at the heart of the EA's flood risk management decision-making chain. The aim was to understand how an operational transition to probabilistic flood forecasts might affect their decision-making activities. Overall, none of the interviewed duty officers mentioned concerns about impacts of this transition on their two roles' interaction. Perceived challenges lie mostly outside of their roles and relate to: communication with emergency responders and the public, translating uncertain information into a binary decision and the speed of the transition. Ten high priority recommendations were made to the EA to ensure a successful transition. They include: i) communicating with all key players in the decision-making chain (as well as emergency responders and the public) to ensure that they are all aware that this transition will become operational practice, ii) facilitating the co-design of the new products by forecasters and users and collecting end-user feedback during a reasonable period of overlap between the two systems, iii) employing an inter-disciplinary approach to translate probabilistic information into meaningful content for communication with emergency responders and the public, and iv) being prepared to adapt the EA's overarching warning and incident planning strategy to reflect this transition. It is vital for these recommendations to be followed to ensure that state-of-the-art science is used to its fullest potential for risk management practice and is not being under- or misused.

Author contributions. H.L.C., L.An. and S.M. posed the original question. L.An., S.M., T.N. and L.W. brought L.Ar. up to speed about the EA and their decision-making practices. T.N. identified the interviewees. L.Ar., H.L.C., T.N. and E.S. designed the interviews. L.Ar. carried out the interviews and analysed the interview transcripts. L.Ar., J.N. and H.L.C. wrote the paper. H.L.C., S.M., J.N. and T.N. commented on the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Disclaimer. The information and findings in this paper are based on interviewees with six EA duty officers. They should not be taken as representing the views or practice of the EA as a whole.

Acknowledgements. This work was funded by the EU Horizon 2020 IMPREX project (www.imprex.eu) (641811) and the joint Flood and Coastal Erosion Risk Management Research and Development Programme. We thank all the interviewees who dedicated some time to this work. We would also like to thank Stuart Hyslop at the EA for the background he provided in preparation for the interviews, and for his support in the organisation of the interviews.

5.2.7 Appendix

Appendix A. Glossary of terms.

Best Estimate	A forecaster's assessment of the most likely rainfall, river and groundwater levels, and coastal conditions, and their impacts.
Catchment characteristics and response	Catchment characteristics are the features that describe a river basin (i.e. the area of land drained by a river), such as its location, size, vegetation cover, soil type and topography. They partially define the catchment response, the catchment's reaction when subjected to a rainfall event (e.g. how fast the water level increases after a rainfall event).
Chaos	The property of a complex system, like the weather, whose behaviour is so unpredictable that it appears random. This is due to the system's sensitivity to small changes in conditions.
Confident	A forecaster's expert judgement of how certain they are that the forecast is right.
Convective rainfall events	The sun heats the ground, warming up the air above it. This causes the air to rise. As the air rises it cools and condenses, forming water droplets that organise into clouds and lead to rainfall. Convective rainfall events can lead to thunderstorms.
Department for Environment, Food and Rural Affairs (Defra)	UK government department responsible for safeguarding the UK's natural environment and supported by 33 agencies and public bodies, including the Environment Agency (EA). www.gov.uk/government/organisations/department-for-environment-food-rural-affairs

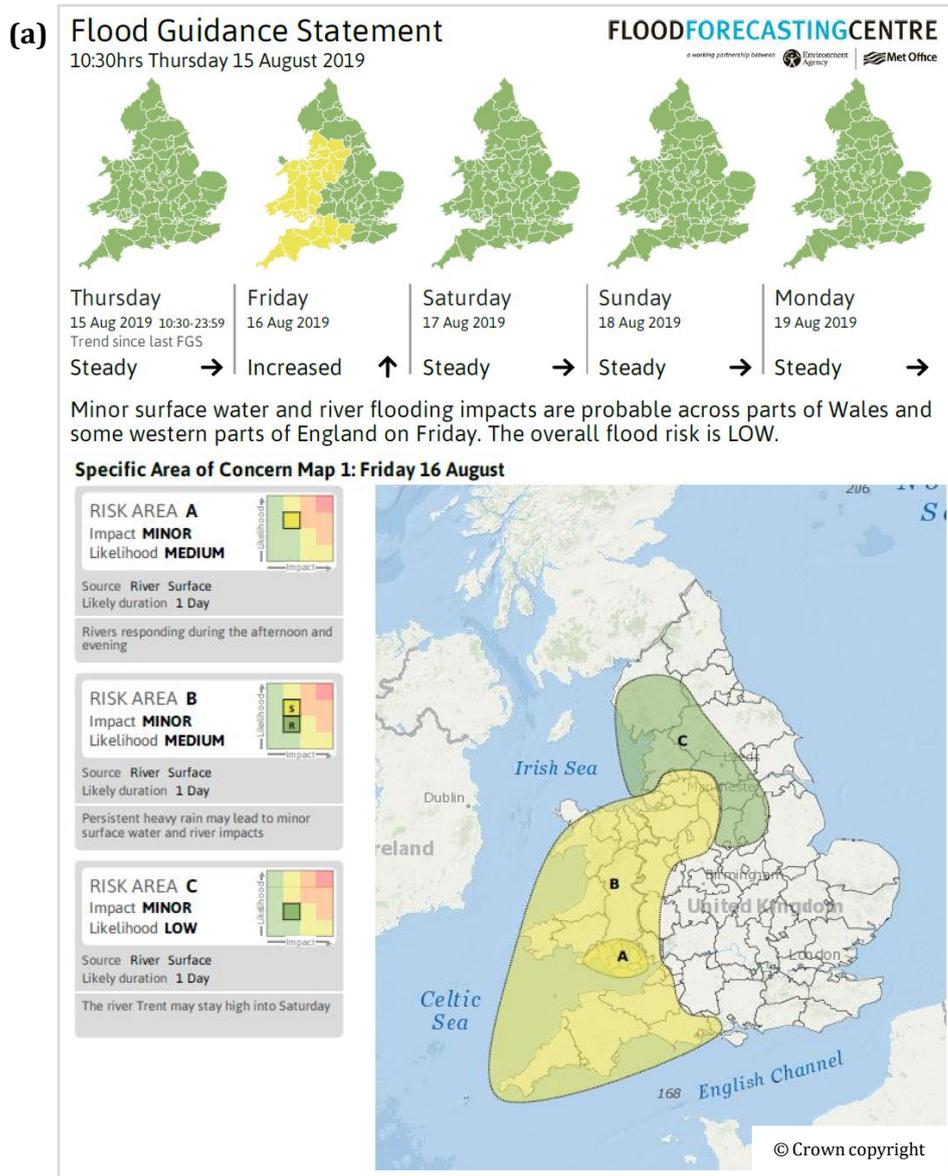
Deterministic forecasts	Refers to a forecast which gives a single possible outcome of the future rainfall, river and groundwater levels and coastal conditions.
Ensemble	Instead of running a single deterministic forecast, computer models can run a forecast several times, using slightly different inputs to account for uncertainties in the forecasting process. The complete set of forecasts is called an 'ensemble', and each individual forecast within it are 'ensemble members'. Each ensemble member represents a different possible scenario of future rainfall, river and groundwater levels and coastal conditions. Each scenario is equally likely to occur.
Environment Agency (EA)	<p>An executive non-departmental public body sponsored by Defra. The EA has an operational responsibility to manage risks of flooding from rivers and the sea in England, by warning and informing the public and businesses about impending floods.</p> <p>www.gov.uk/government/organisations/environment-agency</p>
False alarms	A warning given ahead of an event (e.g. flood) that does not ultimately occur.
Flood action groups	Cores of local people who act as representative voices for their wider community. They work alongside agencies and authorities and meet on a regular basis with the aim of reducing their community's flood risk and improving its resilience to flooding.
Flood Forecasting Centre (FFC)	<p>A partnership between the Environment Agency and the UK Met Office. It provides a UK-wide 24/7 hydro-meteorological service to emergency responders to better prepare for flooding (river, surface water, tidal/coastal and groundwater).</p> <p>www.ffc-environment-agency.metoffice.gov.uk</p>
Flood Guidance Statement (FGS)	A daily flood risk forecast for the UK, produced by the FFC (in collaboration with the EA and Natural Resources Wales) to assist with strategic, tactical and operational planning decisions. It gives a flood risk assessment shown by county and unitary authority across England and Wales over the next five days for all types of

	<p>natural flooding (coastal/tidal, river, groundwater and surface water). The FGS is issued by the FFC every day at 10:30am and at other times, day or night, if the flood risk assessment changes.</p> <p>www.fcc-environment-agency.metoffice.gov.uk/services/FGS_User_Guide.pdf</p>
Flood incident management strategy	An institute's priorities for preparing for and responding to flood events.
Flood management measures	Solutions to reduce the impacts that floods pose to humans and the environment. They can be natural (e.g. planting vegetation to retain extra water in the ground) or engineered (e.g. flood barriers).
Flood preparedness	Measures taken to prepare for and reduce the effects of a flood event.
Flood scenarios	Possible future development of a flood event and its associated likelihood.
Flood wardens	Volunteers from local communities who have the responsibility to monitor watercourses in the area they cover and contact local authorities with up to date information.
Forcing	The action of inputting information into a computer model to produce a forecast.
Forecast accuracy	The level of agreement between the forecast and the truth (i.e. what is observed in reality).
Forecasting product	A comprehensive and tailored overview (i.e. in the form of text, graphics and/or tables, etc.) of the forecast.
Hydraulic model	Mathematical model of the movement of water in a system (e.g. a river).
Hydrological model	Simplified model of a real-world system that describes the water cycle.

Hydro-meteorological observations and forecasts	Hydro-meteorology is a branch of meteorology and hydrology that studies the transfer of water and energy between the land surface and the lower atmosphere. Hydro-meteorological observations include observations of meteorological (e.g. temperature and rainfall) and hydrological variables (e.g. river and groundwater levels). Hydro-meteorological forecasts are forecasts that predict the evolution of meteorological and hydrological variables in time.
Hydro-Meteorological Services	Hydro-meteorological forecasting* products* produced by the FFC and issued daily (Hydro-Meteorological Guidance), twice daily (Forecast Meteorological Data) or whenever required (Heavy Rainfall Alerts).
Lead time	The length of time between when the forecast is made and the occurrence of the event (e.g. flood) being predicted.
Long-range forecasts	Forecasts which cover a period of time from a month to more than a season.
Missed flood events	A flood for which no warning was given ahead of it happening.
Model performance	The level of agreement between the model's outputs and their observations in reality. The difference between a model output and its respective observation is the error. The lower the error, the greater the model performance.
Nowcasting	Extrapolating from the latest observations (e.g. radar rainfall) to forecast the evolution of, for example the weather, in the next couple of hours.
Operational decision-making	Decision-making based on real-time information to resolve imminent situations.
Outlook	Refers to a forecasting product* based on long-range forecasts* (i.e. monthly to seasonal).
Performance measures	Metrics that characterise the quality of a forecast or a model compared to observations.

Probabilistic forecasts	While a deterministic model gives a single possible outcome for an event, a probabilistic model gives a probability distribution as a solution, indicating the likelihood of each scenario to occur. Probabilistic and ensemble forecasts are sometimes used interchangeably (see 'Ensemble').
Rapid-response catchments	Catchments and rivers that respond quickly to rainfall events.
Real-time river gauges	Instruments that measure a river's characteristics (e.g. flow or water level) and communicate these data in real-time remotely.
Reasonable Worst Case	A forecaster's assessment of the potential upper range of rainfall, river and groundwater levels, and coastal conditions, and their impacts.
Risk	A combination of likelihood and impact of an event.
River level correlations	Mathematical characterisation of the river level at one point of the river with respect to another point on the river. This can be used to estimate the river level at a point on the river if the river level upstream is known.
Short-range forecasts	Forecasts which cover a period of time from a couple of a hours to a couple of weeks.
Surface water flooding	Flooding caused when the volume of rainwater falling does not drain away through the river network and other drainage systems, or infiltrate into the ground, but lies on or flows over the ground.
Surge forecasts	Forecasts of the rise of water along coastlines.
Telemetry sites	Sites where instruments collect measurements automatically and transmit it remotely (see 'Real-time river gauges')
Uncertainty	Having limited knowledge or understanding of our environment, it is impossible to characterise and predict its evolution with 100% certainty. All forecasts are uncertain, and that uncertainty amplifies with lead time*. Ensemble* or probabilistic forecasting* can be used to represent the forecast uncertainty.

Appendix B. Visual examples of operational products used by EA MFDOs and FWDs: (a) Flood Guidance Statement, (b) Hydro-Meteorological Guidance, (c) Forecast Meteorological Data, (d) Heavy Rainfall Alert, and (e) National Flood Forecasting System (source: EA). The works published in this journal are distributed under the Creative Commons Attribution 4.0 License. This licence does not affect the Crown copyright work, which is re-usable under the Open Government Licence (OGL). The Creative Commons Attribution 4.0 License and the OGL are interoperable and do not conflict with, reduce or limit each other.



(b) Daily Hazard Assessment

Issued 14:02 on Wednesday, 14 August 2019

The Daily Hazard Assessment is intended to provide an 'at a glance' top level overview only. The links provided to the relevant Partner Organisations should then be used to obtain further

Hazards Five Day Summary – FLOOD: YELLOW, THUNDERSTORM: YELLOW

FLOOD: Significant surface water and river flooding impacts are possible but not expected across central England on Wednesday and into Thursday morning. The overall flood risk is LOW.

THUNDERSTORM: Heavy showers and thunderstorms may cause flooding and transport disruption for parts of central and eastern England on Wednesday afternoon.

Hazards Five Day Summary Maps

Wednesday 14-Aug 1400 – 2359	Thursday 15-Aug 0000 – 2359	Friday 16-Aug 0000 – 2359	Saturday 17-Aug 0000 – 2359	Sunday 18-Aug 0000 – 2359
------------------------------------	-----------------------------------	---------------------------------	-----------------------------------	---------------------------------



(c) **Forecast Meteorological Data**
EA South East Region

FLOODFORECASTINGCENTRE

a working partnership between  Environment Agency |  Met Office

Issued by the Flood Forecasting Centre on 15/08/19 at 05:11 GMT (06:11 local time)
 Unique Reference No. 5786 Version 1 Morning Issue

Precipitation Forecast Days 1 and 2

		Thursday 15/08/19					Friday 16/08/19			
		00-06 (GMT)	06-12	12-18	18-24	Day 1 Total (06-24)	00-06	06-12	12-24	Day 2 Total (00-24)
WT(N)	Ave (mm)		0	0	0	0	0	1	18	19
	Max (mm)		1	0	0	1	0	2	22	23
WT(S)	Ave (mm)		0	0	0	0	0	0	13	13
	Max (mm)		1	0	0	1	0	2	23	24
NET	Ave (mm)		0	0	0	0	0	1	11	12
	Max (mm)		1	0	0	1	0	2	16	16
HIOW	Ave (mm)		0	0	0	0	0	0	17	17
	Max (mm)		0	0	0	0	0	1	24	25
Sussex	Ave (mm)		0	0	0	0	0	0	11	11
	Max (mm)		1	0	0	1	0	1	18	18
KSL	Ave (mm)		0	0	0	0	0	0	8	8
	Max (mm)		1	0	0	1	0	1	12	12

WT(N)	West Thames (North)	WT(S)	West Thames (South)
NET	North East Thames	HIOW	Hampshire & IOW
KSL	Kent and South London		

Notes: All precipitation values are given as rainfall equivalents

Ave: Best estimate of mean rainfall depth over the Area during the period.

Max: Best estimate of maximum rainfall depth at any one location in this time period, this is not an extreme value.

Model Output: A number in brackets shows the original model output value. The number below this, is the FFC's Hydrometeorologists modification.

Daily Summary Days 1 – 5

		Thursday 15/08/19	Friday 16/08/19	Saturday 17/08/19	Sunday 18/08/19	Monday 19/08/19
Precipitation	Ave(mm)	See table above		1	1	0
	Max(mm)	See table above		3	4	2

		Thursday 15/08/19	Friday 16/08/19	Saturday 17/08/19	Sunday 18/08/19	Monday 19/08/19
Temperature	Min(degC)	10	8	12	11	11
	Max(degC)	22	20	23	23	© Crown copyright

(d) **Heavy Rainfall Alert
EA South East Region (Summer)**

FLOODFORECASTINGCENTRE

a working partnership between  Environment Agency |  Met Office

Issued by the Flood Forecasting Centre on 19/07/19 at 16:51 GMT (17:51 local time)
Unique Alert Reference No. 2817_SOUTHEAST_795 Version 1

ORIGINAL

Start of meteorological event: 0800 GMT on 20/07/19

End of meteorological event: 2100 GMT on 20/07/19

Summary of Alert Criteria Met

Alert Criteria	HRA Areas covered	Confidence
10 mm (or more) in 1 hours (or less)	West Thames (North)	L
	West Thames (South), North East Thames, Kent and South London	M
30 mm (or more) in 6 hours (or less)	Sussex, Kent and South London	L
30 mm (or more) in 12 hours (or less)	West Thames (South), North East Thames	L

Notes:

- **Confidence:** The probability of this threshold being achieved anywhere in the specific HRA Area within the time periods outlined by the Heavy Rainfall Alert. H = more than 60%; M = 40 – 60%; and L = 20 – 40%
- Issue of a Heavy Rainfall Alert means the probability of rainfall thresholds being met or exceeded during the meteorological event is within the bands indicated by the confidence levels above.
- All Alert criteria should be defined in this table. If it is predicted that some criteria will not be exceeded, these boxes should be greyed out

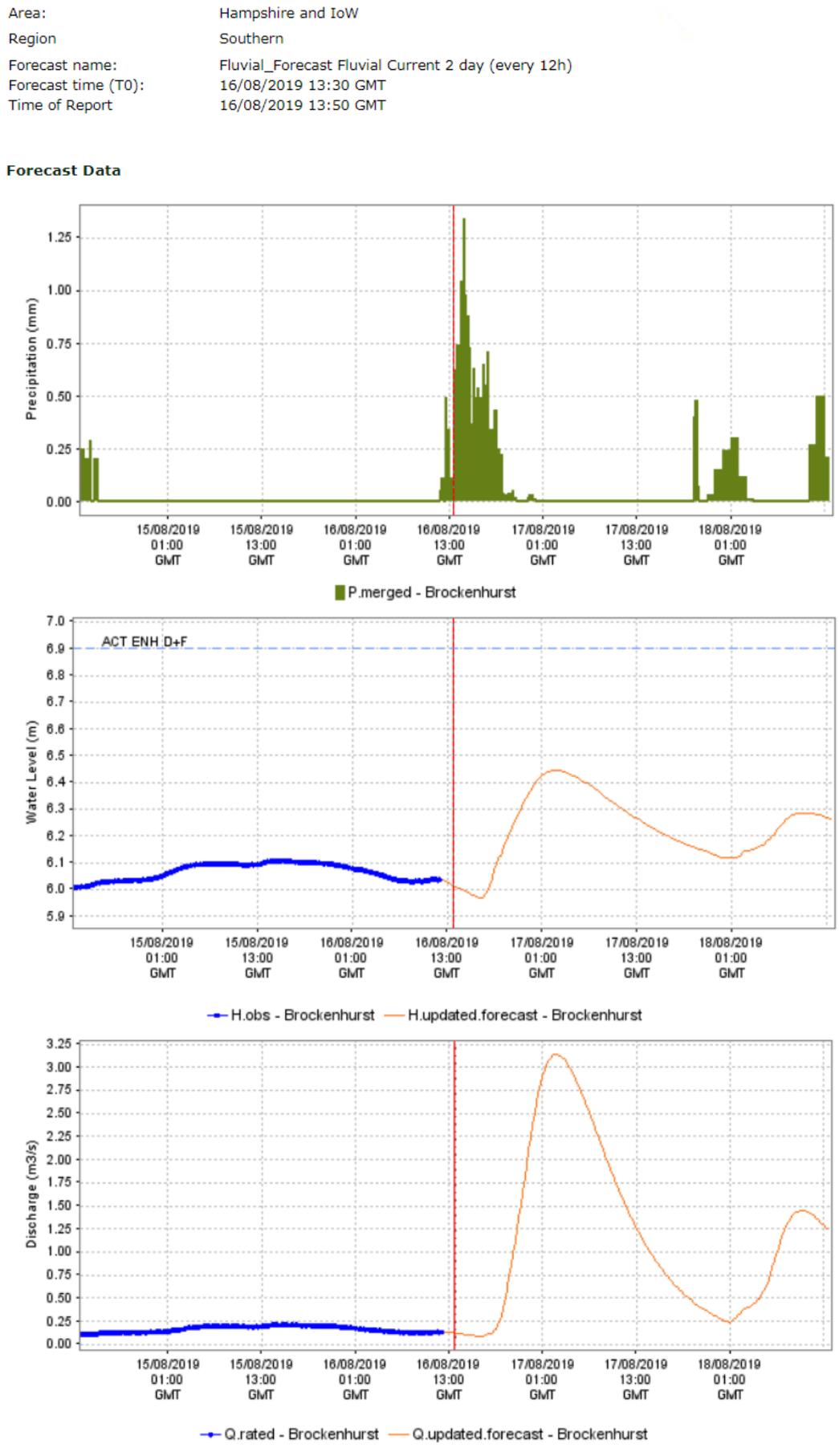


© Crown copyright

Each HRA Area is coloured according to the probability of its threshold being breached:

- Low (20 - 39%)
- Medium (40-59%)
- High (>= 60%)

(e)



Appendix C. Interviewees' quotes in relation to their perceptions (opportunities, neutral and challenges) associated with a transition to probabilistic forecasts.

Opportunities

01: *"If you've got a huge spread then you know that there's a very wide range of impact potentially, but if [...] everything's within a couple of centimetres of each other, it gives you a lot more confidence in saying, no I think we're going, we're not going to see a threshold crossing. So [...] it will help decision making I think" [MFDO3]*

02: *"I think in a good way [...] it will [...] reveal the uncertainty that's hidden by apparent simplicity" [I1]*

03: *"The new flood forecasting system is being developed at the moment so it's going to replace the NFFS. [The] benefits to that I suppose [...] are that if we can look to be more consistent across the country in even simple things like what displays look like [...] we're more interoperable if we need to" [MFDO1]*

04: *"I think in an incident I'm happy that that's [...] a useful range of things to know, like you said, you probably warn for the lowest one and plan for the highest one and we can interpret between them" [FWDO2]*

05: *"We're talking about some of these decisions that have got a long lead time, we're going to move people around the country, we're going to move equipment. It takes a long time to do that" [I1]*

06: *"Between us [duty officers], it's probably OK because we've got that understanding of the roles" [FWDO3]*

07: *"I can see some benefits to it, especially when you've got less experienced staff [...], you're almost [...] showing them the breadth of what a catchment could do given a range of responses" [MFDO2]*

Neutral

N1: *"Uncertainties are very tricky to deal with, whether probabilistic forecasting and a switch to that is going to help?" [MFDO2]*

N2: *"I think the MFDO role won't change, it will still be to communicate a forecast but the [...] wording of the forecast may change slightly" [MFDO1]*

N3: *"I think from our point of view it will just mean a bit more interpretation of forecasts and then [...] just a slightly different way of passing it on [...]. But I don't think it will change the process" [MFDO3]*

Challenges

C1: *“All the comms research we hear about generally says [...] the public message has to be as simple as possible, so that is working the opposite way to any proposal for probabilistic forecasting” [FWDO2]*

C2: *“A lot of local authorities standing their staff up, putting them on standby for a weekend is quite a big budget thing [...]. So [...] if we say, it is going to flood, they can justify the spend on it [...]. If we pass it on as shades of grey, a lot of them, they’ll appreciate the information but some of them would actually resent having the decision forced on them because they will struggle to then justify doing something or they’ll be blamed, either way, blamed for spending money if it doesn’t happen and blamed for not spending enough if it does happen.” [FWDO2]*

C3: *“That would be my concern that it’s even more information and more uncertainty and it’s kind of like, well what do you do with this information? And which bit do you communicate to who?” [FWDO3]*

C4: *“It is something to bear in mind with that if probabilistic forecasting put too much pressure and stress on decision making on the people in these roles, the system probably would just collapse, people would walk away” [FWDO2]*

C5: *“You’re still going to have this overriding issue with fast responding catchment where one scenario says we might need to issue a flood warning but 99 of them say no. Someone has to make a decision” [MFD01]*

C6: *“I think still for a lot of people the question they [...] want answered is am I going to flood?” [I2]*

C7: *“I think my role is going to be the one where it has to stop and it can’t be probabilistic because it [...] does come to a yes or no, issue it, don’t issue it. So to some extent, probabilistic forecasting does feel like everyone else just pushing things down the line saying you make the decision, [...], we have to make the decision because we’re the last ones on the line” [FWDO2]*

C8: *“Having probabilistic forecasting just moves the burden of making a decision further down the tree” [MFD02]*

Chapters 2 to 5 have shown how the combination of various disciplines and tools (e.g. the science of forecasting, decision-making, serious games, sensitivity analyses and interviews) can give a comprehensive overview of a specific topic. In this case: streamflow forecasting over Europe.

Art can be a powerful tool to communicate science, as well as making space for questioning, raising unconventional scientific questions of societal relevance. In Chapter 6, we will see how the combination of science and art can be beneficial to both fields. This chapter will also introduce a science and art exhibition, created as part of this PhD.

Chapter 6

The science and art of predicting floods

6.1 Background

6.1.1 Why (not) combine science and art?

At school, we have all been told that science and art are their own distinct disciplines, and followed classes in both of these, separately. The process behind these disciplines seems quite different. Art is about connecting feelings with a colour, a shape, a texture. The process nourishes itself from creativity and intuition. Science is about logic. About applying known concepts to solve a given problem objectively, trading intuition for hard facts.

“The beauty of art is that it can be provocative, metaphoric, poetic, it can use juxtaposition [...] to express. Art to me is an expressive medium. Whereas a scientist is trained to hold back personal expression. It’s a field that is not subjective, but objective. There to me is where there is a big difference, and that’s good, that’s fine.” (interview with Gayil Nalls¹, founder of the Art & Science Collaborations™ Inc.)

But this has not always been the case; the distinction between science and art is relatively recent. Until the 17th century, art was used to refer to a ‘skill’ or ‘mastery’, and was not differentiated from science. During the 15th century, Leonardo da Vinci was using science to elevate his art, fusing both disciplines to question the world in which he lived. Despite their apparent dissimilarities, science and art have respectively flourished from their combination and from the creativity that scientists and artists have shown. The creative process present in both science and art is perhaps their most common trait.

In the early 20th century, L.C.W. Bonacina believed that the close relationship between aesthetic and scientific problems was vital in reaching a unifying vision of our world. He argued that the description of meteorological events could not be done by numerical measurement and verbal categorisation alone, but needed to be completed by pictorial qualification. The contemplation of beautiful scenes can, for example, raise novel scientific questions, while scientific knowledge can enhance aesthetic appreciation.

¹ www.sciartmagazine.com/straight-talk-cynthia-pannucci.html

“It is difficult, for example, to believe that a person could completely come to terms with the majestic beauty of a thunder cloud, who had not at least a partial understanding of the mode in which water in mid-air is able to pile itself up into mountainous formations with peaks, domes, cliffs and gulleys.” (Bonacina, 1941)

6.1.2 Art as a scientific process

Science depends on the close observation of specimens and phenomena and the accurate recording of information. To this end, science relies on the artistic (or scientific) drawing process. Leonardo da Vinci, one of the greatest inventors of all times, produced an extensive amount of drawings and sketches, ranging from anatomy and physiology to engineering, following an observational approach to science.

In fact, the modern-day classification of clouds would not have existed without the fruitful combination of science and art. Before the 19th century, meteorologists thought each cloud to be unique in every way. We had to wait until 1802 for Luke Howard, amateur meteorologist, to put forward the now widely used clouds classification. Luke Howard spent years monitoring the skies and making sketches and watercolour paintings of clouds (Fig. 1) from which he came up with the three categories and descriptive Latin terms: cirrus (a curl of hair), cumulus (a heap), and stratus (a layer). Historians also believe Howard’s scientific classification of clouds has influenced artists throughout Europe, such as painters John Constable and Joseph M. W. Turner, changing the way clouds were depicted in many 19th century European paintings compared to earlier work².

² blog.sciencemuseum.org.uk/the-man-who-named-the-clouds/



Figure 1. Cloud study by Luke Howard.

Until recently, most weather centres have relied on their meteorologists drawing weather maps by hand as an important step of the operational weather forecast generation. It is a process which requires both scientific knowledge and intuition (Fig. 2).



Evelyn Müller
@meteoemüller

Manual forecast against machine! 100 meteorologists at #UEF2019 estimated the European weather 5 days after a given initial condition. The resulting ensemble of forecasts is quite unconfident!



Figure 2. 'Draw the forecast' science and art activity organised by L. Arnal at the 'Using ECMWF's Forecasts' event 2019 (source: Twitter).

6.1.3 Science as an artistic inspiration

Just like painters John Constable and Joseph Turner, inspired by Howard's scientific clouds classification, artists have been inspired by scientific discoveries and concepts for centuries.

In October 2018, the World Bank's Disaster Risk Finance and Insurance Program (DRFIP), the Global Facility for Disaster Reduction and Recovery (GFDRR), and the Government of Vanuatu co-hosted Understanding Risk Finance Pacific in Vanuatu. The centerpiece of the conference was a five-meter long data sculpture illustrating the range, frequency, and severity of disasters that have affected Pacific Island Countries over the last decades³ (Fig. 3). Based on 25 years (1993-2018) of data on atmospheric and oceanic parameters, extreme events, economic losses, deaths, and people affected by disasters in the Pacific region, the designer and lead artist, Pablo Suarez, wanted to communicate the risks of natural disasters in this part of the world. By engaging with the audience's emotions, the sculpture highlights the need to build resilience to climate risks.

³ blogs.worldbank.org/psd/giving-life-numbers-communicating-risk-through-art



By giving life to numbers, art can play a critical role in communicating #disaster risk. At @UnderstandRisk Finance Pacific, a captivating data sculpture served as a reminder of the importance of innovative financing for disaster #resilience: wrl.d.bg/yPm130nCybz



7:01 PM · Feb 11, 2019 · Hootsuite Inc.

Figure 3. 'Risk & Time: A Data Sculpture on Nature, Disasters, & Finance' (source: Twitter).

Another example is the art installation of UK-based artists Lise Autogena and Joshua Portway, called 'Most Blue Skies'⁴. As part of this installation, a computer runs continuously to find 'the bluest skies' in the world, by measuring the passage of light through particulate matter in the atmosphere and calculating the exact colours of the sky at billions of places on Earth. By combining "atmospheric research, environmental monitoring and sensing technologies with the romantic history of the blue sky and its fragile optimism", this project looks at "our changing relationship to the sky space as the subject for scientific and symbolic representation".

Climate change is a topic that has been widely explored by artists. The work of visual artist Andrecò for the Climate 04-Sea Level Rise project is one example⁵. Inspired by international research about the effects of sea level rise and extreme waves in the Venice lagoon, Andrecò created a meters' long wall painting to bring these scientific concepts and motivations to

⁴ www.autogena.org/mbs.html

⁵ streetartnews.net/2017/10/climate-04-sea-level-rise-project-art-science-andrecò-venice.html

the streets of Venice (Fig. 4). Reaching a wide audience, the painting stimulated public discussions on the causes and effects of climate change.



Figure 4. Climate 04-Sea Level Rise project mural by Andreco (source: Twitter).

While scientists might experience difficulties bringing their findings to the public, art can bridge this gap by touching the public's imagination and emotions. Artists can use juxtaposition, metaphors, poetry, engage with the audience's senses to share information and create a questioning in the audience's mind.

"Artists excel at bringing ideas that you cannot see with a naked eye and conceptualise in them. They can reach not only the heart but the mind." (interview with Gayil Nalls¹)

Another strength of using art to communicate science is the possibility to overcome problems associated with a language barrier (as mentioned in Chapter 5 regarding the communication between scientists and decision-makers).

Science and art collaborations have soared in the last few decades, with the term SciArt coined to describe this rising phenomenon. There are nowadays many SciArt projects on the topic of weather and water. Science and art is also becoming more present amongst scientific institutes and at scientific conferences. The European Commission's Joint

Research Centre (JRC) has been hosting a SciArt summer school for the past few years. This is followed by the Resonances festival, where the co-created work of artists, JRC and invited scientists is presented to trigger conversations on a pre-defined topic. Art is also becoming more present at the European Geosciences Union (EGU) General Assembly, where a SciArt session has been held for the past couple of years for scientists and artists to display their SciArt work on any environmental topic. In 2018 for the first time, the EGU also hosted two artists in residence at the General Assembly (Fig. 5). As part of Pint of Science, a global festival which aims to bring scientific research to local pubs to reach a wider audience, several cities have organised SciArt-themed events ('Creative Reactions'⁶). This year, Reading organised its first Creative Reactions event (of which I was one of the co-organisers; Fig. 6). The event brought together about 30 scientists and artists from around Reading to talk about and create art inspired from a range of scientific topics, displayed during the event.



Figure 5. Artworks by artists in residence at the EGU GA 2019 (source: Twitter).

⁶ pintofscience.co.uk/creativereactions



Figure 6. Reading Pint of Science 'Creative Reactions' 2019 (source: Twitter).

6.2 'Gambling with floods?' A science and art exhibition by L. Arnal

6.2.1 My science and art practice

I am a scientist with a lifelong love of art. Growing up in a family of artists, art has been a go-to-activity and a way for me to express myself since a very young age. But when I started my PhD, I put art aside, thinking that I had to dedicate all my time to science.

Two years through my PhD, I went through a difficult phase, as many PhD students will experience. I picked up art again, capturing my days at work through sketches. Art became, once again, an essential part of my life, and one of the main reasons you are reading this sentence.

I realised that, this once personal tool, could be used to share my passion for the science I do with an audience. This led to my first SciArt piece, 'The science & art of predicting floods'⁷. In this video, I am telling my mother (an artist) what my PhD is about, while she draws what it means to her. Through this project, I wanted to share my passion for my PhD topic with my mum, and an audience like her (with no a priori knowledge of the field). I also wanted to visualise how my mum understood the science I do.

⁷ sciartfloods.wordpress.com/2018/03/03/the-journey-begins/

This video inspired me to work on many more and various SciArt projects (e.g. comics summarising my scientific papers, live-cartoons capturing an event, chaos SciArt experiments, SciArt activities organised for children and adults, collaboration with artists for the JRC's Resonances festival 2019; see my blog⁸ for some examples), all related to my PhD's topic. Through these projects, I wanted to bring together two of my passions, science and art, to reach a wider audience than the scientific community to which discussions around my PhD's topic were usually confined (via the more classical formats of scientific papers, presentations at conferences, and ultimately the PhD thesis).

These projects bolstered my science and art practice through the varied challenges they brought. They also reinforced my passion for SciArt and my aspiration to steer my career towards this emerging field.

6.2.2 Art exhibition

Building on these last couple of years of experience doing SciArt, I wanted to create an exhibition that captures the core of my PhD topic and practice. Having realised multiple static SciArt pieces in the past, I wanted this piece to be dynamic, to recreate an experience metaphorical of my PhD practice and of forecasting (a very dynamic process per se).

Through the creation of this SciArt exhibition, I also wanted to challenge the traditional format of the scientific PhD thesis and pave the way for (and hopefully inspire) future PhD students to use art as a tool and process within their scientific practice.

My PhD's final art piece is a SciArt exhibition called 'Gambling with floods?' (Fig. 7). This exhibition makes space for a wide audience to explore the science and art of flood forecasting, questioning how we, humans, can work with machines to anticipate floods in the noisy chaos of nature. Below is a synopsis of the exhibition.

⁸ sciartfloods.wordpress.com

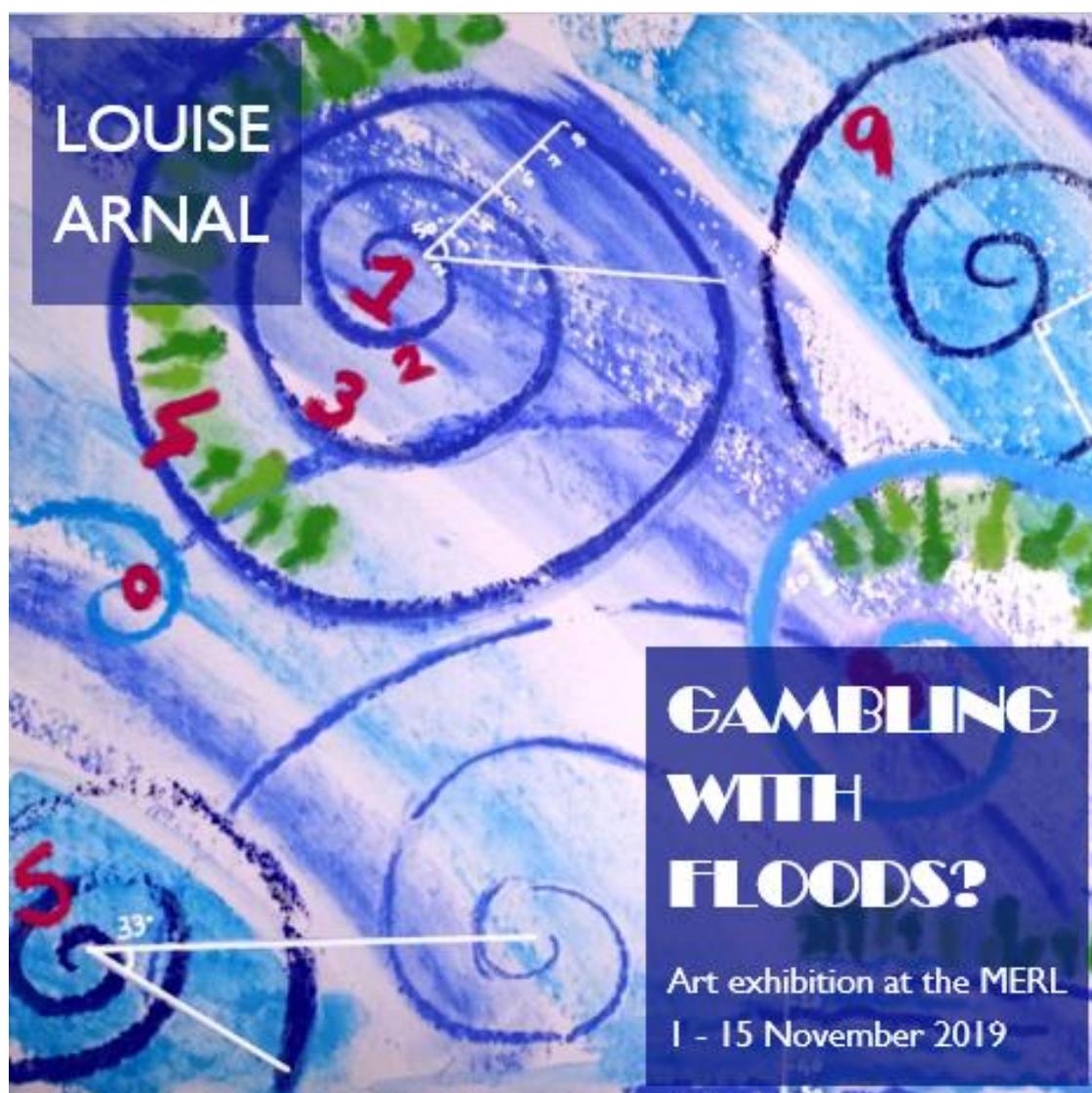


Figure 7. 'Gambling with floods?' art exhibition by L. Arnal.

Floods are expected to become more intense and frequent in the future. With an increasing population at risk, it is vital to predict these events well in advance. Machines are taking an ever-growing role in the prediction and anticipation of floods. But to what extent can they be trusted? How much can be predicted by machines in the noisy chaos of nature? And where is the place of machines in the midst of long-standing traditional methods, based on experience and intuition, such as weather-related proverbs?

These are all questions explored through a multi-sensory immersive science and art installation, in which the spectator becomes actor, a forecaster playing with real data and creating a forecast, at the heart of the forecasting machine. The installation will stimulate the spectators' senses (sight, hearing and touch) through the overlay of flood related sounds and images, filling up the space of a small and dark room from top to bottom. Triggered by a

jackpot machine (controlled by the spectator-actor), these sounds and images will build-up towards a finale, the creation of a forecast.

The spectator will be prompted to activate the jackpot machine to create the forecast, with a single question posed on the wall. The installation will be layered in space and time and partly codified to recreate an atmosphere metaphorical of my PhD's research, and evoke the precision of the 'flood forecasting machine'. Contrastingly, the installation will be controlled by a jackpot machine to symbolise the randomness or chaos in flood forecasts, leading the spectator-actor to wonder: 'can I actually trust the forecast I'm creating?' or 'Will next time I activate the jackpot machine be the right one?' The whole installation is built on the juxtaposition of the plurality of chaos with the singularity of the concept of predictability.

This exhibition mirrors my PhD's topic through its overarching theme, as well as installation set-up. For example, the spectator-actor has to wait a certain amount of time to see the forecast they are creating, raising their emotions and perhaps apprehension of the result to come. This is to reflect the notion of time inherent and important in forecasting. The weather proverbs will remind us of more traditional 'predictions' based on past observations, such as the ESP used as a benchmark in Chapters 3 and 4.

This installation was exhibited at the Museum of English Rural Life (Reading, UK) from 1st to 15th November 2019. In order to engage with the audience, there was additionally a 'meet the artist' event on 6th November 2019.

The exhibition grasps the essence of the scientific field and explores complex scientific themes (e.g. flood probabilities and chaos theory). The beauty of it being art is that it can however leave many of these scientific themes abstract. It is then the spectator-actor's choice whether they want to find out more by reading the exhibition brochure, which provides a more detailed perspective on key scientific aspects of the exhibition and where they sit in the wider context of flood forecasting. For instance, the jackpot machine contains some notion of chaos and forecast probabilities, which the spectator can find out more about in the exhibition brochure after having experienced the installation.

But this exhibition is more than a tool to communicate science to society. By provoking the audience, through the exhibition's name, the installation concept (i.e. the jackpot machine being used to create a flood forecast) and the audience's experience when the installation is running, it aims to deconstruct pre-acquired concepts, to create a moment of confusion, hopefully followed by a self-gained personal new perspective (and perhaps also the will to find out more about the topic at heart). By being linked to a physical space in time, this

exhibition also created a space for science and society, scientists and the public, to get one step closer to one another, bridging the existing gap between these two worlds.

This exhibition generated interest from a diverse audience, from The MERL's regular visitors, to artists, scientists, and decision-makers from the University of Reading, ECMWF and the Environment Agency, to cite a few. These were also all represented at the 'meet the artist' event and could engage further with one another on flood forecasting and decision-making. The Environment Agency exhibited interest to, in the future, use art as part of their communication strategy. The hope is that the next time one of the exhibition visitors (with before no a priori knowledge of the field) sees a flood forecast, they will be able to connect the emotions they felt in the exhibition space (and knowledge gained through these) to a forecast which may have in the past appeared as a cold and blurry scientific fact.

Acknowledgements. L. Arnal would like to acknowledge the valuable help of Stuart Mitchell with the technical installation. L. Arnal would also like to thank the guidance and support during the creative process and planning from: Julius Kreißig, Dominique Caseneuve, Marie Arnal, Hannah Cloke, Florian Pappenberger, Lise Autogena, Joshua Portway, Katie Cooper, Hilda Carr and Katya Dimitrova. L. Arnal would like to thank the twittersphere for their weather sayings and the following persons for their voices and photos: Voices & photos: Hannah Cloke, Rebecca Emerton, Pete Castle, Jess Neumann, Jamie Towner, Sazzad Hossain, Ervin Zsoter, Damien Decremer, Calum Baugh, Siobhan Dolan, Maureen Wanzala, Liz Stephens, Claudia di Napoli, Jamie Towner, Shaun Harrigan, Helen Griffith, Fredrik Wetterhall, Christel Prudhomme and Florian Pappenberger. L. Arnal received funding from the EU Horizon 2020 IMPREX project (www.imprex.eu) (641811) and the University of Reading for this exhibition.

Chapter 7

Conclusions

The aim of this thesis, as part of the IMPREX Horizon 2020 project, was to improve operational flood early warning in Europe, through long-term forecasting and fostering the uptake of state-of-the-art scientific information in practice and outside of the scientific community. As the science and technology is improving, we are able to produce forecasts with longer lead times, such as the seasonal timescale. While seasonal hydro-meteorological forecasts have the potential to inform a range of water sectors, their operationalisation, performance assessment and applicability to flood early warning in Europe has until now been poorly investigated.

This thesis has enriched research and the operational worlds of forecast generation and decision-making, bridging the gap between science and society, combining decision-makers', scientific and artistic perspectives throughout the following objectives:

- 1 Explore decision-makers' requirements when using probabilistic hydrological forecasts for flood early warning.
- 2 Investigate the current capabilities in seasonal hydrological forecasting on the global, continental and basin scales, and implement a global and European operational seasonal streamflow outlook.
- 3 Develop a cost-efficient method for tangible seasonal streamflow forecast improvements and apply it over Europe.
- 4 Suggest ways to facilitate decision-makers' intake of probabilistic flood forecasts.
- 5 Explore art as a tool to bridge the gap between science and society.

This thesis was structured around these five objectives, explored in a variety of ways, including: the co-creation of two serious games and a decision-making activity, the publication of four first-author (of which one is in review) and three co-author peer-reviewed published articles and a first-author peer-reviewed published IMPREX deliverable, the co-development of two operational seasonal streamflow outlooks, research interviews and the creation of an art exhibition.

This chapter summarises the lessons learnt for each objective, discusses and highlights the (scientific) advances of this thesis and proposes directions for future work.

7.1 Lessons learnt

7.1.1 Objective 1: Decision-makers' requirements when using probabilistic hydrological forecasts for flood early warning

We live in a society in which science and decision-making are sometimes disconnected. This disconnect is perhaps more noticeable for high societal stakes, such as the anticipation of floods. The first objective of this thesis was an attempt at reconnecting both worlds, through an understanding of decision-makers' requirements when using probabilistic hydrological forecasts for flood early warning.

This was addressed using results from a HEPEX serious game (Arnal et al., 2016; Chapter 2), an IMPREX decision-making activity (Neumann et al., 2018b; Chapter 2) and research interviews at the UK Environment Agency (EA; Arnal et al., in review; Chapter 5). For the serious game and the decision-making activity, participants were given probabilistic information, based on which they had to make decisions associated with flood anticipation activities. The interviews were carried out with EA decision-makers to understand their perspectives on potential changes associated a future transition from deterministic to probabilistic flood forecasts at the EA.

Serious games, decision-making activities and interviews are very useful tools for a better understanding of complex topics (such as this objective's). They can bring to light important hidden elements, through simple scenarios and open questions, providing directions for future research. They can additionally foster discussions among different communities, such as decision-makers and scientists.

While the HEPEX serious game and the interviews were based on short-term probabilistic flood forecasts, the IMPREX decision-making activity was designed around seasonal hydrological forecasts. Many of the conclusions drawn from these various activities were however similar, applicable for a range of timescales.

Results from the HEPEX serious game and the IMPREX activity showed that **several factors shape a decision-maker's uptake of forecast information and are important for informing decisions. These include: resources available, perceptions of forecast quality, experience of previous flood events, local knowledge and risk appetite.** These results are in line with the research interviews with decision-makers at the EA. These interviews also highlighted institutional practices and external considerations (e.g. the necessity to align flood warnings with key responders) as key factors.

Despite these factors, results hinted that **(probabilistic) forecasts, even biased or uncertain, are a valuable element of the decision-making chain**, defining the timing of warning and response activities at the EA. Uncertainty is not a limitation of probabilistic forecasts per se. While EA decision-makers have access to and use forecast quality information on a daily basis, results from HEPEX and IMPREX activities both emphasised that **forecast quality should be (better) communicated with decision-makers to ensure appropriate use of information and more informed decisions**.

The IMPREX decision-making activity communicated results which, although made with respect to seasonal hydrological forecast, might also apply to shorter- or longer-term forecasts. It identified the **need to communicate forecast information in a combination of different formats** (e.g. maps, texts, hydrographs and tables), **tailored to different user groups, to enhance interpretation and use**. Moreover, it was shown that to be more useful, **(seasonal) hydrological forecasts should be presented at a scale which matches that employed in decision-making**. Finally, local stakeholders and decision-makers expressed the **necessity of improved communication between scientists, providers and users to ensure that users are kept up to date with scientific developments**.

As identified by the IMPREX activity participants and the EA decision-makers, **language can be a barrier to communication and ultimately forecast use**. A lot more work is needed on translating a probabilistic information into a binary decision, communicating the shades of grey.

7.1.2 Objective 2: Current capabilities in seasonal hydrological forecasting on the global, continental and basin scales and operational implementation of a global and European seasonal streamflow outlook

Results from the IMPREX decision-making activity hinted that seasonal hydrological forecasts have the potential to inform a range of flood anticipation decisions and actions. Combining seasonal meteorological forecasts with hydrological models is a recent endeavour in the world of operational hydrological forecasting. Despite its potential to inform decisions and actions in theory, the literature has shown that the predictability of seasonal hydro-meteorological forecasting is still limited, especially over Europe. However, results from the first objective of this thesis highlighted that uncertain forecasts are still valuable for decision-making if the forecast quality is adequately communicated to ensure the appropriate use of information and more informed decisions.

The second objective of this thesis was to assess our current capabilities in seasonal hydrological forecasting, through the co-development and analysis of two operational seasonal hydro-meteorological forecast outlooks: EFAS-Seasonal over Europe (Arnal et al., 2018; Chapter 3) and at the basin scale (Neumann et al., 2018a; Chapter 3) and GloFAS-Seasonal at the global scale (Emerton et al., 2018; Chapter 3).

Working at the ECMWF as part of this PhD led to the **operational implementation of this research as one of the first pan-European seasonal hydro-meteorological forecasting system**, as part of EFAS, **and guidance in the design of the first global scale seasonal hydro-meteorological forecasting system outlook**, as part of GloFAS.

EFAS-Seasonal and GloFAS-Seasonal were assessed in terms of the hindcasts' overall performance, as well as their ability to provide an earlier indication of low and high (as a proxy for floods) streamflows over Europe and for the global river network, respectively. The potential of EFAS-Seasonal for flood early warning was also investigated at the basin scale, for the 2013/14 Thames (UK) river floods. This study also explored the impact of improvements in seasonal meteorological forecasting to the hydrology.

EFAS-Seasonal and GloFAS-Seasonal provide streamflow forecasts out to two months and four months ahead for the European and global river networks, respectively. They are openly available through the EFAS¹ and GloFAS² websites.

Findings from both EFAS-Seasonal and GloFAS-Seasonal communicated the current limitations of seasonal hydro-meteorological forecasts. The analysis of EFAS-Seasonal showed that, on average over Europe, seasonal meteorological forecasts translate to skilful seasonal hydrological forecasts for the first month of lead time only. The performance of the system (in terms of hindcast accuracy, sharpness and overall performance) varies by region and season and EFAS-Seasonal is more skilful on average at predicting autumn and winter streamflows than for the spring and summer. The performance of GloFAS-Seasonal (hindcast skill and reliability) was also found to vary by region and season, over-predicting streamflow in general. These results highlight the **importance of initial hydrological conditions and the land surface memory for seasonal streamflow forecasting.**

However, it was also observed that **hydrological predictability can be gained from seasonal meteorological forecasts** (compared to historical weather information). **EFAS-Seasonal and GloFAS-Seasonal are potentially useful for predicting low and high**

¹ www.efas.eu

² www.globalfloods.eu

streamflow events out to several months ahead in some cases, noticeably in winter for almost 40% of the European regions.

Results from the analysis of EFAS-Seasonal at the European scale and for the Thames river floods of 2013/14 showed that **seasonal meteorological forecast skill does not always translate to more skilful seasonal hydrological forecasts**. There is knowledge to be gained from jointly considering the hydrology and meteorology.

While extreme events such as the 2013/14 Thames river floods are difficult to predict with confidence at seasonal timescales, considering the local hydrogeological context can provide an effective early alert of potentially high impact events, for better preparedness and greater confidence in the forecasts as the event approaches. Seasonal hydrological forecasts could be used as complementary information to existing forecasts on shorter timescales, to provide monitoring and early-warning information for flood preparedness.

It may seem as though we have reached a limit in the predictability of seasonal streamflow forecasting. However, as climate change is likely to generate changes in the drivers of seasonal predictability with time, these limits of predictability will equally shift. **Seasonal predictability is a dynamic field of research which deserves greater future investigation.**

7.1.3 Objective 3: A cost-efficient method for tangible seasonal streamflow forecast improvements

As shown in the second objective and by the literature (Chapter 4), seasonal hydro-meteorological forecast skill is still limited in many parts of the world, including Europe. There exist various methods that can lead to forecast improvements, including the post-processing of seasonal meteorological forecasts and the assimilation of better hydro-meteorological observations. However, these methods are costly, adding to the costs and resources needed to produce and maintain a seasonal hydro-meteorological forecasting system in the first place. Moreover, research has shown that operational streamflow forecast quality has not significantly improved in the last decade, despite the costly research and developments they are receiving. Wood et al. (2016a) proposed a sensitivity analysis method, VESPA, which can disentangle dominant error sources in seasonal streamflow forecasting systems, to guide future system developments. Yet, VESPA is computationally intensive, a limitation to its use.

This motivated the third objective of this thesis, to develop a cost-efficient alternative method to VESPA. **An alternative sensitivity analysis to VESPA was co-developed and tested with Andy Wood during a visit to the National Center for Atmospheric Research (NCAR) in Boulder (US).** This is presented in a first-author publication (Arnal et al., 2017b; Chapter 4). The method was subsequently applied to the EFAS-Seasonal forecasting system in a first-author IMPREX deliverable (Arnal et al., 2017a; Chapter 4). Results from these publications inspired additional research, introducing the novel concept for a cost-efficient seasonal hydro-meteorological forecasting system.

The EPB sensitivity analysis method is a computationally inexpensive alternative to VESPA. It is easy to implement and can identify the dominant predictability source (i.e. initial hydrological conditions and seasonal meteorological forecasts) of seasonal streamflow forecast skill, for a given basin and forecast initialisation date. Using the EPB, forecast skill elasticities can easily be calculated, indicating the degree of influence of changes in either of the predictability sources' skill on the streamflow forecast skill. **The EPB is therefore valuable to guide future developments for tangible seasonal streamflow forecast improvements.**

The EPB was first applied over Europe on the EFAS-Seasonal forecasting system (based on the hindcasts CRPSS). Overall, results showed that **improving the seasonal meteorological forecast skill would lead to the largest seasonal streamflow forecast improvements beyond the first month of lead time**, when the influence of the initial hydrological conditions on the streamflow start to disappear. The second objective of this thesis has indeed shown that currently, seasonal meteorological forecasts translate to skilful seasonal hydrological forecasts for the first month of lead time only over Europe.

However, **larger seasonal streamflow forecast improvements can be obtained from improving the initial hydrological conditions beyond the first month of lead time in regions where streamflow is driven by snowmelt and groundwater.**

Operational seasonal hydro-meteorological forecasting systems do not reflect these forecast predictability findings in their setup. Seasonal hydro-meteorological forecasts are usually computed with a fixed method, which might not be the most suitable (resource-friendly, skilful and of use to decision-makers) for all forecast initialisation dates and river basins. **We should challenge current practices and reconsider the way we produce seasonal hydro-meteorological forecasts operationally.**

7.1.4 Objective 4: Facilitating decision-makers' use of probabilistic flood forecasts

The first objective of this thesis has given us perspectives from the decision-makers on the use of probabilistic forecasts for flood early warning. Subsequently, perspectives were gained from the science (from objectives 2 and 3) on the current limits of predictability in seasonal hydro-meteorological forecasting for earlier flood warnings, challenges in operational forecast development and product design with EFAS-Seasonal and GloFAS-Seasonal. Objectives 1 to 3 all showed that while the science is far from being perfect, it is still valuable for decision-making. These perspectives were unified to suggest ways to promote decision-makers' uptake of probabilistic flood forecasts (and new science in general).

This was addressed through research interviews at the EA, with the aim to understand the context in which EA decision-makers currently operate (Arnal et al., in review; Chapter 5). Based on these findings, **recommendations were made to the EA** (which might apply to other institutes in a similar situation) **to support a successful transition to probabilistic forecasts for flood early warning in England.**

The interviews' analysis showed that, in a flood incident management context, EA decision-makers are influenced by both hard (scientific) facts and soft values (as shown in objective 1), and **combining these sources of uncertain information to make a binary decision is seen as a major challenge.**

More specifically, findings from objective 1 highlighted that even uncertain or biased, forecasts can inform decisions and actions and should be communicated appropriately to forecast users for an appropriate use of the information. Forecast products additionally need to be tailored and communicated at the scale at which decisions are made. And language can be a barrier to forecast use. All of these findings suggest that **a closer interaction between decision-makers and scientists is needed, for a co-leadership on improving science for society, the co-design of forecasting products, and the establishment of a common language.** This more widely calls for an interdisciplinary approach to science, including decision-makers, scientists and other relevant fields (e.g. social scientists).

While we address the weaknesses of the forecasting chain, we should also use its strengths. For example, EA decision-makers (Monitoring and Forecasting Duty Officers (MFDOs) and Flood Warning Duty Officers (FWDOs)) work closely together and do not perceive a transition to probabilistic flood forecasts as a threat to their interaction. The

MFDOs could help FWDOs understand the strengths and intricacies of probabilistic forecasting to reduce their fears of having to make a decision on “visibly uncertain” information.

Throughout this thesis, the term “binary decision” has been used to refer to the ultimate decision taken by a decision-maker (e.g. sending out a flood warning). It was also used by the interviewed EA duty officers. However, **the term “binary decision” might not be the most appropriate as its simplicity hides real-life complexities and the shades of grey associated with and inherent to the decision-making process in the anticipation of extreme events** (as seen in Chapter 5).

7.1.5 Objective 5: Art as a tool to bridge the gap between science and society

As seen in the previous objectives of this thesis (Chapters 2 to 5), there is a disconnect between science and practice for flood anticipation, a topic of high societal relevance. Language is a barrier to communication and the use and uptake of scientific research. Art does not rely on language to engage with its audience. It can use a juxtaposition of methods and engage a range of senses to convey an idea in an abstract and metaphorical way, giving a larger audience the opportunity to be a part of relevant scientific and decision-making developments. Art can also be used to question and challenge the way things are done in science, giving a new perspective on well-established practices. **Art is a powerful tool that can bridge the gap between science and society.**

This motivated the last objective of this thesis, through **the creation of a science and art exhibition, as part of this scientific PhD**. Through this multi-sensory science and art installation, I wanted to give the opportunity to an audience outside of the scientific community to which discussions about my PhD’s topic are usually confined, to experience flood forecasting first-hand. **This installation questions the limit between predictability and chaos in seasonal streamflow forecasting, and the place of computers in the noisy chaos of nature.**

This exhibition was more than a tool to communicate science to society. By provoking the audience, its aim was to deconstruct pre-acquired concepts through confusion in order to instigate a self-gained personal new perspective on the topic at heart. **This exhibition also constituted a physical space for scientists and the public to meet, bridging the existing gap between science and society.** It generated interest from a diverse audience, from the museum’s regular visitors, to artists, scientists, and decision-makers from the University of Reading, ECMWF and the EA.

7.2 Key contributions

This thesis considers the full forecasting chain and uses a diversity of methods, including serious games, interviews, scientific research and art. The specific key contributions of this work are summarised below:

- An IMPREX online serious game³ that allows a broad audience to proactively explore the full forecasting chain (from forecast production to evaluating the situation and making a decision), to understand the challenges and opportunities of flood preparedness based on probabilistic flood forecasts, and to appreciate the flood warning ‘behind-the-scenes’.
- Co-design of some of the first operational seasonal hydro-meteorological outlooks over Europe and for the globe.
- Benchmark of state-of-the-art dynamical seasonal streamflow forecast performance over Europe.
- People have looked at using seasonal hydrological forecasts for many water sector applications, but flood anticipation was until now mostly unexplored. This thesis has shown the potential usefulness of seasonal hydrological forecasts for this field. This work builds on a study by Coughlan de Perez et al. (2017), who looked at the usability of seasonal rainfall forecasts for flood preparedness.
- Paving the way for the uptake of seasonal hydrological forecasts in the Thames river basin. While the full potential of seasonal hydrological forecasts has not been attained yet, making the existing systems operational and familiarising users with their concepts, challenges and opportunities is key to their successful future applicability.
- A novel cost-efficient sensitivity analysis to guide future seasonal hydro-meteorological forecast developments for tangible improvements. This can be applied easily to any forecasting system and has already been used by several IMPREX partners, as well as for decadal predictions of terrestrial water storage (Zhu et al., 2019). This complements work by Hawkins and Sutton (2009; 2010), who have looked at attributing uncertainty in climate predictions of temperature and precipitation change.
- Recommendations were given to the EA, to support a successful transition to operational probabilistic fluvial flood forecasts for flood early warning in England.

³ www.imprex.arctik.tech

- Art (via blog posts, individual and collaborative art pieces, SciArt activities at scientific events, and an exhibition) used to extend scientific discussions and engagement outside of the community of practice.

7.3 Next steps

This research forms an estuary to a sea of future research directions.

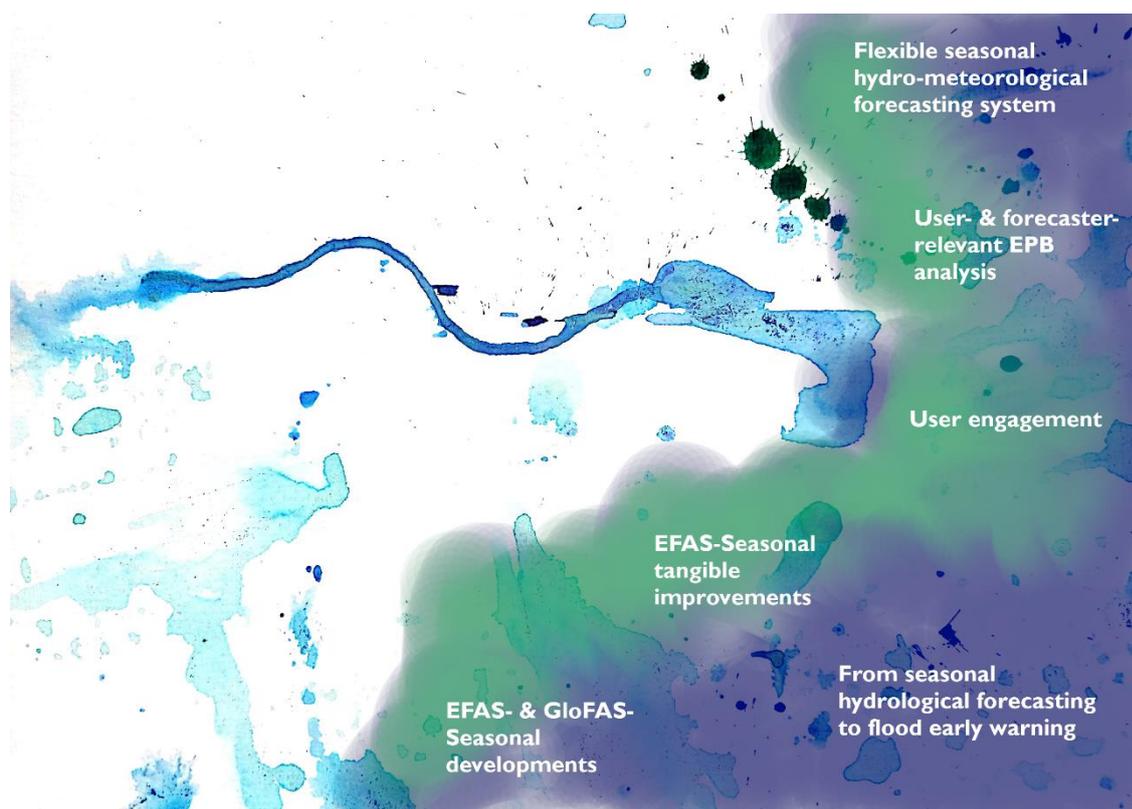


Figure 1. A sea of future research directions.

- **User engagement.** “We conclude that the current level of understanding in the West Thames provides an excellent basis upon which to incorporate future developments of operational forecasts and for facilitating communication and decision-making between water sector partners” (Neumann et al., 2018b). As highlighted here, user engagement is a key process of scientific development (from forecasting system design to forecast analysis and use in practice). To this end, focus groups, carried out under the umbrella of wider funded projects (such as IMPREX), are great initiatives. They should however carry on beyond a project’s end.
- **EFAS- & GloFAS-Seasonal developments.** Engagement with EFAS, GloFAS and West Thames user communities has highlighted decision-makers’ needs for using seasonal hydrological forecasts in practice (e.g. displaying forecast performance,

tailoring products for a range of applications, communicating the forecasts on an actionable scale, etc.). These feedbacks should be used to drive EFAS-Seasonal and GloFAS-Seasonal developments.

- **User- and forecaster-relevant EPB analysis.** Arnal et al. (2017b) showed that the choice of verification score impacts the forecast's sensitivity to improvements made to its predictability sources (Chapter 4, Sect. 4.2). There exist a variety of forecast verification scores, designed to assess different aspects of a forecast and its value for decision-making (Cloke et al., 2017). The EPB sensitivity analysis should be performed with verification scores relevant for decision-makers (e.g. the ROC score to improve the prediction of low and high streamflows; after further research of the limitations of the EPB for certain verification scores – as seen in Chapter 4, Sect. 4.2 – has been conducted). Moreover, to clearly guide system developments by forecasters and model developers, the EPB sensitivity analysis would benefit from further improvements, allowing to pinpoint the specific hydro-meteorological variables (from the initial hydrological conditions and the seasonal meteorological forecasts) affecting skill. The effect of degradations in the temporal and spatial accuracy of the input data could also be of added value.
- **EFAS-Seasonal tangible improvements.** Results of the EPB sensitivity analysis applied to EFAS-Seasonal (Chapter 4, Sect. 4.3) should be used to drive forecast developments. It would be interesting to check whether obtained improvements match expectations.
- **From seasonal hydrological forecasting to flood early warning.** Discussions with the EFAS, GloFAS and West Thames user communities showed that there are still challenges in the operational use of seasonal hydrological forecasts for flood preparedness. More work is needed on integrating these forecasts into existing forecasting and decision-making chains. This could be partly tackled by merging monthly and seasonal forecasts to obtain more skilful forecasts: updated more frequently and at longer lead times (Wetterhall and Di Giuseppe, 2018). The EFAS-Seasonal and GloFAS-Seasonal outlooks could be integrated in a 'Ready-Set-Go!' approach (Coughlan de Perez et al, 2017). Paired with short-term and higher quality forecasts in a "two-stage action approach", seasonal hydro-meteorological forecasts could be used by decision-makers as monitoring information, as a basis to implement no-regret actions. This would additionally help to increase their "portfolio of options at a later stage" (Bischiniotis et al., 2019). In order to make seasonal hydro-meteorological forecasts more actionable at the local scale, more research is needed on linking river basin hydro-climatic characteristics (Knoben et

al., 2018) and seasonal hydro-meteorological forecast performance (assessed against hydro-meteorological observations; Greuell et al., 2018). We must be able to utilise our understanding of links between basins' hydro-climatic characteristics and seasonal predictability to improve forecasting systems for flood early warning. Coughlan de Perez et al. (2017) for example looked at the link between several seasonal rainfall indicators and the likelihood of flooding across sub-Saharan Africa. This approach could be extended to seasonal hydrological forecasting and other parts of the world, for more tailored seasonal hydrological outlooks.

- **Flexible seasonal hydro-meteorological forecasting system.** As shown in Chapter 4, Sect. 4.4, more skilful and resource-friendly forecasts could be obtained through the design of a flexible seasonal forecasting system. Clark et al. (2015a)'s 'SUMMA' (Structure For Unifying Multiple Modeling Alternatives) approach could be extended to forecasting, integrating different forecasting approaches for a point in space and time.

7.4 Closing remarks

This thesis guides the readers through the art of streamflow forecasting over Europe, presenting some of the latest challenges in operational hydro-meteorological forecasting. The EFAS-Seasonal and GloFAS-Seasonal outlooks are some of the first operational seasonal hydrological products in the world, giving users an overview of potential streamflow changes for the next eight and 16 weeks, respectively, over Europe and the globe. While more work is needed to make these products more actionable in practice, they have the potential to inform decision-makers in a range of water sectors, such as flood early warning. Since their scientific application, ensemble (or probabilistic) hydrological forecasts are becoming more integrated in practice. However, efforts are still needed to bridge the gap between science and practice, where the co-development of actionable science is key. Serious games and art can help bridge the gap between scientists, decision-makers and society. It is the crossroads between disciplines that fosters creative thinking and findings and the uptake of science in practice.

References

- Acworth, R. I., 2009: Surface water and groundwater: understanding the importance of their connections, *Australian Journal of Earth Sciences*, 56, 1-2, doi:10.1080/08120090802541853
- Alfieri, L., F. Pappenberger, F. Wetterhall, T. Haiden, D. Richardson, and P. Salamon, 2014: Evaluation of ensemble streamflow predictions in Europe, *J. Hydrol.*, 517, 913-922, doi:10.1016/j.jhydrol.2014.06.035
- Anaman, K. A., S. C. Lellyett, L. Drake, R. J. Leigh, A. Henderson-Sellers, P. F. Noar, P. J. Sullivan, and D. J. Thampapillai, 1998: Benefits of meteorological services: evidence from recent research in Australia, *Meteorol. Appl.*, 5, 103-115, doi:10.1017/S1350482798000668
- Arnal, L., M.-H. Ramos, E. Coughlan de Perez, H. L. Cloke, E. Stephens, F. Wetterhall, S. J. van Andel, and F. Pappenberger, 2016: Willingness-to-pay for a probabilistic flood forecast: a risk-based decision-making game, *Hydrol. Earth Syst. Sci.*, 20, 3109-3128, doi:10.5194/hess-20-3109-2016
- Arnal, L. et al.: "IMPRES D4.2 - The sensitivity of sub-seasonal to seasonal streamflow forecasts to meteorological forcing quality, modelled hydrology and the initial hydrological conditions". Deliverable of EU H2020 project "IMPRES - Improving predictions and management of hydrological extremes" (contract n° 641811), 2017a
- Arnal, L., A. W. Wood, E. Stephens, H. L. Cloke, and F. Pappenberger, 2017b: An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity, *J. Hydrometeorol.*, 18, 1715-1729, doi:10.1175/JHM-D-16-0259.1
- Arnal, L., H. L. Cloke, E. Stephens, F. Wetterhall, C. Prudhomme, J. Neumann, B. Krzeminski and F. Pappenberger, 2018: Skilful seasonal forecasts of streamflow over Europe?, *Hydrol. Earth Syst. Sci.*, 22, 2057-2072, doi:10.5194/hess-22-2057-2018
- Arnal, L., L. Anspoks, S. Manson, J. Neumann, T. Norton, E. Stephens, L. Wolfenden, and H. L. Cloke, 2019: Are we talking just a bit of water out of bank? Or is it Armageddon? Front line perspectives on transitioning to probabilistic fluvial flood forecasts in England, *Geosci. Commun. Discuss.*, doi:10.5194/gc-2019-18, in review

- Arribas, A., M. Glover, A. Maidens, K. Peterson, M. Gordon, C. MacLachlan, R. Graham, D. Fereday, J. Camp, A. A. Scaife, P. Xavier, P. McLean, and A. Colman, 2010: The GloSea4 Ensemble Prediction System for Seasonal Forecasting, *Mon. Weather. Rev.*, 139, 1891–1910, doi:10.1175/2010MWR3615.1
- Aubert, A. H., W. Medema, and A. E. J. Wals, 2019: Towards a Framework for Designing and Assessing Game-Based Approaches for Sustainable Water Governance, *Water*, 11, doi:10.3390/w11040869
- Baroni, G., and S. Tarantola, 2014: A general probabilistic approach for uncertainty and global sensitivity analysis of deterministic models: A hydrological case study, *Environ. Modell. Software*, 51, 26–34, doi:10.1016/j.envsoft.2013.09.022
- Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction, *Nature*, 525, 47–55, doi:10.1038/nature14956
- Bell, V. A., H. N. Davies, A. L. Kay, A. Brookshaw, and A. A. Scaife, 2017: A national-scale seasonal hydrological forecast system: development and evaluation over Britain, *Hydrol. Earth Syst. Sci.*, 21, 4681–4691, doi:10.5194/hess-21-4681-2017
- Bennett, J. C., J. Q. Wang, M. Li, D. E. Robertson, and A. Schepen, 2016: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model, *Water Resour. Res.*, 52, 8238–8259, doi:10.1002/2016WR019193
- Bierkens, M. F., and L. P. van Beek, 2009: Seasonal Predictability of European Discharge: NAO and Hydrological Response Time, *J. Hydrometeorol.*, 10, 953–968, doi:10.1175/2009JHM1034.1
- Bischiniotis, K., B. van den Hurk, E. Coughlan de Perez, T. Veldkamp, G. Guimarães Nobre, and J. Aerts, 2019: Assessing Time, Cost and Quality Trade-Offs in Forecast-Based Action for Floods, *Int. J. Disast. Risk Re.*, 40, doi:10.1016/j.ijdr.2019.101252
- Bonacina, L. C. W., 1941: The scenic approach to meteorology, *Quarterly Journal of the Royal Meteorological Society*, 67, 305–314, doi:10.1002/qj.49706729202
- Boucher, M. A., D. Tremblay, L. Delorme, L. Perreault, and F. Anctil, 2012: Hydro-economic assessment of hydrological forecasting systems, *J. Hydrol.*, 416, 133–144, doi:10.1016/j.jhydrol.2011.11.042

-
- Breidert, C., M. Hahsler, and T. Reutterer, 2006: A review of methods for measuring willingness-to-pay, *Innovative Marketing*, 2, 8–32
- Bruen, M., P. Krahe, M. Zappa, J. Olsson, B. Vehvilainen, K. Kok, and K. Daamen, 2010: Visualizing Flood Forecasting Uncertainty: Some Current European EPS Platforms-COST731 Working Group 3, *Atmos. Sci. Lett.*, 11, 92–99, doi:10.1002/asl.258
- Buizza, R., 2008: The value of probabilistic prediction, *Atmos. Sci. Lett.*, 9, 36–42, doi:10.1002/asl.170
- Burek, P., J. M. Van Der Knijff, and A. De Roo, 2013: LISFLOOD – Distributed Water Balance and Flood Simulation Model – Revised User Manual 2013, EUR – Scientific and Technical Research Reports, Publications Office of the European Union, Luxembourg, 150 pp., doi:10.2788/24719
- Cai, W., A. Santoso, G. Wang, S. W. Yeh, S. I. An, K. M. Cobb, M. Collins, E. Guilyardi, F. F. Jin, J. S. Kug, M. Lengaigne, M. J. McPhaden, K. Takahashi, A. Timmermann, G. Vecchi, M. Watanabe, and L. Wu, 2015: ENSO and greenhouse warming, *Nat. Clim. Chang.*, 5, 9, 849–859, doi:10.1038/nclimate2743
- Candogan Yossef, N., R. van Beek, A. Weerts, H. Winsemius, and M. F. P. Bierkens, 2017: Skill of a global forecasting system in seasonal ensemble streamflow prediction, *Hydrol. Earth Syst. Sci.*, 21, 4103–4114, doi:10.5194/hess-21-4103-2017
- Céron, J.-P., G. Tanguy, L. Franchistéguy, E. Martin, F. Regimbeau, and J.-P. Vidal, 2010: Hydrological seasonal forecast over France: feasibility and prospects, *Atmos. Sci. Lett.*, 11, 78–82, doi:10.1002/asl.256
- Changnon, S. A., R. A. Pielke, D. Changnon, R. T. Sylves, and R. Pulwarty, 2000: Human Factors Explain the Increased Losses from Weather and Climate Extremes*, *B. Am. Meteorol. Soc.*, 81, 437–442, doi:10.1175/1520-0477(2000)081<0437:HFETIL>2.3.CO;2
- Cherry, J., H. Cullen, M. Visbeck, A. Small, and C. Uvo, 2005: Impacts of the North Atlantic Oscillation on Scandinavian hydropower production and energy markets, *Water Resour. Manage.*, 19, 673–691, doi:10.1007/s11269-005-3279-z
- Chiew, F. H., S. L. Zhou, and T. A. McMahon, 2003: Use of Seasonal Streamflow Forecasts in Water Resources Management, *J. Hydrol.*, 270, 135–144, doi:10.1016/S0022-1694(02)00292-5

- Church, J. E., 1935: Principles of snow surveying as applied to forecasting stream flow, edited by: Merrill, M. C., *J. Agric. Res.*, Washington, D. C., Vol. 51, no. 2, 97–130
- Clark, M. P., M. C. Serreze, and G. J. McCabe, 2001: Historical effects of El Niño and La Niña events on the seasonal evolution of the montane snowpack in the Columbia and Colorado River basins, *Water Resour. Res.*, 37, 741–757, doi:10.1029/2000WR900305
- Clark, M. P., et al., 2015a: A unified approach for process-based hydrologic modeling: 1. Modeling concept, *Water Resour. Res.*, 51, 2498–2514, doi:10.1002/2015WR017198
- Clark, M. P., et al., 2015b: A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies, *Water Resour. Res.*, 51, 2515–2542, doi:10.1002/2015WR017200
- Cloke, H. L., and F. Pappenberger, 2009: Ensemble flood forecasting: a review, *J. Hydrol.*, 375, 613–626, doi:10.1016/j.jhydrol.2009.06.005
- Cloke, H. L., F. Pappenberger, and J.-P. Renaud, 2008: Multimethod global sensitivity analysis (MMGSA) for modelling floodplain hydrological processes. *Hydrol. Processes*, 22, 1660–1674, doi:10.1002/hyp.6734
- Cloke, H. L., F. Pappenberger, P. Smith, and F. Wetterhall, 2017: How do I know if I've improved my continental scale flood early warning system? *Environ. Res. Lett.*, 12, doi:10.1088/1748-9326/aa625a
- Cloke, H. L., J. Thielen, F. Pappenberger, S. Nobert, G. Bálint, C. Edlund, A. Koistinen, C. de Saint-Aubin, E. Sprokkereef, C. Viel, P. Salamon, and R. Buizza, 2009: Progress in the Implementation of Hydrological Ensemble Prediction Systems (HEPS) in Europe for Operational Flood Forecasting, ECMWF Newsletter No. 121, Autumn, Reading, UK, 20–24, doi:10.21957/bn6mx5nxfq
- Copernicus: SWICCA, Service for Water Indicators in Climate Change Adaptation, available at: swicca.climate.copernicus.eu/, last access: 12 September 2019
- Coughlan de Perez, E., E. Stephens, K. Bischiniotis, M. van Aalst, B. van den Hurk, S. Mason, H. Nissan, and F. Pappenberger, 2017: Should seasonal rainfall forecasts be used for flood preparedness?, *Hydrol. Earth Syst. Sci.*, 21, 4517–4524, doi:10.5194/hess-21-4517-2017

-
- Crochemore, L., M.-H. Ramos, and F. Pappenberger, 2016: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, *Hydrol. Earth Syst. Sc.*, 20, 3601–3618, doi:10.5194/hess-2016-78
- Crochemore, L., M.-H. Ramos, F. Pappenberger, and C. Perrin, 2017: Seasonal streamflow forecasting by conditioning climatology with precipitation indices, *Hydrol. Earth Syst. Sci.*, 21, 1573–1591, doi:10.5194/hess-21-1573-2017
- Crochemore, L., M. H. Ramos, F. Pappenberger, S. J. van Andel, and A. W. Wood, 2015: An experiment on risk-based decision-making in water management using monthly probabilistic forecasts, *B. Am. Meteorol. Soc.*, 97, 541–551, doi:10.1175/BAMSD-14-00270.1
- Dale, M., Y. Ji, J. Wicks, K. Mylne, F. Pappenberger, and H. L. Cloke, 2013: Applying Probabilistic Flood Forecasting in Flood Incident Management, Environment Agency Technical Report, Project No. SC090032, Bristol, UK, 97 pp.
- Dale, M., J. Wicks, K. Mylne, F. Pappenberger, S. Laeger, and S. Taylor, 2014: Probabilistic flood forecasting and decision-making: an innovative risk-based approach, *Nat. Hazards*, 70.1, 159–172, doi:10.1007/s11069-012-0483-z
- Davies, A., K. Hoggart, and L. Lees, 2014: *Researching Human Geography*, 1st edition, Routledge, London, 384 pp.
- Day, G. N., 1985: Extended streamflow forecasting using NWSRFS, *J. Water Res. Plan. Man.*, 111, 157–170, doi:10.1061/(ASCE)0733-9496(1985)111:2(157)
- De Roo, A. P., C. G. Wesseling, and W. P. Van Deursen, 2000: Physically based river basin modelling within a GIS: the LISFLOOD model, *Hydrol. Process.*, 14, 1981–1992, doi:10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F
- Demargne, J., et al., 2014: The science of NOAA's operational Hydrologic Ensemble Forecast Service. *Bull. Amer. Meteor. Soc.*, 95, 79–98, doi:10.1175/BAMS-D-12-00081.1
- Demeritt, D., H. L. Cloke, F. Pappenberger, J. Thielen, J. Bartholmes, and M.-H. Ramos, 2007: Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting, *Environmental Hazards*, 7, 115–127, doi:10.1016/j.envhaz.2007.05.001
- Demeritt, D., S. Nobert, H. L. Cloke, and F. Pappenberger, 2013: The European Flood Alert System and the Communication, Perception, and Use of Ensemble Predictions for Operational Flood Risk Management, *Hydrol. Process.*, 27, 147–57,

doi:10.1002/hyp.9419

Demeritt, D., S. Nobert, H. L. Cloke, and F. Pappenberger, 2010: Challenges in Communicating and Using Ensembles in Operational Flood Forecasting, *Meteorol. Appl.*, 17, 209–22, doi:10.1002/met.194

Demirel, M. C., M. J. Booij, and A. Y. Hoekstra, 2015: The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models, *Hydrol. Earth Syst. Sci.*, 19, 275–291, doi:10.5194/hess-19-275-2015

Department for Environment Food and Rural Affairs, 2010: Flood and Water Management Act 2010, UK Public General Acts, 1–84

Dessai, S., and M. Hulme, 2004: Climate Policy Does Climate Adaptation Policy Need Probabilities? Does Climate Adaptation Policy Need Probabilities?, *Clim. Policy*, 4, 107–28, doi:10.1080/14693062.2004.9685515

Dettinger, M. D., and H. F. Diaz, 2000: Global characteristics of stream flow seasonality and variability, *J. Hydrometeorol.*, 1, 289–310, doi:10.1175/1525-7541(2000)001<0289:GCOSFS>2.0.CO;2

Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. Rodrigues, 2013: Seasonal climate predictability and forecasting: status and prospects, *WIREs Clim. Change*, 4, 245–268, doi:10.1002/wcc.217

ECMWF, 2017a: EFAS and GloFAS seasonal hydrological outlooks, Summer 2017 Newsletter, available at: www.ecmwf.int/en/newsletter/152/news/efas-and-glofas-seasonal-hydrological-outlooks, last access: 12 September 2019

ECMWF, 2017b: SEAS5 user guide, available at: www.ecmwf.int/sites/default/files/medialibrary/2017-10/System5_guide.pdf, last access: 12 September 2019

Emerton, R. E., E. M. Stephens, F. Pappenberger, T. C. Pagano, A. H. Weerts, A. W. Wood, P. Salamon, J. D. Brown, N. Hjerdt, C. Donnelly, C. Baugh, and H. L. Cloke, 2016: Continental and global scale flood forecasting systems, *Wiley Interdisciplinary Reviews: Water*, 3, 391–418, doi:10.1002/wat2.1137

Emerton, R., E. Zsoter, L. Arnal, H. L. Cloke, D. Muraro, C. Prudhomme, E. M. Stephens, P. Salamon, and F. Pappenberger, 2018: Developing a global operational seasonal hydro-

meteorological forecasting system: GloFAS-Seasonal v1.0, *Geosci. Model Dev.*, 11, 3327-3346, doi:10.5194/gmd-11-3327-2018

Environment Agency, 2018: Creating a Better Place - Our Ambition to 2020, available at: assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/713127/Environment_Agency_our_ambition_to_2020.pdf, last access: 12 September 2019

Ferrell, J., 2009: The Secrets of Weather Forecast Models, Exposed, available at: www.accuweather.com/en/weather-blogs/weathermatrix/why-are-the-models-so-inaccurate/18097, last access: 12 September 2019

Flato, G. M., 2011: Earth system models: An overview, *Wiley Interdiscip. Rev.: Climate Change*, 2, 783–800, doi:10.1002/wcc.148

Flood Forecasting Centre, 2017: Flood Guidance Statement User Guide, available at: www.ffc-environment-agency.metoffice.gov.uk/services/FGS_User_Guide.pdf, last access: 12 September 2019

Flowerdew, J., K. Horsburgh, and K. Mylne, 2009: Ensemble Forecasting of Storm Surges, *Mar. Geod.*, 32, 91–99, doi:10.1080/01490410902869151

Forecast skill metrics: meteoswiss.shinyapps.io/skill_metrics/, last access: 12 September 2019

Funtowicz, S. O., and J. R. Ravetz, 1993: Science for the Post-Normal Age, *Futures*, 25, 739–55, doi:10.1016/0016-3287(93)90022-L

García-Morales, M. B. and L. Dubus, 2007: Forecasting precipitation for hydroelectric power management: how to exploit GCM's seasonal ensemble forecasts, *Int. J. Climatol.*, 27, 1691–1705, doi:10.1002/joc.1608

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness, *J. Roy. Stat. Soc. B.*, 69, 243–268, doi:10.1111/j.1467-9868.2007.00587.x

Gobena, A. K., and T. Y. Gan, 2010: Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system, *J. Hydrol.*, 385, 336–352, doi:10.1016/j.jhydrol.2010.03.002

- Goddard, L., S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane, 2001: Current approaches to seasonal to interannual climate predictions, *Int. J. Climatol.*, 21, 1111–1152, doi:10.1002/joc.636
- Golding, N., C. Hewitt, P. Zhang, P. Bett, X. Fang, H. Hu, and S. Nobert, 2017: Improving User Engagement and Uptake of Climate Services in China, *Climate Services*, 5, 39–45, doi:10.1016/j.cliser.2017.03.004
- Greuell, W., W. H. P. Franssen, H. Biemans, and R. W. A. Hutjes, 2018: Seasonal streamflow forecasts for Europe – Part I: Hindcast verification with pseudo- and real observations, *Hydrol. Earth Syst. Sci.*, 22, 3453–3472, doi:10.5194/hess-22-3453-2018
- Guimarães Nobre, G., B. Jongman, J. Aerts, and P. J. Ward, 2017.: The role of climate variability in extreme floods in Europe, *Environ. Res. Lett.*, 12, doi:10.1088/1748-9326/aa7c22
- Hamlet, A. F., D. Huppert, and D. P. Lettenmaier, 2002: Economic Value of Long-Lead Streamflow Forecasts for Columbia River Hydropower, *J. Water Res. Plan. Man.*, 128, 91–101, doi:10.1061/(ASCE)0733-9496(2002)128:2(91)
- Handmer, J., and B. Proudley, 2007: Communicating uncertainty via probabilities: The case of weather forecasts, *Environmental Hazards*, 7, 79–87, doi:10.1016/j.envhaz.2007.05.002
- Hartmann, H. C., T. Pagano, R. Bales, and S. Sorooshian, 2002: Weather, climate, and hydrologic forecasting for the US Southwest: A survey, *Climate Res.*, 21, 239–258, doi:10.3354/cr021239
- Hashino, T., A. A. Bradley, and S. S. Schwartz, 2007: Evaluation of bias-correction methods for ensemble streamflow volume forecasts, *Hydrol. Earth Syst. Sci.*, 11, 939–950, doi:10.5194/hess-11-939-2007
- Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions, *B. Am. Meteorol. Soc.*, 90, 1095–1108, doi:10.1175/2009BAMS2607.1
- Hawkins, Ed, and R. Sutton, 2010: The potential to narrow uncertainty in projections of regional precipitation change, *Climate Dynamics*, 37, 407–418, doi:10.1007/s00382-010-0810-6

-
- He, Y., F. Wetterhall, H. L. Cloke, F. Pappenberger, M. Wilson, J. Freer, and G. McGregor, 2009: Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions, *Meteorol. Appl.*, 16, 91–101, doi:10.1002/met.132
- Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, *Weather Forecast.*, 15, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2
- Hirpa, F. A., et al., 2018: Global flood forecasting for averting disasters worldwide, *Global Flood Hazard: Applications in Modeling, Mapping, and Forecasting*, 205-228, doi:10.1002/9781119217886.ch12
- HM Government, 2016: National Flood Resilience Review, available at: assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/551137/national-flood-resilience-review.pdf, last access: 12 September 2019
- Horton, R.E., 1940: An approach toward a physical interpretation of infiltration capacity, *Soil Sci. Am. Prec.*, 5, 399-417
- House of Common - Environment Food and Rural Affairs Committee, 2016: Future Flood Prevention - Second Report of Session 2016-17, available at: publications.parliament.uk/pa/cm201617/cmselect/cmenvfru/115/115.pdf, last access: 12 September 2019
- Hurrell, J. W., 1995: Decadal trends in the North Atlantic oscillation: Regional temperatures and precipitation, *Science*, 269, 676–679, doi:10.1126/science.269.5224.676
- Hurrell, J. W., and H. Van Loon, 1997: Decadal Variations in Climate Associated with the North Atlantic Oscillation, in: *Climatic Change at High Elevation Sites*, edited by: Diaz, H. F., Beniston, M., and Bradley, R. S., Springer, Dordrecht, 69–94, doi:10.1007/978-94-015-8905-5_4
- IPCC, 2014: *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp.
- Joslyn, S., and S. Savelli, 2010: Communicating Forecast Uncertainty: Public Perception of Weather Forecast Uncertainty, *Meteorol. Appl.*, 17, 180–95, doi:10.1002/met.190

- Keller, J. D., and A. Hense, 2011: A new non-Gaussian evaluation method for ensemble forecasts based on analysis rank histograms, *Meteorol. Z.*, 20, 107–117, doi:10.1127/0941-2948/2011/0217
- Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter, *Clim. Dynam.*, 39, 2957–2973, doi:10.1007/s00382-012-1364-6
- Kirchhoff, C. J., M. C. Lemos, and N. L. Engle, 2013: What influences climate information use in water management? The role of boundary organizations and governance regimes in Brazil and the U.S., *Environ. Sci. Policy*, 26, 6–18, doi:10.1016/j.envsci.2012.07.001
- Knoben, W., J. M. Ross, A. Woods, and J. E. Freer, 2018: A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data, *Water Resour. Res.*, 54, 5088–5109, doi:10.1029/2018WR022913
- Krzysztofowicz, R., 2001: The case for probabilistic forecasting in hydrology, *J. Hydrol.*, 249, 2–9, doi:10.1016/S0022-1694(01)00420-6
- Kwon, H.-H., C. Brown, K. Xu, and U. Lall, 2009: Seasonal and annual maximum streamflow forecasting using climate information: Application to the Three Gorges Dam in the Yangtze River basin, China, *Hydrol. Sci. J.*, 54, 582–595, doi:10.1623/hysj.54.3.582
- Laio, F., and S. Tamea, 2007: Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrol. Earth Syst. Sci.*, 11, 1267–1277, doi:10.5194/hess-11-1267-2007
- LeClerc, J., and S. Joslyn, 2015: The Cry Wolf Effect and Weather-Related Decision Making, *Risk Anal.*, 35, 385–395, doi:10.1111/risa.12336
- Leviäkangas, P., 2009: Valuing meteorological information, *Meteorol. Appl.*, 16, 315–323, doi:10.1002/met.122
- Li, H., L. Luo, E. F. Wood, and J. Schaake, 2009: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, *J. Geophys. Res.*, 114, doi:10.1029/2008JD010969
- Li, Y., M. Giuliani, and A. Castelletti, 2017: A coupled human-natural system to assess the operational value of weather and climate services for agriculture, *Hydrol. Earth Syst. Sci.*, 21, 4693–4709, doi:10.5194/hess-21-4693-2017

-
- Lilburne, L., and S. Tarantola, 2009: Sensitivity analysis of spatial models, *Int. J. Geogr. Inf. Sci.*, 23, 151–168, doi:10.1080/13658810802094995
- Lins, H. F., 2012: USGS Hydro-Climatic Data Network 2009 (HCDN-2009), USGS Fact Sheet 2012-3047, 4 pp., available at: pubs.usgs.gov/fs/2012/3047/, last access: 12 September 2019
- Linsley, R. K., 1967: The relation between rainfall and runoff, *J. Hydrol.*, 5, 297–311, doi:10.1016/S0022-1694(67)80128-8
- Liu, Y., et al., 2012: Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities, *Hydrol. Earth Syst. Sci.*, 16, 3863–3887, doi:10.5194/hess-16-3863-2012
- Lorenz, E. N., 1963: Deterministic nonperiodic flow, *Journal of the atmospheric sciences*, 20, 130–141, doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2
- Lorenz, E. N., 1969: The Predictability of a Flow Which Possesses Many Scales of Motion, *Tellus*, 21, 289–307, doi:10.3402/tellusa.v21i3.10086
- Lorenzo-Lacruz, J., S. M. Vicente-Serrano, J. I. López-Moreno, J. C. González-Hidalgo, and E. Morán-Tejeda, 2011: The response of Iberian rivers to the North Atlantic Oscillation, *Hydrol. Earth Syst. Sci.*, 15, 2581–2597, doi:10.5194/hess-15-2581-2011
- Luo, L., and E. F. Wood, 2007: Monitoring and predicting the 2007 U.S. drought, *Geophys. Res. Lett.*, 34, doi:10.1029/2007GL031673
- MacLachlan, C., et al., 2015: Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system, *Quarterly Journal of the Royal Meteorological Society*, 141, 1072–1084, doi:10.1002/qj.2396
- MacLeod, D., H. Cloke, F. Pappenberger, and A. Weisheimer, 2016: Evaluating uncertainty in estimates of soil moisture memory with a reverse ensemble approach, *Hydrol. Earth Syst. Sci.*, 20, 2737–2743, doi:10.5194/hess-20-2737-2016
- Magnusson, L., and E. Källén, 2013: Factors Influencing Skill Improvements in the ECMWF Forecasting System, *Mon. Weather Rev.*, 141, 3142–3153, doi:10.1175/MWR-D-12-00318.1
- Magnusson, L., 2017: Diagnostic methods for understanding the origin of forecast errors, *Quart. J. Roy. Meteor. Soc.*, 143, 2129–2142, doi:10.1002/qj.3072

- Maraun, D., et al., 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, *Rev. Geophys.*, 48, doi:10.1029/2009rg000314
- Mason, S. J., and N. E. Graham, 1999: Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels, *Weather Forecast.*, 14, 713–725, doi:10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2
- McEwen, L. J., F. Krause, O. Jones, and J. Garde Hansen, 2012: Sustainable Flood Memories, Informal Knowledge and the Development of Community Resilience to Future Flood Risk, *WIT Trans. Ecol. Envir.*, 159, 253–64, doi:10.2495/FRIAR120211
- Meißner, D., and B. Klein, 2013: The added value of probabilistic forecasts for navigation, available at: hepex.irstea.fr/the-added-value-of-probabilistic-forecasts-for-navigation-2, last access: 12 September 2019
- Meißner, D., B. Klein, and M. Ionita, 2017: Development of a monthly to seasonal forecast framework tailored to inland waterway transport in central Europe, *Hydrol. Earth Syst. Sci.*, 21, 6401–6423, doi:10.5194/hess-21-6401-2017
- Mendler de Suarez, J., Suarez, P., Bachofen, C., Fortugno, N., Goentzel, J., Gonçalves, P., Grist, N., Macklin, C., Pfeifer, K., Schweizer, S., Van Aalst, M., and Virji, H.: Games for a New Climate: Experiencing the Complexity of Future Risks, Pardee Center Task Force Report, The Frederick S. Pardee Center for the Study of the Longer-Range Future, Boston University, Boston, 2012.
- Mendoza, P. A., A. W. Wood, E. Clark, E. Rothwell, M. P. Clark, B. Nijssen, L. D. Brekke, and J. R. Arnold, 2017: An intercomparison of approaches for improving operational seasonal streamflow forecasts, *Hydrol. Earth Syst. Sci.*, 21, 3915–3935, doi:10.5194/hess-21-3915-2017
- Michaels, S., 2014: Probabilistic forecasting and the reshaping of flood risk management, *Journal of Natural Resources Policy Research*, 7, 41–51, doi:10.1080/19390459.2014.970800
- Mo, K. C., and D. P. Lettenmaier, 2014: Hydrologic Prediction over the Conterminous United States Using the National Multi-Model Ensemble, *J. Hydrometeorol.*, 15, 1457–1472, doi:10.1175/JHM-D-13-0197.1

-
- Molteni, F., T. Stockdale, M. Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, T. Palmer, and F. Vitart, 2011: The new ECMWF seasonal forecast system (System 4), ECMWF Tech. Memorandum, 656, 1–49
- Morss, R. E., O. V. Wilhelmi, M. W. Downton, and E. Grunfest, 2005: Flood Risk, Uncertainty, and Scientific Information for Decision Making: Lessons from an Interdisciplinary Project, *B. Am. Meteorol. Soc.*, 86, 1593–1602, doi:10.1175/BAMS-86-11-1593
- Mulder, K. J., M. Lickiss, N. Harvey, A. Black, A. Charlton-Perez, H. Dacre, and R. McCloy, 2017: Visualizing Volcanic Ash Forecasts: Scientist and Stakeholder Decisions Using Different Graphical Representations and Conflicting Forecasts, *Weather Clim. Soc.*, 9, 333–48, doi:10.1175/WCAS-D-16-0062.1
- Neumann, J. L., L. Arnal, L. Magnusson, and H. Cloke, 2018a: The 2013/14 Thames basin floods: Do improved meteorological forecasts lead to more skillful hydrological forecasts at seasonal time scales?, *J. Hydrometeorol.*, 19, 6, 1059-1075, doi:10.1175/JHM-D-17-0182.1
- Neumann, J. L., L. Arnal, R. E. Emerton, H. Griffith, S. Hyslop, S. Theofanidi and H. L. Cloke, 2018b: Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity, *Geosci. Commun.*, 1, 35-57, doi:10.5194/gc-1-35-2018
- Neville, N., 1999: Cognitive Illusions, Heuristics, and Climate Prediction, *B. Am. Meteorol. Soc.*, 80, 1385–97, doi:10.1175/1520-0477(1999)080<1385:CIHACP>2.0.CO;2
- New, M., A. Lopez, S. Dessai, and R. Wilby, 2007: Challenges in using probabilistic climate change information for impact assessments: an example from the water sector, *Philos. T. Roy. Soc. A.*, 365, 2117–2131, doi:10.1098/rsta.2007.2080
- Nobert, S., D. Demeritt, and H. L. Cloke, 2010: Informing Operational Flood Management with Ensemble Predictions: Lessons from Sweden, *J. Flood Risk Manag.*, 3, 72–79, doi:10.1111/j.1753-318X.2009.01056.x
- Orr, P., and C. Twigger-Ross, 2009: Communicating Risk and Uncertainty in Flood Warnings: A Review of Defra/Environment Agency FCERM Literature, Environment Agency Science Report, Project No. SC070060/SR2, Bristol, UK, 60 pp.
- Pagano, T., D. Garen, and S. Sorooshian, 2004a: Evaluation of official western U.S. seasonal water supply outlooks, 1922–2002, *J. Hydrometeorol.*, 5, 896–909, doi:10.1175/1525-7541(2004)005,0896:E00WUS.2.0.CO;2

- Pagano, T. C., and D. C. Garen, 2006: Integration of climate information and forecasts into western US water supply forecasts, *Climate variations, climate change, and water resources engineering*, edited by: Garbrecht, J. D. and Piechota, T. C., American Society of Civil Engineers location, Reston, Virginia, US, 86–103
- Pagano, T. C., H. C. Hartmann, and S. Sorooshian, 2004b: Seasonal Forecasts and Water Management in Arizona: A Case Study of the 1997–98 El Niño Event, 29th Annual Water Resources Planning and Management Conference, 21, 1–11, doi:10.1061/40430(1999)227
- Pagano, T. C., A. W. Wood, M.-H. Ramos, H. L. Cloke, F. Pappenberger, M. P. Clark, M. Cranston, D. Kavetski, T. Mathevet, S. Sorooshian, and J. S. Verkade, 2014: Challenges of operational river forecasting, *J. Hydrometeorol.*, 15, 1692–1707, doi:10.1175/JHM-D-13-0188.1
- Paiva, R. C. D., W. Collischonn, M. P. Bonnet, and L. G. G. de Gonçalves, 2012: On the sources of hydrological prediction uncertainty in the Amazon, *Hydrol. Earth Syst. Sci.*, 16, 3127–3137, doi:10.5194/hess-16-3127-2012
- Pappenberger, F., E. Stephens, J. Thielen, P. Salamon, D. Demeritt, S. J. van Andel, F. Wetterhall, and L. Alfieri, 2013: Visualizing Probabilistic Flood Forecast Information: Expert Preferences and Perceptions of Best Practice in Uncertainty Communication, *Hydrol. Process.*, 27, 132–46, doi:10.1002/hyp.9253
- Pappenberger, F., H. L. Cloke, D. J. Parker, F. Wetterhall, D. S. Richardson, and J. Thielen, 2015: The monetary benefit of early flood warnings in Europe, *Environ. Sci. Policy*, 51, 278–291, doi:10.1016/j.envsci.2015.04.016
- Pappenberger, F., M. Ratto, and V. Vandenberghe, 2010: Review of sensitivity analysis methods. *Modelling Aspects of Water Approach Directive Implementation*, P.A. Vanrolleghem, Ed., IWA Publishing, 191–265
- Parker, D. J., S. J. Priest, and S. M. Tapsell, 2009: Understanding and Enhancing the Public's Behavioural Response to Flood Warning Information, *Meteorol. Appl.*, 114, 103–14, doi:10.1002/met.119
- Pielke, R. A. Jr, 1997.: Asking the Right Questions: Atmospheric Sciences Research and Societal Needs, *B. Am. Meteorol. Soc.*, 78, 255–255, doi:10.1175/1520-0477(1997)078<0255:ATRQAS>2.0.CO;2

-
- Pilling, C., V. Dodds, M. Cranston, D. Price, T. Harrison, and A. How, 2016: Chapter 9 - Flood Forecasting — A National Overview for Great Britain, *Flood Forecasting - A Global Perspective*, edited by: Adams, T. E., III, and Pagano, T. C., Academic Press, 201–47, doi:10.1016/B978-0-12-801884-2.00009-8
- Pitt, M., 2008: The Pitt Review: Lessons learned from the 2007 floods, Cabinet Office (archived by the National Archives), available at: webarchive.nationalarchives.gov.uk/20100812084907/http://archive.cabinetoffice.gov.uk/pittreview/_/media/assets/www.cabinetoffice.gov.uk/flooding_review/pitt_review_full%20pdf.pdf, last access: 12 September 2019
- Prudhomme, C., et al., 2017: Hydrological Outlook UK: an operational streamflow and groundwater level forecasting system at monthly to seasonal time scales, *Hydrolog. Sci. J.*, 62, 2753–2768, doi:10.1080/02626667.2017.1395032
- Ramos, M. H., S. J. van Andel, and F. Pappenberger, 2013: Do probabilistic forecasts lead to better decisions?, *Hydrol. Earth Syst. Sci.*, 17, 2219–2232, doi:10.5194/hess-17-2219-2013
- Ramos, M. R., T. Mathevet, J. Thielen, and F. Pappenberger, 2010: Communicating uncertainty in hydro-meteorological forecasts: mission impossible?, *Meteorol. Appl.*, 17, 223–235, doi:10.1002/met.202
- Red Cross/Red Crescent Climate Centre: Game “Paying for Predictions”, International Federation of Red Cross and Red Crescent Societies (IFRC), The Netherlands, available at: www.climatecentre.org/resources-games/paying-for-predictions, last access: 12 September 2019
- Regonda, S. K., B. Rajagopalan, M. Clark, and E. Zagona, 2006: A multimodel ensemble forecast approach: Application to spring seasonal flows in the Gunnison River basin, *Water Resour. Res.*, 42, doi:10.1029/2005WR004653
- Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks, 2010: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, 46, doi:10.1029/2009WR008328
- Robertson, D. E., and Q. J. Wang, 2012: A Bayesian approach to predictor selection for seasonal streamflow forecasting, *J. Hydrometeorol.*, 13, 155–171. doi:10.1175/JHM-D-10-05009.1

- Rollins, K. S., and J. Shaykewich, 2003: Using willingness-to-pay to assess the economic value of weather forecasts for multiple commercial sectors, *Meteorol. Appl.*, 10, 31–38, doi:10.1017/S1350482703005048
- Saltelli, A., S. Tarantola, and F. Campolongo, 2000: Sensitivity analysis as an ingredient of modeling. *Stat. Sci.*, 15, 377–395, doi:10.1214/ss/1009213004
- Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto, 2004: Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. John Wiley & Sons, 218 pp.
- Scaife, A. A., et al., 2014: Skillful long-range prediction of European and North American winters, *Geophys. Res. Lett.*, 41, 7, 2514–2519, doi:10.1002/2014GL059637
- Schepen, A., T. Zhao, Q. J. Wang, S. Zhou, and P. Feikema, 2016: Optimising seasonal streamflow forecast lead time for operational decision making in Australia, *Hydrol. Earth Syst. Sci.*, 20, 4117–4128, doi:10.5194/hess-20-4117-2016
- Schoenberger, E., 1991: The Corporate Interview as a Research Method in Economic Geography, *The Professional Geographer*, 43, 180–89, doi:10.1111/j.0033-0124.1991.00180.x
- Seifert, I., W. J. W. Botzen, H. Kreibich, and J. C. J. H. Aerts, 2013: Influence of flood risk characteristics on flood insurance demand: a comparison between Germany and the Netherlands, *Nat. Hazards Earth Syst. Sci.*, 13, 1691–1705, doi:10.5194/nhess-13-1691-2013
- Sene, K., A. Weerts, K. Beven, R. J. Moore, C. Whitlow, and P. Young, 2009: Risk-Based Probabilistic Fluvial Flood Forecasting for Integrated Catchment Models - Phase 1 Report, Environment Agency Science Report, Project No. SC080030/SR1, Bristol, UK, 179 pp.
- Sheffield, J., E. F. Wood, N. Chaney, K. Guan, S. Sadri, X. Yuan, L. Olang, A. Amani, A. Ali, S. Demuth, and L. Ogallo, 2013: A Drought Monitoring and Forecasting System for Sub-Saharan African Water Resources and Food Security, *B. Am. Meteorol. Soc.*, 95, 861–882, doi:10.1175/BAMS-D-12-00124.1
- Sherman, L. K., 1932: Streamflow from rainfall by the unit graph method, *Eng. News Rec.*, 108, 501–505

-
- Shi, W., N. Schaller, D. MacLeod, T. N. Palmer, and A. Weisheimer, 2015: Impact of hindcast length on estimates of seasonal climate predictability, *Geophys. Res. Lett.*, 42, 1554–1559, doi:10.1002/2014GL062829
- Shukla, S., and D. P. Lettenmaier, 2011: Seasonal hydrologic prediction in the United States: Understanding the role of initial hydrologic conditions and seasonal climate forecast skill, *Hydrol. Earth Syst. Sci.*, 15, 3529–3538, doi:10.5194/hess-15-3529-2011
- Shukla, S., J. Sheffield, E. F. Wood, and D. P. Lettenmaier, 2013: On the sources of global land surface hydrologic predictability, *Hydrol. Earth Syst. Sci.*, 17, 2781–2796, doi:10.5194/hess-17-2781-2013
- Shuman, F. G., 1989: History of numerical weather prediction at the National Meteorological Center, *Weather and Forecasting*, 4, 286–296, doi:10.1175/1520-0434(1989)004<0286:HONWPA>2.0.CO;2
- Simmons, A. J., and A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction, *Q. J. Roy. Meteor. Soc.*, 128, 647–677, doi:10.1256/003590002321042135
- Singla, S., J.-P. Céron, E. Martin, F. Regimbeau, M. Déqué, F. Habets, and J.-P. Vidal, 2012: Predictability of soil moisture and river flows over France for the spring season, *Hydrol. Earth Syst. Sci.*, 16, 201–216, doi:10.5194/hess-16-201-2012
- Sivle, A. D., S. D. Kolstø, P. J. K. Hansen, and J. Kristiansen, 2014: How Do Laypeople Evaluate the Degree of Certainty in a Weather Report? A Case Study of the Use of the Web Service Yr.No., *Weather Clim. Soc.*, 6, 399–412, doi:10.1175/WCAS-D-12-00054.1
- Slater, L. J., G. Villarini, and A. A. Bradley, 2016: Evaluation of the skill of North-American Multi-Model Ensemble (NMME) global climate models in predicting average and extreme precipitation and temperature over the continental USA, *Climate Dyn.*, doi:10.1007/s00382-016-3286-1
- Slater, L. J., G. Villarini, A. A. Bradley, and G. A. Vecchi, 2017: A dynamical statistical framework for seasonal streamflow forecasting in an agricultural watershed, *Clim. Dynam.*, 1–17, doi:10.1007/s00382-017-3794-7
- Slater, L. J., and G. Villarini. 2018: Enhancing the predictability of seasonal streamflow with a statistical-dynamical approach, *Geophysical Research Letters*, 45, 6504–6513, doi:10.1029/2018GL077945

- Smith, P., F. Pappenberger, F. Wetterhall, J. Thielen, B. Krzeminski, P. Salamon, D. Muraro, M. Kalas, and C. Baugh, 2016: On the operational implementation of the European Flood Awareness System (EFAS), ECMWF Tech. Memorandum, 778, 1–34
- Smith, K. A., R. L. Wilby, C. Broderick, C. Prudhomme, T. Matthews, S. Harrigan, and C. Murphy, 2018: Navigating Cascades of Uncertainty — As Easy as ABC? Not Quite..., *Journal of Extreme Events*, 5, doi:10.1142/S2345737618500070
- Soares, M. B. and S. Dessai, 2016: Barriers and enablers to the use of seasonal climate forecasts amongst organisations in Europe, *Climatic Change*, 137, 89–103, doi:10.1007/s10584-016-1671-8
- Staudinger, M., and J. Seibert, 2014: Predictability of low flow—An assessment with simulation experiments, *J. Hydrol.*, 519, 1383–1393, doi:10.1016/j.jhydrol.2014.08.061
- Steirou, E., L. Gerlitz, H. Apel, and B. Merz, 2017: Links between large-scale circulation patterns and streamflow in Central Europe: A review, *J. Hydrol.*, 549, 484–500, doi:10.1016/j.jhydrol.2017.04.003
- Stephens, E., and H. L. Cloke, 2014: Improving flood forecasts for better flood preparedness in the UK (and beyond), *Geogr. J.*, 180, 310–316, doi:10.1111/geoj.12103
- Stephens, E., J. J. Day, F. Pappenberger, and H. Cloke, 2015: Precipitation and floodiness, *Geophys. Res. Lett.*, 42, 10316–10323, doi:10.1002/2015GL066779
- Stephens, E. M., T. L. Edwards, and D. Demeritt, 2012: Communicating Probabilistic Information from Climate Model Ensembles—Lessons from Numerical Weather Prediction, *WIREs Clim. Change*, 3, 409–26, doi:10.1002/wcc.187
- Thielen, J., J. Bartholmes, M.-H. Ramos, and A. de Roo, 2009: The European Flood Alert System – Part 1: Concept and development, *Hydrol. Earth Syst. Sci.*, 13, 125–140, doi.org/doi:10.5194/hess-13-125-2009
- Thielen, J., J. Bartholmes, and M.-H. Ramos, 2006: The Benefit of Probabilistic Flood Forecasting on European Scale – Results of the European Flood Alert System for 2005/2006, European Commission, Ispra, Italy, 99 pp.
- Troccoli, A., 2010: Seasonal climate forecasting, *Meteorol. Appl.*, 17, 251–268, doi:10.1002/met.184

-
- Turner, S. W. D., J. C. Bennett, D. E. Robertson, and S. Galelli, 2017: Complex relationship between seasonal streamflow forecast skill and value in reservoir operations, *Hydrol. Earth Syst. Sci.*, 21, 4841–4859, doi:10.5194/hess-21-4841-2017
- Twedt, T. M., J. C. Schaake, and E. L. Peck, 1977: National Weather Service extended streamflow prediction, Proceedings Western Snow Conference, Albuquerque, New Mexico, 52–57
- UNISDR, 2009: Disaster Statistics in Europe - Floods, droughts and storms: a major threat for European countries, [press release], 04/03, available at: www.unisdr.org/files/8867_pr200903DisasterStatisticsEurope.pdf, last access: 12 September 2019
- van den Hurk, B. J. J. M., et al., 2016: Improving predictions and management of hydrological extremes through climate services: www.imprex.eu, *Climate Services*, 1, 6–11, doi:10.1016/j.cliser.2016.01.001
- Van Der Knijff, J. M., J. Younis, and A. P. De Roo, 2010: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, *Int. J. Geogr. Inf. Sci.*, 24, 189–212, doi:10.1080/13658810802549154
- van Dijk, A. I. J. M., J. L. Peña-Arancibia, E. F. Wood, J. Sheffield, and H. E. Beck, 2013: Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide, *Water Resour. Res.*, 49, 2729–2746, doi:10.1002/wrcr.20251
- Verkade, J. S., and M. G. F. Werner, 2011: Estimating the benefits of single value and probability forecasting for flood warning, *Hydrol. Earth Syst. Sci.*, 15, 3751–3765, doi:10.5194/hess-15-3751-2011
- Viel, C., A.-L. Beuland, J.-M. Soubeyroux, and J.-P. Céron, 2016: How seasonal forecast could help a decision maker: an example of climate service for water resource management, *Adv. Sci. Res.*, 13, 51–55, doi:10.5194/asr-13-51-2016
- Wanders, N., et al., 2019: Development and Evaluation of a Pan-European Multimodel Seasonal Hydrological Forecasting System, *J. Hydrometeorol.*, 20, 99–115, doi:10.1175/JHM-D-18-0040.1
- Weisheimer, A., and T. N. Palmer, 2014: On the reliability of seasonal climate forecasts, *Journal of the Royal Society Interface*, 11, 96, doi:10.1098/rsif.2013.1162

- Welles, E., S. Sorooshian, G. Carter, and B. Olsen, 2007: Hydrologic verification: A call for action and collaboration, *Bull. Amer. Meteor. Soc.*, 88, 503–511, doi:10.1175/BAMS-88-4-503
- Werner, M., M. Cranston, T. Harrison, D. Whitfield, and J. Schellekens, 2009: Recent Developments in Operational Flood Forecasting in England, Wales and Scotland, *Meteorol. Appl.*, 16, 13–22, doi:10.1002/met.124
- Wetherald, R. T., and S. Manabe, 2002: Simulation of hydrologic changes associated with global warming, *J. Geophys. Res.*, 107, 4379, doi:10.1029/2001JD001195
- Wetterhall, F., and F. Di Giuseppe, 2018: The benefit of seamless forecasts for hydrological predictions over Europe, *Hydrol. Earth Syst. Sci.*, 22, 3409–3420, doi:10.5194/hess-2017-527
- White, C. J., et al., 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions, *Meteorol. Appl.*, 24, 315–325, doi:10.1002/met.1654
- WMO, 2011: Manual on Flood Forecasting and Warning, World Meteorological Organisation, Geneva, Switzerland, Unesco, WMO No. 1072, 142 pp.
- Wood, A. W., A. Kumar, and D. P. Lettenmaier, 2005: A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States, *J. Geophys. Res.-Atmos.*, 110, doi:10.1029/2004JD004508
- Wood, A. W., and D. P. Lettenmaier, 2006: A test bed for new seasonal hydrologic forecasting approaches in the western United States, *Bull. Amer. Meteor. Soc.*, 87, 1699–1712, doi:10.1175/BAMS-87-12-1699
- Wood, A. W., and D. P. Lettenmaier, 2008: An ensemble approach for attribution of hydrologic prediction uncertainty, *Geophys. Res. Lett.*, 35, doi:10.1029/2008GL034648
- Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States, *J. Geophys. Res.-Atmos.*, 107, doi:10.1029/2001JD000659
- Wood, A. W., T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark, 2016a: Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill, *J. Hydrometeor.*, 17, 651–668, doi:10.1175/JHM-D-14-0213.1

-
- Wood, A. W., T. Pagano, and M. Roos, 2016b: Tracing the origins of ESP, available at: hepex.irstea.fr/tracing-the-origins-of-esp, last access: 12 September 2019
- Yossef, N. C., H. Winsemius, A. Weerts, R. van Beek, and M. F. P. Bierkens, 2013: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing, *Water Resour. Res.*, 49, 4687–4699, doi:10.1002/wrcr.20350
- Yuan, X., J. K. Roundy, E. F. Wood, and J. Sheffield, 2015a: Seasonal forecasting of global hydrologic extremes: system development and evaluation over GEWEX basins, *B. Am. Meteorol. Soc.*, 96, 1895–1912, doi:10.1175/BAMS-D-14-00003.1
- Yuan, X., E. F. Wood, and Z. Ma, 2015b: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, *Wiley Interdisciplinary Reviews: Water*, 2, 523–536, doi:10.1002/wat2.1088
- Yuan, X., E. F. Wood, N. W. Chaney, J. Sheffield, J. Kam, M. Liang, and K. Guan, 2013a: Probabilistic Seasonal Forecasting of African Drought by Dynamical Models, *J. Hydrometeorol.*, 14, 1706–1720, doi:10.1175/JHM-D-13-054.1
- Yuan, X., 2016: An experimental seasonal hydrological forecasting system over the Yellow River basin—Part 2: The added value from climate forecast models, *Hydrol. Earth Syst. Sci.*, 20, 2453–2466, doi:10.5194/hess-20-2453-2016
- Yuan, X., E. F. Wood, L. Luo, and M. Pan, 2011: A first look at Climate Forecast System version 2 (CFSv2) for hydrological seasonal prediction, *Geophys. Res. Lett.*, 38, doi:10.1029/2011GL047792
- Yuan, X., E. F. Wood, J. K. Roundy, and M. Pan, 2013b: CFSv2-based seasonal hydroclimatic forecasts over the conterminous United States, *J. Climate*, 26, 4828–4847, doi:10.1175/JCLI-D-12-00683.1
- Yuan, X., F. Ma, L. Wang, Z. Zheng, Z. Ma, A. Ye, and S. Peng, 2016: An experimental seasonal hydrological forecasting system over the Yellow River basin—Part 1: Understanding the role of initial hydrological conditions, *Hydrol. Earth Syst. Sci.*, 20, 2437–2451, doi:10.5194/hess-20-2437-2016
- Zajac, Z., M. Zambrano-Bigiarini, P. Salamon, P. Burek, A. Gentile, and A. Bianchi, 2013: Calibration of the lisflood hydrological model for europe – calibration round 2013, Joint Research Centre, European Commission

Zhu, E., X. Yuan, and A. W. Wood, 2019: Benchmark decadal forecast skill for terrestrial water storage estimated by an elasticity framework, *Nature communications*, 10, doi:10.1038/s41467-019-0924

Appendix

This appendix contains the typeset versions of each of the published chapters presented in this thesis, alongside further publications co-authored during this PhD. All author contribution statements (provided in the respective chapters or in the Appendix) have been approved by Professor Hannah Cloke, supervisor.

Hannah L. Cloke

A1: Willingness-to-pay for a probabilistic flood forecast: a risk-based decision-making game

This paper presents the published version of Chapter 2, Sect. 2.2 of this thesis, with the following reference:

Arnal, L., M.-H. Ramos, E. Coughlan de Perez, H. L. Cloke, E. Stephens, F. Wetterhall, S. J. van Andel, and F. Pappenberger, 2016: Willingness-to-pay for a probabilistic flood forecast: a risk-based decision-making game, *Hydrol. Earth Syst. Sci.*, 20, 3109-3128, 10.5194/hess-20-3109-2016*

* ©2016. The Authors. Hydrology and Earth System Sciences, a journal of the European Geosciences Union published by Copernicus. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.



Willingness-to-pay for a probabilistic flood forecast: a risk-based decision-making game

Louise Arnal^{1,2}, Maria-Helena Ramos³, Erin Coughlan de Perez^{4,5,6}, Hannah Louise Cloke^{1,7}, Elisabeth Stephens¹, Fredrik Wetterhall², Schalk Jan van Andel⁸, and Florian Pappenberger^{2,9}

¹Department of Geography and Environmental Science, University of Reading, Reading, UK

²ECMWF, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, UK

³IRSTEA, Catchment Hydrology Research Group, UR HBAN, Antony, France

⁴Red Cross/Red Crescent Climate Centre, The Hague, the Netherlands

⁵Institute for Environmental Studies, VU University Amsterdam, the Netherlands

⁶International Research Institute for Climate and Society, Palisades, New York, USA

⁷Department of Meteorology, University of Reading, Reading, UK

⁸UNESCO-IHE Institute for Water Education, Delft, the Netherlands

⁹School of Geographical Sciences, University of Bristol, Bristol, UK

Correspondence to: Louise Arnal (l.l.s.arnal@pgr.reading.ac.uk)

Received: 13 January 2016 – Published in Hydrol. Earth Syst. Sci. Discuss.: 19 January 2016

Revised: 24 May 2016 – Accepted: 25 May 2016 – Published:

Abstract. Probabilistic hydro-meteorological forecasts have over the last decades been used more frequently to communicate forecast uncertainty. This uncertainty is twofold, as it constitutes both an added value and a challenge for the forecaster and the user of the forecasts. Many authors have demonstrated the added (economic) value of probabilistic over deterministic forecasts across the water sector (e.g. flood protection, hydroelectric power management and navigation). However, the richness of the information is also a source of challenges for operational uses, due partially to the difficulty in transforming the probability of occurrence of an event into a binary decision. This paper presents the results of a risk-based decision-making game on the topic of flood protection mitigation, called “How much are you prepared to pay for a forecast?”. The game was played at several workshops in 2015, which were attended by operational forecasters and academics working in the field of hydro-meteorology. The aim of this game was to better understand the role of probabilistic forecasts in decision-making processes and their perceived value by decision-makers. Based on the participants’ willingness-to-pay for a forecast, the results of the game show that the value (or the usefulness) of a forecast depends on several factors, including the way users

perceive the quality of their forecasts and link it to the perception of their own performances as decision-makers.

1 Introduction

In a world where hydrological extreme events, such as droughts and floods, are likely to be increasing in intensity and frequency, vulnerabilities are also likely to increase (WMO, 2011; Wetherald and Manabe, 2002; Changnon et al., 2000). In this context, building resilience is a vital activity. One component of building resilience is establishing early warning systems, of which hydrological forecasts are key elements.

Hydrological forecasts suffer from inherent uncertainties, which can be from diverse sources, including the model structure, the observation errors, the initial conditions (e.g. snow cover, soil moisture, reservoir storages) and the meteorological forecasts of precipitation and temperature (Verkade and Werner, 2011; He et al., 2009). The latter variables are fundamental drivers of hydrological forecasts and are therefore major sources of uncertainty. In order to capture some of this uncertainty, there has been a gradual adoption of prob-

abilistic forecasting approaches, with the aim of providing forecasters and forecast users with additional information not contained in the deterministic forecasting approach. Whereas “a deterministic forecast specifies a point estimate of the predictand (the variate being forecasted)”, “a probabilistic forecast specifies a probability distribution function of the predictand” (Krzysztofowicz, 2001). For operational forecasting, this is usually achieved by using different scenarios of meteorological forecasts following the ensemble prediction approach (Buizza, 2008; Cloke and Pappenberger, 2009).

Many authors have shown that probabilistic forecasts provide an added (economic) value compared to deterministic forecasts (Buizza, 2008; Verkade and Werner, 2011; Pappenberger et al., 2015). This is due, for example, to the quantification of uncertainty by probabilistic forecasting systems, their ability to better predict the probability of occurrence of an extreme event and the fact that they issue more consistent successive forecasts (Dale et al., 2014; Cloke and Pappenberger, 2009). This probability of occurrence makes the probabilistic forecasts useful in the sense that they provide information applicable to different decision thresholds, essential since not all forecast users have the same risk tolerance (Michaels, 2015; Buizza, 2008; Cloke and Pappenberger, 2009). Probabilistic forecasts therefore enable the quantification of the potential risk of impacts (New et al., 2007) and, as a result, they can lead to more optimal decisions for many hydrological operational applications, with the potential to realise benefits from better predictions (Verkade and Werner, 2011; Ramos et al., 2013). These applications are, for example, flood protection (Stephens and Cloke, 2014; Verkade and Werner, 2011), hydroelectric power management (García-Morales and Dubus, 2007; Boucher et al., 2012) and navigation (Meissner and Klein, 2013). Moreover, the continuous increase in probabilistic forecast skill is very encouraging for the end-users of the probabilistic forecasts (Bauer et al., 2015; Magnusson and Källén, 2013; Simmons and Hollingsworth, 2002; Ferrell, 2009).

However, the communication of uncertainty through probabilistic forecasts and the use of uncertain forecasts in decision-making are also challenges for their operational use (Cloke and Pappenberger, 2009; Ramos et al., 2010; Michaels, 2015; Crochemore et al., 2015). One of the reasons why the transition from deterministic to probabilistic forecasts is not straightforward is the difficulty in transforming a probabilistic value into a binary decision (Dale et al., 2014; Demeritt et al., 2007; Pappenberger et al., 2015). Moreover, decision-makers do not always understand probabilistic forecasts the way forecasters intend them to (Handmer and Proudley, 2007). This is why it is essential to bridge the gap between forecast production and hazard mitigation, and to foster communication between the forecasters and the end-users of the forecasts (Cloke and Pappenberger, 2009; Michaels, 2015).

As (Michaels, 2015) notes, “the extent to which forecasts shape decision making under uncertainty is the true measure of the worth of a forecast”. The potential added value of the forecast can furthermore only be entirely realised with full buy-in from the decision-makers. However, how much are users aware of this added value? How much are they ready to pay for a forecast? These are questions that motivated the work presented in this paper. In order to understand how users perceive the value of probabilistic forecasts in decision-making, we designed a risk-based decision-making game – called “How much are you prepared to pay for a forecast?” – focusing on the use of forecasts for flood protection. The game was played during the European Geophysical Union (EGU) General Assembly meeting 2015 (Vienna, Austria), at the Global Flood Partnership (GFP) workshop 2015 (Boulder, Colorado), as well as at Bristol University (BU) in 2015. Games are increasingly promoted and used to convey information of scientific relevance. They foster learning, dialogue and action through real-world decisions, which allow the study of the complexities hidden behind real-world decision-making in an entertaining and interactive set-up (Mendler de Suarez et al., 2012).

This paper presents the details of the game and the results obtained from its different applications. The participants’ perceived forecast value is analysed by investigating the way participants use the forecasts in their decisions and their willingness-to-pay (WTP) for a probabilistic forecast. The WTP is the amount an individual is inclined to disburse to acquire a good or a service, or to avoid something undesirable (Breidert et al., 2006; Leviäkangas, 2009). It is a widely and very commonly adopted method to make perceived value assessments and its use has been demonstrated in a meteorological context (Leviäkangas, 2009; Anaman et al., 1998; Rollins and Shaykewich, 2003; Breidert et al., 2006). (Breidert et al., 2006) present a complete overview of the methods available, organised by data collection types. According to their classification, there exist two main WTP measuring approaches: the “revealed preference” and the “stated preference”. The former describes price-response methods (such as market data analysis, laboratory experiments and auctions, amongst others), while the latter refers to surveys in general. This experiment combines both “revealed preference” and “stated preference” methods. The design of the game is described in Sect. 2 and justified in terms of the purpose and contribution of the different components of the game to its main aim. The results and the discussion promoted by the latter are subsequently presented in Sects. 3 and 4 respectively.

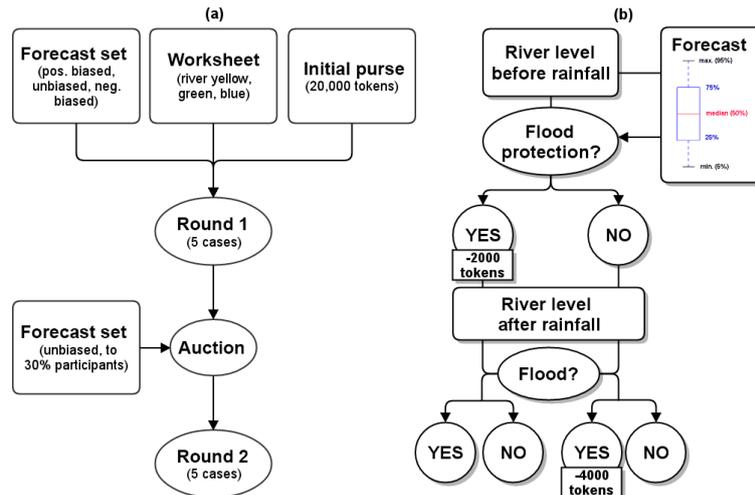


Figure 1. (a) Experiment set-up and (b) flow diagram of the game decision problem for one case.

2 Set-up of the decision-making game

2.1 Experimental design

This game was inspired by the table game “Paying for Predictions”, designed by the Red Cross/Red Crescent Climate Centre (<http://www.climatecentre.org/resources-games/paying-for-predictions>). Its focus is however different. Here, our aim is to investigate the use of forecasts for flood protection and mitigation. Also, we strongly adapted the game to be played during conferences and with large audiences.

The set-up of the game (illustrated in Fig. 1a) was the following: participants were told that they were competing for the position of head of the flood protection team of a company. Their goal was to protect inhabitants of a fictitious town bordering a fictitious river against flood events, while spending as little money as possible during the game. The participant with the highest amount of money at the end of the game was chosen as head of the flood protection team. Each participant was randomly assigned a river (river yellow, river blue or river green) for the entire duration of the game. Each river had distinct initial river levels and rates of flood occurrences (see Table 1). Participants worked independently and had a worksheet to take notes (see Appendix A). An initial purse of 20 000 tokens was given to each player to be used throughout the game.

Based on this storyline, the participants were presented the following sequence of events (illustrated in Fig. 1b): after being given their river’s initial level (ranging from 10 to 60 included), each participant was asked to make use of a probabilistic forecast (see Fig. 1b) of their river level increment after rainfall (ranging from 10 to 80 included) to decide whether they wanted to pay for flood protection or not. The cost of flood protection was 2000 tokens. They were informed, prior to the start of the game, that a flood occurred if

Table 1. Number of flood events for each round of the game and each river.

Round	River		
	Yellow	Green	Blue
1	1	2	3
2	3	2	1
Total	4	4	4

the sum of the initial river level and the river level increment after rainfall (i.e. the actual river level after rainfall) reached a given threshold of 90. The probabilistic forecasts were visualised using boxplot distributions. They had a spread of about 10–20, and indicated the 5th and 95th percentiles as well as the median (i.e. 50th percentile) and the lower and upper quartiles (i.e. 25th and 75th percentiles respectively) of the predicted river level increment after rainfall. Forecasts were given to participants case by case (i.e. when playing the first case, they could only see the boxplot distribution of forecast river increment for case 1). Once the participants had made their decisions using both pieces of information (i.e. river level before rainfall and forecast of river level increment), they were given the observed (actual) river level increment after rainfall for their rivers. If a flood occurred and the participant had not bought flood protection, a damage cost (i.e. price paid when no protection was bought against a flood that actually happened) of 4000 tokens had to be paid.

The monetary values (initial purse, price of flood protection and damage cost) were deliberately chosen. The price of a protection was set to 2000 tokens such that if a participant decided to buy flood protection every time during the game (i.e. two rounds of five cases each, thus ten times) they would have no tokens left in their purse at the end of the game. This

was done in order to discourage such a behaviour. The damage cost was set to twice the flood protection cost as this was estimated to be a realistic relation between the two prices based on (Pappenberger et al., 2015). The latter states that the avoided damages due to early flood warning amount to a total of about 40 %. Here, for simplicity, we used a percentage of 50 %.

Once the context was explained, the participants were then told that they would first play one round of five independent cases, which would each be played exactly according to the sequence of events presented, and for which they would have to record their decisions on the worksheet they were provided (see Appendix A). The game had a total of two rounds of five cases each. This specific number of cases and rounds was chosen because of the time constraint to play the game during conferences (the game should last around 20–30 min only). Table 1 presents the total number of flood events for each round and each river. The number of flood events was different for every river for each round as river level values were randomly generated for the purpose of the game. This allowed the exploration of the influence of different flood frequencies in round 1 on the participants' WTP for a second forecast set. The number of flood events was however sampled to some extent in order to obtain decreasing (increasing) numbers of flood events between the two rounds for the blue (yellow) river, or constant throughout the two rounds for the green river. This was done to investigate the effect of the change (or not) in flood frequency between rounds 1 and 2 on the participants' strategies throughout the game.

During the first round of the game, the participants had forecasts of river level increments to help their decisions. These forecasts were however not available for all participants in the second round, but were sold between the two rounds through an auction. The purpose and set-up of each round and the auction are explained in the following paragraphs.

2.1.1 Round 1

The objective of the first round was to familiarise the participants with the probabilistic forecasts they were given to help them in their decisions, and to create a diversity amongst the decision-makers in terms of

- their river behaviour: which is why different rivers, each with different flood frequencies and different initial levels, were assigned to the participants;
- the money they would spend during this round and have in hand for the ensuing auction (before round 2);
- the quality of their forecasts in the first round: to this end, different forecast sets were distributed to the players for round 1.

This diversity was triggered in round 1 in order to analyse whether or not the WTP for a second forecast set, measured

in the auction performed before round 2, was dependent on any of the factors inherent to the first round (i.e. river-specific flood frequency, money left in purse, or quality of the forecasts).

Before the start of the first round each participant was given a forecast set containing probabilistic forecasts of their river level increment after rainfall for the five cases of round 1. Participants were however not aware that three different forecast sets were produced for each of the rivers. One set had only forecasts with a positive bias (forecast sets 1), the second set had only unbiased forecasts (forecast sets 2) and the third set only forecasts with a negative bias (forecast sets 3). There were therefore nine different sets of forecasts which were distributed randomly amongst the audience prior to the start of the game. The three different forecast types were obtained by varying the position of the observation inside the forecast distribution. The unbiased forecasts had the observations fall between the lower and upper quartiles of their distributions, while the biased forecasts had the observations fall outside of the lower and upper quartiles of their distributions, leading to over- (positively biased forecast sets) or under-predictions (negatively biased forecast sets) of the observations.

The quality of each forecast set can be represented in terms of the number of correct forecast flood events (given a forecast percentile threshold) with respect to the number of observed flood events. For each forecast set type and each river, the number of forecast flood events during the first round was calculated by adding the median of the forecast river level increment to the initial river level for each case. A forecast is referred to as a false alarm if this sum forecasts a flood (i.e. it exceeds the flood threshold) but the flood is subsequently not observed. It is referred to as a hit if the sum forecasts the flood and the flood is subsequently observed. A miss is an observed flood that was not forecast. The numbers of hits, misses and false alarms are usually gathered in a contingency table as a matrix (e.g. Table 2): hits are placed on top, left, misses on bottom, left, and false alarms on top, right. The place on bottom, right is usually not considered in the evaluation of forecasts as it represents situations of low interest to a forecaster (i.e. when floods are neither forecast nor observed). Table 2 displays the nine contingency tables we obtain considering each forecast set type and each river. Each participant would find themselves in one of the contingency tables represented. We can see the higher number of total misses (false alarms) considering all rivers together in negatively (positively) biased forecast sets, and the absence of these in the unbiased forecast sets.

After all the five cases of round 1 were played, participants were asked to rate their performance as a decision-maker and the quality of their forecast set for round 1 on a scale from “very bad” to “very good” (the option “I don't know” was also available) (see Appendix A).

Table 2. Contingency table for each river and forecast set type for the first round (considering the 50th percentile, i.e. the median forecast). The numbers for a specific river-forecast set type represent, clockwise from the top left, hits (*italics*), false alarms (**bold**), correct negatives (–) and misses (regular).

Forecast set type	River					
	Yellow		Green		Blue	
Positively biased	<i>1</i>	1	<i>2</i>	1	<i>3</i>	2
	0	–	0	–	0	–
Unbiased	<i>1</i>	0	<i>2</i>	0	<i>3</i>	0
	0	–	0	–	0	–
Negatively biased	<i>0</i>	0	<i>2</i>	0	<i>2</i>	0
	1	–	0	–	1	–

2.1.2 Auction

The auction was carried out after round 1 in order to measure the participants' WTP for a second forecast set and to evaluate its dependencies on any of the elements of the game in round 1. The auction was implemented as follows.

At the end of the first round participants were asked to transfer the remaining tokens from round 1 to the second round. They were then told that the forecasting centre distributing the probabilistic forecasts now wanted the decision-makers to pay for the forecast sets if they wanted to have access to them for the second round. Furthermore, they were informed that only 30 % of them could get a second forecast set for this round. This percentage was chosen in order to restrict the number of participants that could buy a forecast set (and create a competitive auction), while keeping a high enough number of participants playing with a forecast set in round 2 for the analysis of the results.

Participants were then asked to make a sealed bid, writing down on their worksheets the number of tokens they were willing to disburse from their final purse of round 1 to obtain a set of probabilistic forecasts for all five cases of round 2. After the bids were made, a forecast set was distributed to the participants within the highest 30 % of the bids. This was done through an auction. It was carried out by asking the participants whether any of them wrote down a bid superior or equal to 10 000 tokens. If any participants did, they raised their hands, after which a forecast set – for the same river as the river assigned to them at the beginning of the game – was given to them. The auction continued by lowering the number of tokens stated to the participants until all forecast sets for round 2 were distributed. Each participant having bought a forecast set for round 2 was then asked to disburse the number of tokens they paid for this forecast set from their remaining purse from round 1.

We note that participants were not told that the forecasts for the second round were all unbiased forecasts. Once again,

the quality of the forecasts was kept secret in order for the participants to assign a value to the second forecast set that would strictly be related to the conditions under which they played the first round.

2.1.3 Round 2

The second round was played in order to measure the added value of an unbiased forecast set, compared to no forecast set at all, to the decisions of the participants on protecting or not against floods. Moreover, as the winner of the game was determined by the number of tokens left in their purse at the end of the game, this round would give a chance to participants who bought a second forecast set to make up for the money spent with the auction, during round 2.

The second round developed similarly to the first round, with five independent cases of decision-making, with the exception that only participants who bought a second forecast set could use it to make their decisions. Participants who did not buy a second forecast set did not have any forecasts on which to base their decisions.

After the five cases were played, the participants were asked to once again answer a set of questions (see Appendix A). They were asked to rate their performance as a decision-maker in the second round, on a scale from “very bad” to “very good” (the option “I don't know” was also available). Participants without a second forecast set were invited to provide a justification for not purchasing a set of forecasts for this round. Participants who had bought a second forecast set were also asked to rate the quality of their forecast set for round 2 (on a scale from “very bad” to “very good”; the option “I don't know” was also available) and whether those were worth the price they had paid for them. If not, they were asked to provide a new price that they would have rather paid.

The winner was finally determined by finding the player with the largest number of tokens in their purse at the end of the game.

2.2 Objectives and evaluation strategy

The main aim of this paper is to investigate the participants' WTP for a probabilistic forecast set in the context of flood protection, following the game experiment designed as presented in the previous paragraphs. It unfolds into two objectives that were pursued in the analysis of the results:

1. to analyse how participants used the information they were provided (probabilistic forecast sets) in this risk-based decision-making context, and
2. to characterise the participants' WTP for a probabilistic forecast set for flood protection.

We assess these objectives through six questions, which are presented below, together with the evaluation strategy implemented.

2.2.1 Did the participants use their forecasts and, in this case, follow the 50th percentile of their forecast during the decision-making process?

This first question was investigated using the results of the first round. We first wanted to know whether the players were actually using their forecasts to make their decisions. Moreover, we searched for clues indicating that the participants were following the 50th percentile (i.e. the median) of the probabilistic forecasts. This was done in order to see whether the 50th percentile was considered by the players as the optimal value to use for the decision-making process under this specific flood risk experiment. Additionally, this question relates to an intrinsic characteristic of the use of probabilistic forecasts for decision-making, which is the difficulty in transforming the probabilistic values into a binary decision (Dale et al., 2014; Demeritt et al., 2007; Pappenberger et al., 2015). The way in which probabilistic flood forecasts are used depends on attitudes of decision-makers towards risk, the uncertainty and the error in the information provided to them (Demeritt et al., 2007; Ramos et al., 2013), and decisions can vary from one participant to the next provided the same information (Crochemore et al., 2015).

Question one was explored by looking at the worksheets collected in order to infer from the decisions taken by the participants whether or not they most probably used the median of their forecasts to consider whether the river level would be above, at or under the flood threshold. In cases where the decisions did not coincide with what the median forecast indicated, other factors that could also influence the decisions were considered, such as (a) the flood frequency of each river and their initial river levels, (b) the forecast set type each participant had (i.e. biased – positively or negatively – or unbiased) and (c) the familiarity of the participants with probabilistic forecasts and decision-making (given their occupation and years of experience).

2.2.2 Was there a correspondence between the way participants perceived the quality of their forecasts in round 1 and their “true” quality?

A well-known effect, called the “cry wolf”, was studied for weather-related decision-making by LeClerc and Joslyn (2015). It describes the reluctance of users to comply with future alarms when confronted in the past with false alarms. This leads to the second question which was explored in this paper: was there a correspondence between the way participants perceived the quality of their forecasts in round 1 and their “true” quality? Our aim here is to investigate whether the participants were more sensitive to false alarms or misses. The participants’ answers to the question on their forecast set quality for the first round (see Appendix A) were analysed against their “true” quality. The latter was measured in terms of forecast bias, calculated from the hits, false alarms and misses presented in Table 2. A bias value was computed

for each forecast set type of each river (i.e. each contingency table; there were therefore nine different bias values in total) with the following equation:

$$\text{Bias} = \frac{\text{hits} + \text{false alarms}}{\text{hits} + \text{misses}}. \quad (1)$$

A bias value equal to one is a perfect value (which corresponds to unbiased forecasts), and a value less than (superior to) one indicates under- (over-)prediction.

2.2.3 Did the participants’ perceptions of their own performance coincide with their “true” performance?

We also looked at the perception the participants had of their own performance. The answers to the question “How was your performance as a decision-maker” (see Appendix A) were assessed against the participants’ “true” performances (in rounds 1 and 2), which were calculated in terms of the money participants spent as a consequence of their decisions. The following general formula (n being the round number) was used:

$$\text{Performance} = \frac{\text{Money spent round } n}{\text{Optimal}}. \quad (2)$$

The performance is expressed relatively to an optimal performance, which is the minimum amount a participant could have spent, given the river they were assigned, defined as

$$\text{Optimal} = \text{Protection cost} \times \text{Number of floods in round } n. \quad (3)$$

A performance value of one indicates an optimal performance. Performance values greater than one indicate that participants spent more money than the minimum amount necessary to protect the city from the observed floods. The greater the value, the higher the amount of money unnecessarily spent.

2.2.4 What was the participants’ willingness-to-pay for a probabilistic forecast set?

The auction was incorporated into the experiment in order to explore the WTP of participants for a probabilistic forecast set, considering the risk-based decision-making problem proposed by the game. To characterise this WTP, the bids were analysed and their relationships with several other aspects of the game were explored to explain the differences (if any) in the bids. These aspects were the following.

- The way participants used the forecasts. Here we try to learn about the effectiveness of the information on the user, which is an attribute of the value of information (Leviäkangas, 2009). It is assumed that a participant is not expected to be willing to disburse any money for information they are not using. The answers to question one (i.e. “Did the participants use their forecasts and, in this case, follow the 50th percentile of their forecast during the decision-making process?”) are used here.

- The money available to participants after round 1 to make their bids. As participants were informed at the beginning of the game that the winner would be the player with the highest number of tokens in purse at the end of the game, the tokens they had in hand for the auction (after round 1) may have restricted them in their bids. The bids are thus also explored relative to the number of tokens in hand at the time of the auction.
- The forecast set type. The bias of the forecasts during round 1 could also have been a potential determinant of participants' WTP for a forecast set in round 2.
- The river flood frequency. This was different for all the rivers in the first round and could be an element of the relevance of the information, another attribute of the value of information (Leviäkangas, 2009). Indeed, one could ask: "If my river never floods, why should I pay for forecasts?".
- The years of experience and occupation. This might influence the familiarity participants may have with the use of probabilistic forecasts for decision-making.

2.2.5 Did participants with a forecast set perform better than those without?

Round 2 was led by a central question: did participants with a forecast set perform better than those without? It was investigated by looking at the performance of participants in round 2, calculated from Eq. (2). While we expect players with more (unbiased) information to make better decisions, other factors could have influenced the trust participants had in the information during round 2, such as, for instance, the quality of the forecasts experienced by participants in round 1 or the flood events observed in the river in round 2, compared to the experience participants had previously had in round 1.

2.2.6 What were the winning and losing strategies (if any)?

Finally, from the final results of the game, a question arose: what were the winning and losing strategies (if any)? This question was explored by looking at the characteristics (e.g. river assigned, forecast set type in round 1, performances in both rounds, purchase of a second forecast set) and decisions of the participants during the game, in order to distinguish common attributes for the winning and losing strategies.

Furthermore, an "avoided cost" was calculated for each river based on the difference between the tokens spent by participants without a second forecast set and the tokens spent by participants with a second forecast set, during round 2. It represents the average number of tokens participants without a second forecast set lost by protecting when a flood did not occur or by not protecting when a flood did occur, compared

Table 3. Distribution of the 129 worksheets collected for the analysis per river (yellow, green and blue) and forecast set type (positively biased, unbiased and negatively biased).

Forecast set type	River			Total
	Yellow	Green	Blue	
Positively biased	15	11	18	44
Unbiased	13	21	9	43
Negatively biased	11	19	12	42
Total	39	51	39	129

to participants with a second forecast set. This "avoided cost" was measured and compared to the average bid of participants for each river in order to evaluate participants' estimation of the value of the forecasts compared to their "true" value in terms of the money they enabled the participants with a second forecast set to save in the second round. An average "new bid" was also calculated by replacing the bids of participants who had said that their forecast set in the second round was not worth the price they had paid initially, with the new bids they would have rather paid (see Appendix A). This average "new bid" was compared to the "avoided cost" and the actual average bid obtained from the auction.

3 Results

The results are based on the analysis of 129 worksheets from the 145 worksheets collected. The remaining 16 worksheets were either incomplete or incorrectly completed and were thus not used. Table 3 shows the distribution of the 129 worksheets among the three forecast set types and the three rivers.

The game was played at the different events mentioned in the introduction. The participants present at those events displayed a diversity in terms of their occupation and years of experience. This was surveyed at the beginning of the game and is presented in Fig. 2, for all the participants as well as for each river and forecast set type separately. Participants were mainly academics (postdoctoral researchers, PhDs, research scientists, lecturers, professors and students), followed by professionals (forecasters, operational hydrologists, scientists, engineers and consultants). The majority had less than 5 years of experience.

3.1 Participants were using the forecasts, but consistent patterns of use are difficult to detect

Figure 3 presents, on the one hand, the final purses of all the participants at the end of round 1, according to their river and forecast set type (columns and rows respectively), and, on the other hand, the final purses that participants would have had if they had made their decisions according to the median of their forecasts. Participants in charge of the yellow

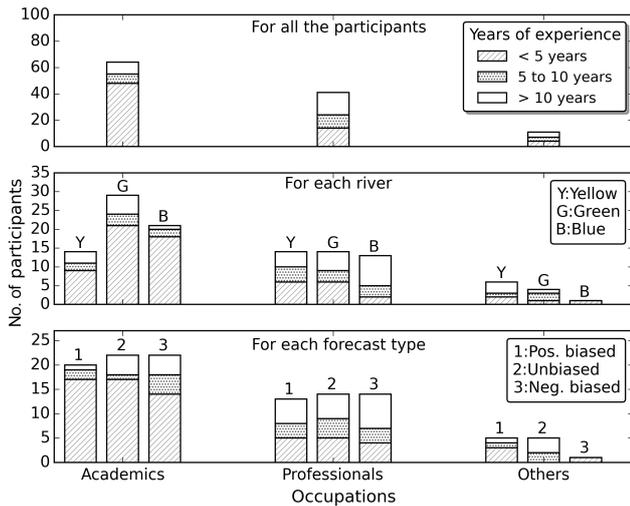


Figure 2. Number of participants according to occupation and years of experience. The categories of occupations are academics (post-doctoral researchers, PhDs, research scientists, lecturers, professors and students), professionals (forecasters, operational hydrologists, scientists, engineers and consultants) and others. Top: overall participant distribution; middle: distribution according to their river; bottom: distribution according to the forecast quality types (1: positively biased, 2: unbiased and 3: negatively biased).

river (first column) ended the first round with, on average, more tokens than the others. Participants playing with the blue river (last column) are those who ended round 1 with less money in purse, on average. This is due to the higher number of flood events for the blue river in round 1 (see Table 1). There are also differences in terms of final purses for the participants assigned the same river but given a different forecast set type. Overall, participants who had unbiased forecasts (middle row) ended the first round with on average more money than the other players. These results are an indication that the participants were using their forecasts to make their decisions.

In order to see whether the participants were using the median values of the forecasts, a forecast final purse was computed considering the case where the participants followed the median of their forecasts for all the cases of the first round (red vertical lines shown in Fig. 3). If the participants had followed the median values of the forecasts during the entire first round, their final purses would have been equal to this value. Although this is almost the case for participants with unbiased forecast sets (for all rivers), for participants with the yellow river and positively biased forecast sets and the green river and negatively biased forecast sets, it is not an overall generally observed behaviour.

Could some participants have discovered the bias in their forecasts and adjusted them for their decisions? Although it is hard to answer this question from the worksheets only, some of the decisions taken seem to support this idea. Figure 4 presents in more detail the results for the blue river in

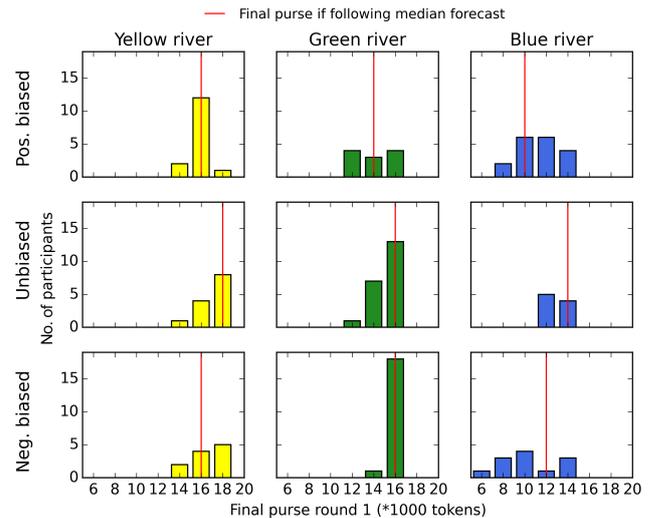


Figure 3. Participants' round 1 final purses for each river (from the leftmost to the rightmost column: the yellow, the green and the blue river) and for each forecast set type (from the top to the bottom row: positively biased, unbiased and negatively biased). The red lines show the final purses that the participants of a given river-forecast set type group would have gotten if they had followed the median of their forecasts for all five cases of the first round.

the first round. The forecast final levels are shown as box-plots for each forecast set type and for each of the five cases of round 1. These are the levels the river would reach if the initial level is added to the percentiles of the forecasts for each case. The bars at the bottom of the figure show the percentages of participants whose decisions differed from what the median of their forecast final level indicated (i.e. participants who bought (or did not buy) protection while no flood (or a flood) was predicted by the median of their forecast).

When comparing cases 1 and 4, for which the initial river levels and the observed and forecast final river levels were the same, we would not expect any changes in the way participants were using their forecasts. This is however not true. Figure 4 shows that the percentages of participants not following their forecast median differs between the two cases. For instance, about 80 % of the participants with negatively biased forecast sets (under-predicting the increment of the river level) did not follow the median forecast in case 1, and did not protect against the predicted flood by their median forecast, while this percentage drops to about 20 % in case 4. The fact that they were not consistently acting the same way may be an indication that they found out the bias in the forecasts and tried to compensate for it throughout round 1. We can also see that, in general, the lowest percentages of participants not following the median forecast are for the unbiased forecast set. This is especially observed in the cases where the forecast final levels given by the median forecast are well above or below the flood threshold (cases 1, 2, 4, 5). The fact that from case 1 to case 4, for unbiased forecast sets,

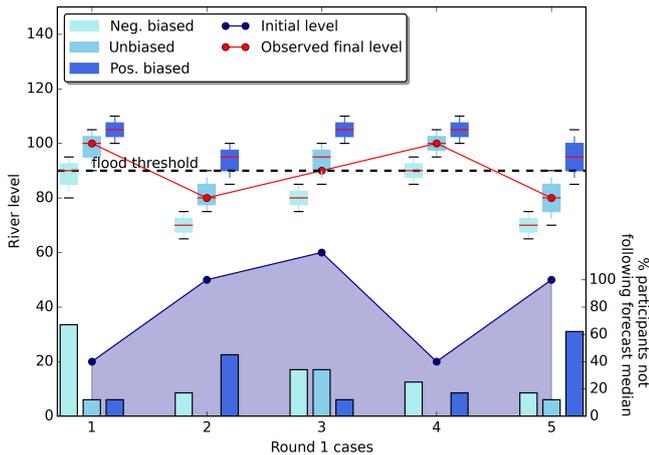


Figure 4. Observed initial and final river levels for the blue river for each case of the first round. The boxplots show the forecast final river levels by each forecast set type (negatively biased, unbiased and positively biased). The bars display the percentages of participants whose decisions did not correspond to what their forecast median indicated.

we moved from about 10 % of participants not following the median forecast to 0 %, may also indicate that they built confidence in their forecasts (at least in the median value) along round 1, by perceiving that the median forecast could be a good indication of possible flooding or not in their river.

Figure 4 also shows that some participants with unbiased forecasts did not always follow the median of their forecasts (for instance, cases 1, 3 and 5). Additional factors may therefore have influenced the way participants used their forecasts. A number of worksheets indicated that the distance of the initial river level to the flood threshold could have been influential. In a few cases where the median forecast clearly indicated a flood, while the initial river level was low, some players did not purchase any flood protection. This can be observed in Fig. 4 for case 1, for example, for participants with positively biased or unbiased forecast sets. The inverse situation (i.e. the initial river level was high, but the river level forecast by the median was low, below the flood threshold) was also observed and is illustrated in Fig. 4 for case 2 and negatively biased forecast sets. Hence, in some cases, the initial river level seemed to also play a role in the decisions taken.

There are indications that the participants could also have used other percentiles of the forecast to make their decisions, especially in cases where the median of the forecast was marginally above or below the flood threshold. For example, in case 4, the entire unbiased forecast lies above the flood threshold and all the participants chose the same and correct action. In cases where the 5th or 95th percentiles of the forecast fell above or below the flood threshold, the participants showed less consistent decisions (e.g. case 3 for unbiased forecast sets).

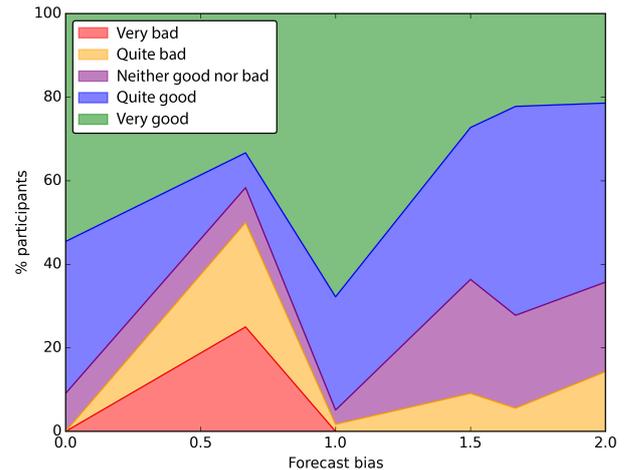


Figure 5. Cumulative percentages of participants who rated their forecast quality from “very bad” to “very good”, as a function of the forecast set bias (“true” forecast quality; Eq. 1) in round 1. A bias equal to one indicates perfect forecasts; a bias less than (superior to) one indicates under- (over-)prediction.

Other possible influencing factors, such as occupation and years of experience, were also investigated (not shown). No strong indications that these factors could have played a role in the participants’ decision-making were however found.

3.2 Participants were overall less tolerant to misses than to false alarms in round 1

Figure 5 displays the cumulative percentages of participants having answered that the quality of their forecast set in round 1 (see Appendix A) was “very bad” to “very good”, as a function of the “true” quality of the corresponding forecasts, measured by the forecast set bias (Eq. 1). While participants with forecast sets for which the bias equalled one (perfect value) mostly rated their forecasts “quite good” or “very good”, the percentage of negative perceptions of the quality of the forecasts increases with increasing or decreasing forecast bias.

It is interesting to note that participants with forecasts biased towards over-prediction never rated their forecasts as “very bad”. Also noteworthy is the very good rating given by participants with the most negatively biased forecasts (bias of 0). These participants belonged to the yellow river and had negatively biased forecasts in round 1. There was only one flood event for river yellow in the first round, which occurred at the end of the round and which was missed by the negatively biased forecasts. During the analysis of the results, it was observed that only about 25 % of the yellow river participants given the negatively biased forecasts did not purchase flood protection for this flood. An explanation for this low percentage could be that participants had time to learn about their forecasts’ quality until the occurrence of the flood at the end of the first round. This low number of participants who

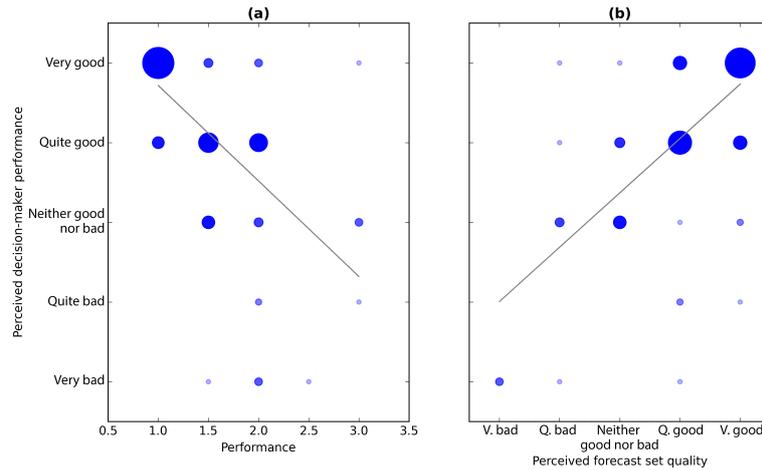


Figure 6. Number of participants having rated their performance as a decision-maker from “very bad” to “very good” in round 1, as a function of (a) their “true” performance (calculated from Eq. 2), and (b) their perceived forecast set quality. A performance value of one denotes a “true” performance equal to the optimal performance (Eq. 3). The larger the performance value, the more distant from optimal the decisions were during round 1. The size and the colour of the point indicate the number of participants that fall into a specific perceived–actual performance combination or perceived performance–forecast set quality combination.

actually suffered from their negative bias and the presence of only one miss out of the five cases of round 1 could therefore justify the good rating of their forecasts by those participants.

Overall, forecasts exhibiting under-prediction seem to be less appreciated by the participants. This could be an indication that participants were less tolerant to misses, while they accepted better forecasts leading to false alarms (over-predictions). This is contrary to the “cry wolf” effect, and could be explained by the particular game set-up for which the damage cost (4000 tokens) was twice the protection cost (2000 tokens).

3.3 Participants had a good perception of their good (or bad) performance during the game and related it to the quality of their forecasts

Figure 6a illustrates the answers to the question “How was your performance as a decision-maker in round 1?” as a function of the participants’ “true” performance (calculated from Eq. (2), i.e. the ratio to an optimal performance). The figure shows the distribution of participants across all perceived–actual performance combinations, for all rivers and forecast set types combined. The perceived decision-maker performance is presented on a scale from “very bad” to “very good”. An overall positive relationship between the participants’ perceived performance and their “true” performance is observed: the best performances (i.e. performance values of one or close to one) are indeed associated with a very good perception of the performance by the decision-makers and vice versa. The same analysis carried out for the answers concerning round 2 (not displayed) showed similar results: the ratings participants gave to their performance were similarly close to their “true” performance.

Figure 6b looks at the relationship between the perceived decision-maker performance and the rating the decision-makers gave to their forecast set quality in round 1. A positive relationship can also be seen: the majority rated their performance and the quality of their forecast set as “quite good” and “very good”, while those who rated their performance “very bad” also considered their forecast set “very bad”. The rating participants gave to their performance was therefore closely connected to the rating they gave to their forecast set quality. This also contributes to the evidence that participants were using their probabilistic forecast sets to make their decisions. It is furthermore an indication that participants linked good forecast quality to good performance in their decision-making and vice versa.

3.4 Several factors may influence the WTP for a forecast, including forecast quality and economic situation

Given the evidence that most participants were using their forecasts to make their decisions in round 1 (see Sect. 3.1), we now investigate their willingness-to-pay (WTP) for a new forecast set to be used in round 2.

Figure 7 shows the bids participants wrote on their worksheets prior to the auction, for a second forecast set, as a function of the number of tokens they had in their purses at the end of round 1. All bids are plotted and those from participants who succeeded in buying a second forecast set are displayed as red triangles in the figure. On average, participants were willing to pay 4566 tokens, which corresponds to 32 % of the average number of tokens left in their purses. The minimum bid was zero tokens (i.e. no interest in buying forecasts for round 2), which was made by 10 % of the

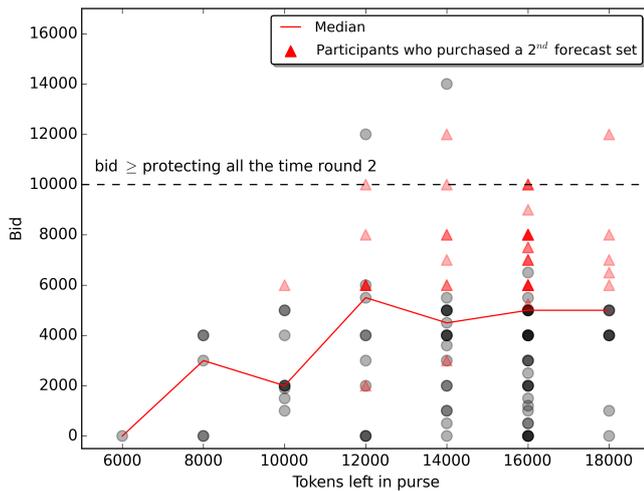


Figure 7. Bids declared by participants to purchase a forecast set for round 2, as a function of the number of tokens they had left in their purse at the end of round 1. The colour of the points indicates the number of participants that fall into a specific bid–tokens left in purse combination.

players. Half of these players were participants who were assigned the blue river (the river for which players ended the first round with on average the lowest number of tokens in purse). The only three participants who never bought flood protection in the first round (i.e. who could be seen as “risk-seeking” players) made bids of zero, 3000 and 4000 tokens. The highest bid made was 14 000 tokens, corresponding to 100 % of the tokens left in that participant’s purse. However, this participant did not raise their hand during the auction to purchase a second forecast set. Nine participants (less than 10 % of the total number of players) made a bid of 10 000 tokens or above, corresponding to, on average, 77 % of the tokens they had left in their purses. The total cost of protecting all the time for round 2 being 10 000 tokens, as indicated in Fig. 7 by the dashed black line, bidding 10 000 tokens or more for a second forecast set was clearly pointless. Half of these participants were players to which the yellow river was assigned (the river that experienced the least number of floods in round 1 and for which participants thus ended the first round with on average the highest number of tokens left in their purse) and eight out of these nine participants had a forecast set with a bias during the first round. These nine participants, who paid 10 000 tokens or more for the second forecast set, were removed from the subsequent analyses of the auction results, as their bids suggest that they have not understood the stakes of the game.

From Fig. 7, there is a clear positive relationship between the maximum bids within each value of tokens left in purse and the tokens left in purse, as the participants did not disburse more tokens than they had left in their purse during the auction. When we look at the evolution of the median of the bids with the number of tokens in purse, in general, the more

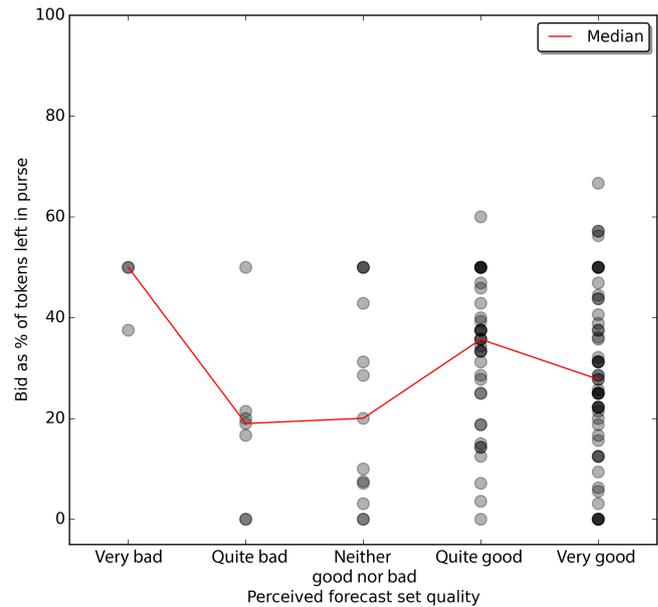


Figure 8. Participants’ % bids, bids expressed as a percentage of the tokens participants had left in their purse at the time of the auction, as a function of the rating they gave to their forecast set quality in round 1 (from “very bad” to “very good”). The colour of the points indicates the number of participants that fall into a specific bid–perceived forecast set quality combination.

tokens one had left in purse, the higher their WTP for a forecast set. Nonetheless, the WTP seems to have a limit. It can be seen that from a certain number of tokens left in purse, the median value of the bids remains almost constant (in our game case, at about a bid of 6000 tokens for participants with 12 000 tokens or more in their purse). The number of tokens that the participants had in hand therefore only influenced to a certain extent their WTP for a second probabilistic forecast set.

We also investigated whether the way participants perceived the quality of their forecast set in the first round was a plausible determinant of their WTP for another forecast set to be used in round 2. Figure 8 shows the % bids (i.e. bids expressed as a percentage of the tokens participants had left in their purse at the time of the auction) as a function of the rating participants gave to their forecast set quality in round 1 (from “very bad” to “very good”; see Appendix A). Firstly, it is interesting to observe that three participants judged their first forecast set to have been of “very bad” quality but were nonetheless willing to disburse on average 50 % of the tokens they had left in purse. Those bids were however quite low, 4000 tokens on average. Moreover, players who rated their first forecast set from “quite good” to “very good” were on average willing to disburse a larger percentage of their tokens than candidates who rated their previous forecast set from “quite bad” to “neither good nor bad”. Therefore, the way participants rated the quality of their first forecast set

Table 4. Distribution of the 44 forecast sets sold during the auction, per river (yellow, green and blue) and forecast set type (positively biased, unbiased and negatively biased).

Forecast set type	River			Total
	Yellow	Green	Blue	
Positively biased	5	2	6	13
Unbiased	6	9	3	18
Negatively biased	3	10	0	13
Total	14	21	9	44

was to a certain degree influential on their WTP for a second forecast set.

During the auction following the closed bids, 44 forecast sets were distributed to the participants who made the highest bids, in order to be used in round 2. Table 4 shows that participants who purchased these second forecast sets were quite well distributed among the different forecast set types of round 1, with a slightly higher frequency of buyers among participants who had played round 1 with unbiased forecasts; 42 % of all participants with unbiased forecasts purchased a second forecast set, while 30 % (31 %) of participants with positively biased (negatively biased) forecasts bought a second forecast set. Buyers also belonged more often to the group assigned river green (48 %, or 41 % of all green river participants), followed by rivers yellow (32 %, or 36 % of all yellow river participants) and blue (20, or 23 % of all blue river participants). The higher percentage of green river participants buying a second forecast set could have been due to a combination of the river green flood frequency in round 1 (not as low as for the yellow river, making it more relevant for green river participants to buy a second forecast set) and of money left in purse (on average, not as low as for the blue river participants). The buyers of the second forecast sets are displayed as red triangles in Fig. 7. We note that these red triangles are not necessarily the highest bid values in the figure, since we plot results from several applications of the game (in one unique application, they would coincide with the highest bids, unless a participant had a high bid but had not raised their hand during the auction to buy a second forecast set). Differences in the highest bids among the applications of the game could be an indication that the size (or type) of the audience might have had an impact on the bids (i.e. the WTP for a probabilistic forecast). Our samples were however not large enough to analyse this aspect.

Participants who did not purchase a second probabilistic forecast set (85 players in total) stated their reason for doing so. The majority of them (66 %, or 56 players) said that the price was too high (which means, in other words, that the bids made by the other participants were too high, preventing them from purchasing a second forecast set during the auction). Ten participants (12 %) argued that the model did not

seem reliable. Most of these participants were among those who had indeed received a forecast set with a bias in the first round. The rest of the candidates who did not purchase a second forecast set (22 %, or 19 players) wrote down on their worksheet the following reasons.

- *Low flood frequency in the first round* – a participant assigned the yellow river wrote: “Climatology seemed probability of flood = 0.2”.
- *Assessment of the value of the forecasts difficult* – a participant wrote: “No information for the initial bidding line”; and another wrote: “Wrong estimation of the costs versus benefits”.
- *Preference for taking risks* – “Gambling” was a reason given by a player.
- *Enough money left in purse to protect all the time during round 2* – which can be an indication of risk-averse behaviour coupled with economic wealth and no worries of false alarms.
- *Not enough money left in purse to bid successfully* – a participant wrote: “The purse is empty due to a lot of floods”.

3.5 Decisions are better when they are made with the help of unbiased forecasts, compared to having no forecasts at all

The analysis of the results of round 2 allowed us to compare the performance of participants with and without a forecast set. Overall, participants without a second forecast set had an average “true” performance value of 3.1, computed as shown in Eq. (2) and over the five cases of round 2. The best performance was equal to the optimal performance (“true” performance value equal to 1) and the worst performance reached a value of 6. Comparatively, participants with a second forecast set had an average “true” performance of 1.2, thus much closer to the optimal performance than the average performance of participants without a second forecast set. The best performance in this group also equalled the optimal performance, while the worst performance value was 2.5, much lower (i.e. thus much closer to the optimal value) than the worst performance value of participants making their decisions without any forecasts. These numbers clearly indicate that the possession of a forecast set in the second round led to higher performances and to a lower spread in performances within the group of players with a second probabilistic forecast set (compared to players without forecasts in round 2).

Does this conclusion however depend on the participants’ performances in round 1? Do you need to be a good decision-maker to benefit from the forecasts in hand? Our results suggest otherwise. All the participants with a bad performance in the first round and a forecast set in round 2 had a good performance in the second round. This indicates that even if

those participants had a bad performance in round 1, they took advantage of the forecasts and had a good performance in round 2. Additionally, 57 out of 59 participants with a good performance in round 1 and no forecasts in round 2 had a bad performance in the second round. This therefore indicates that no matter how well the participants performed in round 1, the possession of a forecast set led to better decisions in round 2.

All the participants without a second forecast set who were assigned the yellow river missed the first two floods in the second round. Some of these participants purchased flood protection for all or some of the subsequent cases, while the others never bought any protection. It could have been due to the low flood frequency of their river in the first round (see Table 1). This behaviour was not observed for the green river participants without a second forecast set, for which a very diverse sequence of decisions was seen in the second round. As for the blue river participants without any second forecast set, most of them missed the first flood event that occurred in round 2 and, subsequently, purchased flood protection for a few cases where no flood actually occurred. These decision patterns were not observed for participants with a second forecast set within each river, who took more consistently right decisions.

The large majority of participants with a second forecast set in round 2 (41 out of 44) rated their forecasts as either “quite good” or “very good”, which was expected since all the forecasts were unbiased in round 2. The three remaining participants said that their second forecast set was “neither good nor bad” or “quite bad”. These participants all had biased forecasts in the first round and their behaviour during round 2 suggested that they might have been influenced by the bias in their forecasts for round 1.

3.6 Overall winning strategies would combine good performance with an accurate assessment of the value of the forecasts

The average final purse at the end of round 2 was 3149 tokens (3341 tokens for participants without a second forecast set and 2778 tokens for participants with a second forecast set), remaining from the 20 000 tokens initially given to each participant. The minimum final purses observed were zero tokens or less. Twenty-five participants, out of the total of 129 players, finished the game with such amounts of tokens. Out of these 25 participants, 22 had received a biased forecast set in the first round. From the analysis of the game worksheets, we could detect three main losing strategies followed by these 25 participants who finished with zero tokens or less in purse.

1. Eighteen participants, most of them blue river players, had an “acceptable to bad” performance in round 1 (performances ranging between 1.3 and 3), did not purchase a second forecast set, and performed badly in round 2 (performances ranging between 2.3 and 6).

Table 5. Average values of “avoided cost” for round 2, average bid for a second forecast set and average “new bid” if forecasts were considered not worth the price originally paid. Values are in tokens and for the three different rivers.

River	Average “avoided cost”	Average bid	Average “new bid”
Yellow	7251	7929	7083
Green	5829	7083	6224
Blue	5711	6889	5875

2. Four players, mostly in charge of the yellow river, had a “good to bad” performance in round 1 (performances ranging between 1 and 3), purchased a second forecast set for 10 000 tokens or higher, and performed very well in round 2 (performances of 1).
3. Three participants, all green river players, had a “good to acceptable” performance in round 1 (performances ranging between 1 and 1.5), bought a second forecast set for 6000–8000 tokens, but performed badly in round 2 (performances ranging between 2 and 2.5).

The winners of the game, six players in total, finished round 2 with 8000 or 12 000 tokens in their purse. Half of these participants were assigned the green river and the other half the blue river. Apart from one participant, all had received a biased forecast set in the first round. Most participants had a “good to acceptable” performance in the first round (performances ranging between 1 and 1.7), did not purchase any forecast set and had a “good to bad” performance in the second round (performances ranging between 1 and 3). Their performance in round 2 did not lead to large money losses, as it did for yellow river participants, which can be explained by the fact that they did not have so many flood events in this round (see Table 1).

The average “avoided cost”, the average bid for a second forecast set and the average “new bid” are presented in Table 5 for each river. By comparing the “avoided cost” with the average bid for each river, it is noticeable that the average bid was larger than the “avoided cost” of each river. On average participants paid 1000 tokens more for their second forecast set than the benefit, in terms of tokens spared in the second round, that they derived from having this forecast set. This could explain why none of the winners of the game had a forecast set in the second round. From the average “new bid”, it is evident that participants would have liked to pay less on average than what they originally paid for their second forecast set. For all the rivers, the average ‘new bid’ is closer to the “avoided cost” than the average bid of participants during the auction.

4 Discussion

4.1 Experiment results and implications

It was clear during the game that most participants had used the probabilistic forecasts they were given at the beginning of the game to help them in their decisions. This was an important issue in our game since it was an essential condition to then be able to evaluate how the participants were using their forecasts and to understand the links between the way they perceived the quality of their forecasts and the way they rated their performance at the end of a round. There was evidence that participants were mostly using the 50th percentile of the forecast distributions, but, interestingly, the median alone could not explain all the decisions made. Other aspects of the game might have also shaped the participants' use of the information, such as the discovery, during the first round, of the forecast set bias (i.e. two out of three forecast sets were purposely biased for round 1). This was also mentioned by some participants at the end of some applications of the game, who said that the fact of noticing the presence of a bias (or suspecting it, since they were not told beforehand that the forecasts were biased) led them to adjust the way they were using the information. This could suggest that forecasts, even biased, can still be useful for decision-making, compared to no forecasts at all, if users are aware of the bias and know how to consider it before making a decision.

Interestingly, in the analysis of the worksheets, there was an indication that the players had, however, different tolerances to the different biases. Indeed, a lower tolerance for under-predictive forecasts than for over-predictive forecasts was identified. Biased forecasts were hence problematic for the users and indicative of the manner in which the information was used. This strongly indicates that there is an important need for probabilistic forecasts to be bias-corrected previously to decision-making, a crucial aspect for applications such as flood forecasting, for instance (Hashino et al., 2007; Pitt, 2008).

There was additionally evidence that, in a few cases, some participants with unbiased forecasts did not use their forecasts (when considering the 50th percentile as key forecast information). The analysis suggested that the players' risk perception, triggered by the initial river level or the proximity of the forecast median to the flood threshold, might have been a reason for this. This led to less consistent actions, where participants based their decisions on extremes of the forecast distribution (other percentiles of the forecast) or on no apparent information contained in the forecast distribution. A similar finding was reported by Kirchoff et al. (2013) through a case study in America, where it was found that the perception of a risk was a motivational driver of a water manager's use of climate information. There is a constant effort from forecasters to produce and provide state-of-the-art probabilistic forecasts to their users. However, it was seen here that even participants with unbiased forecasts did not al-

ways use them. This is an indication that further work needs to be done on fostering communication between forecasters and users, to promote an enhanced use of the information contained in probabilistic forecasts.

From the results, it also appeared that the participants had an accurate perception of their decision-maker performance and related it to the quality of their forecasts. This implies that participants viewed their forecasts as key elements of their decision-making. This result is very encouraging for forecasters and also bears important implications for the real world. It could indeed suggest that decision-makers forget that their own interpretation of the forecasts is as important as the information held in the forecast itself, as there is a myriad of ways to interpret and use probabilistic forecasts for decision-making. The choice of the percentile on which the decisions are based is an example of such an interpretation. This could potentially mean that decision-makers will tend to blame (thank) the forecast providers for their own wrong (good) decisions.

Many papers have shown, through different approaches, the expected benefits of probabilistic forecasts vs. deterministic forecasts for flood warning (e.g. Buizza, 2008; Verkade and Werner, 2011; Pappenberger et al., 2015; Ramos et al., 2013). However, many challenges still exist in the operational use of probabilistic forecasting systems and the optimisation of decision-making. This paper is a contribution to improve our understanding of the way the benefits of probabilistic forecasts are perceived by the decision-makers. It proposes to investigate it from a different perspective, by allowing, through a game experiment, decision-makers to bid for a probabilistic forecast set during an auction. The auction was used in this paper as an attempt to characterise and understand the participants' WTP for a probabilistic forecast in the specific flood protection risk-based experiment designed for this purpose. Our results indicate that the WTP displays dependencies on various aspects.

The bids were to a certain extent influenced by the participants' economic situation. They were on average positively related to the money available to participants during the auction. Nonetheless, this was mainly a factor for participants who had little money left in their purses at the time of the auction. The participants' perceived forecast quality was also a factor influencing their WTP for another forecast set. Players who had played the first round with biased forecasts were less prone to disburse money for another forecast set for the second round. There was moreover an indication that the flood frequency of the river might have influenced the WTP for a forecast set. Some players in charge of a river with only one flood event in the first round (i.e. low flood risk) did not consider beneficial the purchase of a forecast set for the second round. The participants' risk perception was therefore an important element of their WTP for a probabilistic forecast. The more risk-averse participants did not buy a second forecast set as they had enough money to protect all the time; "gambling" was also stated as a reason for not buying a sec-

ond forecast set. Seifert et al. (2013) have similarly shown that “the demand for flood insurance is strongly positively related to individual risk perceptions”.

These results show that the perceived benefit of probabilistic forecasts as a support of decision-making in a risk-based context is multifaceted, and varies not only with the quality of the information and its understanding, but also with the relevance and the risk tolerance of the user. This further demonstrates that more work is needed not solely to provide guidance on the use of probabilistic information for decision-making, but also to develop efficient ways to communicate the actual relevance and evaluate the long-term economic benefits of probabilistic forecasts for improved decisions in various applications of probabilistic forecasting systems within the water sector. This could additionally provide insights into bridging the gap between the theoretical or expected benefit of probabilistic forecasts in a risk-based decision-making environment and the perceived benefits by key users.

4.2 Game limitations and further developments

This paper aimed to depict behaviours in the flood forecasting and protection decision-making context. Although game experiments offer a flexible assessment framework, compared to real operational configurations, it is however extremely complex to search for general explanatory behaviours in such a context. This is partially due to the uniqueness of individuals and the interrelated factors that might influence decisions, which are both aspects that are difficult to evaluate when playing a game with a large audience. A solution to overcome this, as proposed by Crochemore et al. (2015), could be to prolong the game by incorporating a discussion with the audience or with selected individuals, aiming at understanding the motivations hidden underneath their decisions during the game. Having more time available to apply the game would also allow one to play more cases in each round, bringing additional information to the analysis and clarifying key aspects of the game, such as the effect of the bias on the participants’ use of the forecasts and on their WTP for more forecasts. Co-designing such an experiment with social anthropologists could bring to light many more insights into participants’ decision-making behaviours.

Being set up as a game, this study also presents some limitations. As mentioned by Breidert et al. (2006), a source of bias in such studies is their artificial set-up. Indeed, under those circumstances, participants are not directly affected by their decisions, as they neither use their own money nor is the risk a real one. This might lead them to make decisions which they would normally not make in real life or in operational forecasting contexts.

Moreover, in our game, the costs given to both flood protection and flood damages were not chosen to represent the real costs that one encounters in real environments. First, real costs in integrated flood forecasting and protection sys-

tems are difficult to assess, given the complexity of flood protection and its consequences. Secondly, the external imposed conditions for playing our game (i.e. the fact that we wanted to play it during oral talks in conferences, workshops or teaching classes, with expected eclectic audiences of variable sizes, having a limited amount of time, and using paper worksheets to be collected at the end of the game for the analysis) were not ideal to handle any controversy on the realism (or absence of realism) of the game scenario.

It is however arguable whether the game results could be a reflection of the experiment set-up, and hence of the parameters of the game (the protection and damage costs, the number of flood events, etc.). For instance, the higher damage costs might have influenced the participants’ tolerance to misses and false alarms. Further developments could include testing the influence of the parameters of this experiment on its results as a means of analysing the sensitivity of flood protection mitigation to a specific decision-making setting.

Additionally, the small sample size of this experiment limited the statistical significance of its results. Replicating it could ascertain some of the key points discussed, leading to more substantial conclusions, and improve our understanding of the effect of the professional background of the participants on their decisions.

Finally, the experiment’s complex structure was its strength as well as its weakness. When analysing the game results, the chicken and egg situation arose. Several factors of the participants’ use of the forecasts and of their WTP for a forecast set were identified, but it was not possible to measure causalities. It would therefore be interesting to carry out further work in this direction, together with behavioural psychologists, by, for instance, testing the established factors separately.

5 Conclusions

This paper presented the results of a risk-based decision-making game, called “How much are you prepared to pay for a forecast?”, played at several workshops and conferences in 2015. It was designed to contribute to the understanding of the role of probabilistic forecasts in decision-making processes and their perceived value by decision-makers for flood protection mitigation.

There were hints that participants’ decisions to protect (or not) against floods were made based on the probabilistic forecasts and that the forecast median alone did not account for all the decisions made. Where participants were presented with biased forecasts, they adjusted the manner in which they were using the information, with an overall lower tolerance for misses than for false alarms. Participants with unbiased forecasts also showed inconsistent decisions, which appeared to be shaped by their risk perception; the initial river level and the proximity of the forecast median to the flood threshold both led the participants to base their decisions on ex-

tremes of the forecast distribution or on no apparent information contained in the forecast.

The participants' willingness-to-pay for a probabilistic forecast, in a second round of the game, was furthermore influenced by their economic situation, their perception of the forecasts' quality and the river flood frequency.

Overall, participants had an accurate perception of their decision-making performance, which they related to the quality of their forecasts. However, there appeared to be difficulties in the estimation of the added value of the probabilistic forecasts for decision-making, thus leading the participants who bought a second forecast set to end the game with a lower amount of money in hand.

The use and perceived benefit of probabilistic forecasts as a support of decision-making in a risk-based context is a complex topic. The paper has shown the factors that need to be considered when providing guidance on the use of probabilistic information for decision-making and developing efficient ways to communicate their actual relevance for improved decisions for various applications. Games such as this one are useful tools for better understanding and discussing decision-making among forecasters and stakeholders, as well as highlighting potential factors that influence decision-makers and that deserve further research.

6 Resources

This version of the game is licensed under CC BY-SA 4.0 (Creative Commons public license). It is part of the activities of HEPEX (Hydrologic Ensemble Prediction Experiment) and is freely available at www.hepex.org. This game was inspired by the Red Cross/Red Crescent Climate Centre game "Paying for Predictions" (<http://www.climatecentre.org/resources-games/paying-for-predictions>).

Appendix A: Example of a worksheet distributed to the game participants (here for river blue and the set 1 of positively biased forecasts: BLUE-1)

BLUE-1



How much are you prepared to PAY for a forecast?

Occupation (student, PhD candidate, scientist, operational hydrologist, forecaster, professor, lecturer, other):

.....

How many years of experience do you have? < 5 years 5 to 10 years > 10 years

Flood protection = -2,000 tokens; flood without protection = -4,000 tokens
 Flood occurs at 90 or above

Round	Case	River level before rainfall (10-60)	Flood protection?	River level increment (10-80)	River level after increment	Flood? (≥ 90)	Tokens spent	Purse (20,000)
1	1		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	2		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	3		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	4		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	5		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		

- How was your forecast set in Round 1?
 - very bad quite bad neither good nor bad quite good very good I don't know
- How was your performance as a decision-maker in Round 1?
 - very bad quite bad neither good nor bad quite good very good I don't know



Do not forget to transfer your final Round 1 purse to Round 2 (in the brackets under ‘Purse’)

Round	Case	River level before rainfall (10-60)	Flood protection?	River level increment (10-80)	River level after increment	Flood? (≥ 90)	Tokens spent	Purse (.....)
<i>Bid: tokens. Did you buy a probabilistic forecast set? YES / NO</i> <i>If yes, deduct the money you paid for it here:</i>								
2	1		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	2		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	3		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	4		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		
	5		Yes <input type="checkbox"/> No <input type="checkbox"/>			Yes <input type="checkbox"/> No <input type="checkbox"/>		

- How was your performance as a decision-maker in Round 2?
 very bad quite bad neither good nor bad quite good very good I don't know
- For the people who DID NOT buy a forecast set:
 - Why didn't you buy a forecast set?
 The model did not seem reliable
 The price was too high
 Other reason (explain):
- For the people who DID buy a forecast set:
 - How was your forecast set in Round 2?
 very bad quite bad neither good nor bad quite good very good I don't know
 - Were the forecasts worth what you paid for them? Yes No
 - If not, how many tokens would you now pay for them? tokens

Please return this worksheet into the envelope and give it to one of the assistants before you leave.
 Thank you for your participation!
 We hope you enjoyed it!

The Supplement related to this article is available online at doi:10.5194/hess-20-1-2016-supplement.

Acknowledgements. The authors gratefully acknowledge financial support from the Horizon 2020 IMPREX project (grant agreement no. 641811) (project IMPREX: www.imprex.eu). The authors would like to thank the participants of the game who very enthusiastically took part in this experiment. Furthermore, we would like to acknowledge L. Crochemore, A. Ficchi, C. Poncelet and P. Brigode for their valuable help with the game preparation and worksheet distribution at EGU 2015. Finally, we would like to thank C. Bachofen and everyone who tested and gave suggestions to improve the game during its development.

Edited by: S. Illingworth

References

- Anaman, K. A., Lelleyett, S. C., Drake, L., Leigh, R. J., Henderson-Sellers, A., Noar, P. F., Sullivan, P. J., and Thampapillai, D. J.: Benefits of meteorological services: evidence from recent research in Australia, *Meteorol. Appl.*, 5, 103–115, doi:10.1017/S1350482798000668, 1998.
- Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, *Nature*, 525, 47–55, doi:10.1038/nature14956, 2015.
- Boucher, M. A., Tremblay, D., Delorme, L., Perreault, L., and Anctil, F.: Hydro-economic assessment of hydrological forecasting systems, *J. Hydrol.*, 416–417, 133–144, doi:10.1016/j.jhydrol.2011.11.042, 2012.
- Breidert, C., Hahsler, M., and Reutterer, T.: A review of methods for measuring willingness-to-pay, *Innovative Marketing*, 2, 8–32, 2006.
- Buizza, R.: The value of probabilistic prediction, *Atmos. Sci. Lett.*, 9, 36–42, doi:10.1002/asl.170, 2008.
- Changnon, S. A., Pielke, R. A., Changnon, D., Sylves, R. T., and Pulwarty, R.: Human Factors Explain the Increased Losses from Weather and Climate Extremes*, *B. Am. Meteorol. Soc.*, 81, 437–442, 2000.
- Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: a review, *J. Hydrol.*, 375, 613–626, 2009.
- Crochemore, L., Ramos, M. H., Pappenberger, F., van Andel, S. J., and Wood, A. W.: An experiment on risk-based decision-making in water management using monthly probabilistic forecasts, *B. Am. Meteorol. Soc.*, 97, 541–551, doi:10.1175/BAMS-D-14-00270.1, 2015.
- Dale, M., Wicks, J., Mylne, K., Pappenberger, F., Laeger, S., and Taylor, S.: Probabilistic flood forecasting and decision-making: an innovative risk-based approach, *Nat. Hazards*, 70.1, 159–172, 2014.
- Mendler de Suarez, J., Suarez, P., Bachofen, C., Fortugno, N., Goentzel, J., Gonçalves, P., Grist, N., Macklin, C., Pfeifer, K., Schweizer, S., Van Aalst, M., and Virji, H.: Games for a New Climate: Experiencing the Complexity of Future Risks, Pardee Center Task Force Report, The Frederick S. Pardee Center for the Study of the Longer-Range Future, Boston University, Boston, 2012.
- Demeritt, D., Cloke, H. L., Pappenberger, F., Thielen, J., Bartholmes, J., and Ramos, M. H.: Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting, *Environmental Hazards*, 7, 115–127, 2007.
- Ferrell, J.: The Secrets of Weather Forecast Models, Exposed, available at: <http://www.accuweather.com/en/weather-blogs/weathermatrix/why-are-the-models-so-inaccurate/18097>, last access: 10 June 2016, 2009.
- García-Morales, M. B. and Dubus, L.: Forecasting precipitation for hydroelectric power management: how to exploit GCM's seasonal ensemble forecasts, *Int. J. Climatol.*, 27, 1691–1705, doi:10.1002/joc.1608, 2007.
- Handmer, J. and Proudley, B.: Communicating uncertainty via probabilities: The case of weather forecasts, *Environmental Hazards*, 7, 79–87, 2007.
- Hashino, T., Bradley, A. A., and Schwartz, S. S.: Evaluation of bias-correction methods for ensemble streamflow volume forecasts, *Hydrol. Earth Syst. Sci.*, 11, 939–950, doi:10.5194/hess-11-939-2007, 2007.
- He, Y., Wetterhall, F., Cloke, H. L., Pappenberger, F., Wilson, M., Freer, J., and McGregor, G.: Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions, *Meteorol. Appl.*, 16, 91–101, doi:10.1002/met.132, 2009.
- Kirchhoff, C. J., Lemos, M. C., and Engle, N. L.: What influences climate information use in water management? The role of boundary organizations and governance regimes in Brazil and the U.S., *Environ. Sci. Policy*, 26, 6–18, doi:10.1016/j.envsci.2012.07.001, 2013.
- Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, *J. Hydrol.*, 249, 2–9, doi:10.1016/S0022-1694(01)00420-6, 2001.
- LeClerc, J. and Joslyn, S.: The Cry Wolf Effect and Weather-Related Decision Making, *Risk Anal.*, 35, 385–395, 2015.
- Leviäkangas, P.: Valuing meteorological information, *Meteorol. Appl.*, 16, 315–323, doi:10.1002/met.122, 2009.
- Magnusson, L. and Källén, E.: Factors Influencing Skill Improvements in the ECMWF Forecasting System, *Mon. Weather Rev.*, 141, 3142–3153, doi:10.1175/MWR-D-12-00318.1, 2013.
- Meissner, D. and Klein, B.: The added value of probabilistic forecasts for navigation, available at: <http://hepex.irstea.fr/the-added-value-of-probabilistic-forecasts-for-navigation-2/>, last access: 10 June 2016, 2013.
- Michaels, S.: Probabilistic forecasting and the reshaping of flood risk management, *Journal of Natural Resources Policy Research*, 7, 41–51, doi:10.1080/19390459.2014.970800, 2015.
- New, M., Lopez, A., Dessai, S., and Wilby, R.: Challenges in using probabilistic climate change information for impact assessments: an example from the water sector, *Philos. T. Roy. Soc. A*, 365, 2117–2131, doi:10.1098/rsta.2007.2080, 2007.
- Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S., and Thielen, J.: The monetary benefit of early flood warnings in Europe, *Environ. Sci. Policy*, 51, 278–291, doi:10.1016/j.envsci.2015.04.016, 2015.
- Pitt, M.: The Pitt Review: Lessons learned from the 2007 floods, Cabinet Office (archived by the National Archives), June 2008.
- Ramos, M. H., van Andel, S. J., and Pappenberger, F.: Do probabilistic forecasts lead to better decisions?, *Hydrol. Earth Syst. Sci.*, 17, 2219–2232, doi:10.5194/hess-17-2219-2013, 2013.

- Ramos, M. R., Mathevet, T., Thielen, J., and Pappenberger, F.: Communicating uncertainty in hydro-meteorological forecasts: mission impossible?, *Meteorol. Appl.*, 17, 223–235, doi:10.1002/met.202, 2010.
- Red Cross/Red Crescent Climate Centre: Game “Paying for Predictions”, International Federation of Red Cross and Red Crescent Societies (IFRC), The Netherlands, available at: <http://www.climatecentre.org/resources-games/paying-for-predictions>, last access: June 2016.
- Rollins, K. S. and Shaykewich, J.: Using willingness-to-pay to assess the economic value of weather forecasts for multiple commercial sectors, *Meteorol. Appl.*, 10, 31–38, doi:10.1017/S1350482703005048, 2003.
- Seifert, I., Botzen, W. J. W., Kreibich, H., and Aerts, J. C. J. H.: Influence of flood risk characteristics on flood insurance demand: a comparison between Germany and the Netherlands, *Nat. Hazards Earth Syst. Sci.*, 13, 1691–1705, doi:10.5194/nhess-13-1691-2013, 2013.
- Simmons, A. J. and Hollingsworth, A.: Some aspects of the improvement in skill of numerical weather prediction, *Q. J. Roy. Meteor. Soc.*, 128, 647–677, doi:10.1256/003590002321042135, 2002.
- Stephens, E. and Cloke, H. L.: Improving flood forecasts for better flood preparedness in the UK (and beyond), *Geogr. J.*, 180, 310–316, 2014.
- Verkade, J. S. and Werner, M. G. F.: Estimating the benefits of single value and probability forecasting for flood warning, *Hydrol. Earth Syst. Sci.*, 15, 3751–3765, doi:10.5194/hess-15-3751-2011, 2011.
- Wetherald, R. T. and Manabe, S.: Simulation of hydrologic changes associated with global warming, *J. Geophys. Res.*, 107, 4379, doi:10.1029/2001JD001195, 2002.
- WMO: Manual on Flood Forecasting and Warning, World Meteorological Organisation, Geneva, Switzerland, Unesco, WMO No. 1072, 142 pp., 2011.

A2: Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity

This paper presents a co-author contribution arising through collaboration during this PhD, summarised in Chapter 2, Sect. 2.4, and has the following reference:

Neumann, J. L., L. Arnal, R. E. Emerton, H. Griffith, S. Hyslop, S. Theofanidi and H. L. Cloke, 2018b: Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity, *Geosci. Commun.*, 1, 35-57, doi:10.5194/gc-1-35-2018*

* ©2018. The Authors. Geoscience Communication, a journal of the European Geosciences Union published by Copernicus. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.



Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity

Jessica L. Neumann¹, Louise Arnal^{1,2}, Rebecca E. Emerton^{1,2}, Helen Griffith¹, Stuart Hyslop³, Sofia Theofanidi¹, and Hannah L. Cloke^{1,4,5}

¹Department of Geography and Environmental Science, University of Reading, Reading, UK

²European Centre for Medium-Range Weather Forecasts (ECWMF), Reading, UK

³Environment Agency, Kings Meadow House, Reading, UK

⁴Department of Meteorology, University of Reading, Reading, UK

⁵Department of Earth Sciences, Uppsala, Sweden

Correspondence: Jessica L. Neumann (j.l.neumann@reading.ac.uk)

Received: 17 July 2018 – Discussion started: 25 July 2018

Revised: 23 October 2018 – Accepted: 26 October 2018 – Published: 6 December 2018

Abstract. While this paper has a hydrological focus (a glossary of terms highlighted by asterisks in the text is included in Appendix A), the concept of our decision-making activity will be of wider interest and applicable to those involved in all aspects of geoscience communication.

Seasonal hydrological forecasts (SHF) provide insight into the river and groundwater levels that might be expected over the coming months. This is valuable for informing future flood or drought risk and water availability, yet studies investigating how SHF are used for decision-making are limited. Our activity was designed to capture how different water sector users, broadly flood and drought forecasters, water resource managers, and groundwater hydrologists, interpret and act on SHF to inform decisions in the West Thames, UK. Using a combination of operational and hypothetical forecasts, participants were provided with three sets of progressively confident and locally tailored SHF for a flood event in 3 months' time. Participants played with their “day-job” hat on and were not informed whether the SHF represented a flood, drought, or business-as-usual scenario. Participants increased their decision/action choice in response to more confident and locally tailored forecasts. Forecasters and groundwater hydrologists were most likely to request further information about the situation, inform other organizations, and implement actions for preparedness. Water resource managers more consistently adopted a “watch and wait” approach. Local knowledge, risk appetite, and experience of previous flood events were important for inform-

ing decisions. Discussions highlighted that forecast uncertainty does not necessarily pose a barrier to use, but SHF need to be presented at a finer spatial resolution to aid local decision-making. SHF information that is visualized using combinations of maps, text, hydrographs, and tables is beneficial for interpretation, and better communication of SHF that are tailored to different user groups is needed. Decision-making activities are a great way of creating realistic scenarios that participants can identify with whilst allowing the activity creators to observe different thought processes. In this case, participants stated that the activity complemented their everyday work, introduced them to ongoing scientific developments, and enhanced their understanding of how different organizations are engaging with and using SHF to aid decision-making across the West Thames.

1 Introduction

There has been a recent shift away from the conventional linear model of science, where research is carried out within the scientific community with the expectation that users will be able to access and apply the information, towards co-production and stakeholder-led initiatives that bring together scientists and decision-makers to frame and deliver “actionable research” (Asrar et al., 2012; Lemos et al., 2012; Meadow et al., 2015). Regular and clear communication between scientists and policy-makers and practitioners in

workshops, focus groups, consultations, and interviews, and through the development of games, activities, and interactive media, is imperative for ensuring that projects deliver impact outside of the academic environment. Here, we share findings from an activity that explored the use of seasonal hydrological forecasts* for local decision-making. This was conducted as part of an IMPREX (IMproving PRedictions and management of hydrological Extremes) stakeholder focus group for the West Thames, UK (van den Hurk et al., 2016; IMPREX, 2018a), co-organized by the University of Reading (UoR), UK, Environment Agency (EA) and supported by the European Centre for Medium-Range Weather Forecasts (ECMWF).

Seasonal hydrological forecasts (SHF) have the ability to predict principal changes in the hydrological environment such as river flows and groundwater levels weeks or months in advance. This has the potential to benefit humanitarian action and economic decision-making, e.g. to provide early warning of potential flood and drought events, assist with water quality monitoring, and ensure optimal management and use of water resources for public water supply, agriculture, and industry (Chiew et al., 2003; Arnal et al., 2017; Li et al., 2017; Meißner et al., 2017; Turner et al., 2017). SHF systems covering a range of spatial scales have been developed – Hydrological Outlook UK forecasts at a national level (Prudhomme et al., 2017; CEH, 2018) – while the Copernicus European and Global Flood Awareness Systems (EFAS and GloFAS) provide operational forecasts over larger scales (JRC, 2018a, b). Recent research has demonstrated improvements in SHF quality*, including increased accuracy out to 4 months for high-flow events during the winter in Europe (Arnal et al., 2018; Emerton et al., 2018).

There is growing interest in SHF amongst policy-makers and practitioners; however, in many cases, there is limited information about whether SHF products are *actually* being used. Research output has focused largely on technical system development and improvements to forecast skill* (see the review by Yuan et al., 2015), with relatively fewer studies exploring how users engage with and apply SHF to inform decisions (see Crochemore et al., 2015; Viel et al., 2016). Many seasonal forecasting studies, including those investigating the application of seasonal meteorological forecasts* (which provide information about future weather variables, rather than hydrology more specifically), have identified forecast uncertainty*, whereby forecast skill and sharpness* decrease with increasing lead time* (Wood and Lettenmaier, 2008; Soares and Dessai, 2015), and how this uncertainty can be communicated effectively as key barriers to use (Arnal et al., 2016; Vaughan et al., 2016). Non-technical factors, including the level of knowledge and training required to interpret and apply SHF information effectively (Bolson et al., 2013; Soares and Dessai, 2016), the visualization, format, and compatibility of the information provided (Fry et al., 2017; Soares et al., 2018), and the level of communication between different users in the water sector and between

research developers and practitioners (Golding et al., 2017), have all been found to act as both barriers and enablers, depending on the user group in question.

The potential for SHF to meet the needs of the water sector is recognized by a host of UK environmental organizations, including the EA, the Met Office, and research centres (see Prudhomme et al., 2017). The West Thames specifically is underlain by a slowly responding, largely groundwater-driven hydrogeological system (Mackay et al., 2015), meaning that there is potential for extreme hydrological events such as the drought of 2010–2012 (Bell et al., 2013) and winter floods of 2013–2014 (Neumann et al., 2018) to be detected weeks or months in advance. It also has a dense population and high demands for water which require effective long-term management of resources for public drinking supply, industry, agriculture, and wastewater treatment (further details about the West Thames can be found in Sect. 2.2). The value of using SHF in the West Thames is of particular interest to the EA; however, information on the level of understanding, uptake, and application is currently unknown. We therefore aimed to develop a clearer understanding about how different professional water sector users – broadly forecasters, groundwater hydrologists, and water resource managers – are currently engaging with SHF in the West Thames using a decision-making activity.

In the context of flood science communication with experts, real-time activities such as simulation exercises (that imitate real-world processes and behaviours) or roleplay (where participants engage with real-world scenarios but take on personas and positionalities that differ from their own) are known to be effective when engaging with stakeholders who bring a range of scientific ideas and perspectives to the table (McEwen et al., 2014). Such activities encourage participants to apply their knowledge to realistic situations and to reflect on issues and the perspectives of other stakeholders (Pavey and Donoghue, 2003, p. 7). They are also valuable for understanding decision-making processes, e.g. for environmental hazards and conflicting community views (Harrison, 2002), for capacity building in response to new water legislation (Farolfi et al., 2004), and for understanding climate forecasts and decision-making (Ishikawa et al., 2011). Our decision-making activity provided an interactive and entertaining platform that encouraged participants to engage with real-world scenarios whilst fostering discussions about the barriers and enablers to use of SHF. Using three activity stages, participants were provided with sets of progressively confident and locally tailored SHF for the next 3 to 4 months. The SHF were produced using output from operational systems including Hydrological Outlook UK and the European Flood Awareness System (EFAS), and hypothetical forecasts generated through scientific research (see Neumann et al., 2018). Participants were asked to play in real time, i.e. as if receiving the forecasts on the day for the next 3 to 4 months. They did not know in advance whether the SHF represented a flood, drought, or business-as-usual scenario

and had to use their knowledge and experiences to make informed decisions based on the maps, hydrographs*, tables, and text provided. In reality, all three sets of SHF represented the same time period: winter 2013–2014 (a period of extensive flooding nationwide that occurred at the end of 2 years of drought conditions in the UK). Between December 2013 and February 2014 the West Thames experienced extreme flooding from fluvial and groundwater sources which had knock-on impacts for local water quality, sewage treatment, and water resource management – opening up discussions for all participants. Given that issues relating to flood and drought risk, water quality, and water resource management in the West Thames are generally managed by local and regional-area authorities (Thames Water, 2010), the activity focused on whether SHF can be used to support decision-making at the local level. To the best of our knowledge, this scale of practical application has yet to be explored, we suspect mainly due to the lower skill of seasonal meteorological forecasts in Europe, particularly with respect to precipitation, which is a key variable of interest for hydrology (Arribas et al., 2010; Doblas-Reyes et al., 2013). A brief overview of the focus group is provided in Sect. 2, the full activity set-up is detailed in Sect. 3, and the findings and the discussion are presented in Sects. 4 and 5.

2 Overview of the focus group

2.1 Aims of the focus group

The focus group was developed in collaboration with the EA and in line with the objectives of the IMPREX project. The aims were the following.

- Introduce and discuss current SHF projects, products, and initiatives for the UK and Europe.
- Engage with participants' experiences and knowledge of using SHF.
- Learn how SHF are being applied in the West Thames and recognize how different users in the water sector approach and apply SHF information for decision-making.
- Identify limitations and barriers to use.
- Identify future opportunities for SHF application and research.

These aims were delivered through a series of four interactive sessions designed to actively engage participants to share their knowledge and experiences of SHF, and short presentations that introduced the main topics surrounding SHF and informed participants about current SHF projects and developments in the scientific research. While this paper focuses on the decision-making activity (interactive session 2), discussions from the other sessions are also presented where

relevant. An outline of the focus group programme is provided in Supplement 1 and a full report of the activities is available; see Neumann et al. (2017).

2.2 The West Thames in southern England

2.2.1 Physical geography

The West Thames refers to the non-tidal portion of the Thames River Basin*, from its source in the Cotswolds in the west of England to 230 km downstream at Teddington Lock in western London (Fig. 1). It covers an area of 9857 km² (the Thames basin is 16 980 km²) and comprises 10 river catchments* that are the tributaries* that feed directly into the River Thames (Fig. 1). The western catchments are predominantly rural; land use is a mix of agriculture and woodland with rolling hills and wide, flat floodplains (elevation up to 350 m a.s.l.). Towards the centre and east, the region becomes increasingly urbanized, encompassing the towns of Reading and Slough and outskirts of Greater London (elevation 4 m a.s.l. at Teddington Lock). Lithology* varies markedly across the West Thames. Catchments overlaying the Cotswolds (upstream) and the Chilterns (middle sections) are dominated by chalk and limestone aquifers* with high baseflow*, while a band of less-permeable clays and mudstones separates these two areas. Sandstones, mudstones, and clays are also prevalent towards London (downstream) – these catchments have higher levels of surface runoff* and can exhibit a flashier* response to storm events (Bloomfield et al., 2011; EA, 2009).

2.2.2 Water demands, risk, and management – why the West Thames is of interest

The West Thames is a highly pressured environment – 15 million people and a substantial part of the UK's economy rely directly on its water supply (EA, 2015). There are more than 2000 licensed abstraction points in the chalk aquifers and superficial alluvium and river terrace gravel deposits; 90 % of abstractions are for public water supply, the rest providing water for agriculture, aquaculture, and industry (Thames Water, 2010). There are 12 000 registered wastewater discharge points; pollution from sewage treatment works, transport, and urban areas affects more than 45 % of rivers, water bodies, and aquifers, largely towards London. Diffuse pollution and sedimentation from agricultural and forestry practice are the main contributors to poor water quality in the upper catchments, especially during times of high rainfall (EA, 2015).

Urbanization and land-use change in combination with more varied rainfall patterns have seen the region affected by a number of extreme drought and flood events in recent years (EA, 2009; Parry et al., 2015; Muchan et al., 2015). Across the Thames Basin, 200 000 properties are at risk from a 1 : 100*-year fluvial flood, with 10 000 at risk from a 1 : 5*-year event (EA, 2009). Low and high river flows also pose



Figure 1. Location and lithology of the West Thames and its 10 main river catchments.

risks to navigation and management of the canal network which is highly important for recreation, local living, and the economy (Wells and Davis, 2016).

2.3 Participants

2.3.1 Who took part?

SHF have the potential for wide-ranging application and it was important to capture the different perspectives of the West Thames water sector. The organizers agreed that the focus group would work well with a relatively small number of participants (up to 12) so that all perspectives could be heard. Based on discussions held between the organizers, individuals from local organizations working in established (i.e. long-term/permanent/leadership) roles relevant to SHF in the West Thames were invited; many but not all participants had previously collaborated with the University of Reading and/or EA. In some cases, an invitee was unable to attend due to prior commitments or because they had a colleague who they felt would be a better fit for the focus group. A total of 17 participants were invited from six organizations – 12 accepted and 11 took part on the day. They were responsible for flood and drought forecasting ($F \times 3$), groundwater modelling and hydrogeology ($GH \times 2$), navigation ($N \times 1$), water resource and reservoir management ($WR \times 2$), public water supply ($WS \times 2$), and wastewater modelling and operations ($WW \times 1$). They represented five organizations: two non-departmental public bodies (sponsored by government agencies), two science and research centres, one water service company, and one non-for-profit organization (Table 1).

2.3.2 Current engagement with SHF

By inviting local stakeholders we ensured that participants represented a range of different water sector personas and were familiar with the West Thames environment. We did not assume that participants had any prior knowledge of SHF and invitees were encouraged to attend even if they were unfamiliar with the concept as this would be an important indicator of the state of play in the West Thames (invite poster; see Supplement 1).

All 11 focus group participants were familiar with the concept of seasonal hydrological forecasting and 10 regularly used SHF in their everyday job (according to results from interactive session 1 – “What are seasonal hydrological forecasts?”). Using post-its, participants noted that Hydrological Outlook UK (CEH, 2018) and the associated raw forecasts from the analogue, hydrological, and meteorological models (produced by the UK Met Office, Centre for Ecology and Hydrology, British Geological Survey, EA, Natural Resources Wales, Scottish Environment Protection Agency, and Rivers Agency Northern Ireland) were the main sources of SHF information currently being used, primarily for flood and drought outlook, groundwater monitoring, and river flow projection purposes. Scientific research, operational planning, and sharing of information with other organizations in the water sector were also listed as reasons for engaging with SHF. It is important to note that no prior definitions or information were provided and no restrictions or guidance were placed on what participants should write down. This suggests that many in the water sector are using SHF to obtain an insight into whether the upcoming season will be drier or wetter than normal, but that they also believe SHF *potentially*

Table 1. Breakdown of participants who took part in the activity.

Job title	Organization type	Role in the activity
Modelling and Forecasting Team Leader	Public body/government agency (1)	Flood and drought forecaster
Chief Hydrometeorologist	Public body/government agency (2)	Flood and drought forecaster
Climate Scientist (Professor)	Science and research centre (1)	Flood and drought forecaster
Thames Water Resources Technical Specialist	Public body/government agency (1)	Groundwater modelling and hydrogeology
Groundwater Research Directorate	Science and research centre (2)	Groundwater modelling and hydrogeology
Principal Hydrologist for Water Management	Not-for-profit (charitable trust)	Navigation
Water Resources, Environment and Business Directorate	Public body/government agency (1)	Water resource and reservoir management
Abstraction and Transfers Analyst	Water service company	Water resource and reservoir management
Water Strategy and Resources Modeller	Water service company	Public water supply
Thames Region Hydrologist	Public body/government agency (1)	Public water supply
Wastewater Modelling Specialist	Water service company	Wastewater modelling and operations

have the capability to forecast possible flood and drought risk, which could be used to support decision-making and provide better preparedness. This is an encouraging starting point, although many participants noted that this potential is not currently being realized due to the uncertainty and coarse spatio-temporal resolution of SHF; e.g. Hydrological Outlook UK forecasts are only published monthly for the main UK river basins.

3 Set-up of the decision-making activity

3.1 Background

Our activity was inspired by the success of previous decision-making activities and games run by the HEPEX (Hydrological Ensemble Prediction EXperiment) community (e.g. Ramos et al., 2013; Crochemore et al., 2015; Arnal et al., 2016). The aim was to better understand how different water sector users in the West Thames interpret and act on SHF by providing them with hydrological context, maps, and forecasts for the region. The activity was designed for the West Thames so that we could capture the relationship between local stakeholders and the environment in which they work.

3.2 Activity design

3.2.1 Overview of the set-up

The set-up of the activity (illustrated in Fig. 2) had the following structure: Choose groups > Define the Objectives > Background Context > Stage 1 > Stage 2 > Stage 3.

Participants divided themselves into three groups based on their area of expertise and where they felt they could best contribute to the discussions. There were three flood and drought “forecasters” and two “groundwater hydrologists”. The remaining participants (navigation, water resource and reservoir management, public water supply and wastewater operations) grouped themselves as “water resource managers”. While the results and discussions focus on these three broad groups, individual perspectives are also included to

capture the variety of water sector personas present. There were also three research facilitators and three note-takers whose role it was to capture and record the key discussion points.

Groups were first provided with background context to the West Thames to set the scene, followed by three sets of progressively confident SHF for the next 3 to 4 months (Stages 1–3). Stage 1 forecasts were from Hydrological Outlook UK, Stage 2 were from EFAS-Seasonal (European Flood Awareness System) and Stage 3 were “improved” output from EFAS-Seasonal (Fig. 2 and Sect. 3.4). Participants were asked to discuss the information presented in their groups and make informed decisions about each of the 10 West Thames catchments (Fig. 1 and Sect. 3.3.2). All groups were provided with exactly the same information and discussion was encouraged. The activity took around 2 h and timings were only loosely controlled.

SHF at all three stages of the activity represented the same time period – dating from 1 November 2013 to 28 February 2014 (or 31 January 2014 for Hydrological Outlook UK, which only extends to 3 months; CEH, 2018). These dates captured a period of severe and widespread river and groundwater flooding in the West Thames (Huntingford et al., 2014; Kendon and McCarthy, 2015; Muchan et al., 2015). *Participants did not know the dates of the forecasts, nor were they informed whether the situation being forecasted was a high flow (flood), low flow (drought) or a business-as-usual scenario.* Dates were removed from all information, and streamflow- and groundwater-level units were removed from the Stage 2 and Stage 3 EFAS hydrographs, although exceedance thresholds were provided for context. The decision to remove units was advised by the EA. The concern was that participants familiar with average and high-flow values for specific catchments would deduce that the SHF must represent the 2013–2014 floods, which would bias their decision-making based on their previous experience and memories. No information on forecast skill or quality was given and participants were asked to treat all information as

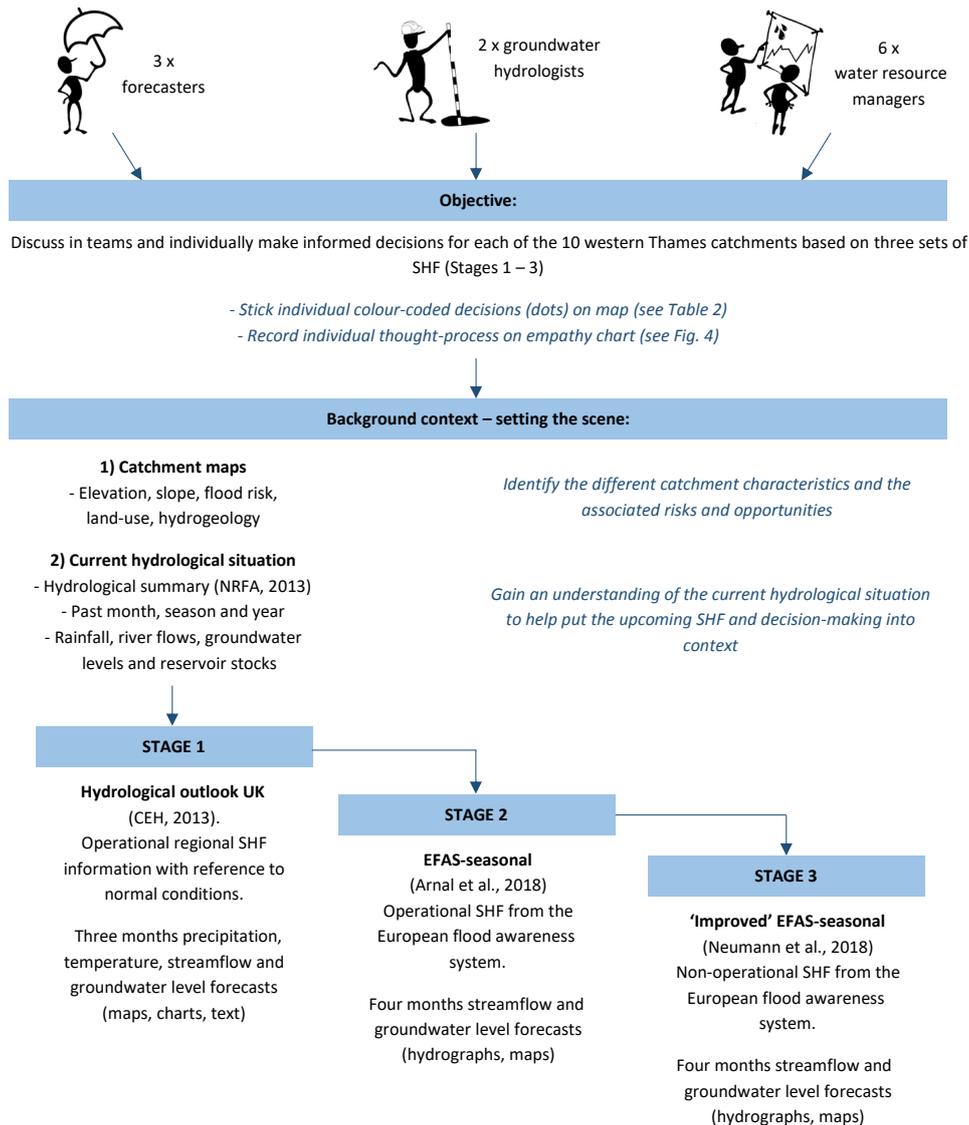


Figure 2. Set-up of the activity.

being “current”, i.e. as if receiving the SHF today, for the next 3–4 months to create a realistic forecasting scenario.

3.2.2 Recording the decisions

In real life, a user’s decision process can encompass a range of possible actions and associated consequences (Crochemore et al., 2015). Decisions can be controlled by providing participants with a set of options to choose from, e.g. to deploy temporary flood defences or not – the consequences of which usually determine the outcome of a game or activity. In this case, participants were asked to select from a broad range of colour-coded options (Table 2), but specific decisions were not defined as these had the potential to differ greatly between participants and might prompt unreal-

istic answers. At each stage, the colour-coded options were discussed by the three groups, simulating conversations that could happen in real life, but it was stressed that *the colour chosen was to be representative of what an individual participant, or their organization, would do with the SHF information in each catchment*. This was recorded on an A1 map using coloured sticky dots marked with the participant’s initials ($n \sim 110$ dots per map (11 participants, 10 catchments)) (Fig. 3). In cases where participants were not familiar with all catchments, or did not feel able to make an informed decision, they did not place a dot. It was important to gather a written record explaining how and why the decisions were reached, and so participants were also asked to complete an A4 empathy map at each stage (Fig. 4). Originally designed as a collaborative tool to be used in business and marketing,

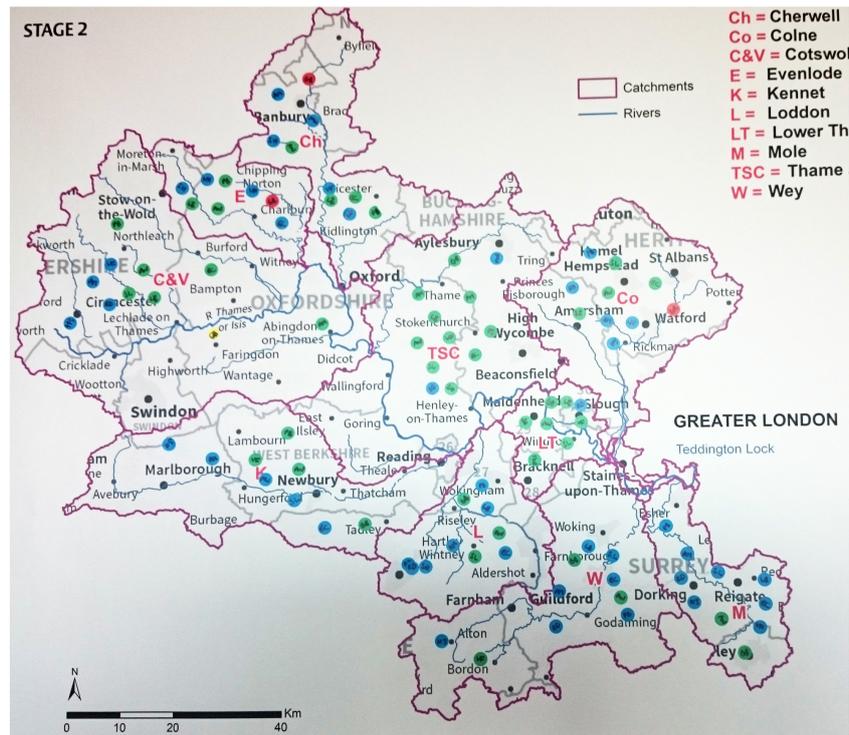


Figure 3. Participants’ individual colour-coded decisions recorded on an A1 map.

Table 2. Colour codes and corresponding action or decision to be taken.

Decision to be made or action to be taken	
●	Ignore the SHF information: wait for the more skilful forecasts with shorter lead times (e.g. a 7–10-day forecast).
●	Look at the SHF information: decide there is no notable risk and do nothing at this point.
●	Look at the SHF information: discuss or pass the information on to relevant colleagues/departments in your organization and agree to keep an eye on the situation.
●	Look at the SHF information: discuss or pass the information on to relevant colleagues/departments in your organization <i>but also</i> external partners – actively request further information about the situation or seek advice on possible actions.
●	Look at the SHF information: decide to implement or set in motion action(s) in a catchment, e.g. to help with drought preparedness, early warning, repairs, or maintenance to flood defences.

empathy maps aim to gain a deeper understanding about an external user’s experiences and decisions (Gray, 2017). Here, we adapted the traditional use by asking individuals to reflect on their own decisions based on their real-life experiences and discussions with other group members. This allowed us to capture individuals’ thought processes, influences, discussions, and the potential risks and gains associated with their decision (Fig. 4). By combining the information recorded on

empathy maps for each group, we also gathered an overview of the shared understanding between forecasters, groundwater hydrologists, and water resource managers and how their SHF needs and expectations match and differ when it comes to decision-making.

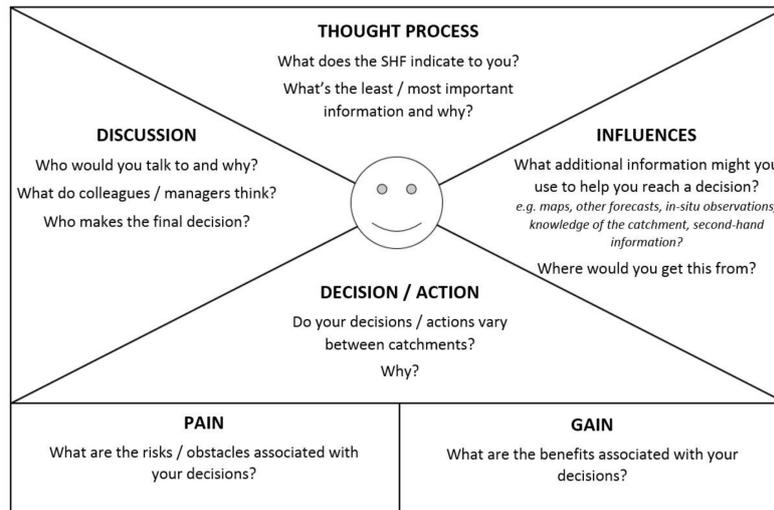


Figure 4. Empathy map completed by each participant during Stages 1–3.

3.3 Background context

Groups were given information about the West Thames catchment characteristics and “current” hydrological conditions (units and dates removed) to place the upcoming SHF into context and aid interpretation.

3.3.1 Catchment characteristics – driving factors, risks and opportunities

Five maps (Supplement 2) that provided a visual representation and a numerical breakdown of the characteristic differences between each catchment were given to participants.

- Hydrogeology* – dominant geological type (sandstone, chalk, clay)
- Elevation – minimum, maximum and mean elevation (m a.s.l.)
- Slope – minimum, maximum and standard deviation of slope angle (degrees)
- Land cover – dominant land use (urban, woodland, agricultural, semi-natural)
- Flood risk – flood warning and flood alert areas and an indication of “urban flood risk”

Participants were asked to discuss and identify the key differences between catchments and highlight the associated risks and opportunities. As some participants were more familiar with specific areas/catchments based on their day job, the maps provided a wider view of where catchment characteristics differ across the West Thames region.

3.3.2 Current hydrological situation

To help set the scene with respect to initial conditions, i.e. the “current” levels of water contained in the soil, groundwater, rivers, and reservoirs, groups were provided with information from the Hydrological Summary (NRFA, 2018) for the last month, past season, and past year (October 2013, June to September 2013, and November 2012 to October 2013 with dates removed). The Hydrological Summary (Supplement 3) focuses on rainfall, river flows, groundwater levels, and reservoir stocks and places the events of each month, and the conditions at the end of the month, into a historical context. In the real world, decision-makers are already prepared with this information; thus, providing evidence about whether hydrological conditions were wet, dry, or normal at the point of receiving the forecasts was an important piece of information for the participants to consider.

3.4 Activity Stages 1–3: the seasonal hydrological forecasts

3.4.1 Stage 1 – Hydrological Outlook UK

The first set of SHF information provided to participants was the Hydrological Outlook UK (from 1 November 2013 to 31 January 2014, with dates removed) (CEH, 2013). This provided regional information for the next 3 months with reference to normal conditions for precipitation, temperature, river flows and groundwater levels. Hydrological Outlook UK uses observations, ensemble models and expert judgement (CEH, 2018) to produce the seasonal forecasts. Information is publicly available and consists of text, graphs, tables and regional maps (examples are shown in Fig. 5 and the full set of forecasts provided to participants are in Supplement 4).

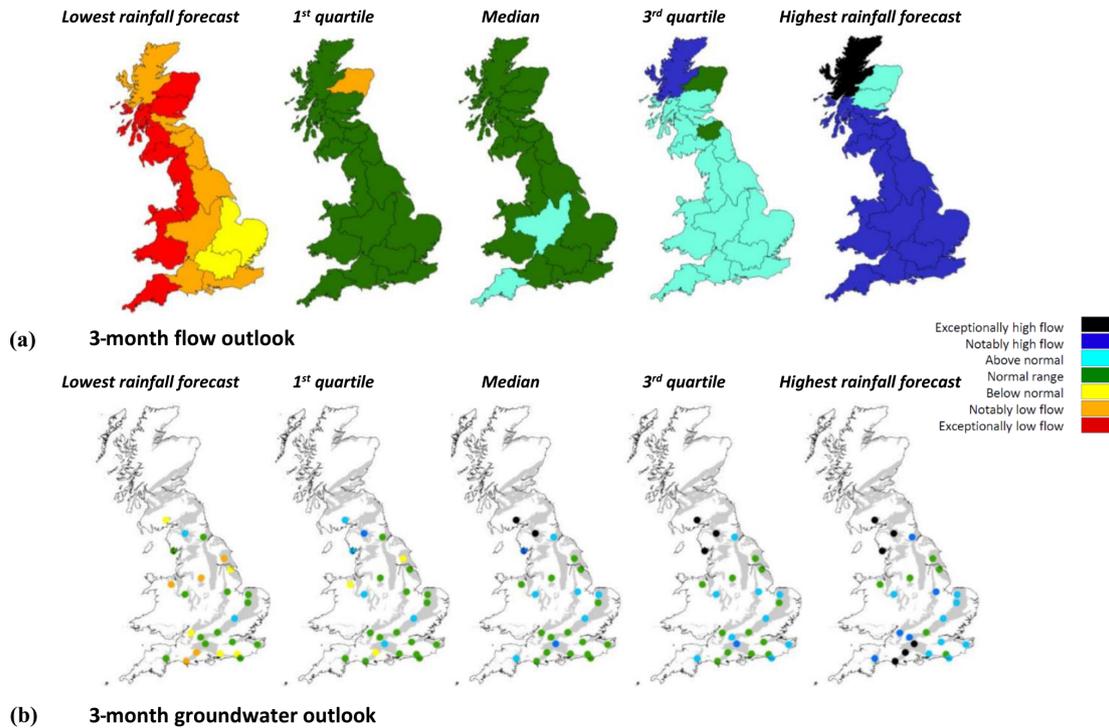


Figure 5. UK 3-month outlook maps from November 2013 (colours based on the percentile range of historical observed values). **(a)** Regional river flow forecasts created from climate forecasts. **(b)** Groundwater level forecasts at 25 UK boreholes created from climate forecasts (CEH, 2013).

3.4.2 Stage 2 – EFAS-Seasonal

EFAS-Seasonal (European Flood Awareness System) is an operational system that monitors and forecasts streamflow* across Europe, with the potential to predict higher than normal streamflow events up to 2 months ahead in an operational capacity, and up to 7 months in practice (JRC, 2018a; Arnal et al., 2018). It runs on a 5 km × 5 km grid and uses the LISFLOOD hydrological model (Van der Knijff et al., 2010; Alfieri et al., 2014). Seasonal ensemble* meteorological forecasts from the ECMWF’s “System 4” operational meteorological forecasting system (Molteni et al., 2011) are used as input to LISFLOOD, from which seasonal ensemble hydrological forecasts are generated on the first day of each month (see Arnal et al., 2018, for details).

For the activity, SHF were produced from 1 November 2013 out to 4 months to focus on the period of extreme stormy weather and flooding experienced. As EFAS-Seasonal is designed to run at the scale of large river basins (i.e. the whole Thames basin), GIS shapefiles were used to extract forecast information for the 10 West Thames catchments using Python v3.5. This provided more locally tailored forecasts compared with Hydrological Outlook UK (Stage 1).

To ascertain whether participants had a preference for how SHF information is presented, the Stage 2 forecasts were presented as both hydrographs and choropleth* maps (Fig. 6).

Ensemble hydrographs for streamflow ($\text{m}^3 \text{s}^{-1}$) and groundwater levels (mm) indicated the predicted trajectory of the hydrological conditions for the next 4 months in each of the 10 catchments (n.b. the greater the spread, the more uncertain the forecast) (Fig. 6a). Units and dates were removed; however, exceedance thresholds*, based on daily observed streamflow and groundwater records between 1994 and 2014 for each of the catchments, were provided for context (EA, 2017; NRFA, 2017). Q50 (median) indicated average streamflow and groundwater conditions for the catchment. Q10 (90th percentile) indicated high streamflow/high groundwater level conditions – 90% of all recorded observations over the previous 20-year period fell below this line.

The choropleth maps showed the maximum probability that the full forecast ensemble for a catchment exceeded the Q10 (90th percentile) threshold in a given month (Fig. 6b), thus providing a snapshot of the probability of potentially extreme conditions at catchment level. The full set of EFAS-Seasonal SHF provided to participants can be found in Supplement 5.

3.4.3 Stage 3 – “Improved” EFAS-Seasonal

Stage 3 followed the exact same set-up and provided the same style output (Fig. 7a, b) as Stage 2 – the only difference being that the seasonal meteorological forecasts used as input to LISFLOOD were taken from a set of atmospheric re-

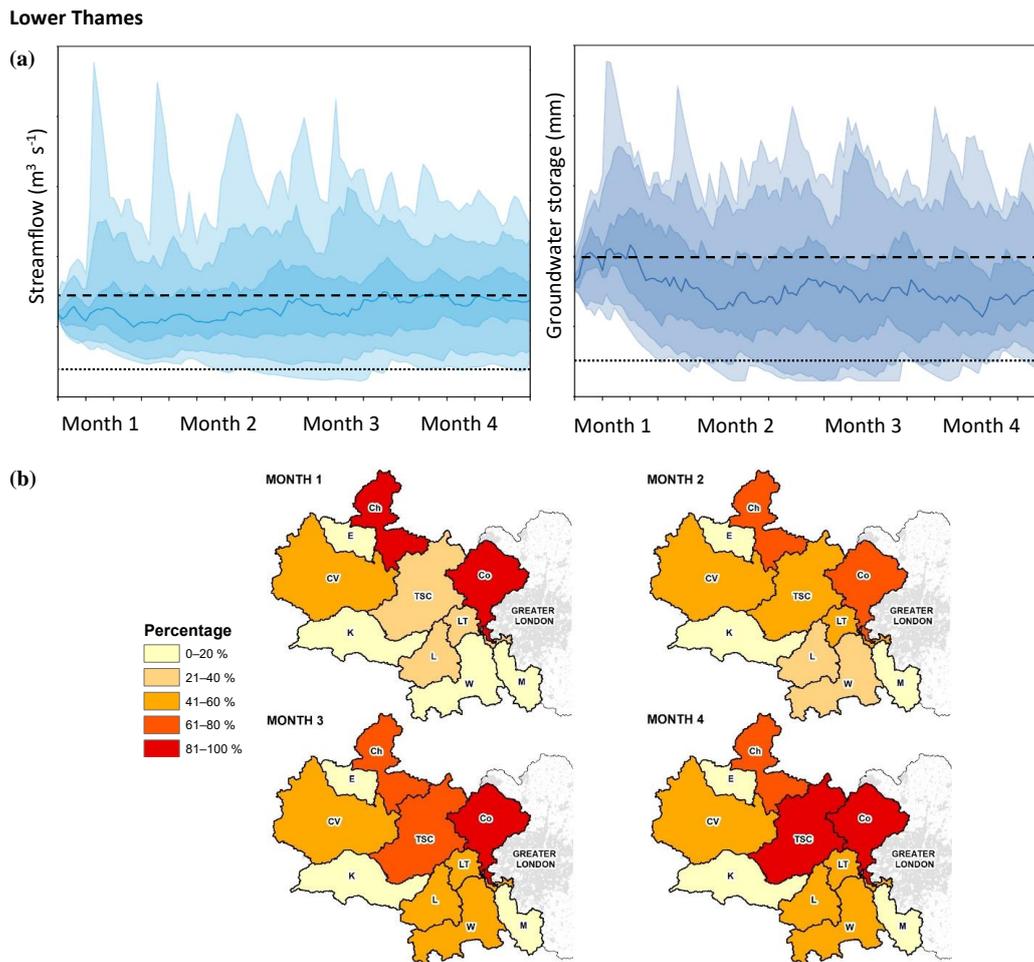


Figure 6. Four-month hydrological forecasts from EFAS-Seasonal (Stage 2). (a) Ensemble hydrographs for streamflow (light blue) and groundwater levels (dark blue) for the Lower Thames (LT) catchment. Exceedance thresholds (based on records from 1994 to 2014) are shown as Q10 (dashed line) and Q50 (dotted line). (b) Choropleth map shows the maximum probability that the full hydrograph ensemble for a catchment exceeds the Q10 streamflow threshold in a given month.

laxation experiments* conducted as part of a scientific study in the West Thames (see Neumann et al., 2018) rather than the operational seasonal meteorological forecasts from “System 4”.

Atmospheric relaxation experiments were conducted by the ECMWF in late 2014 *after* the extreme weather and flooding (Rodwell et al., 2015). The aim was to recreate the atmospheric conditions that prevailed between November 2013 and February 2014, so that the ECMWF could better understand how weather anomalies across the globe contributed to the flooding experienced in the West Thames (Neumann et al., 2018). The SHF at Stage 3 represented near “perfect” forecasts as they were produced *once the floods had happened and the weather conditions were known*. The hydrographs are thus much sharper and more accurate than those presented to the participants at Stage 2 (Fig. 7, Supplement 6). It is important to note that this is not something that

can be achieved by operational systems currently, but does represent the theoretical upper level of forecast skill that may be available to water sector users in the future.

4 Results

4.1 Background context

4.1.1 Catchment differences – “hydrogeology is the driving factor of risks and opportunities”

All groups recognized spatial variability between the catchments and general consensus was that hydrogeology was the most important factor determining flood risk, drought risk, and water availability in the West Thames (Supplement 2). All groups were interested in the persistence, hydrological memory, and slower response of the groundwater-driven catchments upstream (e.g. the Evenlode, Thames, and

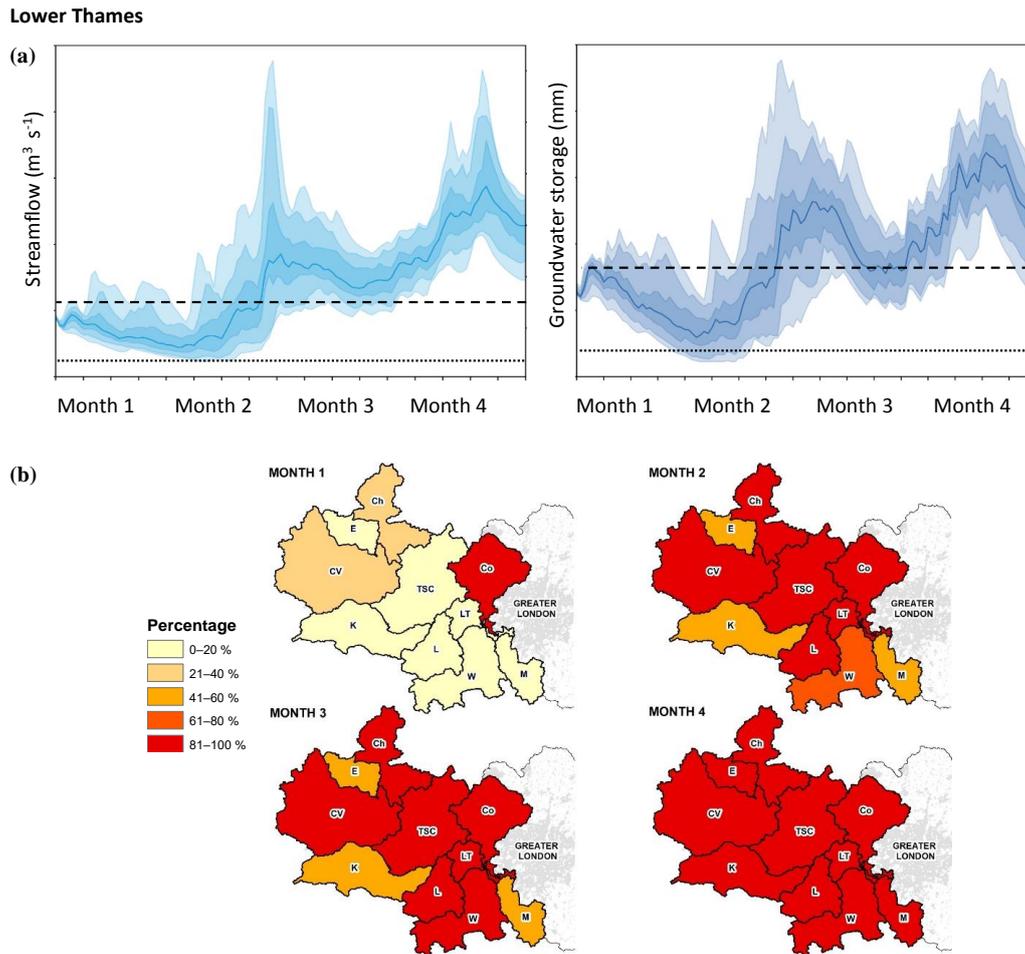


Figure 7. Four-month hydrological forecasts from the “Improved” EFAS-Seasonal (Stage 3). **(a)** Ensemble hydrographs for streamflow (light blue) and groundwater levels (dark blue) for the Lower Thames (LT) catchment. Exceedance thresholds (based on records from 1994 to 2014) are shown as Q10 (dashed line) and Q50 (dotted line). **(b)** Choropleth map shows the maximum probability that the full hydrograph ensemble for a catchment exceeds the Q10 streamflow threshold in a given month.

South Chilterns and Kennet) as these provided the greatest opportunity for water supply but also increased risk of local groundwater flooding and widespread fluvial flooding further downstream. Forecasters also highlighted the risks posed by impermeable catchments (e.g. the Cherwell and Lower Thames) that have a flashier response to rainfall. Water resource managers stated that upstream reservoirs were at increased risk of pollution (from agriculture), whilst dry weather (drought) was a greater issue towards London.

4.1.2 Current hydrological situation – “normal”

Hydrological Summary placed the “current” hydrological conditions for river flows, groundwater levels, and reservoir stocks within the “normal” range (Supplement 3). Maps indicated that rainfall was below average over the past season but above average the previous month. All groups were happy

with the current hydrological situation (no risks currently), although water resource managers stated that rainfall deficiency in the background should be kept in mind due to future drought potential.

4.2 Participant responses from Stages 1 to 3

The findings from each stage of the activity are presented below. At no point did participants ignore the SHF information (no black stickers were placed on the maps), which matched previous discussions about organizations’ current use of SHF (Sect. 2.3.2). Colour-coded decisions made by all participants (calculated by counting the stickers on the A1 catchment maps) are represented as pie charts. An accompanying bar chart details the breakdown of choices made by each participant and their specific role in the water sector (Fig. 8a–c). Quotes and information in the text are taken from discussions

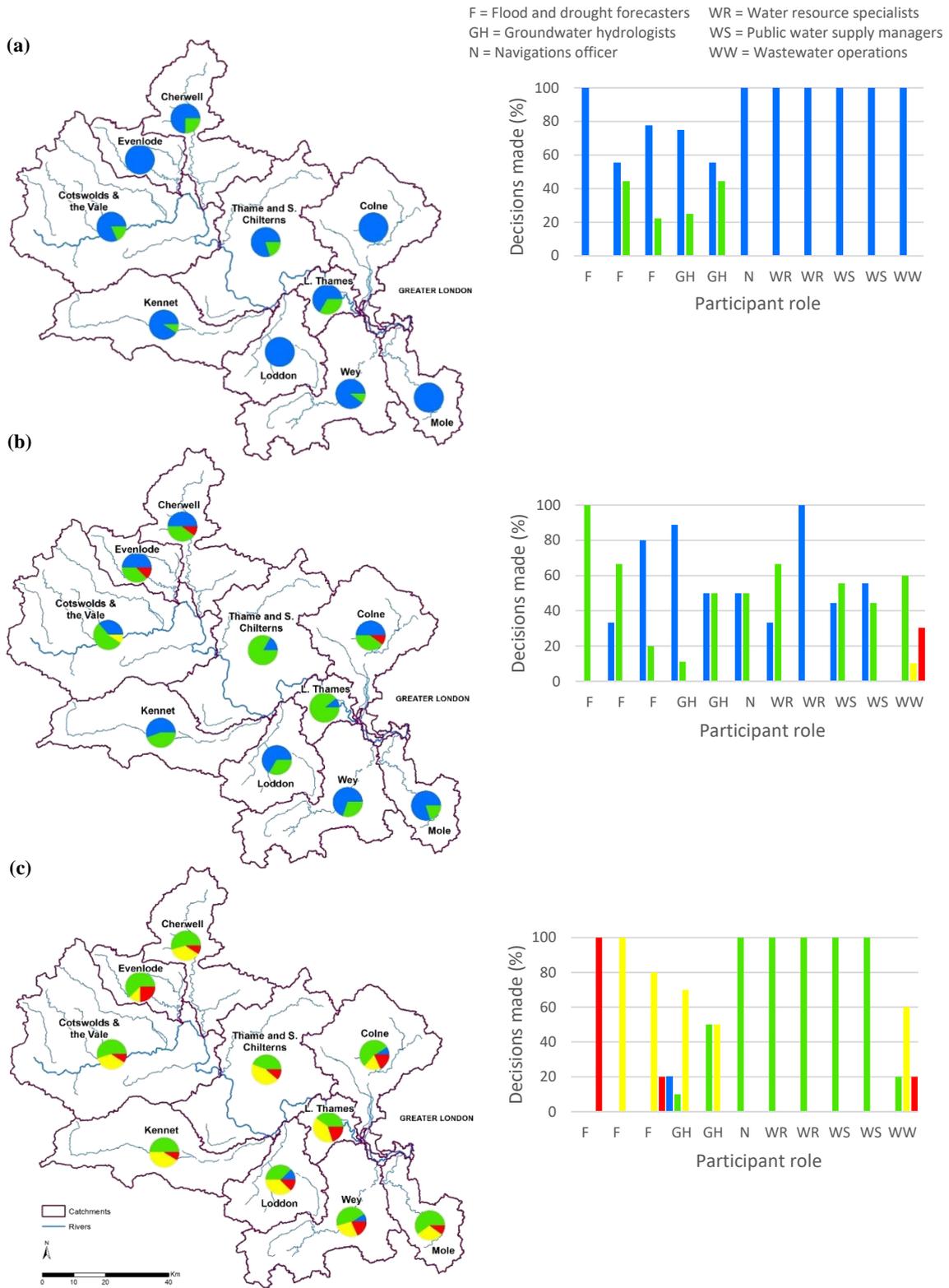


Figure 8. Summary of decisions and actions taken by different water sector personas based on (a) Hydrological Outlook UK; (b) EFAS-Seasonal; and (c) “Improved” EFAS-Seasonal. Blue – no notable risk; green – discuss internally; yellow – discuss externally and seek advice; red – implement action. Refer to Table 2 for full colour code descriptors.

recorded on the day and empathy maps – these are presented for the three groups (forecasters, groundwater hydrologists, and water resource managers).

4.2.1 Stage 1 – Hydrological Outlook UK

General consensus was for normal or above-normal conditions over the next 3 months; however, the information was “too vague to be actionable”. Forecasters and groundwater hydrologists were more likely to discuss the situation with colleagues and keep an eye on the situation (green/blue), although there was some disagreement about the level of risk. Those involved in water resources, water supply, navigation, and wastewater operations (water resource managers) identified no risks requiring action (blue) (Fig. 8a).

Key statements:



“**Analogy with the summer 2007 floods*** suggests that **there’s a risk that might be worth communicating internally**. Political influences e.g. known flooding hotspots might also be singled out for further engagement. However, there’s not much evidence to divert from a normal pattern of preparedness.”

*The UK suffered extensive flooding during June and July 2007 (the West Thames was flooded in late July). Thirteen people died and damages exceeded 3.2 billion GBP nationwide (Chatterton et al., 2010).



“**No major issues currently** but there is a **signal for rising groundwater levels**, potentially leading to flood risk – discuss with colleagues and keep an eye on borehole observations and new forecasts.”



“Conditions are **favourable from a water resources perspective** – possibly heading more towards flood than drought conditions but currently **no notable risk and no concerns**. Discussions may arise during regular business briefings, but unlikely to be pursued unless changes are observed.”

4.2.2 Stage 2 – EFAS-Seasonal

General consensus was for above-average streamflow and groundwater levels. Although the SHF provided more detail compared with Hydrological Outlook UK (Stage 1), clarity remained an issue. There was a general shift towards more internal communication (green), although actions were taken

by the wastewater operations manager in the water resource managers’ group (yellow/red) (Fig. 8b).

Key statements:



“**Repeated rainfall events can lead to accumulated flood risk** in the Lower Thames and Thame and South Chilterns. Streamflow appears to convey more risk than groundwater levels. Would discuss in general terms with colleagues and internal decision-makers to avoid an over-reaction at senior level.”



“**A moderate risk of groundwater flooding** (especially if the time period is for autumn – winter) but river flows do not appear to contribute much to groundwater risk at this stage and the forecasts are uncertain. Our **attention is focused on the chalk catchments and Thames gravels**; no direct actions are taken at the moment but we’d keep an eye on the situation and discuss at monthly meetings.”



“**No significant concerns** from a water resources or navigation perspective however, there is **potential for localised flood risk which may impact on water supply and turbidity**. Not all catchments are affected so focus attention on Cotswolds and the Vale, Cherwell, Thame and South Chilterns and Colne where maps indicate high probability of Q10 exceedance. Discuss at internal briefings.”

4.2.3 Stage 3 – “Improved” EFAS-Seasonal

General consensus was for confident forecasts that showed a high risk of streamflow and groundwater flooding in approximately 6 weeks’ time. At this stage, forecasters and groundwater hydrologists were looking to verify the reliability and quality of the forecasts. Internal discussion and wider communication (green/yellow) were actively explored, although forecasters and groundwater hydrologists were still more likely to act on the information compared with water resource managers (Fig. 8c).

Key statements:



“Compared with our previous experiences of SHF these are very **sharp with a strong signal** and we would actively seek expert guidance as to the quality of the forecasts. If credible, our concern is that the signal is likely to **represent a nationwide flood risk** (not just the West Thames). **Low-consequence actions that deliver a measured message** should be implemented – e.g., identifying and locating resources and stocks, movement of temporary flood defences to high risk areas, completing projects, careful media release, strategic planning and staff briefing.”

“There’s **high probability of substantially exceeding the Q10 threshold**. Catchment characteristics are important to identify areas most at risk of groundwater flooding (chalk and gravels). **Drawing on previous experiences** we’d discuss the situation, obtain regular updates from partner organisations, use localised groundwater models to verify forecasts and consider communication via press release.”



“These are **confident forecasts that give a good overview of magnitude and sequencing of possible flood events and subsequent knock-on effects to water quality**. Expect issues in 2–4 months so any actions taken would depend on how regularly forecasts are updated. We’d keep an eye on groundwater levels, hold internal briefings and discuss with groundwater team members to ensure they are kept informed and prepared. For navigation and wastewater operations where impacts can directly affect the public, we’d consider some open discussion with customers who will want to know how long an event might last.”



represented an extreme flood event between November 2013 and February 2014. There was clear evidence that more confident (sharper) and locally tailored forecasts led to increased levels of decision and action, although water sector users did not respond uniformly. Forecasters and groundwater hydrologists were most likely to inform other organizations, request further information about the situation, and implement action, while water resource managers more consistently adopted a “watch and wait” approach. In this section, the results are discussed in more detail and the findings are placed into the wider context of policy, practice, and next steps based on discussions captured during the focus group.

5.1 Operational SHF systems can support decision-making and uncertainty is expected

Throughout the focus group, participants expressed positively the potential for SHF to deliver better preparedness and early warning of flood and drought events, and the benefits associated with more consistent management of water resources, whilst recognizing that low skill and coarse resolution are current barriers to use (see also Soares and Desai, 2015, 2016; Vaughan et al., 2016; Soares et al., 2018). These benefits and barriers were demonstrated during the activity as participants increased their level of decision-making in response to the more confident and locally tailored forecasts presented: Stage 1 Hydrological Outlook UK > Stage 2 EFAS-Seasonal > Stage 3 “Improved” EFAS-Seasonal.

Hydrological Outlook UK is the first operational SHF system for the UK and was the product that participants were most familiar with, likely due to its partnership set-up (Prudhomme et al., 2017). All groups indicated that the regional focus of the maps, i.e. the whole Thames basin, and lack of resolution and certainty as to the trajectory of the upcoming hydrological conditions, limited their ability to make informed decisions. No participants however ignored or dismissed the information despite there being no perceived risk. All agreed that on a day-to-day basis, Hydrological Outlook UK serves as a useful outlook tool when supplemented with additional sources of information including water situation reports (UK Gov, 2018) and other hydro-meteorological forecasts. As of 2017, exactly how the water sector uses Hydrological Outlook UK in practice had yet to be assessed (Bell et al., 2017), and here we provide a first step towards answering this question.

Stage 2 (EFAS-Seasonal) also represented an operational forecasting system designed to run at the scale of the whole Thames basin akin to Hydrological Outlook UK. The forecasts however were presented at a catchment level on a month-by-month basis to provide a more localized outlook. This finer spatio-temporal resolution allowed participants to supplement the SHF with their knowledge of local hydrogeology and other risk factors to identify those catchments where attention would likely be most needed. This led to increased levels of communication within organizations,

5 Discussion

Our decision-making activity was designed to help understand how different water sector users engage with and act on SHF at a local level. The SHF for the three activity stages rep-

even though the overall hydrological outlook was very similar to that observed at Stage 1 (uncertain but with indication towards normal–high flows). The use of large-scale (regional or global) operational forecasting products that trigger worthwhile actions at the local level has been demonstrated at shorter lead times (e.g. Coughlan de Perez et al., 2016). While the development of higher-resolution seasonal meteorological forecasts and better representation of the coupled system and initial conditions are expected to lead to improvements in SHF (Lewis et al., 2015; Bell et al., 2017; Arnal et al., 2018), we pose the open question: do operational systems such as Hydrological Outlook UK *already* have the potential to support better communication and decision-making if they could be presented at a more local scale? This would require careful communication of the uncertainty, reliability, and skill of the forecast, and how to do this effectively is a topic of current interest in meteorological and hydrological forecasting (e.g. Ramos et al., 2013; Vaughan et al., 2016; Fry et al., 2017). Although communicating uncertainty was not a specific focus of our activity, one key message from the focus group was that “uncertainty is expected” with SHF and water sector users would engage with a local forecast, even if they chose not to act on it. As pointed out by Viel et al. (2016), “low skill” is not the same as “no skill”, and SHF which may have minimal value from the perspective of a scientific researcher can sometimes elicit significant interest from the view of a water sector user who is familiar with the area. Importantly, it should also be noted that although no measures of forecast skill and quality were included in our activity, participants only expressed a need to verify the quality of the forecasts at Stage 3. In discussions as to why this was the case, the forecasters and groundwater hydrologists stated that holding internal briefings and increasing awareness of “at risk” catchments are suitable low-cost actions when dealing with SHF that indicate some degree of risk, even if the information is uncertain and unverified. At Stage 3, to obtain such confident SHF was well beyond current operational standards; thus, its reliability was questioned. Participants did agree however that even in the absence of information on forecast quality, a sharper, more confident forecast that indicated high potential flood risk would be more likely to provoke a response than a dispersive one, even if the maximum of the forecast ensemble indicated values of comparable magnitude in both cases.

5.2 Interactions with SHF are user-specific and should be tailored accordingly

The manner in which users approached and used SHF differed markedly depending on the perceived severity of the flood event; the responsibilities and risk appetite of an organization; and the local knowledge and experiences possessed by the individual (see also Kirchhoff et al., 2013; Golding et al., 2017). Forecasters and groundwater hydrologists displayed the lowest risk appetite, admitting that they

were likely to err on the side of caution to avoid negative media impacts, economic damages, and loss of trust by the public.

“Analogy with the summer floods of 2007 ... my previous experience makes me think that the risk is worth communicating...” – forecaster at Stage 1/2.

“A much stronger and more coherent signal regarding river flows and groundwater levels, but the forecasts indicate that the potential impact isn’t right now ... we’ll keep an eye on the situation” – water resource manager at Stage 3.

While a flood event is less of an immediate issue for water resource managers, secondary effects relating to closure of canals (navigation), turbidity, and sewer surcharge (wastewater operations) did invoke action where there was potential to impact on the public. Participants were notably proactive where they had had previous experience of extreme events, e.g. forecasters’ analogies with the 2007 floods (Chatterton et al., 2010), or had been witness to poor management; e.g. the wastewater operations manager recognized high potential for groundwater flooding and sewer surcharge at 1 month’s lead time in the Evenlode, Cherwell, and Colne (Fig. 7).

“Based on previous operational issues, I’d advise pre-emptive actions such as the cleaning and maintenance of pumping stations for these catchments” – Wastewater operations manager at Stage 2/3.

This highlights the value of retaining institutional memory where possible (see also McEwen et al., 2012) and being aware of organizations’ or individuals’ pre-determined positions or perceived self-interests which may largely be founded on previous experiences (Ishikawa et al., 2011).

It is important to note that while this activity focused on a flood event, decisions made by the groups would almost certainly have differed if the SHF had indicated drought conditions. The impacts of drought have the potential to affect larger areas, for longer (Bloomfield and Marchant, 2013), notably with respect to agriculture (Li et al., 2017), reservoir management (Turner et al., 2017) and navigation (Meißner et al., 2017). The difference in response between water sector users supports the notion that tailoring SHF information to specific user groups will improve uptake and ability to inform decision-making (Jones et al., 2015; Lorenz et al., 2015; Vaughan et al., 2016; Soares et al., 2018), an area currently being explored by the IMPREX Risk Outlook (IMPREX, 2018b).

5.3 Communication is both a barrier and enabler to decision-making

Communication is one of the most frequently identified barriers when it comes to uptake and use of seasonal meteorological and hydrological forecasts (Soares and Dessai, 2015;

Vaughan et al., 2016; Golding et al., 2017; Soares et al., 2018). Discussions captured during the focus group and indicated on some empathy maps identified two key communication barriers in the West Thames: (1) between water sector users themselves and how they interpret and communicate SHF information and (2) a disconnect between scientists developing the forecasts and those involved in policy, practice and decision-making.

All groups said they felt better able to interpret and communicate the messages when presented with a range of complementary forms of SHF information including maps, hydrographs, and text, with maps being of particular value. This supports findings by Lorenz et al. (2015), who identified clear differences in users' comprehension of and preference for visualizations of climate information. Mapping information was also found to be important in the survey by Vaughan et al. (2016), while numerical representations were preferred over text and graphics in the study by Soares et al. (2018). Many participants said they would feel better prepared and able to discuss upcoming hydrological conditions if SHF information was visualized in a variety of ways and regular engagement was made a routine part of their job (see Sect. 5.4).

A number of participants also felt that scientific improvements and developments to SHF are not being adequately communicated to those involved in policy and practice. General consensus was that knowledge exchange events and information sharing services through projects such as IMPREX are an excellent way of addressing this disconnect. Presentations during the focus group shared findings from other projects, including the European Provision Of Regional Impacts Assessments on Seasonal and Decadal Timescales (EUPORIAS) (Met Office, 2018), the End-to-end Demonstrator for improved decision-making in the water sector in Europe (EDgE), Service for Water Indicators in Climate Change Adaptation (SWICCA) (Copernicus, 2017a, b), and Improving Predictions of Drought for User Decision Making (IMPETUS) (Prudhomme et al., 2015) – much of which was new knowledge to some participants. It was further expressed that stakeholder events yield maximum benefit for both the scientist and the user when they are co-produced with an organization that is involved in receiving, tailoring, and distributing SHF information (Rapley et al., 2014). Importantly, we do not want to be in the position whereby SHF skill has improved but the credibility and reliability of the information is questioned by decision-makers who have not been kept up to date with developments. The potential for this disconnect was demonstrated by both forecasters and groundwater hydrologists at Stage 3 (“Improved” EFAS-Seasonal) whereby decisions would only be made if the accuracy of the forecast could be verified.

“Forecast signal is implausibly strong but, if valid, gives a clear signal for disturbed conditions”

“Surprised at forecast and the strength of the signal... IF credible, then actions need to be taken”

“Would definitely talk to the Environment Agency and search for other monitoring data to verify the forecast” – forecasters and groundwater hydrologists at Stage 3.

In this case, the SHF at Stage 3 were hypothetical and no information on forecast quality was given; however, the forecasts provided a good representation of what scientists hope to achieve with operational seasonal forecasting systems in the future (Neumann et al., 2018). This emphasizes the need to keep water sector users informed of scientific developments (see also Bolson et al., 2013), and to build awareness and knowledge around interpreting and using forecast quality information, as it is becoming more widely adopted in seasonal forecasting (see Copernicus, 2017a; Fry et al., 2017).

5.4 Implications for future policy and decision-making

The EA is the public body responsible for managing flood risk in the UK. They focus on maintaining a certain level of preparedness whilst recognizing that particular conditions and types of flooding/drought are more likely at different times of year. Currently, the EA use SHF predominantly as supporting information and rely on shorter-range forecasts for action. As co-developers of this focus group, the EA recognized the following points for future consideration.

1. To upskill and help staff interpret SHF information received.
2. To identify suitable low-consequence actions that could be taken based on SHF.
3. To move beyond the current position of using SHF for information only, to making conscious decisions as part of routine incident management strategies (relies on 1 and 2).

“Regular review and discussion of extended outlooks (5–30 days) and the 1–3 months forecasts during weekly handover between the incoming and outgoing flood duty teams would improve familiarity of long range forecast products and dealing with the uncertainty that they present. This would be an excellent way of considering the possible conditions and the potential for disruption going forward.” – EA activity co-developer.

In short, more engagement with SHF and improved clarity for easier interpretation by different users will ensure that SHF have a valuable role to play in future decision-making at the local scale.

5.5 Learning outcomes and future considerations

Encouragingly, we identified that SHF are being used, and participants agreed that the decision-making activity was an

entertaining platform for fostering discussions which complemented their everyday work and general understanding of SHF. From the participants' perspective, learning outcomes included knowing more about the ongoing scientific developments in SHF and a better understanding of how different organizations in the West Thames water sector are using SHF. Many also stated that the activity and focus group discussions enhanced their ability to think about possible decisions and actions that may be taken in the future. As the activity developers, we found that the group discussions stimulated participants' motivations and interests more so than would have been achieved by asking participants to engage on an individual basis. We also advocate the use of empathy maps or other forms of obtaining a written record of participant thought processes in addition to their decision choices.

Our activity was designed to provide a first insight into the current state of play regarding SHF in the West Thames. Although 11 participants was a small sample size, they represented an important and well-balanced mix of water sector decision-makers in the West Thames. The only exception was the agricultural sector, which could not attend, and thus it would be interesting to capture this perspective with ongoing research (e.g. Li et al., 2017). We also recognize the possibility that those who took part had a vested interest in SHF; however, we did encourage participants to attend even where they had no background knowledge or experience of SHF. Finally, we advocate that others conducting a similar activity may wish to consider whether participant interpretation can be subconsciously influenced by the information provided. For example, flood risk maps were provided as part of the background context, but may have inadvertently led participants to consider the upcoming forecasts with respect to high-flow events. Likewise, there is potential that the

3-month SHF (Stage 1) may have been interpreted differently to the 4-month forecasts (Stage 2 and Stage 3) and we do not know the degree to which individuals may have been swayed to place a particular colour on the map based on the conversations they had with their group members (and how big an influence such conversations play in real life). Discussions with the participants at the end of the activity with respect to these points would have been helpful.

6 Conclusions

Key findings were that engagement is user-specific and SHF have the potential to be more useful if they could be presented at a scale which matches that employed in decision-making. The ability to interpret messages is aided by complementary forms of SHF visualization that provide a wider overview of the upcoming hydrological outlook, with maps being of particular value. However, improved communication between scientists, providers, and users is required to ensure that users are kept up to date with developments. We conclude that the current level of understanding in the West Thames provides an excellent basis upon which to incorporate future developments of operational forecasts and for facilitating communication and decision-making between water sector partners.

Data availability. All data/graphs/information that were used by participants for the focus group activity are included in the Supplement. Individual participant results are not publicly available in order to protect anonymity. If readers require further information, this may be provided by contacting the corresponding author.

Appendix A: Glossary

Aquifer	underground layer of water-bearing permeable rock which can occur at various depths.
Atmospheric relaxation experiments	are used by meteorologists once an extreme weather event has happened. Put simply, when a seasonal forecast predicts the wrong weather, scientists “force” the conditions in the atmosphere so that they can try to recreate the extreme weather conditions and better understand what happened.
Baseflow	the portion of the river flow (streamflow) that is sustained between rainfall events and is fed into streams and rivers by delayed shallow subsurface flow. Not to be confused with “groundwater” which is water which has entered an aquifer, or “groundwater flow” where water enters a river having been in an aquifer.
Choropleth map	uses differences in shading, patterning or colouring in proportion to the value of a given variable in areas of interest.
Exceedance threshold	a user-defined threshold (e.g. 90 %) that is based on river flow or groundwater level observations (measurements) from the previous 20 years. E.g. if an exceedance threshold is set to the 90th percentile, this means that 90 % of all recorded observations over the past 20 years fell below this level.
Flashy	ivers and catchments that respond quickly to rainfall events.
Forecast ensemble	instead of running a single forecast (known as a deterministic forecast that has one outcome), computer models can run a forecast several times using slightly different starting conditions (to account for uncertainties in the forecasting process). The complete set of forecasts is referred to as the ensemble, and the individual forecasts are known as ensemble members. Each ensemble member represents a different possible scenario, and each scenario is equally likely to happen.
Forecast quality	the SHF is compared to, or verified against, a corresponding observation of what actually happened, or a good estimate of the true outcome. SHF quality describes the degree to which the forecast corresponds to what actually happened (see also “forecast skill”).
Forecast sharpness	describes the spread or variability among the different ensemble members of a forecast (the different forecast values). The more concentrated (close together) the ensemble members are, the sharper the forecast is, and vice versa. Importantly, a forecast can be sharp even if it is wrong i.e. far from what actually happened. (See also “forecast ensemble”).
Forecast skill	the SHF quality can be compared to the quality of a benchmark or reference, usually another forecast. The relative quality of the SHF over this reference forecast is the SHF skill (see also “forecast quality”).
Forecast uncertainty	the skill and accuracy of SHF tends to decrease with increasing lead time due to factors such as variations in weather conditions, how the hydrological model has been set-up to represent complex processes, and how well the hydrological model has captured the real-world hydrologic conditions at the time the forecast is started (e.g. how wet is the soil or how much water is currently in the river?). There is an element of uncertainty in all forecasts that can amplify with time. Ensemble forecasting is one way of representing forecast uncertainty. (See also “forecast ensemble”).
Hydrogeology	the area of geology that deals with the distribution and movement of below-ground water in the soil, rocks and aquifers.
Hydrograph	a graph showing how river and groundwater levels are expected to change over time at a specific location. Ensemble hydrographs show the full spread of the forecast ensemble.
Lead time	the length of time between when the SHF is started (initiated) and the occurrence of the phenomena (e.g. flood) being predicted. Can also be used to represent the point at which the SHF is started and the beginning of the forecast validity period (e.g. from 3 weeks).
Lithology	the general physical characteristics of rocks.
River basin	the largest and total area of land drained by a major river (in this case the River Thames) and all its tributaries. (See also “river catchment”).
River catchment	the area of land drained by a river. “Catchment” and “basin” are sometimes used interchangeably. Here catchments represent the drainage areas of the River Thames main tributaries, of which there are 10 in the West Thames.

Seasonal hydrological forecasts (SHF)	provide information about the hydrological conditions e.g. streamflow (river flows), ground-water levels and soil moisture levels, that might be expected over the next few months (e.g. from 3 weeks out to 7 months).
Seasonal meteorological forecasts	provide information about the weather conditions e.g. rainfall, air temperature, humidity, pressure, wind, that might be expected over the next few months (e.g. from 3 weeks out to 7 months).
Streamflow	the flow of water in a stream or river. Also known as river flow.
Surface runoff	the flow of water that occurs when water from excess rainfall, meltwater or drainage systems flows over the Earth's surface and not into the ground.
Tributary	a river or stream that flows into a larger stream, river or lake. Tributaries do not flow into the sea.
1 : 100-year flood event	a 100-year flood is a flood event that has a 1 % chance of occurring in any given year.
1 : 5-year flood event	a 1-in-5-year flood is a flood event that has a 20 % chance of occurring in any given year.

Information about the Supplement

- Supplement 1: Invitation flyer and programme for the focus group
- Supplement 2: West Thames catchment characteristic maps
- Supplement 3: Hydrological Summary: October 2013, June–September 2013 and November 2012–October 2013
- Supplement 4: Stage 1 Hydrological Outlook UK: November 2013–January 2014
- Supplement 5: Stage 2 EFAS-Seasonal: November 2013–February 2014
- Supplement 6: Stage 3 “Improved” EFAS-Seasonal: November 2013–February 2014

Supplement. The supplement related to this article is available online at: <https://doi.org/10.5194/gc-1-35-2018-supplement>.

Author contributions. JLN and LA designed the decision-making activity. JLN, LA, SH, and HLC co-organized the set-up of the focus group. All the authors took part in delivering the focus group, including as note-takers, organizers, and presenters of their scientific research. JLN wrote the manuscript with input from all the authors.

Competing interests. The authors declare that they have no conflict of interest.

Disclaimer. The information and findings in this paper are based on discussions and actions captured during the decision-making activity. They should not be taken as representing the views or practice of particular organizations or institutions.

Acknowledgements. This work was funded by the EU Horizon 2020 IMPREX project (<http://www.imprex.eu/>, last access: 21 May 2018) (641811) with additional financial support provided by the University of Reading’s Endowment Fund. Support-in-kind was also provided by the NERC LANDWISE project (<https://landwise-nfm.org/about/>, last access: 10 July 2018) (NE/R004668/1). We would like to express our sincere thanks to all participants who shared their knowledge and experience relating to seasonal hydrological forecasting and to their organizations who enabled their participation. We would especially like to thank Stuart Hyslop and Simon Lewis at the EA for their support in the organization of the day and also Len Shaffrey (Department of Meteorology, University of Reading) for his input on the day.

Edited by: Katharine Welsh

Reviewed by: two anonymous referees

References

- Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., and Salamon, P.: Evaluation of ensemble streamflow predictions in Europe, *J. Hydrol.*, 517, 913–922, <https://doi.org/10.1016/j.jhydrol.2014.06.035>, 2014.
- Arnal, L., Ramos, M.-H., Coughlan de Perez, E., Cloke, H. L., Stephens, E., Wetterhall, F., van Andel, S. J., and Pappenberger, F.: Willingness-to-pay for a probabilistic flood forecast: a risk-based decision-making game, *Hydrol. Earth Syst. Sci.*, 20, 3109–3128, <https://doi.org/10.5194/hess-20-3109-2016>, 2016.
- Arnal, L., Wood, A. W., Stephens, E., Cloke, H., and Pappenberger, F.: An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity, *J. Hydrometeorol.*, 18, 1715–1729, <https://doi.org/10.1175/JHM-D-16-0259.1>, 2017.
- Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B., and Pappenberger, F.: Skilful seasonal forecasts of streamflow over Europe?, *Hydrol. Earth Syst. Sci.*, 22, 2057–2072, <https://doi.org/10.5194/hess-22-2057-2018>, 2018.
- Arribas, A., Glover, M., Maidens, A., Peterson, K., Gordon, M., MacLachlan, C., Graham, R., Fereday, D., Camp, J., Scaife, A. A., Xavier, P., McLean, P., and Colman, A.: The GloSea4 Ensemble Prediction System for Seasonal Forecasting, *Mon. Weather. Rev.*, 139, 1891–1910, <https://doi.org/10.1175/2010MWR3615.1>, 2010.
- Asrar, G. R., Hurrell, J. W., and Busalacchi, A. J.: A need for “actionable” climate science and information: summary of WCRP open science conference, *B. Am. Meteorol. Soc.*, 94, ES8–ES12, <https://doi.org/10.1175/BAMS-D-12-00011.1>, 2012.
- Bell, V. A., Davies, H. N., Kay, A. L., Marsh, T. J., Brookshaw, A., and Jenkins, A.: Developing a large-scale water-balance approach to seasonal forecasting: application to the 2012 drought in Britain, *Hydrol. Process.*, 27, <https://doi.org/10.1002/hyp.9863>, 2013.
- Bell, V. A., Davies, H. N., Kay, A. L., Brookshaw, A., and Scaife, A. A.: A national-scale seasonal hydrological forecast system: development and evaluation over Britain, *Hydrol. Earth Syst. Sci.*, 21, 4681–4691, <https://doi.org/10.5194/hess-21-4681-2017>, 2017.
- Bloomfield, J. P. and Marchant, B. P.: Analysis of ground-water drought building on the standardised precipitation index approach, *Hydrol. Earth Syst. Sci.*, 17, 4769–4787, <https://doi.org/10.5194/hess-17-4769-2013>, 2013.
- Bloomfield, J. P., Bricker, S. H., and Newell, A. J.: Some relationships between lithology, basin form and hydrology: A case study from the Thames basin, UK, *Hydrol. Process.*, 25, 2518–2530, <https://doi.org/10.1002/hyp.8024>, 2011.
- Bolson, J., Martinez, C., Breuer, N., Srivastava, P., and Knox, P.: Climate information use among southeast US water managers: beyond barriers and toward opportunities, *Reg. Environ. Change*, 13, 141–151, <https://doi.org/10.1007/s10113-013-0463-1>, 2013.
- CEH: Hydrological Outlook – Further Information for November 2013, available at: <http://www.hydoutuk.net/archive/2013/november-2013/further-information-november-2013/> (last access: 25 April 2018), 2013.
- CEH: Hydrological Outlook UK, available at: <http://www.hydoutuk.net/>, last access: 9 April 2018.

- Chatterton, J., Viavattene, C., Morris, J., Penning-Rowsell, E., and Tapsell, S.: The costs of the summer 2007 floods in England, Environment Agency Report SC070039, Rio House, Bristol, UK, 2010.
- Chiew, F. H. S., Zhou, S. L., and McMahon, T. A.: Use of seasonal streamflow forecasts in water resources management, *J. Hydrol.*, 270, 135–144, 2003.
- Copernicus: EDgE, Climate Change Service, available at: <http://edge.climate.copernicus.eu/> (last access: 31 May 2018), 2017a.
- Copernicus: SWICCA: Service for Water Indicators in Climate Change Adaptation, SMHI, available at: <http://swicca.climate.copernicus.eu/> (last access: 31 May 2018), 2017b.
- Coughlan de Perez, E., van den Hurk, B., van Aalst, M. K., Amuron, I., Bamanya, D., Hauser, T., Jongma, B., Lopez, A., Mason, S., Mendler de Suarez, J., Pappenberger, F., Rueth, A., Stephens, E., Suarez, P., Wagemaker, J., and Zsoter, E.: Action-based flood forecasting for triggering humanitarian action, *Hydrol. Earth Syst. Sci.*, 20, 3549–3560, <https://doi.org/10.5194/hess-20-3549-2016>, 2016.
- Crochemore, L., Ramos, M.-H., Pappenberger, F., van Andel, S. J., and Wood, A. W.: An experiment on risk-based decision-making in water management using monthly probabilistic forecasts, *B. Am. Meteorol. Soc.*, 97, 541–551, 2015.
- Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R. L.: Seasonal climate predictability and forecasting: status and prospects, *WIREs Clim. Change*, 4, 245–268, <https://doi.org/10.1002/wcc.217>, 2013.
- EA (Environment Agency): Thames Catchment Flood Management Plan – Managing Flood Risk, Summary Report December 2009, EA, Kings Meadow House, Reading, 2009.
- EA (Environment Agency): The costs and impacts of the winter 2013 to 2014 floods, Technical Report SC140025, Defra/Environment Agency Joint R&D programme, 2015.
- EA (Environment Agency): Groundwater Level Measurements (AFA075), Data contains Environment Agency information[©] Environment Agency and/or database right, All rights reserved, Data sourced under Environment Agency Conditional Licence, 2017.
- Emerton, R., Zsoter, E., Arnal, L., Cloke, H. L., Muraro, D., Prudhomme, C., Stephens, E. M., Salamon, P., and Pappenberger, F.: Developing a global operational seasonal hydro-meteorological forecasting system: GloFAS-Seasonal v1.0, *Geosci. Model Dev.*, 11, 3327–3346, <https://doi.org/10.5194/gmd-11-3327-2018>, 2018.
- Farolfi, S., Hassan, R., Perret, S., and MacKay, H.: A role-playing game to support multi-stakeholder negotiations related to water allocation in South Africa: First applications and potential developments, *Midrand: Water Resources as Ecosystems: Scientists, Government and Society at the Crossroads*, 2004.
- Fry, M., Smith, K., Sheffield, J., Watts, G., Wood, E., Cooper, J., Prudhomme, C., and Rees, G.: Communication of uncertainty in hydrological predictions: a user-driven example web service for Europe, *Geophys. Res. Abstr.*, EGU2017-16474, EGU General Assembly 2017, Vienna, Austria, 2017.
- Golding, N., Hewitt, C., Zhang, P., Bett, P., Fang, X., Hu, H., and Nobert, S.: Improving user engagement and uptake of climate services in China, *Climate Services*, 5, 39–45, 2017.
- Gray, D.: Gamestorming – Empathy Map, available at: <http://gamestorming.com/empathy-mapping/> (last access: 1 May 2018), 2017.
- Harrison, J.: Flood hazard management: Using an alternative community-based approach, *Planet*, 4, 5–6, 2002.
- Huntingford, C., Marsh, T., Scaife, A. A., Kendon, E. J., Hanaford, J., Kay, A. L., Lockwood, M., Prudhomme, C., Reynar, N. S., Parry, S., Lowe, J. A., Screen, J. A., Ward, H. C., Roberts, M., Stott, P. A., Bell, V. A., Bailey, M., Jenkins, A., Legg, T., Otto, F. E. L., Massey, N., Schaller, N., Slingo, J., and Allen, M. A.: Potential influences on the United Kingdom’s floods of winter 2013/14, *Nat. Clim. Change*, 4, 769–777, <https://doi.org/10.1038/nclimate2314>, 2014.
- Ishikawa, T., Barnson, A. G., Kastens, K. A., and Louchouart, P.: Understanding, evaluation, and use of climate forecast data by environmental policy students, in: *Qualitative inquiry in geoscience education research*, edited by: Feig, A. D. and Stokes, A., Geological Society of America Special Paper 474, 153–170, Geol. Soc. Am., Denver, CO, 2011.
- IMPRES: Thames River Basin, available at: <http://impres.eu/thames-river-basin> (last access: 8 April 2018), 2018a.
- IMPRES: Risk Outlook Tool, available at: <http://www.impres.eu/innovation/risk-outlook> (last access: 21 May 2018), 2018b.
- Jones, L., Dougill, A., Jones, R. G., Steynor, A., Watkiss, P., Kane, C., Koelle, B., Moufouma-Okia, W., Padgham, J., Ranger, N., Roux, J.-P., Suarez, P., Tanner, T., and Vincent, K.: Ensuring climate information guides long-term development, *Nat. Clim. Change*, 5, 812–814, <https://doi.org/10.1038/nclimate2701>, 2015.
- JRC: European Flood Awareness System, available at: <https://www.efas.eu/> (last access: 9 April 2018), 2018a.
- JRC: Global Flood Awareness System, available at: http://www.globalfloods.eu/user-information/seasonal_outlook/ (last access: 9 April 2018), 2018b.
- Kendon, M. and McCarthy, M.: The UK’s wet and stormy winter of 2013/2014, *Weather*, 70, 40–47, <https://doi.org/10.1002/wea.2465>, 2015.
- Kirchhoff, C. J., Lemos, M. C., and Engle, N. L.: What influences climate information use in water management? The role of boundary organizations and governance regimes in Brazil and the U.S., *Environ. Sci. Policy*, 26, 6–18, <https://doi.org/10.1016/j.envsci.2012.07.001>, 2013.
- Lemos, M. C., Kirchhoff, C. J., and Ramprasad, V.: Narrowing the climate information usability gap, *Nat. Clim. Change*, 2, 789–794, 2012.
- Lewis, H., Mittermaier, M., Mylne, K., Norman, K., Scaife, A., Neal, R., Pierce, C., Harrison, D., Jewell, S., Kendon, M., Saunders, R., Brunet, G., Golding, B., Kitchen, M., Davies, P., and Pilling, C.: From months to minutes – exploring the value of high-resolution rainfall observation and prediction during the UK winter storms of 2013/2014, *Meteorol. Appl.*, 22, 90–104, 2015.
- Li, Y., Giuliani, M., and Castelletti, A.: A coupled human–natural system to assess the operational value of weather and climate services for agriculture, *Hydrol. Earth Syst. Sci.*, 21, 4693–4709, <https://doi.org/10.5194/hess-21-4693-2017>, 2017.
- Lorenz, S., Dessai, S., Forster, P., and Paavola, J.: Tailoring the visual communication of climate projections for local adaptation practitioners in Germany and the

- United Kingdom, Philos. T. Roy. Soc. A, 373, 20140457, <https://doi.org/10.1098/rsta.2014.0457>, 2015.
- Mackay, J. D., Jackson, C. D., Brookshaw, A., Scaife, A. A., Cook, J., and Ward, R. S.: Seasonal forecasting of groundwater levels in principal aquifers of the United Kingdom, *J. Hydrol.*, 530, 815–828, <https://doi.org/10.1016/j.jhydrol.2015.10.018>, 2015.
- McEwen, L. J., Krause, F., Jones, O., and Garde Hansen, J.: Sustainable flood memories, informal knowledge and the development of community resilience to future flood risk, *Transactions on Ecology and The Environment*, 159, 253–263, 2012.
- McEwen, L., Stokes, A., Crowley, K., and Roberts, C.: Using role-play for expert science communication with professional stakeholders in flood risk management, *J. Geogr. Higher Educ.*, 38, 277–300, <https://doi.org/10.1080/03098265.2014.911827>, 2014.
- Meadow, A., Ferguson, D., Guido, Z., Horangic, A., Owen, G., and Wall, T.: Moving toward the deliberate co-production of climate science knowledge, *Weather, Clim. Soc.*, 7, 179–191, <https://doi.org/10.1175/WCAS-D-14-00050.1>, 2015.
- Meißner, D., Klein, B., and Ionita, M.: Development of a monthly to seasonal forecast framework tailored to inland waterway transport in central Europe, *Hydrol. Earth Syst. Sci.*, 21, 6401–6423, <https://doi.org/10.5194/hess-21-6401-2017>, 2017.
- Met Office.: EUPORIAS Project, available at: <https://www.metoffice.gov.uk/research/collaboration/euporias>, last access: 16 June 2018.
- Molteni, F., Stockdale, T., Alonso-Balmaseda, M., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T. N., and Vitart, F.: The new ECMWF seasonal forecast system (System 4), *ECMWF Tech. Memo.*, 656, 1–49, 2011.
- Muchan, K., Lewis, M., Hannaford, J., and Parry, S.: The winter storms of 2013/2014 in the UK: hydrological responses and impacts, *Weather*, 70, 55–61, <https://doi.org/10.1002/wea.2469>, 2015.
- Neumann, J. L., Arnal, L., Emerton, R., Griffith, H., Theofanidi, S., and Cloke, H.: Supporting the integration and application of seasonal hydrological forecasts in the West Thames, Technical Report for IMPREX, <https://doi.org/10.13140/RG.2.2.19905.25447>, 2017.
- Neumann, J. L., Arnal, L. L. S., Magnusson, L., and Cloke, H. L.: The 2013/14 Thames basin floods: Do improved meteorological forecasts lead to more skilful hydrological forecasts at seasonal timescales?, *J. Hydrometeorol.*, 19, 1059–1075, <https://doi.org/10.1175/JHM-D-17-0182.1>, 2018.
- NRFA (National River Flow Archive): Search for Gauging Stations, available at: <http://nrfa.ceh.ac.uk/data/search>, last access: 10 July 2017.
- NRFA (National River Flow Archive): Monthly Hydrological Summaries, available at: <https://nrfa.ceh.ac.uk/monthly-hydrological-summary-uk?page=5>, last access: 22 May 2018.
- Parry, S., Prudhomme, C., Wilby, R., and Wood, P.: Chronology of drought termination for long records in the Thames catchment, in: *Drought: Research and Science-Policy Interfacing*, edited by: Andreu, J., Solera, A., Paredes-Arquiola, J., Haro-Monteagudo, D., and van Lanen, H., London, Taylor & Francis (CRC Press), 165–170, 2015.
- Pavey, J. and Donoghue, D.: The use of role play and VLEs in teaching environmental management, *Planet*, 10, 7–10, 2003.
- Prudhomme, C., Shaffrey, L. C., Woolings, T., Jackson, C. R., Fowler, H. J., and Anderson, B.: IMPETUS: Improving predictions of drought for user decision-making, in: *Drought: Research and Science-Policy Interfacing*, edited by: Andreu, J., Solera, A., Paredes-Arquiola, J., Haro-Monteagudo, D., and van Lanen, H., CRC Press, <https://doi.org/10.1201/b18077-47>, 2015.
- Prudhomme, C., Hannaford, J., Harrigan, S., Boorman, D., Knight, J., Bell, V., Jackson, C., Svensson, C., Parry, S., Bachiller-Jareno, N., Davies, H., Davis, R., Mackay, J., McKenzie, A., Rudd, A., Smith, K., Bloomfield, J., Ward, R., and Jenkins, A.: Hydrological Outlook UK: an operational streamflow and groundwater level forecasting system at monthly to seasonal time scales, *Hydrolog. Sci. J.*, 62, 2753–2768, <https://doi.org/10.1080/02626667.2017.1395032>, 2017.
- Ramos, M. H., van Andel, S. J., and Pappenberger, F.: Do probabilistic forecasts lead to better decisions?, *Hydrol. Earth Syst. Sci.*, 17, 2219–2232, <https://doi.org/10.5194/hess-17-2219-2013>, 2013.
- Rapley, C. G., de Meyer, K., Carney, J., Clarke, R., Howarth, C., Smith, N., Stilgoe, J., Youngs, S., Brierley, C., Haugvaldstad, A., Lotto, B., Michie, S., Shipworth, M., and Tuckett, D.: Time for Change? Climate Science Reconsidered, Report of the UCL Policy Commission on Communicating Climate Science, 2014.
- Rodwell, M. J., Ferranti, L., Magnusson, L., Weisheimer, A., Rabier, F., and Richardson, D.: Diagnosis of northern hemispheric regime behaviour during winter 2013/14, *ECMWF Tech. Memo.*, 769, 1–12, 2015.
- Soares, M. B. and Dessai, S. J.: Exploring the use of seasonal climate forecasts in Europe through expert elicitation, *Climate Risk Management*, 10, 8–16, 2015.
- Soares, M. B. and Dessai, S. J.: Barriers and enablers to the use of seasonal climate forecasts amongst organisations in Europe, *Climatic Change*, 137, 89–103, <https://doi.org/10.1007/s10584-016-1671-8>, 2016.
- Soares, M. B., Alexander, M., and Dessai, S. J.: Sectoral use of climate information in Europe: A synoptic overview, *Climate Services*, 9, 5–20, 2018.
- Thames Water: Hydrological Context for Water Quality And Ecology Preliminary Impact Assessments, Technical Appendix B, Thames Water Utilities Ltd 2W0H Lower Thames Operating Agreement (Cascade Consulting), 2010.
- Turner, S. W. D., Bennett, J. C., Robertson, D. E., and Galelli, S.: Complex relationship between seasonal streamflow forecast skill and value in reservoir operations, *Hydrol. Earth Syst. Sci.*, 21, 4841–4859, <https://doi.org/10.5194/hess-21-4841-2017>, 2017.
- UK Gov: Water Situation Reports, available at: <https://www.gov.uk/government/collections/water-situation-reports-for-england>, last access: 5 May 2018.
- Van der Knijff, J. M., Younis, J., and De Roo, A. P. J.: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, *Int. J. Geogr. Inf. Sci.* 24, 189–212, <https://doi.org/10.1080/13658810802549154>, 2010.
- van den Hurk, B. J. J. M., Bouwer, L. M., Buontempo, C., Döschner, R., Ercin, E., Hananel, C., Hunink, J. E., Kjellström, E., Klein, B., Manez, M., Pappenberger, F., Pouget, L., Ramos, M.-H., Ward, P. J., Weerts, A. H., and Wijngaard, J. B.: Improving predictions and management of hydrological extremes through climate services, *Climate Services*, 1, 6–11, 2016.

- Vaughan, C., Buja, L., Kruczkiewicz, A., and Goddard, L.: Identifying research priorities to advance climate services, *Climate Services* 4, 65–74, 2016.
- Viel, C., Beaulant, A.-L., Soubeyroux, J.-M., and Céron, J.-P.: How seasonal forecast could help a decision maker: an example of climate service for water resource management, *Adv. Sci. Res.*, 13, 51–55, <https://doi.org/10.5194/asr-13-51-2016>, 2016.
- Wells, M. and Davis, H.: Water transfer for public water supply via the CRT canal network, presentation Black and Veatch, 2016.
- Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, *Geophys. Res. Lett.*, 35, L14401, <https://doi.org/10.1029/2008GL034648>, 2008.
- Yuan, X., Wood, E. F., and Ma, Z.: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, *WIREs Water*, 2, 523–536, <https://doi.org/10.1002/wat2.1088>, 2015.

A3: Skilful seasonal forecasts of streamflow over Europe?

This paper presents the published version of Chapter 3, Sect. 3.2 of this thesis, with the following reference:

Arnal, L., H. L. Cloke, E. Stephens, F. Wetterhall, C. Prudhomme, J. Neumann, B. Krzeminski and F. Pappenberger, 2018: Skilful seasonal forecasts of streamflow over Europe?, *Hydrol. Earth Syst. Sci.*, 22, 2057-2072, doi:10.5194/hess-22-2057-2018*

* ©2018. The Authors. Hydrology and Earth System Sciences, a journal of the European Geosciences Union published by Copernicus. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.



Skilful seasonal forecasts of streamflow over Europe?

Louise Arnal^{1,2}, Hannah L. Cloke^{1,3,4,5}, Elisabeth Stephens¹, Fredrik Wetterhall², Christel Prudhomme^{2,6,7}, Jessica Neumann¹, Blazej Krzeminski², and Florian Pappenberger²

¹Department of Geography and Environmental Science, University of Reading, Reading, RG6 6AB, UK

²European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG6 9AX, UK

³Department of Meteorology, University of Reading, Reading, RG6 6BB, UK

⁴Department of Earth Sciences, Uppsala University, Uppsala, 752 36, Sweden

⁵Centre of Natural Hazards and Disaster Science, CNDS, Uppsala, 752 36, Sweden

⁶Department of Geography, Loughborough University, Loughborough, LE11 3TU, UK

⁷NERC Centre for Ecology & Hydrology, Wallingford, OX10 8BB, UK

Correspondence: Louise Arnal (l.l.s.arnal@pgr.reading.ac.uk, louise.arnal@ecmwf.int)

Received: 10 October 2017 – Discussion started: 24 October 2017

Revised: 26 February 2018 – Accepted: 27 February 2018 – Published: 3 April 2018

Abstract. This paper considers whether there is any added value in using seasonal climate forecasts instead of historical meteorological observations for forecasting streamflow on seasonal timescales over Europe. A Europe-wide analysis of the skill of the newly operational EFAS (European Flood Awareness System) seasonal streamflow forecasts (produced by forcing the Lisflood model with the ECMWF System 4 seasonal climate forecasts), benchmarked against the ensemble streamflow prediction (ESP) forecasting approach (produced by forcing the Lisflood model with historical meteorological observations), is undertaken. The results suggest that, on average, the System 4 seasonal climate forecasts improve the streamflow predictability over historical meteorological observations for the first month of lead time only (in terms of hindcast accuracy, sharpness and overall performance). However, the predictability varies in space and time and is greater in winter and autumn. Parts of Europe additionally exhibit a longer predictability, up to 7 months of lead time, for certain months within a season. In terms of hindcast reliability, the EFAS seasonal streamflow hindcasts are on average less skilful than the ESP for all lead times. The results also highlight the potential usefulness of the EFAS seasonal streamflow forecasts for decision-making (measured in terms of the hindcast discrimination for the lower and upper terciles of the simulated streamflow). Although the ESP is the most potentially useful forecasting approach in Europe, the EFAS seasonal streamflow forecasts appear more potentially useful than the ESP in some regions and for certain

seasons, especially in winter for almost 40 % of Europe. Patterns in the EFAS seasonal streamflow hindcast skill are however not mirrored in the System 4 seasonal climate hindcasts, hinting at the need for a better understanding of the link between hydrological and meteorological variables on seasonal timescales, with the aim of improving climate-model-based seasonal streamflow forecasting.

1 Introduction

Seasonal streamflow forecasts predict the likelihood of a difference from normal conditions in the following months. Unlike forecasts at shorter timescales, which aim to predict individual events, seasonal streamflow forecasts aim at predicting long-term (i.e. weekly to seasonal) averages. The predictability in seasonal streamflow forecasts is driven by two components of the Earth system, the initial hydrological conditions (IHC; i.e. of snowpack, soil moisture, streamflow and reservoir levels, etc.) and large-scale climate patterns, such as the El Niño–Southern Oscillation (ENSO), the North Atlantic Oscillation (NAO), the Pacific–North American (PNA) pattern and the Indian Ocean Dipole (IOD) (Yuan et al., 2015b).

The first seasonal streamflow forecasting method, based on a regression technique developed around 1910–1911 in the United States, harnessed the predictability from accurate IHC of snowpacks to derive streamflow for the following

summer (Church, 1935). This statistical method recognized antecedent hydrological conditions and land surface memory as key drivers of streamflow generation for the following months.

Alongside the physical understanding of streamflow generation processes came technical developments, such as the creation of the first hydrological models and the acquisition of longer observed meteorological time series, which led to the creation of the first operational model-based seasonal streamflow forecasting system. This system, called extended streamflow prediction (ESP; i.e. note that ESP nowadays stands for ensemble streamflow prediction, although it refers to the same forecasting method), was developed by the United States National Weather Service (NWS) in the 1970s (Twedt et al., 1977; Day, 1985). The ESP forecasts are produced by forcing a hydrological model, initialized with the current IHC, with the observed historical meteorological time series available. The output is an ensemble streamflow forecast (where each year of historical data is a streamflow trace) for the following season(s) (Twedt et al., 1977; Day, 1985). The quality of the ESP forecasts can be high in basins where the IHC dominate the surface hydrological cycle for several months (the exact forecast quality depending on the time of year and the basin's physiographic characteristics; Wood and Lettenmaier, 2008).

In basins where the meteorological forcings drive the predictability, however, the lack of information on the future climate is a limitation of the ESP forecasting method and might result in unskilful ESP forecasts. This drawback led to the investigation of the use of seasonal climate forecasts, in place of the historical meteorological inputs, to feed hydrological models and extend the predictability of hydrological variables on seasonal timescales (Pagano and Garen, 2006). This investigation was made possible by technical and scientific advances. Scientifically, seasonal climate forecasts were improved greatly by the understanding of ocean–atmosphere–land interactions and the identification of large-scale climate patterns as drivers of the hydro-meteorological predictability (Goddard et al., 2001; Troccoli, 2010). This was technically implementable with the increase in computing resources, making it possible to run dynamical coupled ocean–atmosphere–land general circulation models on the global scale at high spatial and temporal resolutions (Doblas-Reyes et al., 2013). An additional technical challenge, the coarse spatial resolution of seasonal climate forecasts compared to the finer resolution of hydrological models, had to be addressed. To tackle this issue, many authors have explored different ways of downscaling climate variables for hydrological applications (Maraun et al., 2010, and references therein).

While climate-model-based seasonal streamflow forecasting experiments are more common outside of Europe, for example for the United States (Wood et al., 2002, 2005; Mo and Lettenmaier, 2014), Australia (Bennett et al., 2016), or Africa (Yuan et al., 2013), they remain limited in Europe, with a few examples in France (Céron et al., 2010; Singla et al., 2012;

Crochemore et al., 2016), in central Europe (Demirel et al., 2015; Meißner et al., 2017), in the United Kingdom (Bell et al., 2017; Prudhomme et al., 2017) and at the global scale (Yuan et al., 2015a; Candogan Yossef et al., 2017). This is because, although the quality of seasonal climate forecasts has increased over the past decades, there remains limited skill in seasonal climate forecasts for the extra-tropics, particularly for the variables of interest for hydrology, notably precipitation and temperature (Arribas et al., 2010; Doblas-Reyes et al., 2013).

In Europe, the NAO is one of the strongest predictability sources of seasonal climate forecasts; it is associated with changes in the surface westerlies over the North Atlantic and Europe, and hence with changes in temperature and precipitation patterns over Europe (Hurrell, 1995; Hurrell and Van Loon, 1997). It was shown to affect streamflow predictability, especially during winter (Dettinger and Diaz, 2000; Bierkens and van Beek, 2009; Steirou et al., 2017), in addition to the IHC and the land surface memory. It was furthermore shown to be an indicator of flood damage and occurrence in parts of Europe (Guimarães Nobre et al., 2017).

As the quality and usefulness of seasonal streamflow forecasts increase, their usability for decision-making has lagged behind. Translating the quality of a forecast into an added value for decision-making and incorporating new forecasting products into established decision-making chains are not easy tasks. This has been explored for many water-related applications, such as navigation (Meißner et al., 2017), reservoir management (Viel et al., 2016; Turner et al., 2017), drought-risk management (Sheffield et al., 2013; Yuan et al., 2013; Crochemore et al., 2017), irrigation (Chiew et al., 2003; Li et al., 2017), water resource management (Schepen et al., 2016) and hydropower (Hamlet et al., 2002), but seasonal streamflow forecasts have yet to be adopted by the flood preparedness community.

The European Flood Awareness System (EFAS) is at the forefront of seasonal streamflow forecasting, with one of the first operational pan-European seasonal hydrological forecasting systems. The aim of this paper is to bridge the current gap in pan-European climate-model-based seasonal streamflow forecasting studies. Firstly, the setup of the newly operational EFAS climate-based seasonal streamflow forecasting system is presented. A Europe-wide analysis of the skill of this forecasting system compared to the ESP forecasting approach is then presented, in order to identify whether there is any added value in using seasonal climate forecasts instead of historical meteorological observations for forecasting streamflow on seasonal timescales over Europe. Subsequently, the potential usefulness of the EFAS seasonal streamflow forecasts for decision-making is assessed.

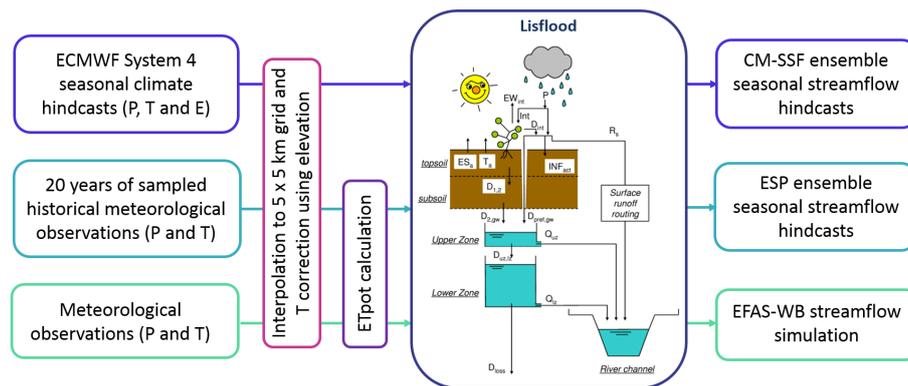


Figure 1. Schematic of the EFAS-WB streamflow simulation and of the CM-SSF and ESP seasonal streamflow hindcast generation, where P is precipitation, T is temperature, E is evaporation and ET_{pot} is potential evapotranspiration. The Lisflood model diagram was taken from Burek et al. (2013).

2 Data and methods

2.1 EFAS hydrological simulation and seasonal hindcasts

The data used in this paper include a streamflow simulation and two seasonal streamflow hindcasts (Fig. 1). Further information on these datasets is given below.

2.1.1 Hydrological modelling and streamflow simulation

The Lisflood model was used to produce all the simulations and hindcasts used in this paper. Lisflood is a GIS-based hydrological rainfall–runoff–routing distributed model written in the PCRaster Dynamic Modelling Language, which enables it to use spatially distributed maps (i.e. both static and dynamic) as input (De Roo et al., 2000; Van Der Knijff et al., 2010). The Lisflood model was calibrated to produce pan-European parameter maps. The calibration was performed for 693 basins from 1994 to 2002 using the Standard Particle Swarm Optimisation 2011 (SPSO-2011) algorithm. The calibration was carried out for parameters controlling snowmelt, infiltration, preferential bypass flow through the soil matrix, percolation to the lower groundwater zone, percolation to deeper groundwater zones, residence times in the soil and subsurface reservoirs, river routing and reservoir operations for a few basins. The results were validated with the Nash–Sutcliffe efficiency (NSE) for the validation period 2003–2012. In validation (calibration), Lisflood obtained a median NSE of 0.57 (0.62). Basins with large discrepancies between the observed and simulated flow statistics were situated mainly on the Iberian Peninsula and on the Baltic coasts (see Zajac et al., 2013, and Smith et al., 2016, for further details).

The Lisflood model is run operationally in EFAS, with the simulation domain covering Europe at a 5×5 km resolution. A reference simulation, called the EFAS water bal-

ance (EFAS-WB), is available on a daily time step starting from February 1990. Lisflood simulates the hydrological processes within a basin (most of which are mentioned above), starting from the previous day's IHC (e.g. snow cover, storage in the upper and lower zones, soil moisture, initial streamflow, reservoir filling) and forced with the most recent observed meteorological fields (i.e. of precipitation, potential evapotranspiration and temperature; provided by the EFAS meteorological data collection centres). The observed meteorological fields are daily maps of spatially interpolated point measurements of precipitation (from more than 6000 stations) and temperature (from more than 4000 stations) at the surface level. These same data are used to produce interpolated potential evapotranspiration maps from the Penman–Monteith method (Alfieri et al., 2014). All meteorological variables are interpolated on a 5×5 km grid using an inverse distance weighting scheme and the temperature is first corrected using the elevation (Smith et al., 2016).

The EFAS-WB is the best estimate of the hydrological state at a given time and for a given grid point in EFAS and is thus used as initial conditions from which the seasonal hydrological forecasts are started.

2.1.2 Ensemble seasonal streamflow hindcasts

In this paper, two types of ensemble seasonal streamflow hindcasts are used: the ensemble streamflow prediction (ESP) hindcast (hereafter referred to as ESP) and the System 4-driven seasonal streamflow hindcast (hereafter referred to as CM-SSF (climate-model-based seasonal streamflow forecast), following the notation from Yuan et al. (2015b)).

They are both initialized from the EFAS-WB, on the first day of each month, to produce a new ensemble streamflow forecast up to a lead time of 7 months (215 days), with a daily time step. Both hindcasts are generated from February 1990 for the same European domain as the EFAS-WB, at the same 5×5 km resolution. The unique difference between the

ESP and the CM-SSF is the meteorological forcing used to drive the hydrological model, described below.

The ESP is produced by driving the Lisflood model with 20 (the number of years of data available at the time the hindcast was produced) randomly sampled years of historical meteorological observations (i.e. the same as the meteorological observations used to produce the EFAS-WB, excluding the year of meteorological observations corresponding to the year that is being forecasted). A new 20-member ESP is thus generated at the beginning of each month and for the next 7 months.

The CM-SSF is produced by driving the Lisflood model with the ECMWF System 4 seasonal climate hindcast (Sys4, i.e. of precipitation, evaporation and temperature). Sys4 has a spatial horizontal resolution of about 0.7° (approximately 70 km). It is re-gridded to the Lisflood spatial resolution using an inverse distance weighting scheme and the temperature is first corrected using the elevation. Sys4 is made up of 15 ensemble members, extended to 51 every 3 months (Molteni et al., 2011). From 2011 onwards the Sys4 forecasts were run in real time and all contained 51 ensemble members. A new 15- to 51-member CM-SSF is hence produced at the beginning of each month and for the next 7 months. Operationally, the CM-SSF forecasts are currently used in EFAS to generate a seasonal streamflow outlook for Europe at the beginning of every month.

2.2 Hindcast evaluation strategy

For this study, monthly region specific discharge averages of the hindcasts (CM-SSF and ESP) and EFAS-WB were used. The specific discharge is the discharge per unit area of an upstream basin. For this paper, the gridded daily specific discharge was calculated by dividing the gridded daily discharge output maps (of the hindcasts and the EFAS-WB) by the Lisflood gridded upstream area static map. Subsequently, the gridded daily specific discharge maps were used to calculate daily region averaged specific discharges (for each region in Fig. 2) by summing up the daily specific discharge values of each grid cell within a region, divided by the number of grid cells in that region. Finally, monthly specific discharge region averages were calculated for each calendar month.

The regions displayed in Fig. 2 were created by merging several basins together (basins used operationally in EFAS for the shorter timescale forecasts), while respecting hydroclimatic boundaries. They were chosen for the analysis presented in this paper for two main reasons. Firstly, they are the regions used operationally to display the EFAS seasonal streamflow outlook. Secondly, they were created in order to capture large-scale variability in the weather.

The analysis of the hindcasts was performed on monthly specific discharge (hereafter referred to as streamflow) region averages for hindcast starting dates spanning February 1990 to November 2016 (included; approximately 27 years of data), with 1 to 7 months of lead time. In this paper, 1

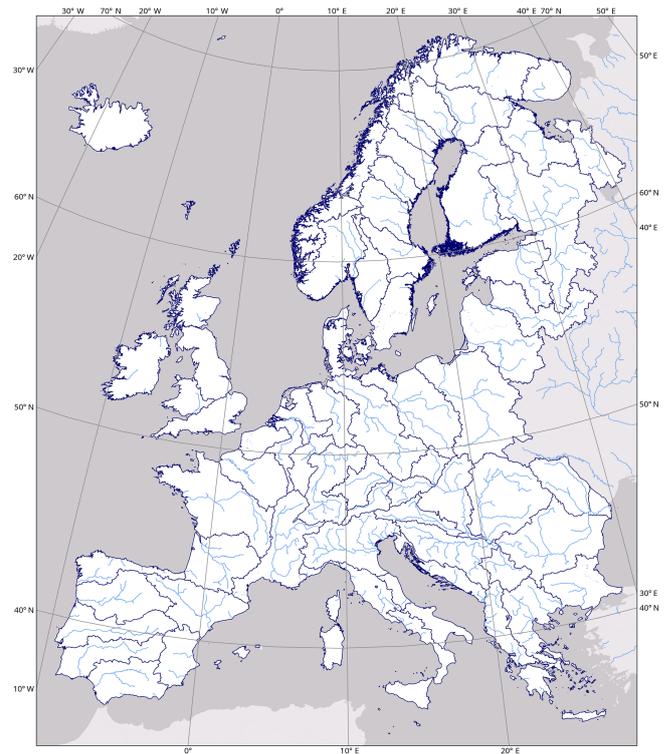


Figure 2. Map of the 74 European regions (dark blue outlines) selected for the analysis of the CM-SSF and the ESP.

month of lead time refers to the first month of the forecast (e.g. the January 2017 streamflow for a forecast made on 1 January 2017). Two months of lead time is the second month of the forecast (e.g. the February 2017 streamflow for a forecast made on 1 January 2017), etc. Monthly averages were selected for the analysis presented in this paper as it is a valuable aggregation time step for decision-makers for many water-related applications (as shown in the literature for applications such as, for example, navigation (Meißner et al., 2017), reservoir management (Viel et al., 2016; Turner et al., 2017), drought-risk management (Yuan et al., 2013), irrigation (Chiew et al., 2003; Li et al., 2017) and hydropower (Hamlet et al., 2002)).

Several verification scores were selected in order to assess the hindcasts' quality. These verification scores were chosen to cover a wide range of hindcast attributes (i.e. accuracy, sharpness, reliability, overall performance and discrimination). All of these verification scores, except for the verification score selected to look at hindcast discrimination, are the same as chosen in Crochemore et al. (2016), and are described below. The EFAS-WB streamflow simulations were used as a proxy for observation against which the seasonal streamflow hindcasts were evaluated, hence minimizing the impact of model errors on the hindcasts' quality.

2.2.1 Hindcast accuracy

Both hindcasts (CM-SSF and ESP) were assessed in terms of their accuracy, the magnitude of the errors between the hindcast ensemble mean and the “truth” (i.e. the EFAS-WB). For this purpose, the mean absolute error (MAE) was calculated for each region, target month (i.e. the month that is being forecast) and lead time (i.e. 1 to 7 months). The lower the MAE, the more accurate the hindcast.

2.2.2 Hindcast sharpness

Both hindcasts were also assessed in terms of their sharpness, an attribute of the hindcast only, which is a measure of the spread of the ensemble members of a hindcast. In this paper, the 90 % interquartile range (IQR; i.e. the difference between the 95th and 5th percentiles of the hindcast distribution) was calculated for each region, target month and lead time. The lower the IQR, the sharper the hindcast.

2.2.3 Hindcast reliability

Both hindcasts were additionally assessed in terms of their reliability, the statistical consistency between the hindcast probabilities and the observed frequencies. For this purpose, the probability integral transform (PIT) diagram was calculated for each region, target month and lead time (Gneiting et al., 2007). The PIT diagram is the cumulative distribution of the PIT values as a function of the PIT values. The PIT values measure where the “truth” (i.e. EFAS-WB) falls relative to the percentiles of the hindcast distribution. For a perfectly reliable hindcast, the “truth” should fall uniformly in each percentile of the hindcast distribution, giving a PIT diagram that falls exactly on the 1-to-1 diagonal. A hindcast that systematically under- (over-) predicts the “truth” will have a PIT diagram below (above) the diagonal. A hindcast that is too narrow (i.e. underdispersive; hindcast distribution smaller than the distribution of the observations) (large (i.e. overdispersive; hindcast distribution greater than the distribution of the observations)) will have a transposed S-shaped (S-shaped) PIT diagram (Laio and Tamea, 2007).

In order to compare the reliability across all regions, target months and lead times, the area between the PIT diagram and the 1-to-1 diagonal was computed for all PIT diagrams (Renard et al., 2010). The smaller this area, the more reliable the hindcast.

Furthermore, to disentangle the causes of poor reliability, the spread and bias of the hindcasts were calculated for all PIT diagrams, using two measures first introduced by Keller and Hense (2011): β -score and β -bias, respectively. By definition, a perfectly reliable hindcast (with regards to its spread) will have a β -score of zero (to which a tolerance interval of ± 0.09 was added), whereas a hindcast that is too narrow (large) will have a negative (positive) β -score (outside of the tolerance interval). A perfectly reliable hindcast

(with regards to its bias) will have a β -bias of zero (to which a tolerance interval of ± 0.09 was added), whereas a hindcast that systematically under- (over-) predicts the “truth” will have a negative (positive) β -bias (outside of the tolerance interval).

2.2.4 Hindcast overall performance

The hindcasts were furthermore assessed in terms of their overall performance from the continuous rank probability score (CRPS), calculated for each region, target month and lead time (Hersbach, 2000). The CRPS is a measure of the difference between the hindcast and the observed (i.e. EFAS-WB) cumulative distribution functions. The lower the CRPS, the better the overall performance of the hindcast.

In this paper, the skill of the CM-SSF is benchmarked with respect to the ESP in order to identify whether there is any added value in using Sys4 instead of historical meteorological observations for forecasting the streamflow on seasonal timescales over Europe. To this end, skill scores were calculated for the MAE, IQR, PIT diagram area and CRPS, using the following equation:

$$\text{Skill score} = 1 - \frac{\text{score}_{\text{CM-SSF}}}{\text{score}_{\text{ESP}}}. \quad (1)$$

Skill scores were calculated for each region, target month and lead time and will be referred to as MAESS, IQRSS, PITSS and CRPSS, respectively. Skill scores larger (smaller) than zero indicate more (less) skill in the CM-SSF compared to the ESP. A skill score of zero means that the CM-SSF is as skilful as the ESP. Note that as the ESP is not a “naive” forecast, using it as a benchmark might lead to lower skill than benchmarking the CM-SSF against, for example, climatology.

2.2.5 Hindcast potential usefulness

For decision-making, the ability of a seasonal forecasting system to predict the right category of an event (e.g. above or below normal conditions) months ahead is of great importance (Gobena and Gan, 2010). In this paper, the potential usefulness of the CM-SSF and the ESP to forecast lower and higher than normal streamflow conditions within their hindcasts is assessed.

To do so, the relative operating characteristic (ROC) score, a measure of hindcast discrimination (Mason and Graham, 1999), was calculated. The thresholds selected to calculate the ROC are the lower and upper terciles of the EFAS-WB climatology for each season. They were calculated for the simulation period (February 1990 to May 2017), by grouping together EFAS-WB monthly streamflows for each month falling in a season (SON: September–October–November, DJF: December–January–February, MAM: March–April–May and JJA: June–July–August). For each season and each region a lower and upper tercile streamflow value was ob-

tained, subsequently used as thresholds against which to calculate the probability of detection (POD) and the false alarm rate (FAR; with 0.1 probability bins) for both hindcasts, and for each region, season and lead time. Finally, the area under the ROC curve, i.e. the ROC score, was calculated for both hindcasts, for each region, season and lead time. The ROC score ranges from 0 to 1, with a perfect score of 1. A hindcast with a ROC score ≤ 0.5 is unskilful, i.e. less good than the long-term average climatology which has a ROC of 0.5, and is therefore not useful.

Because the ROC score was calculated from a low number of events (i.e. approximately $27 \text{ years} \times 3 \text{ months}$ in each season $\times 1/3$ (lower or upper tercile) = 27 simulated events), the hindcasts were judged skilful and useful when their ROC score ≥ 0.6 instead of 0.5. Moreover, the CM-SSF was categorized as more useful than the ESP when the CM-SSF's ROC score was at least 10 % larger than the ESP's ROC score.

3 Results

3.1 Overall skill of the CM-SSF

In the first part of the results, the skill of the CM-SSF (benchmarked with respect to the ESP) is presented, in terms of the accuracy (MAESS), sharpness (IQRSS), reliability (PITSS) and overall performance (CRPSS) in the hindcast datasets. This will benchmark the added value of using Sys4 against the use of historical meteorological observations for forecasting the streamflow on seasonal timescales over Europe.

As shown by the MAESS boxplots (Fig. 3), the CM-SSF appears on average more accurate than the ESP for the first month of lead time only, for all seasons excluding spring (MAM). Beyond 1 month of lead time, the CM-SSF becomes on average as or less accurate than the ESP. There are however noticeable differences between the different seasons. The CM-SSF shows the largest improvements in the average accuracy compared to the ESP in winter (DJF) and for the first month of lead time. For longer lead times (i.e. 2 to 7 months), the accuracy of the CM-SSF is on average quite similar to that of the ESP in autumn (SON) and winter, and on average lower in spring and summer (JJA). The boxplots for the CRPSS look very similar to the MAESS boxplots, the main difference being the lower average scores for 2 to 7 months of lead time in autumn and winter (Fig. 3).

The boxplots of the IQRSS show that the CM-SSF predictions are on average as sharp as those of the ESP for the first month of lead time (slightly sharper in autumn; Fig. 3). For 2 to 7 months of lead time, in autumn and winter, the CM-SSF predictions are on average sharper than those of the ESP, whereas in spring and summer, the CM-SSF predictions are on average slightly less sharp than the ESP predictions.

As shown by the boxplots of the PITSS (Fig. 3), the CM-SSF predictions are less reliable than the ESP predictions for

all seasons and months of lead time. For the first month of lead time and all seasons, 10–20 % of the ESP hindcasts and less than 5 % of the CM-SSF hindcasts are reliable (Fig. 4). About 40–60 % of the ESP hindcasts are not reliable for the first month of lead time and all seasons due to the ensemble spread. Approximately half of these hindcasts are too large, while the other half (slightly more in autumn and winter) are too narrow. Furthermore, 50–80 % of the ESP hindcasts under-predict the simulated streamflow for the first month of lead time and all seasons. The percentage of reliable (unreliable) ESP hindcasts increases (decreases) with lead time, as the effect of the IHC fades away. About 70–90 % of the CM-SSF hindcasts are too narrow for the first month of lead time and all seasons. With increasing lead time, the percentage of CM-SSF hindcasts that are too narrow (large) decreases (increases), especially in spring. Approximately 40–50 % of the CM-SSF hindcasts over-predict the simulated streamflow in spring and summer for the first month of lead time (and increasingly over-predict with longer lead times). In autumn and winter, about 70 % of the CM-SSF hindcasts under-predict the simulated streamflow for the first month of lead time (and increasingly under-predict with longer lead times).

For all verification scores, the boxplots for autumn and winter are slightly smaller than for spring and summer, hinting at a smaller variability in the verification scores amongst regions and target months in autumn and winter than in spring and summer. Furthermore, the presence of the boxplots above the zero line (i.e. no skill line) for all lead times suggests that the CM-SSF is more skilful than the ESP for some regions and target months, beyond the first month of lead time.

3.2 Potential usefulness of the CM-SSF

In the second part of the results, the potential usefulness of the CM-SSF compared to the ESP is described for decision-making. Here, potential usefulness is defined as the ability of the forecasting systems to predict lower or higher streamflows than normal, as measured with the ROC score.

Generally, either of the two forecasting systems (CM-SSF or ESP) is capable of predicting skilfully whether the streamflow will be anomalously low or high in the coming months (Fig. 5). However, for a few seasons and regions, none of the two forecasting systems is skilful at predicting lower and/or higher streamflows than normal. This is especially noticeable in winter.

For most seasons and regions, the ESP is more skilful than the CM-SSF at predicting lower and higher streamflows than normal. However, in winter for most regions and during other seasons for several regions, the CM-SSF appears more skilful than the ESP. Regions where the CM-SSF best predicts lower and higher streamflows than normal at most lead times are summarized in Table 1 for all four seasons and the lower and upper terciles of the simulated streamflow.

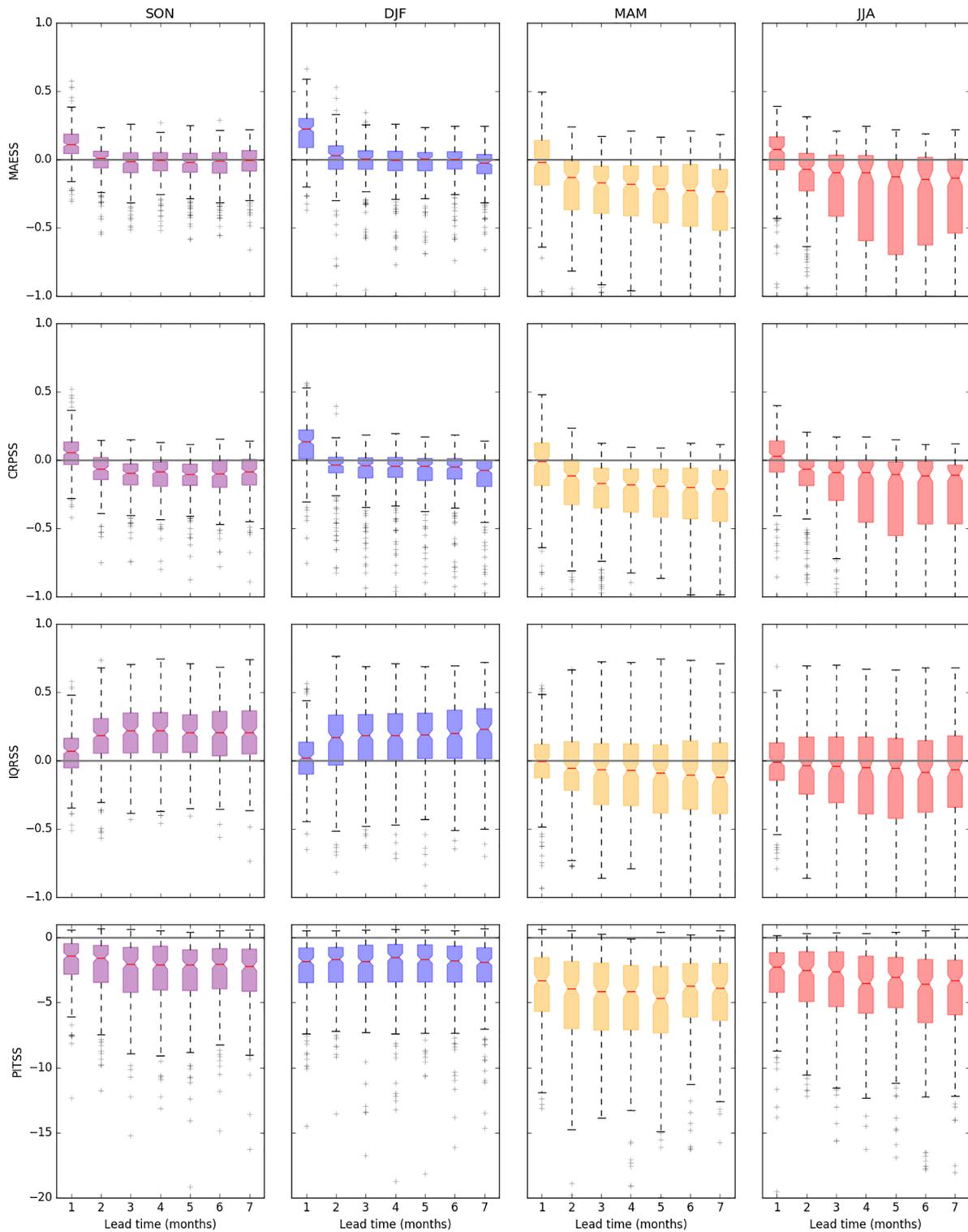


Figure 3. Boxplots of the MAESS, CRPSS, IQRSS and PITSS (from the top to bottom rows) for all four seasons (SON, DJF, MAM and JJA from the left-most to right-most columns) as a function of lead time (i.e. 1 to 7 months). The boxplots contain the scores for all target months falling in a given season and all 74 European regions. For all scores, values larger (smaller) than zero indicate that the CM-SSF is more (less) skilful than the ESP (benchmark). Where the skill is zero, the CM-SSF is as skilful as the ESP for the hindcast period. Note that the PITSS plots have a different y-axis scale.

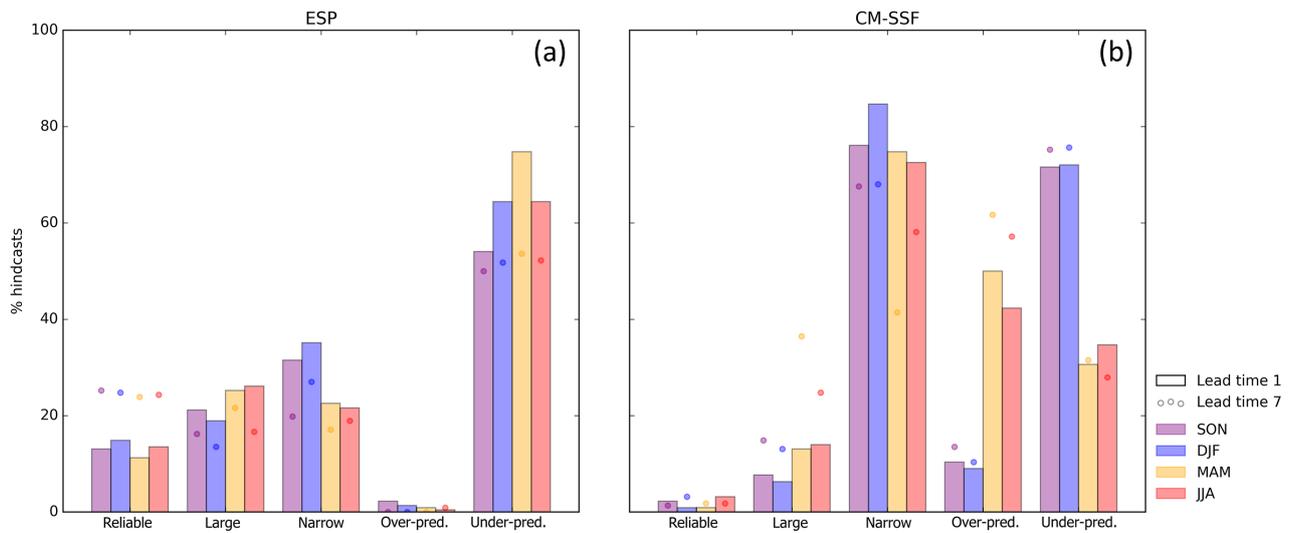


Figure 4. Plots of the percentage of the ESP (a) and the CM-SSF (b) hindcasts falling into each reliability category (reliable – in terms of both spread and bias, too large, too narrow, over-predicting and under-predicting) for all four seasons (SON, DJF, MAM and JJA from the left-most to right-most bars in each reliability category). The results are shown as bar charts for the first month of lead time and as circles for the seventh month of lead time. These lead times were selected for display to highlight the evolution of reliability between the first and last months of the hindcast. The percentages were calculated from hindcasts for all target months falling in a given season and all 74 European regions.

Table 1. Regions where the CM-SSF is more skilful than the ESP at predicting anomalously low (lower tercile; first column) or high (upper tercile; second column) streamflows for all four seasons (SON, DJF, MAM and JJA from the top to bottom rows). This is a summary of the information displayed in Fig. 5.

	Lower tercile	Upper tercile
SON	<ul style="list-style-type: none"> – Few regions in Fennoscandia – Po River basin (northern Italy) – Elbe River basin (south of Denmark) – Upstream of the Rhine River basin – Upstream of the Danube River basin – Duero River basin (Iberian Peninsula) 	<ul style="list-style-type: none"> – Few regions in Fennoscandia – Iceland – Parts of the Danube River basin – Segura River basin (Iberian Peninsula)
DJF	<ul style="list-style-type: none"> Many regions except – in most of Fennoscandia north of the Baltic Sea, – parts of central Europe. 	Same as lower tercile.
MAM	<ul style="list-style-type: none"> – Few regions on the Iberian Peninsula – Few regions in the western part of central Europe 	Same as lower tercile.
JJA	<ul style="list-style-type: none"> – Few regions in the United Kingdom (UK) – Ireland – North-western edge of the Iberian Peninsula – Regions in Fennoscandia around the Baltic Sea – Regions south of the North Sea 	<ul style="list-style-type: none"> – Northern part of the UK – Ireland – North-western edge of the Iberian Peninsula – Regions in Fennoscandia around the Baltic Sea – Around the Elbe River basin – Upstream of the Danube River basin – Along the Adriatic Sea in Italy

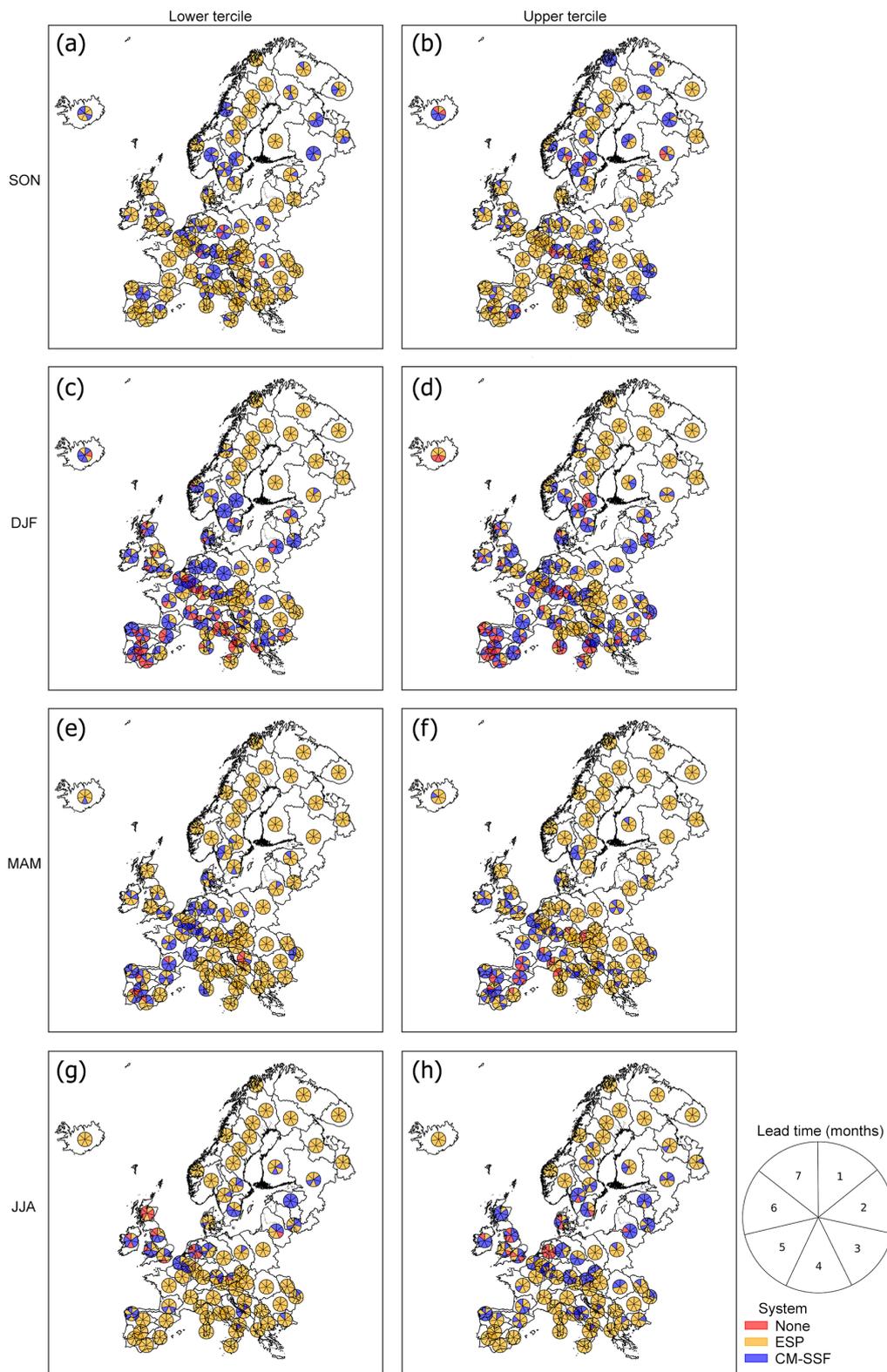


Figure 5. Maps of the best system (as measured with the ROC score) for all four seasons (SON, DJF, MAM and JJA) and the lower and upper simulated streamflow seasonal terciles (left-most and right-most columns, respectively) in each region from (a) to (h). The pie charts display the best system for each lead time (i.e. 1 to 7 months), as shown in the example pie chart on the bottom right of this figure. There are three possible cases: (1) neither the ESP nor the CM-SSF is skilful (red colours), (2) the ESP is skilful and better than the CM-SSF (yellow colours), and (3) the CM-SSF is skilful and better than the ESP (blue colours).

4 Discussion

4.1 Does seasonal climate information improve the predictability of seasonal streamflow forecasts over Europe?

On average over Europe and across all seasons, the CM-SSF is skilful (in terms of hindcast accuracy, sharpness and overall performance, using the ESP as a benchmark) for the first month of lead time only. This means that, on average, Sys4 improves the predictability over historical meteorological information for pan-European seasonal streamflow forecasting for the first month of lead time only. At longer lead times, historical meteorological information becomes as good as or better than Sys4 for seasonal streamflow forecasting over Europe. Crochemore et al. (2016) and Meißner et al. (2017) similarly found positive skill in the seasonal streamflow forecast (Sys4 forced hydrological model compared to an ESP) for the first month of lead time, after which the skill faded away for basins in France and central Europe, respectively. Additionally, on average over Europe and across all seasons, the CM-SSF is less reliable than the ESP for all lead times. This is due to a combination of too narrow and biased CM-SSF hindcasts, where the bias depends on the season that is being forecasted. As mentioned in the methods section of this paper, the ESP is not a “naive” benchmark, which might partially explain the limited predictability gained from Sys4.

The predictability varies per season and the CM-SSF predictions are on average sharper than and as accurate as the ESP predictions in autumn and winter beyond the first month of lead time (and increasingly sharper with longer lead times). The CM-SSF however tends to systematically under-predict the autumn and winter simulated streamflow (and increasingly under-predicts with longer lead times). In spring and summer, the CM-SSF predictions are on average less sharp and less accurate than the ESP predictions, and they tend to systematically over-predict the simulated streamflow (and increasingly over-predict with longer lead times).

The added predictability gained from Sys4 was shown to lead to skilful CM-SSF predictions of lower and higher streamflows than normal for specific seasons and regions. The CM-SSF is more skilful at predicting anomalously low and high streamflows than the ESP in certain seasons and regions, and noticeably in winter in almost 40 % of the European regions, mostly clustered in rainfall-dominated areas of western and central Europe. Several authors have discussed the higher winter predictability over (parts of) Europe, with examples in basins in France (Crochemore et al., 2016), central Europe (Steirou et al., 2017), the UK (Bell et al., 2017) and the Iberian Peninsula (Lorenzo-Lacruz et al., 2011). Bierkens and van Beek (2009) additionally showed that there was a higher winter predictability in Scandinavia, the Iberian Peninsula and around the Black Sea. Our results are mostly consistent with these findings, except for Scandinavia, where the ESP is more skilful than the CM-SSF

in winter. Bierkens and van Beek (2009) produced the seasonal streamflow forecast analysed in their paper by forcing a hydrological model with resampled years of historical meteorological information based on their winter NAO index. However, Sys4 has difficulties in forecasting the NAO over Europe (Kim et al., 2012), which could have led to these inconsistent results with the ones presented by Bierkens and van Beek (2009).

In spring, the CM-SSF is more skilful than the ESP at predicting lower and higher streamflows than normal beyond 1 month of lead time in approximately 15 % of the European regions, and mostly in regions of western Europe. This could be due to a persistence of the skill from the previous winter through the land surface memory (i.e. groundwater-driven streamflow or snowmelt-driven streamflow), as highlighted by Bierkens and van Beek (2009) for Europe, Singla et al. (2012) for parts of France, Lorenzo-Lacruz et al. (2011) for the Iberian Peninsula and Meißner et al. (2017) for the Rhine. Moreover, it could be that most of the gained predictability occurs in March, a transition month between the more predictable winter (as mentioned above) and spring, as discussed by Steirou et al. (2017). The ESP is overall more skilful than the CM-SSF at predicting the spring streamflow in snow-dominated regions (e.g. most of Fennoscandia and parts of central and eastern Europe). This hints at the importance of the IHC (i.e. of snowpack) and the land surface memory for forecasting the spring streamflow in snow-dominated regions in Europe.

The added predictability from Sys4 for forecasting lower and higher streamflows than normal is limited in summer and autumn for most regions. The CM-SSF is more skilful at predicting anomalously low and high streamflows than the ESP in about 10–20 % of the European regions during those seasons. Other studies have found similar patterns for (parts of) Europe; these include less skill in summer than in winter overall for basins in France (Crochemore et al., 2016), less skill for the low flow season (July to October) for basins in central Europe (Meißner et al., 2017), negative correlations in summer and autumn seasonal streamflow forecasts in central Europe as the influence of the winter NAO fades away (Steirou et al., 2017), and less skill overall in summer than in winter in Europe (Bierkens and van Beek, 2009). The lower CM-SSF skill for predicting lower and higher streamflows than normal in summer could additionally be due to the convective storms in summer over Europe, which are hard to predict, and to the fact that it is the dry season in most of Europe, where rivers are groundwater fed. Therefore, in this season, the quality of the IHC controls the streamflow predictability.

While the CM-SSF is most skilful (in terms of hindcast accuracy, sharpness and overall performance, using the ESP as a benchmark) in autumn and winter and most potentially useful in winter, this does not appear to correlate with high performance in the Sys4 precipitation and temperature hindcasts (as seen on the maps of correlation for Sys4 precipitation and temperature for all four seasons (SON, DJF,

MAM and JJA) and with 2 months of lead time (as identified in this paper); available at https://meteoswiss.shinyapps.io/skill_metrics/, Forecast skill metrics, 2017). Over Europe, the Sys4 precipitation and temperature hindcasts are the most skilful in summer and the least skilful in autumn and winter. Moreover, the regions of high CM-SSF skill for predicting lower and upper streamflows than normal do not clearly correspond to regions of high performance in the Sys4 precipitation and temperature hindcasts. These differences could be partially induced by the different benchmark used to evaluate the skill of the CM-SSF (i.e. the ESP) compared to the one used to look at the performance of the Sys4 precipitation and temperature hindcasts (i.e. ERA-Interim). However, these results clearly indicate that looking at the performance of the Sys4 precipitation and temperature hindcasts only does not give a good indication of the skill and potential usefulness of the seasonal streamflow hindcasts over Europe, and that marginal performance in seasonal climate forecasts can translate through to more predictable seasonal streamflow forecasts, and vice versa. The added predictability in the CM-SSF could be due to the combined predictability in the precipitation and temperature hindcasts, as well as a lag in the predictability from the land surface memory.

In most regions and for most seasons, at least one of the two forecasting systems (CM-SSF or ESP) is able to predict lower or higher streamflows than normal. However, in winter, the number of regions and lead times for which none of the forecasting systems are skilful increases. This could be because in winter, many regions experience weather-driven high streamflows and the performance of Sys4 is limited at this time of year (as mentioned above). In those regions, the seasonal streamflow forecasts could be improved either by improving the IHC, through for example data assimilation, or by improving the seasonal climate forecasts.

Overall, the ESP appears very skilful at forecasting lower or higher streamflows than normal, showing the importance of IHC and the land surface memory for seasonal streamflow forecasting (Wood and Lettenmaier, 2008; Bierkens and van Beek, 2009; Yuan et al., 2015b).

4.2 What is the potential usefulness and usability of the EFAS seasonal streamflow forecasts for flood preparedness?

What appears like little added skill does not necessarily mean no skill for the forecast users and can in fact be a large added value for decision-making (Viel et al., 2016). The ability of a seasonal streamflow forecasting system to predict the right category of an event months ahead is valuable for many water-related applications (e.g. navigation, reservoir management, drought-risk management, irrigation, water resource management, hydropower and flood preparedness). From the results presented in this paper, it appears that either of the two forecasting systems (CM-SSF or ESP) is capable of predicting lower or higher streamflows than normal

months in advance, thanks to the predictability gained from the IHC, the land surface memory and the seasonal climate hindcast in some regions and for certain seasons.

However, as highlighted by White et al. (2017), there is currently a gap between usefulness and usability of seasonal information. What is a useful scientific finding does not automatically translate into usable information which will fit into any user's decision-making chain (Soares and Dessai, 2016). While several authors have already investigated the usability of seasonal streamflow forecasts for applications such as navigation (Meißner et al., 2017), reservoir management (Viel et al., 2016; Turner et al., 2017), drought-risk management (Sheffield et al., 2013; Yuan et al., 2013; Crochemore et al., 2017), irrigation (Chiew et al., 2003; Li et al., 2017), water resource management (Schepen et al., 2016) and hydropower (Hamlet et al., 2002), its application to flood preparedness is still left mostly unexplored. One exception being Neumann et al. (in review), who look at the use of the CM-SSF to predict the 2013/14 Thames basin floods. This is partially due to the complex nature of flood generating mechanisms, still poorly studied on seasonal timescales beyond snowmelt-driven spring floods, as well as the fact that seasonal forecasts reflect the likelihood of abnormal seasonal streamflow totals, but without much skilful information on the exact timing, location and severity of the impact of individual flood events within that season. Coughlan de Perez et al. (2017) looked at the usefulness of seasonal rainfall forecasts for flood preparedness in Africa and highlighted the complexities behind using these forecasts as a proxy for floodiness (for a discussion on floodiness, see Stephens et al., 2015). Furthermore, decision-makers in the navigation, reservoir management, drought-risk management, irrigation, water resource management and hydropower sectors are familiar with working on long timescales (i.e. several weeks to months ahead). In contrast, the flood preparedness community is currently mostly used to working on timescales of hours to a couple of days.

The Red Cross Red Crescent Climate Centre has recently designed a new approach that harnesses the usefulness of seasonal climate information for decision-making for disaster management. This approach, called "Ready-Set-Go!", is made up of three stages. The "Ready" stage is based on seasonal forecasts, where they are used as monitoring information to drive contingency planning (e.g. volunteer training). The "Set" stage is triggered by sub-seasonal forecasts, used as early-warning information to alert volunteers. Finally, the "Go!" stage is based on short-range forecasts and consists in the evacuation of people and the distribution of aid (White et al., 2017). Using a similar approach, seasonal streamflow forecasts could complement existing forecasts at shorter timescales and provide monitoring and early-warning information for flood preparedness. Such an approach however requires the use of consistent forecasts from short to seasonal timescales. In this context, moving to seamless forecasting is becoming vital (Wetterhall and Di Giuseppe, in review).

Soares and Dessai (2016) also identified the accessibility to the information, enhanced by collaborations and ongoing relationships between users and producers, as a key enabler of the usability of seasonal information. International projects, such as the Horizon 2020 IMPREX (IMproving PRedictions and management of hydrological EXtremes) project (van den Hurk et al., 2016), alongside promoting scientific progress on hydrological extremes forecasting from short to seasonal timescales over Europe, gather together forecasters and decision-makers and can effectively demonstrate the added value of the integration of seasonal information in decision-making chains. The Hydrologic Ensemble Prediction EXperiment (HEPEX) is another international initiative that brings together researchers and practitioners in the field of ensemble prediction for water-related applications. It is an ideal environment for collaboration and fosters communication and outreach on topics such as the usefulness and usability of seasonal information for decision-making.

4.3 Aspects for future work

In this paper, terciles of the simulated streamflow are used. However, and because the application of the EFAS seasonal streamflow forecasts is of particular relevance for flood preparedness, the evaluation of the hindcasts for lower and higher streamflow extremes (for example the 5th and 95th percentiles, respectively) would be more relevant and might give very different results. This was not done in this paper as the time period covered by the seasonal streamflow hindcasts (i.e. approximately 27 years) was not long enough for statistically reliable results for lower and higher streamflow extremes. The limited hindcast length is a common problem in seasonal predictability studies. Increasing the hindcast length back in time could lead to more stable Sys4 hindcasts and hence to more stable and potentially skilful seasonal streamflow hindcasts (Shi et al., 2015).

Furthermore, in this paper, the hindcasts were analysed against simulated streamflow, used as a proxy for observed streamflow. This is necessary because it enables an analysis of the quality of the hindcasts over the entire computation domain, rather than at non-evenly spaced stations over the same domain (Alfieri et al., 2014). Further work could however include carrying out a similar analysis for selected river stations in Europe, in order to account for model errors in the hindcast evaluation.

The calculation of the verification scores (excluding the ROC) was made by randomly selecting 15 ensemble members from the 51 ensemble members of the CM-SSF hindcasts, for starting dates for which the ensemble varies between 15 and 51 members (i.e. hindcasts made on 1 January, March, April, June, July, September, October and December; this is due to the split between 15 and 51 ensemble members in the Sys4 hindcasts, as described in Sect. 2.1.2 of this paper). In order to investigate the potential impact of this evaluation strategy on the results presented in this paper, the

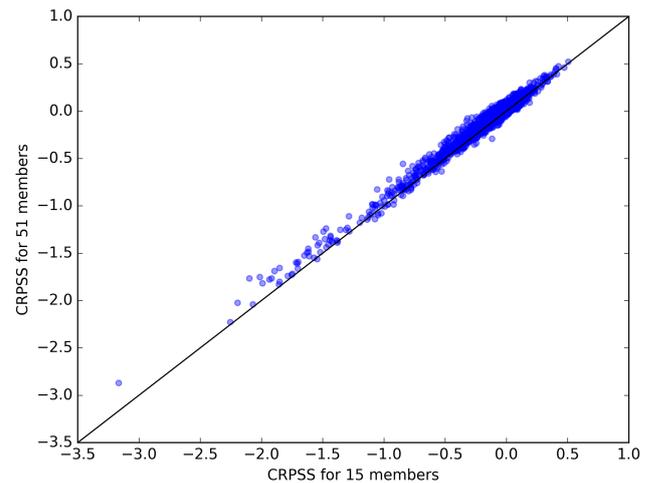


Figure 6. CRPSS calculated for the CM-SSF against the ESP (benchmark) for hindcasts made on 1 February, May, August and November, all lead times (i.e. 1 to 7 months) and all 74 European regions. The *x*-axis (*y*-axis) contains the CRPSS calculated from 15 (all 51) ensemble members of the CM-SSF.

CRPSS was calculated for 15 and 51 ensemble members of the CM-SSF hindcasts for starting dates for which 51 ensemble members are available for the full hindcast period (i.e. hindcasts made on 1 February, May, August and November). This is displayed in Fig. 6 for all hindcast starting dates, lead times (i.e. 1 to 7 months) and regions combined. Overall, it is apparent that the impact of this evaluation strategy on the results presented in this paper should be minimal, as all points align themselves approximately with the 1-to-1 diagonal.

The next version of the ECMWF seasonal climate forecast, SEAS5, was released in November 2017. Future work could include forcing the Lisflood model with SEAS5 and comparing the obtained seasonal streamflow hindcasts to the CM-SSF presented in this paper. This should indicate whether developments to the seasonal climate forecast translate through to better pan-European seasonal streamflow forecasts, which is of particular interest for regions and seasons when neither the ESP nor the CM-SSF is currently skilful.

The operational EFAS medium-range streamflow forecasts are currently post-processed as a means to improve their reliability (Smith et al., 2016, and references therein). Results from this paper have shown that the CM-SSF is mostly unreliable (with regards to the EFAS-WB) and could hence benefit from post-processing of the seasonal climate forecast. However, post-processing techniques used for the EFAS medium-range streamflow forecasts might not be suitable for the CM-SSF, as the seasonal climate forecast used for the latter should be post-processed in terms of its seasonal anomalies rather than for errors in the timing, volume and magnitude of specific events. This is currently being considered for operational implementation within EFAS and is an active area of discussion within the EFAS user community.

For the analysis presented in this paper, the CM-SSF was benchmarked against the ESP. Several other techniques exist for seasonal streamflow forecasting, such as statistical methods using predictors ranging from climate indices to antecedent observed precipitation and crop production metrics, to mention a few (e.g. Mendoza et al., 2017; Slater et al., 2017). Further analysis could include benchmarking the CM-SSF against one or multiple statistical methods, to assess the relative benefits of various seasonal streamflow forecasting techniques.

In this paper, the ability of both systems (CM-SSF and ESP) to forecast lower and higher streamflows than normal was explored, with several hypotheses made to link the streamflow predictability to regions' hydro-climatic processes. This includes the higher potential usefulness of the ESP in forecasting the spring streamflow in snow-dominated regions and the summer streamflow in regions where rivers are groundwater fed. In these regions and for these seasons, the IHC and the land surface memory drive the predictability. The CM-SSF provides an added potential usefulness in winter in the rainfall-dominated regions of central and western Europe, where the skill appears to persist through to spring due to the land surface memory (i.e. groundwater-driven streamflow and snowmelt-drive streamflow). While further exploration of these hypotheses is outside of the scope of this paper, future work is required to disentangle the links between the added predictability from Sys4 and the basins' hydro-climatic characteristics, for example, understanding the predictability in snow-dominated basins, arid regions and temperate groundwater-fed basins.

In this context, additional work to further disentangle and quantify the contribution of both predictability sources (seasonal climate forecasts versus IHC) to seasonal streamflow forecasting quality over Europe could be carried out by using the EPB (end point blending) method (Arnal et al., 2017).

5 Conclusions

In this paper, the newly operational EFAS seasonal streamflow forecasting system (producing the CM-SSF forecasts by forcing the Lisflood model with the ECMWF System 4 seasonal climate forecasts (Sys4)) was presented and benchmarked against the ESP forecasting approach (ESP forecasts produced by forcing the Lisflood model with historical meteorological observations) for the hindcast period 1990 to 2017. On average, Sys4 improves the predictability over historical meteorological information for pan-European seasonal streamflow forecasting for the first month of lead time only (in terms of hindcast accuracy, sharpness and overall performance). However, the predictability varies per season and the CM-SSF is more skilful on average at predicting autumn and winter streamflows than spring and summer streamflows. Additionally, parts of Europe exhibit a longer predictability, up to 7 months of lead time, for certain months

within a season. In terms of hindcast reliability, the CM-SSF is on average less skilful than the ESP for all lead times, due to a combination of too narrow and biased CM-SSF hindcasts, where the bias depends on the season that is being forecasted.

Subsequently, the potential usefulness of the two forecasting systems (CM-SSF and ESP) was assessed by analysing their skill in predicting lower and higher streamflows than normal. Overall, at least one of the two forecasting systems is capable of predicting those events months in advance. The ESP appears the most skilful on average, showing the importance of IHC and the land surface memory for seasonal streamflow forecasting. Nevertheless, for certain regions and seasons the CM-SSF is the most skilful at predicting anomalously low or high streamflows beyond 1 month of lead time, noticeably in winter for almost 40 % of the European regions. This potential usefulness could be harnessed by using seasonal streamflow forecasts as complementary information to existing forecasts at shorter timescales, to provide monitoring and early-warning information for flood preparedness.

Overall, patterns in skill in the CM-SSF are however not mirrored in the Sys4 precipitation and temperature hindcasts. This suggests that using seasonal climate forecast performance as a proxy for seasonal streamflow forecasting skill is not adequate and that more work is needed to understand the link between meteorological and hydrological variables on seasonal timescales over Europe.

Data availability. The data from the European Flood Awareness System are available to researchers upon request (subject to licensing conditions). Please visit www.efas.eu for more details.

Competing interests. The authors declare that they have no conflict of interest.

Special issue statement. This article is part of the special issue "Sub-seasonal to seasonal hydrological forecasting". It is not associated with a conference.

Acknowledgements. Louise Arnal, Hannah L. Cloke and Jessica Neumann gratefully acknowledge financial support from the Horizon 2020 IMPREX project (grant agreement 641811) (project IMPREX: www.imprex.eu). Louise Arnal's time was additionally partly funded by a University of Reading PhD scholarship. Fredrik Wetterhall, Christel Prudhomme and Blazej Krzeminski's work was supported by the EFAS computational centre in support to the Copernicus Management Service/Early Warning Systems (Flood) (contract no. 198702 from JRC-IES). Elisabeth Stephens is thankful for support from the Natural Environment Research Council and Department for International Development (grant number NE/P000525/1) under the Science for Humanitarian Emergencies and Resilience (SHEAR) research programme.

Edited by: Ilias Pechlivanidis

Reviewed by: two anonymous referees

References

- Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., and Salamon, P.: Evaluation of ensemble streamflow predictions in Europe, *J. Hydrol.*, 517, 913–922, <https://doi.org/10.1016/j.jhydrol.2014.06.035>, 2014.
- Arnal, L., Wood, A. W., Stephens, E., Cloke, H. L., and Pappenberger, F.: An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity, *J. Hydrometeorol.*, 18, 1715–1729, <https://doi.org/10.1175/JHM-D-16-0259.1>, 2017.
- Arribas, A., Glover, M., Maidens, A., Peterson, K., Gordon, M., MacLachlan, C., Graham, R., Fereday, D., Camp, J., Scaife, A. A., Xavier, P., McLean, P., and Colman, A.: The GloSea4 Ensemble Prediction System for Seasonal Forecasting, *Mon. Weather. Rev.*, 139, 1891–1910, <https://doi.org/10.1175/2010MWR3615.1>, 2010.
- Bell, V. A., Davies, H. N., Kay, A. L., Brookshaw, A., and Scaife, A. A.: A national-scale seasonal hydrological forecast system: development and evaluation over Britain, *Hydrol. Earth Syst. Sci.*, 21, 4681–4691, <https://doi.org/10.5194/hess-21-4681-2017>, 2017.
- Bennett, J. C., Wang, J. Q., Li, M., Robertson, D. E., and Schepen, A.: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model, *Water Resour. Res.*, 52, 8238–8259, <https://doi.org/10.1002/2016WR019193>, 2016.
- Bierkens, M. F. and van Beek, L. P.: Seasonal Predictability of European Discharge: NAO and Hydrological Response Time, *J. Hydrometeorol.*, 10, 953–968, <https://doi.org/10.1175/2009JHM1034.1>, 2009.
- Burek, P., Van Der Knijff, J. M., and De Roo, A.: LISFLOOD – Distributed Water Balance and Flood Simulation Model – Revised User Manual 2013, EUR – Scientific and Technical Research Reports, Publications Office of the European Union, Luxembourg, 150 pp., <https://doi.org/10.2788/24719>, 2013.
- Candogan Yossef, N., van Beek, R., Weerts, A., Winsemius, H., and Bierkens, M. F. P.: Skill of a global forecasting system in seasonal ensemble streamflow prediction, *Hydrol. Earth Syst. Sci.*, 21, 4103–4114, <https://doi.org/10.5194/hess-21-4103-2017>, 2017.
- Céron, J.-P., Tanguy, G., Franchistéguy, L., Martin, E., Regimbeau, F., and Vidal, J.-P.: Hydrological seasonal forecast over France: feasibility and prospects, *Atmos. Sci. Lett.*, 11, 78–82, <https://doi.org/10.1002/asl.256>, 2010.
- Chiew, F. H., Zhou, S. L., and McMahon, T. A.: Use of Seasonal Streamflow Forecasts in Water Resources Management, *J. Hydrol.*, 270, 135–144, [https://doi.org/10.1016/S0022-1694\(02\)00292-5](https://doi.org/10.1016/S0022-1694(02)00292-5), 2003.
- Church, J. E.: Principles of snow surveying as applied to forecasting stream flow, edited by: Merrill, M. C., *J. Agric. Res.*, Washington, D. C., Vol. 51, no. 2, 97–130, 1935.
- Coughlan de Perez, E., Stephens, E., Bischiniotis, K., van Aalst, M., van den Hurk, B., Mason, S., Nissan, H., and Pappenberger, F.: Should seasonal rainfall forecasts be used for flood preparedness?, *Hydrol. Earth Syst. Sci.*, 21, 4517–4524, <https://doi.org/10.5194/hess-21-4517-2017>, 2017.
- Crochemore, L., Ramos, M.-H., and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, *Hydrol. Earth Syst. Sci.*, 20, 3601–3618, <https://doi.org/10.5194/hess-2016-78>, 2016.
- Crochemore, L., Ramos, M.-H., Pappenberger, F., and Perrin, C.: Seasonal streamflow forecasting by conditioning climatology with precipitation indices, *Hydrol. Earth Syst. Sci.*, 21, 1573–1591, <https://doi.org/10.5194/hess-21-1573-2017>, 2017.
- Day, G. N.: Extended streamflow forecasting using NWSRFS, *J. Water Res. Plan. Man.*, 111, 157–170, [https://doi.org/10.1061/\(ASCE\)0733-9496\(1985\)111:2\(157\)](https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157)), 1985.
- De Roo, A. P., Wesseling, C. G., and Van Deursen, W. P.: Physically based river basin modelling within a GIS: the LISFLOOD model, *Hydrol. Process.*, 14, 1981–1992, [https://doi.org/10.1002/1099-1085\(20000815/30\)14:11/12<1981::AID-HYP49>3.0.CO;2-F](https://doi.org/10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F), 2000.
- Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models, *Hydrol. Earth Syst. Sci.*, 19, 275–291, <https://doi.org/10.5194/hess-19-275-2015>, 2015.
- Dettinger, M. D. and Diaz, H. F.: Global characteristics of stream flow seasonality and variability, *J. Hydrometeorol.*, 1, 289–310, [https://doi.org/10.1175/1525-7541\(2000\)001<0289:GCOSFS>2.0.CO;2](https://doi.org/10.1175/1525-7541(2000)001<0289:GCOSFS>2.0.CO;2), 2000.
- Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R.: Seasonal climate predictability and forecasting: status and prospects, *WIREs Clim. Change.*, 4, 245–268, <https://doi.org/10.1002/wcc.217>, 2013.
- Forecast skill metrics: https://meteoswiss.shinyapps.io/skill_metrics/, last access: 3 October 2017.
- Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, *J. Roy. Stat. Soc. B*, 69, 243–268, <https://doi.org/10.1111/j.1467-9868.2007.00587.x>, 2007.
- Gobena, A. K. and Gan, T. Y.: Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system, *J. Hydrol.*, 385, 336–352, <https://doi.org/10.1016/j.jhydrol.2010.03.002>, 2010.
- Goddard, L., Mason, S. J., Zebiak, S. E., Ropelewski, C. F., Basher, R., and Cane, M. A.: Current approaches to seasonal to interannual climate predictions, *Int. J. Climatol.*, 21, 1111–1152, <https://doi.org/10.1002/joc.636>, 2001.
- Guimarães Nobre, G., Jongman, B., Aerts, J., and Ward, P. J.: The role of climate variability in extreme floods in Europe, *Environ. Res. Lett.*, 12, 084012, <https://doi.org/10.1088/1748-9326/aa7c22>, 2017.
- Hamlet, A. F., Huppert, D., and Lettenmaier, D. P.: Economic Value of Long-Lead Streamflow Forecasts for Columbia River Hydropower, *J. Water Res. Plan. Man.*, 128, 91–101, [https://doi.org/10.1061/\(ASCE\)0733-9496\(2002\)128:2\(91\)](https://doi.org/10.1061/(ASCE)0733-9496(2002)128:2(91)), 2002.
- Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, *Weather Forecast.*, 15, 559–570, [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2), 2000.

- Hurrell, J. W.: Decadal trends in the North Atlantic oscillation: Regional temperatures and precipitation, *Science*, 269, 676–679, <https://doi.org/10.1126/science.269.5224.676>, 1995.
- Hurrell, J. W. and Van Loon, H.: Decadal Variations in Climate Associated with the North Atlantic Oscillation, in: *Climatic Change at High Elevation Sites*, edited by: Diaz, H. F., Beniston, M., and Bradley, R. S., Springer, Dordrecht, 69–94, https://doi.org/10.1007/978-94-015-8905-5_4, 1997.
- Keller, J. D. and Hense, A.: A new non-Gaussian evaluation method for ensemble forecasts based on analysis rank histograms, *Meteorol. Z.*, 20, 107–117, <https://doi.org/10.1127/0941-2948/2011/0217>, 2011.
- Kim, H.-M., Webster, P. J., and Curry, J. A.: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter, *Clim. Dynam.*, 39, 2957–2973, <https://doi.org/10.1007/s00382-012-1364-6>, 2012.
- Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrol. Earth Syst. Sci.*, 11, 1267–1277, <https://doi.org/10.5194/hess-11-1267-2007>, 2007.
- Li, Y., Giuliani, M., and Castelletti, A.: A coupled human-natural system to assess the operational value of weather and climate services for agriculture, *Hydrol. Earth Syst. Sci.*, 21, 4693–4709, <https://doi.org/10.5194/hess-21-4693-2017>, 2017.
- Lorenzo-Lacruz, J., Vicente-Serrano, S. M., López-Moreno, J. I., González-Hidalgo, J. C., and Morán-Tejeda, E.: The response of Iberian rivers to the North Atlantic Oscillation, *Hydrol. Earth Syst. Sci.*, 15, 2581–2597, <https://doi.org/10.5194/hess-15-2581-2011>, 2011.
- Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themessl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., and Thiele-Eich, I.: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, *Rev. Geophys.*, 48, Rg3003, <https://doi.org/10.1029/2009rg000314>, 2010.
- Mason, S. J. and Graham, N. E.: Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels, *Weather Forecast.*, 14, 713–725, [https://doi.org/10.1175/1520-0434\(1999\)014<0713:CPROCA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2), 1999.
- Meißner, D., Klein, B., and Ionita, M.: Development of a monthly to seasonal forecast framework tailored to inland waterway transport in central Europe, *Hydrol. Earth Syst. Sci.*, 21, 6401–6423, <https://doi.org/10.5194/hess-21-6401-2017>, 2017.
- Mendoza, P. A., Wood, A. W., Clark, E., Rothwell, E., Clark, M. P., Nijssen, B., Brekke, L. D., and Arnold, J. R.: An inter-comparison of approaches for improving operational seasonal streamflow forecasts, *Hydrol. Earth Syst. Sci.*, 21, 3915–3935, <https://doi.org/10.5194/hess-21-3915-2017>, 2017.
- Mo, K. C. and Lettenmaier, D. P.: Hydrologic Prediction over the Conterminous United States Using the National Multi-Model Ensemble, *J. Hydrometeorol.*, 15, 1457–1472, <https://doi.org/10.1175/JHM-D-13-0197.1>, 2014.
- Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T., and Tart, F.: The new ECMWF seasonal forecast system (System 4), *ECMWF Tech. Memorandum*, 656, 1–49, 2011.
- Neumann, J. L., Arnal, L., Magnusson, L., and Cloke, H.: The 2013/14 Thames basin floods: Do improved meteorological forecasts lead to more skilful hydrological forecasts at seasonal timescales?, *J. Hydrometeorol.*, in review, 2018.
- Pagano, T. C. and Garen, D. C.: Integration of climate information and forecasts into western US water supply forecasts, *Climate variations, climate change, and water resources engineering*, edited by: Garbrecht, J. D. and Piechota, T. C., American Society of Civil Engineers location, Reston, Virginia, US, 86–103, 2006.
- Prudhomme, C., Hannaford, J., Harrigan, S., Boorman, D., Knight, J., Bell, V., Jackson, C., Svensson, C., Parry, S., Bachiller-Jareno, N., Davies, H. N., Davis, R., Mackay, J., Mackenzie, A., Rudd, A. C., Smith, K., Bloomfield, J., Ward, R., and Jenkins, A.: Hydrological Outlook UK: an operational streamflow and groundwater level forecasting system at monthly to seasonal time scales, *Hydrolog. Sci. J.*, 62, 2753–2768, <https://doi.org/10.1080/02626667.2017.1395032>, 2017.
- Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, 46, W05521, <https://doi.org/10.1029/2009WR008328>, 2010.
- Schepen, A., Zhao, T., Wang, Q. J., Zhou, S., and Feikema, P.: Optimising seasonal streamflow forecast lead time for operational decision making in Australia, *Hydrol. Earth Syst. Sci.*, 20, 4117–4128, <https://doi.org/10.5194/hess-20-4117-2016>, 2016.
- Sheffield, J., Wood, E. F., Chaney, N., Guan, K., Sadri, S., Yuan, X., Olang, L., Amani, A., Ali, A., Demuth, S., and Ogallo, L.: A Drought Monitoring and Forecasting System for Sub-Saharan African Water Resources and Food Security, *B. Am. Meteorol. Soc.*, 95, 861–882, <https://doi.org/10.1175/BAMS-D-12-00124.1>, 2013.
- Shi, W., Schaller, N., MacLeod, D., Palmer, T. N., and Weisheimer, A.: Impact of hindcast length on estimates of seasonal climate predictability, *Geophys. Res. Lett.*, 42, 1554–1559, <https://doi.org/10.1002/2014GL062829>, 2015.
- Singla, S., Céron, J.-P., Martin, E., Regimbeau, F., Déqué, M., Habets, F., and Vidal, J.-P.: Predictability of soil moisture and river flows over France for the spring season, *Hydrol. Earth Syst. Sci.*, 16, 201–216, <https://doi.org/10.5194/hess-16-201-2012>, 2012.
- Slater, L. J., Villarini, G., Bradley, A. A., and Vecchi, G. A.: A dynamical statistical framework for seasonal streamflow forecasting in an agricultural watershed, *Clim. Dynam.*, 1–17, <https://doi.org/10.1007/s00382-017-3794-7>, 2017.
- Smith, P., Pappenberger, F., Wetterhall, F., Thielen, J., Krzeminski, B., Salamon, P., Muraro, D., Kalas, M., and Baugh, C.: On the operational implementation of the European Flood Awareness System (EFAS), *ECMWF Tech. Memorandum*, 778, 1–34, 2016.
- Soares, M. B. and Dessai, S.: Barriers and enablers to the use of seasonal climate forecasts amongst organisations in Europe, *Climatic Change*, 137, 89–103, <https://doi.org/10.1007/s10584-016-1671-8>, 2016.
- Steirou, E., Gerlitz, L., Apel, H., and Merz, B.: Links between large-scale circulation patterns and streamflow in Central Europe: A review, *J. Hydrol.*, 549, 484–500, <https://doi.org/10.1016/j.jhydrol.2017.04.003>, 2017.

- Stephens, E., Day, J. J., Pappenberger, F., and Cloke, H.: Precipitation and floodiness, *Geophys. Res. Lett.*, 42, 10316–10323, <https://doi.org/10.1002/2015GL066779>, 2015.
- Troccoli, A.: Seasonal climate forecasting, *Meteorol. Appl.*, 17, 251–268, <https://doi.org/10.1002/met.184>, 2010.
- Turner, S. W. D., Bennett, J. C., Robertson, D. E., and Galelli, S.: Complex relationship between seasonal streamflow forecast skill and value in reservoir operations, *Hydrol. Earth Syst. Sci.*, 21, 4841–4859, <https://doi.org/10.5194/hess-21-4841-2017>, 2017.
- Twedt, T. M., Schaake, J. C., and Peck, E. L.: National Weather Service extended streamflow prediction, Proceedings Western Snow Conference, Albuquerque, New Mexico, 52–57, April 1977.
- van den Hurk, B. J. J. M., Bouwer, L. M., Buontempo, C., Döscher, R., Ercin, E., Hananel, C., Hunink, J., Kjellström, E., Klein, B., Manez, M., Pappenberger, F., Pouget, L., Ramos, M.-H., Ward, P. J., Weerts, A., and Wijngaard, J.: Improving predictions and management of hydrological extremes through climate services: www.imprex.eu, *Climate Services*, 1, 6–11, <https://doi.org/10.1016/j.cliser.2016.01.001>, 2016.
- Van Der Knijff, J. M., Younis, J., and De Roo, A. P.: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, *Int. J. Geogr. Inf. Sci.*, 24, 189–212, <https://doi.org/10.1080/13658810802549154>, 2010.
- Viel, C., Beaulant, A.-L., Soubeyroux, J.-M., and Céron, J.-P.: How seasonal forecast could help a decision maker: an example of climate service for water resource management, *Adv. Sci. Res.*, 13, 51–55, <https://doi.org/10.5194/asr-13-51-2016>, 2016.
- Wetterhall, F. and Di Giuseppe, F.: The benefit of seamless forecasts for hydrological predictions over Europe, *Hydrol. Earth Syst. Sci. Discuss.*, <https://doi.org/10.5194/hess-2017-527>, in review, 2017.
- White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J., Lazo, J. K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A. J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A. P., Eggen, B., Graham, R., Kjellström, E., Becker, E., Pegion, K. V., Holbrook, N. J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T. J., Street, R., Jones, L., Remenyi, T., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B., and Zebiak, S. E.: Potential applications of subseasonal-to-seasonal (S2S) predictions, *Meteorol. Appl.*, 24, 315–325, <https://doi.org/10.1002/met.1654>, 2017.
- Wood, A. W., Kumar, A., and Lettenmaier, D. P.: A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States, *J. Geophys. Res.-Atmos.*, 110, D04105, <https://doi.org/10.1029/2004JD004508>, 2005.
- Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, *Geophys. Res. Lett.*, 35, L14401, <https://doi.org/10.1029/2008GL034648>, 2008.
- Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D. P.: Long-range experimental hydrologic forecasting for the eastern United States, *J. Geophys. Res.-Atmos.*, 107, 4429, <https://doi.org/10.1029/2001JD000659>, 2002.
- Yuan, X., Roundy, J. K., Wood, E. F., and Sheffield, J.: Seasonal forecasting of global hydrologic extremes: system development and evaluation over GEWEX basins, *B. Am. Meteorol. Soc.*, 96, 1895–1912, <https://doi.org/10.1175/BAMS-D-14-00003.1>, 2015a.
- Yuan, X., Wood, E. F., and Ma, Z.: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, *Wiley Interdisciplinary Reviews: Water*, 2, 523–536, <https://doi.org/10.1002/wat2.1088>, 2015b.
- Yuan, X., Wood, E. F., Chaney, N. W., Sheffield, J., Kam, J., Liang, M., and Guan, K.: Probabilistic Seasonal Forecasting of African Drought by Dynamical Models, *J. Hydrometeorol.*, 14, 1706–1720, <https://doi.org/10.1175/JHM-D-13-054.1>, 2013.
- Zajac, Z., Zambrano-Bigiarini, M., Salamon, P., Burek, P., Gentile, A., and Bianchi, A.: Calibration of the lisflood hydrological model for europe – calibration round 2013, Joint Research Centre, European Commission, 2013.

A4: Developing a global operational seasonal hydro-meteorological forecasting system: GloFAS-Seasonal v1.0

This paper presents a co-author contribution arising through collaboration during this PhD, summarised in Chapter 3, Sect. 3.3.2, and has the following reference:

Emerton, R., E. Zsoter, L. Arnal, H. L. Cloke, D. Muraro, C. Prudhomme, E. M. Stephens, P. Salamon and F. Pappenberger, 2018: Developing a global operational seasonal hydro-meteorological forecasting system: GloFAS-Seasonal v1.0, *Geosci. Model Dev.*, 11, 3327-3346, doi:10.5194/gmd-11-3327-2018*

* ©2018. The Authors. Geoscientific Model Development, a journal of the European Geosciences Union published by Copernicus. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.



Developing a global operational seasonal hydro-meteorological forecasting system: GloFAS-Seasonal v1.0

Rebecca Emerton^{1,2}, Ervin Zsoter^{2,1}, Louise Arnal^{1,2}, Hannah L. Cloke^{1,3}, Davide Muraro⁶, Christel Prudhomme^{2,4,5}, Elisabeth M. Stephens¹, Peter Salamon⁷, and Florian Pappenberger²

¹Department of Geography & Environmental Science, University of Reading, Reading, UK

²European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK

³Department of Earth Sciences, Uppsala University, Uppsala, Sweden

⁴Centre for Ecology and Hydrology (CEH), Wallingford, UK

⁵Department of Geography and Environment, University of Loughborough, Loughborough, UK

⁶Image Recognition Integrated Systems (IRIS), Ispra, Italy

⁷European Commission, Joint Research Centre (JRC), Ispra, Italy

Correspondence: Rebecca Emerton (r.e.emerton@pgr.reading.ac.uk)

Received: 27 April 2018 – Discussion started: 14 May 2018

Revised: 7 August 2018 – Accepted: 9 August 2018 – Published: 21 August 2018

Abstract. Global overviews of upcoming flood and drought events are key for many applications, including disaster risk reduction initiatives. Seasonal forecasts are designed to provide early indications of such events weeks or even months in advance, but seasonal forecasts for hydrological variables at large or global scales are few and far between. Here, we present the first operational global-scale seasonal hydro-meteorological forecasting system: GloFAS-Seasonal. Developed as an extension of the Global Flood Awareness System (GloFAS), GloFAS-Seasonal couples seasonal meteorological forecasts from ECMWF with a hydrological model to provide openly available probabilistic forecasts of river flow out to 4 months ahead for the global river network. This system has potential benefits not only for disaster risk reduction through early awareness of floods and droughts, but also for water-related sectors such as agriculture and water resources management, in particular for regions where no other forecasting system exists. We describe the key hydro-meteorological components and computational framework of GloFAS-Seasonal, alongside the forecast products available, before discussing initial evaluation results and next steps.

1 Introduction

Seasonal meteorological forecasts simulate the evolution of the atmosphere over the coming months. They are designed to provide an early indication of the likelihood that a given variable, for example precipitation or temperature, will differ from normal conditions weeks or months ahead. Will a particular region be warmer or cooler than normal during the next summer? Or will a river have higher or lower flow than normal next winter? Seasonal forecasts of river flow have the potential to benefit many water-related sectors from agriculture and water resources management to disaster risk reduction and humanitarian aid through earlier indications of floods or droughts.

Many operational forecasting centres produce long-range (seasonal) global forecasts of meteorological variables, such as precipitation (Weisheimer and Palmer, 2014). However, at present, operational seasonal forecasts of hydrological variables, particularly for large or global scales, are few and far between. A number of continental-scale seasonal hydro-meteorological forecasting systems have begun to emerge around the globe over the past decade (Yuan et al., 2015a), using seasonal meteorological forecasts as input to hydrological models to produce forecasts of hydrological variables. These include the European Flood Awareness System (EFAS; Arnal et al., 2018; Cloke et al., 2013), the European Service for Water Indicators in Climate Change Adapta-

tion (SWICCA; Copernicus, 2018b), the Australian Government Bureau of Meteorology Seasonal Streamflow Forecasts (Bennett et al., 2017; BoM, 2018), and the USA's National Hydrologic Ensemble Forecast Service (HEFS; Demargne et al., 2014; Emerton et al., 2016). There are also various ongoing research efforts using seasonal hydro-meteorological forecasting systems for forecast applications and research purposes at regional (Bell et al., 2017; Bennett et al., 2016; Crochemore et al., 2016; Meißner et al., 2017; Mo et al., 2014; Prudhomme et al., 2017; Wood et al., 2002, 2005; Yuan et al., 2013) and global (Candogan Yossef et al., 2017; Yuan et al., 2015b) scales. In addition to the ongoing research into improved seasonal hydro-meteorological forecasts at the global scale, an operational system providing consistent global-scale seasonal forecasts of hydrological variables could be of great benefit in regions where no other forecasting system exists and to organisations operating at the global scale (Coughlan De Perez et al., 2017).

Often, in the absence of hydrological forecasts, seasonal precipitation forecasts are used as a proxy for flooding. It has been shown that forecasts of seasonal total rainfall, the most often used seasonal precipitation forecasts, are not necessarily a good indicator of seasonal floodiness (Stephens et al., 2015), and other measures of rainfall patterns, or seasonal hydrological forecasts, would be better indicators of potential flood hazard (Coughlan De Perez et al., 2017).

While it seems a natural next step to produce global-scale seasonal hydro-meteorological forecasts, this is not a simple task, not only due to the complexities of geographical variations in rainfall–run-off processes and river regimes across the globe, but also due to the computing resources required and huge volumes of data that must be efficiently processed and stored and the challenge of effectively communicating forecasts for the entire globe. Indeed, global-scale forecasting for medium-range timescales has only become possible in recent years due to the integration of meteorological and hydrological modelling capabilities, improvements in data, satellite observations, and land-surface hydrology modelling, and increased resources and computer power (Emerton et al., 2016). In addition to continued improvements in computing capabilities, the recent move towards the development of coupled atmosphere–ocean–land models means that it is now becoming possible to produce seasonal hydro-meteorological forecasts for the global river network.

Despite the chaotic nature of the atmosphere (Lorenz, 1963), which introduces a limit of predictability (generally accepted to be ~ 2 weeks), seasonal predictions are possible as they rely on components that vary on longer timescales and are themselves predictable to an extent. This “second type predictability” (Lorenz, 1993) for seasonal river flow forecasts comes from the initial conditions and large-scale modes of climate variability. The most prominent pattern of climate variability is the El Niño–Southern Oscillation (ENSO; McPhaden et al., 2006), which is known to affect river flow and flooding across the globe (Chiew and McMa-

hon, 2002; Emerton et al., 2017; Guimarães Nobre et al., 2017; Ward et al., 2014a, b, 2016). Other teleconnections also influence river flow in various regions of the globe, such as the North Atlantic Oscillation (NAO), Southern Oscillation (SOI), Indian Ocean Dipole (IOD), and Pacific Decadal Oscillation (PDO), and contribute to the seasonal predictability of hydrologic variables (Yuan et al., 2015a). Coupled atmosphere–ocean–land models are key in representing these large-scale modes of variability in order to produce seasonal hydro-meteorological forecasts.

This motivates the development of an operational global-scale seasonal hydro-meteorological forecasting system as an extension of the Global Flood Awareness System (GloFAS; Alfieri et al., 2013), with openly available forecast products. GloFAS is developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) and the European Commission Joint Research Centre (JRC) and has been producing probabilistic flood forecasts out to 30 days for the entire globe since 2012. In 2016, work began in collaboration with the University of Reading to implement a seasonal outlook in GloFAS, aiming to provide forecasts of both high and low river flow for the global river network up to several months in advance. On 10 November 2017, the first GloFAS seasonal river flow forecast was released. This paper introduces the modelling system, its implementation, and the available forecast products and provides an initial evaluation of the potential usefulness and reliability of the forecasts.

2 Implementation

The GloFAS seasonal outlooks are produced by driving a hydrological river routing model with meteorological forecasts from ECMWF. The forecasts are run operationally on the ECMWF computing facilities. This section provides an overview of the computing facilities, introduces the key hydro-meteorological components of the modelling platform (the meteorological forecast input, hydrological model, and reference climatology), and describes the computational framework of GloFAS-Seasonal.

2.1 ECMWF High-Performance Computing Facility

ECMWF's current High-Performance Computing Facility (HPCF) has been in operation since June 2016 and is used for both forecast production and research activities. The HPCF comprises two identical Cray XC40 supercomputers, each of which is self-sufficient with their own storage and each with equal access to the storage of the other. Each Cray XC40 consists of 20 cabinets of compute nodes and 13 storage nodes. One compute node has two Intel Broadwell processors, each with 18 cores, giving 192 nodes (6912 cores) per cabinet. The Cray Aries interconnect is used to connect the processing power. The majority of the nodes of the HPCF are run using the high-performance Cray Linux Environment, a stripped-

down version of Linux, as reducing the number of operating system tasks is critical for providing a highly scalable environment.

In terms of storage, each Cray XC40 has ~ 10 PB of storage, and the data handling system (DHS) also comprises two main applications: the Meteorological Archive and Retrieval System (MARS), which stores and provides access to meteorological data collected or produced by ECMWF, and ECFS, which stores data that are not suitable for storing on MARS. The DHS holds over 210 PB of primary data, and the archive increases by ~ 233 TB per day. The reader is referred to the ECMWF website at <https://www.ecmwf.int/> for further information on the HPCF and DHS.

In addition to the Cray XC40s, the ECMWF computing facility also includes four Linux clusters consisting of 60 servers and 1 PB of storage. The Linux clusters are currently used to run the river routing model used in GloFAS and to produce the forecast products, while the meteorological forcing and ERA5 reanalysis are produced on the HPCF. All data related to GloFAS-Seasonal are stored on the MARS and ECFS archives.

2.2 Hydro-meteorological components

2.2.1 Meteorological forcing

The first model component of the seasonal outlook is the meteorological forecast input from the ECMWF Integrated Forecast System (IFS, cycle 43r1; ECMWF, 2018b). GloFAS-Seasonal makes use of SEAS5, which is the latest version of ECMWF's long-range ensemble forecasting system made operational in November 2017 (ECMWF, 2017a; Stockdale et al., 2018). SEAS5 consists of 51 ensemble members (50 perturbed members and 1 unperturbed control member) and has a horizontal resolution of ~ 36 km (T_{CO319}). The system, which comprises a data assimilation system and a global circulation model, is run once a month, producing forecasts out to 7 months ahead. Initial pre-implementation testing of SEAS5 has suggested that in comparison to the previous version (System 4), SEAS5 better simulates sea surface temperatures (SSTs) in the Pacific Ocean, leading to improved forecasts of the El Niño–Southern Oscillation (ENSO; Stockdale et al., 2018), which is closely linked to river flow across the globe and can provide added predictability.

SEAS5 is a configuration of the ECMWF IFS (cycle 43r1), including atmosphere–ocean coupling to the NEMO ocean model. SEAS5 is run operationally on the HPCF. Each ensemble member is a complex, HPC-intensive, massively parallel code written in Fortran (version F90). In addition, further complex scripting systems are required to control, prepare, run, post-process, and archive all IFS forecasts. The data assimilation systems used to prepare the initial conditions for the forecasts also make use of Fortran and run on

the HPCF. For further information, the reader is referred to the IFS documentation (ECMWF, 2018b).

2.2.2 Land surface component

Within the IFS, which includes SEAS5, the Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land, HTESSEL (Balsamo et al., 2011), is used to compute the land surface response to atmospheric forcing. HTESSEL simulates the evolution of soil temperature, moisture content, and snowpack conditions through the forecast horizon to produce a corresponding forecast of surface and subsurface run-off. This component allows for each grid box to be divided into tiles, with up to six tiles per grid box (bare ground, low and high vegetation, intercepted water, and shaded and exposed snow) describing the land surface. For a given precipitation, the scheme distributes the water as surface run-off and drainage, with dependencies on orography and soil texture. An interception layer accumulates precipitation until saturation is reached, with the remaining precipitation partitioned between surface run-off and infiltration. HTESSEL also accounts for frozen soil, redirecting the rainfall and snowmelt to surface run-off when the uppermost soil layer is frozen, and incorporates a snow scheme. Four soil layers are used to describe the vertical transfer of water and energy, with subsurface water fluxes determined by Darcy's law, and each layer has a sink to account for root extraction in vegetated areas. A detailed description of the hydrology of HTESSEL is provided by Balsamo et al. (2011).

HTESSEL comprises a Fortran library of $\sim 20\,000$ lines of code, using both F77 and F90 Fortran versions, and is implemented modularly. While HTESSEL can be run on diverse architectures from a workstation PC to the HPCF, operationally it is run on the HPCF.

2.2.3 River routing model

As HTESSEL does not simulate water fluxes through the river network, Lisflood (Van Der Knijff et al., 2010), driven by the surface and subsurface run-off output from HTESSEL interpolated to the 0.1° (~ 10 km) spatial resolution of Lisflood is used to simulate the groundwater (subsurface water storage and transport) processes and routing of the water through the river network. The initial conditions used to start the Lisflood model are taken from the ERA5-R river flow reanalysis (see Sect. 2.2.4).

Lisflood is a spatially distributed hydrological model, including a 1-D channel routing model. Groundwater processes are modelled using two linear reservoirs, the upper zone representing a quick run-off component, including subsurface flow through soil macropores and fast groundwater, and the lower zone representing a slow groundwater component fed by percolation from the upper zone. The routing of surface run-off to the outlet of each grid cell, and the routing of run-off produced by every grid cell from the surface, upper,

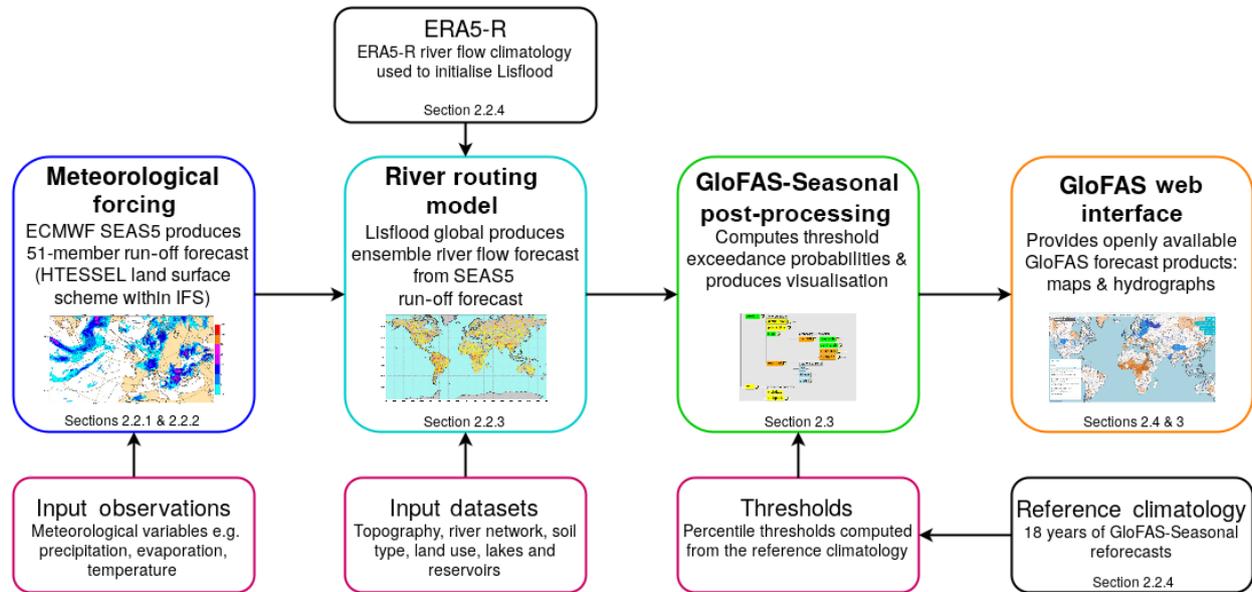


Figure 1. Flowchart depicting the key GloFAS-Seasonal forecasting system components.

and lower groundwater zones through the river network, is done using a four-point implicit finite-difference solution of the kinematic wave equations (Chow et al., 2010). The river network used is that of HydroSHEDS (Lehner et al., 2008), again interpolated to a 0.1° spatial resolution using the approach of Fekete et al. (2001). For a detailed account of the Lisflood model set-up within GloFAS, the reader is referred to Alfieri et al. (2013).

Lisflood is implemented using a combination of PCRaster GIS and Python and is currently run operationally on the Linux cluster at ECMWF.

2.2.4 Generation of reforecasts and reference climatology

In order to generate a reference climatology for GloFAS-Seasonal, the latest of ECMWF's reanalysis products, ERA5, was used. Reanalysis datasets combine historical observations of the atmosphere, ocean, and land surface with a data assimilation system; global models are used to “fill in the gaps” and produce consistent global best estimates of the atmosphere, ocean, and land state. ERA5 represents the current state of the art in terms of reanalysis datasets, providing a much higher spatial and temporal resolution (30 km, hourly) compared to ERA-Interim (79 km, 3-hourly) and better representations of precipitation, evaporation, and soil moisture (ECMWF, 2017b). In order to produce a river flow reanalysis (ERA5-R) for the global river network, the ERA5 surface and subsurface run-off variables were interpolated to 0.1° (~ 10 km) resolution and used as input to the Lisflood model (see Sect. 2.2.3). ERA5 is currently still in production, and while it will cover the period from 1950 to present

when completed, the full dataset will not be available until 2019. ERA5 is being produced in three “streams” in parallel; at the time of producing the ERA5-R reanalysis, 18 years of ERA5 data were available across the three streams (1990–1992, 2000–2007, and 2010–2016). In addition to the historical climatology, ERA5 is also produced in near real time, with a delay of just ~ 3 days, allowing its use as initial conditions for the river routing component of the GloFAS-Seasonal forecasts. The ERA5-R reanalysis is thus updated every month prior to producing the forecast. Figure 2 provides an overview of all datasets used in and produced for the development of GloFAS-Seasonal.

Once the ERA5-R reanalysis was obtained, a set of GloFAS-Seasonal reforecasts was produced. From the 25-ensemble-member SEAS5 reforecasts produced by ECMWF, the surface and subsurface run-off variables were used to drive the Lisflood model with initial conditions from ERA5-R. This generated 18 years of seasonal river flow reforecasts (one forecast per month out to 4 months of lead time, with 25 ensemble members at 0.1° resolution). It is the weekly averaged river flow from this reforecast dataset which is used as a reference climatology, including to calculate the high and low flow thresholds used in the real-time forecasts (described in Sect. 2.3).

2.3 GloFAS-Seasonal computational framework

The GloFAS-Seasonal real-time forecasts are implemented and run operationally on the ECMWF computing facilities using ecFlow (Bahra, 2011; ECMWF, 2012), an ECMWF work package used to run large numbers of programmes with dependencies on each other and on time. An ecFlow suite

is a collection of tasks and scheduling instructions with a user interface allowing for the interaction and monitoring of the suite, the code behind it, and the output. The GloFAS-Seasonal suite is run once per month and is used to retrieve the raw SEAS5 forecast data. It runs this through Lisflood and produces the final forecast products and visualisations using the newly developed GloFAS-Seasonal post-processing code.

The GloFAS-Seasonal suite performs tasks (detailed below) such as retrieving data, running Lisflood, computing weekly averages and forecast probabilities from the raw Lisflood river flow forecast data, and producing maps and hydrographs for the interface. It is primarily written in Python (version 2.7), with some elements written in R (version 3.1) and shell scripts incorporating climate data operators (CDOs). The code was developed and tested on OpenSUSE Leap 42 systems.

When a new SEAS5 forecast becomes available (typically on the 5th of the month at 00:00 UTC), the GloFAS-Seasonal ecFlow suite is automatically deployed. The structure of and tasks within the ecFlow suite are shown in Fig. 3. Each “task” represents one script from the GloFAS-Seasonal code. The suite first retrieves the latest raw SEAS5 forecast surface and subsurface variables for all 51 ensemble members (*stagefc* and *getfc* tasks), alongside the river flow reference climatology (see Sect. 2.2.4) for the corresponding month of the forecast (*copywb* task). The Lisflood river routing model (described in Sect. 2.2.3) is then run for each of the 51 ensemble members (*lisflood* task). Lisflood is initialised using the ERA5-R river flow reanalysis (see Sect. 2.2.4) and driven with the SEAS5 surface and subsurface run-off forecast to produce the 4-month ensemble river flow forecast at a daily time step, from which the weekly averaged ensemble river flow forecast is obtained (*average* task). The weekly averages are computed for every Monday–Sunday starting from the first Monday of each month so that the weekly averages correspond from one forecast to the next. While SEAS5 provides forecasts out to 7 months ahead, the first version of GloFAS-Seasonal uses only the first 4 months. This is in order to reduce the data volumes required and to allow for the assessment of the forecast skill out to 4 months ahead before possible extension of the forecasts out to 7 months ahead in the future.

Once the weekly averaging is complete, the *forecast product* section of the suite is deployed, which post-processes the raw forecast output to produce the final forecast products displayed on the web interface. The code behind the *forecast product* section is provided in the Supplement. For a full description of the forecast products, including examples, see Sect. 3. The suite computes the full forecast distribution (*distribution* task), followed by the probability of exceedance for each week of the forecast and for every grid point (*probability* task) based on the number of ensemble members exceeding the high flow threshold or falling below the low flow threshold. The high and low flow thresholds are defined

as the 80th and 20th percentiles of the reference climatology for the week of the year corresponding to the forecast week to use thresholds based on time of year of the forecast. From these weekly exceedance probabilities, the maximum probability of exceedance across the 4-month forecast horizon is calculated for each grid point (*maxprob* task). Basin-averaged maximum probabilities are also produced (*basinprob* task) by calculating the mean maximum probability of exceedance across every grid point at which the upstream area exceeds 1500 km² in each of the 306 major world river basins used in GloFAS-Seasonal (see Sect. 3.1). A minimum upstream area of 1500 km² is chosen, as the current resolution of the global model is such that reliable forecasts for very small rivers are not feasible.

These probabilities are used to produce the forecast visualisation for the web interface (Sect. 3). Firstly, the *map* task produces colour-coded maps of both the river network, again for grid points at which the upstream area exceeds 1500 km², and the major world river basins. The *reppoint* task then produces an ensemble hydrograph and persistence diagrams for a subset of grid points (the “reporting points”) across the globe. Further details on the location of reporting points are given in Sect. 3.3. Finally, the *web* task collates and subsequently transfers all data required for the web interface.

This process, from the time a new SEAS5 forecast becomes available, takes ~ 4 h on average to complete, with up to 10 tasks running in parallel (for example, running Lisflood for 10 ensemble members at the same time). It is possible to speed up this process by running more ensemble members in parallel; however, the speed is sufficient so that it is not necessary to use further resources to produce the forecast more quickly. GloFAS-Seasonal forecast products are typically produced by the 5th of the month at 05:00 UTC and made available via the web interface on the 10th of the month at 01:00 UTC. This is the earliest that the GloFAS-Seasonal forecasts can be provided publicly under the Copernicus licence agreement. Data are automatically archived at ECMWF as the suite runs in real time; ~ 285 GB of data from each SEAS5 forecast are used as input for GloFAS-Seasonal. Each GloFAS-Seasonal forecast run produces an additional ~ 1.8 TB of data and makes use of the ~ 18 TB reference climatology.

2.4 GloFAS web interface

The GloFAS website is based on a user-centred design (UCD), meaning that user needs are core to the design principles (ISO13407). The website uses Web 2.0 concepts such as simplicity, joy of use, and usability that are synonymous with engaging users. It is a rich internet application (RIA) aiming to provide the same level of interactivity and responsiveness as desktop applications. The website is designed for those engaged in flood forecasting and water resources, as users can browse various aspects of the current forecast or past forecasts in a simple and intuitive way, with spatially distributed

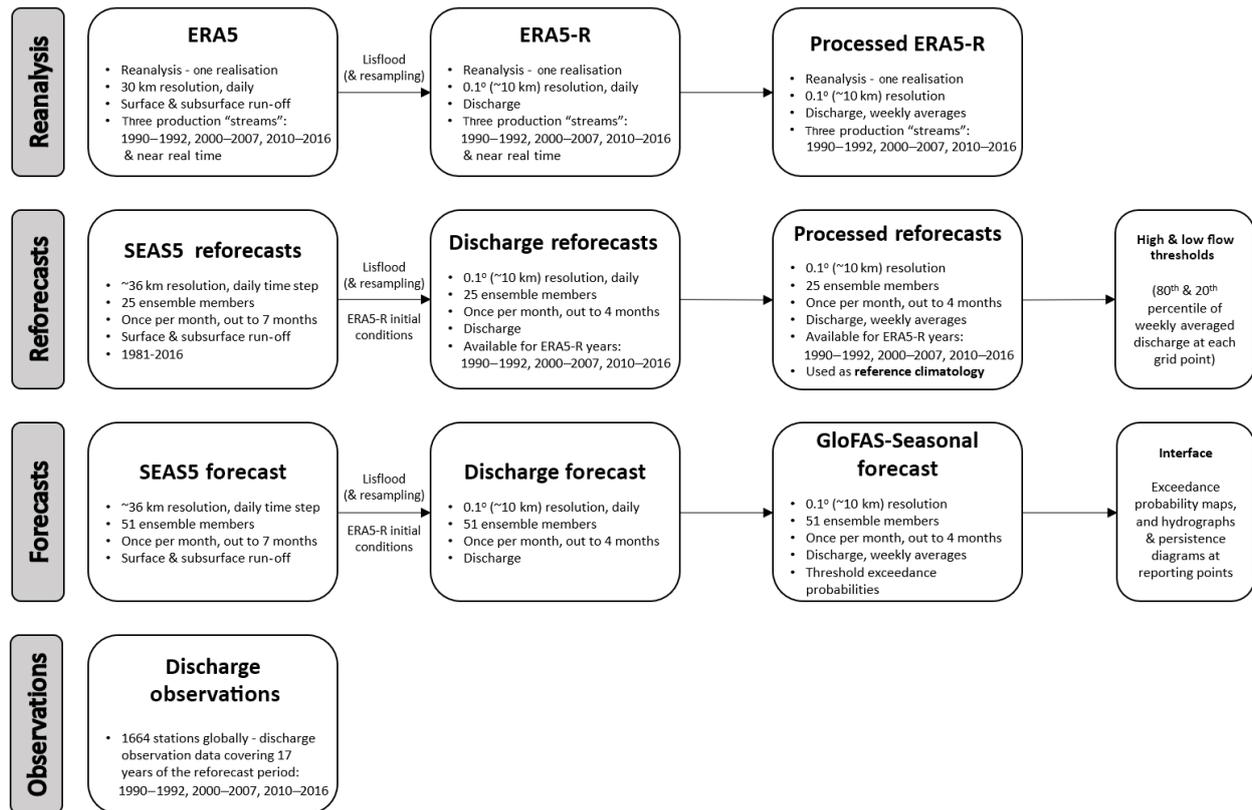


Figure 2. All datasets used and produced for GloFAS-Seasonal, including reanalysis, reforecasts, real-time forecasts, and observations.

information. Map layers containing different information, e.g. flood probabilities for different flood severities, precipitation forecasts, and seasonal outlooks, can be activated and the user can also choose to overlay other information such as land use, urban areas, or flood hazard maps. The interface consists of three principal modules: MapServer, GloFAS Web Map Service Time, and the Forecast Viewer. These are outlined below.

2.4.1 MapServer

MapServer (Open Source Geospatial Foundation, 2016) is an open source development environment for building spatially enabled internet applications developed by the University of Minnesota. MapServer has built-in functionality to support industry standard data formats and spatial databases, which is significant to this project, and the support of popular Open Geospatial Consortium (OGC) standards including WMS. In order to exploit the potential of asynchronous data transfer between server and client, the GloFAS raster data have to be divided into a grid of adequate dimensions and an optimal scale sequence.

2.4.2 GloFAS Web Map Service Time

The OpenGIS Web Map Service (WMS) is a standard protocol for serving geo-referenced map images over the internet. A web map service time (WMS-T) is a web service that produces maps in several raster formats or in vector format that may come simultaneously from multiple remote and heterogeneous sources. A WMS server can provide support to temporal requests (WMS-T) by providing a TIME parameter with a time value in the request.

The WMS specification (OGC, 2015) describes three HTTP requests; *GetCapabilities*, *GetMap*, and *GetFeatureInfo*. *GetCapabilities* returns an XML document describing the map layers available and the server's capabilities (i.e. the image formats, projections, and geographic bounds of the server). *GetMap* returns a raster map image. The request arguments, such as the layer ID and image format, should match those listed as available in the *GetCapabilities* return document. *GetFeatureInfo* is optional and is designed to provide WMS clients with more information about features in the map images that were returned by earlier *GetMap* requests. The response should contain data relating to the features nearest to an image coordinate specified in the *GetFeatureInfo* request. The structure of the data returned is not defined in the specification and is left up to the WMS server

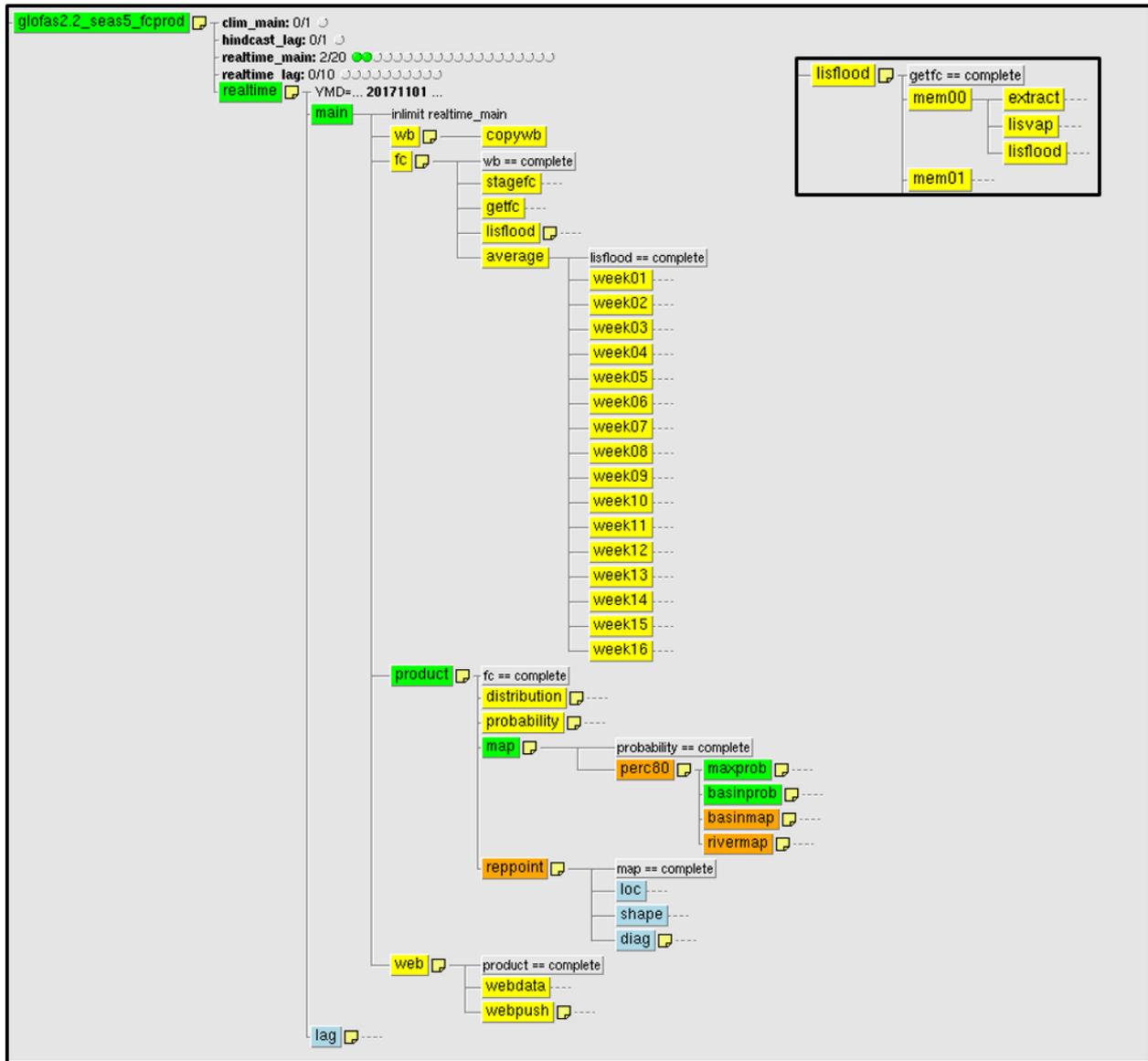


Figure 3. The GloFAS-Seasonal ecFlow suite. The inset image shows the sub-tasks within the Lisflood task for 1 of the 51 ensemble members. Colours indicate the status of each task. Yellow: complete, green: active, orange: suspended, pale blue: waiting, turquoise (not shown): queued, and red (not shown): aborted or failed. Grey boxes indicate dependencies; for example, “lisflood = complete” indicates that the Lisflood task and all Lisflood sub-tasks must have successfully completed in order for the average task to run.

implementation. The GloFAS WMS-T (GloFAS, 2018b) can be freely used, allowing access to the GloFAS layers in any GIS environment, such as QGIS (QGIS Development Team, 2017) or ArcMAP (Environmental Systems Research Institute, 2018). The user manual for the GloFAS WMS-T is available via the GloFAS website (GloFAS, 2018a).

2.4.3 Forecast Viewer

The GloFAS Forecast Viewer is based on the model view controller (MVC) architectural pattern used in software engineering. The pattern isolates “domain logic” (the applica-

tion logic for the user) from input and presentation (user interface, UI), permitting the independent development, testing, and maintenance of each. A fundamental part of this is the AJAX (asynchronous JavaScript and XML) technology used to enhance user-friendly interfaces for web mapping applications. AJAX technologies have a number of benefits; the essential one is removing the need to reload and refresh the whole page after every event. Careful application design and component selection results in a measurably smaller web server load in geodata rendering and publishing, as there is no need to link and send the whole html document, just the relevant part that needs to be changed.

GloFAS uses OpenLayers (OpenLayers, 2018) as a WMS client. OpenLayers is a JavaScript-based web mapping toolkit designed to make it easy to put a dynamic map on any web page. It does not depend on the server technology and can display a set of vector data, such as points, with aerial photographs as backdrop maps from different sources. Closely coupled to the map widget is a layer manager that controls which layers are displayed with facilities for adding, removing, and modifying layers. The new layers associated with GloFAS-Seasonal are described in the following section.

3 Forecast products

The GloFAS seasonal outlook is provided as three new forecast layers in the GloFAS Forecast Viewer: the basin overview, river network, and reporting point layers. Each of the three layers represents a different forecast product described in the following sections. Information on each of the layers is also provided for end users of the forecasts under the dedicated “Seasonal Outlook” page of the GloFAS website.

3.1 Basin overview layer

The first GloFAS seasonal outlook product is designed to provide a quick global overview of areas that are likely to experience unusually high or low river flow over the coming 4 months. The “basin overview” layer displays a map of 306 major world river basins colour coded according to the maximum probability of exceeding the high (blue) or low (orange) flow thresholds (the 80th and 20th percentiles of the reference climatology, respectively) during the 4-month forecast horizon. This value is calculated for each river basin by taking the average of the maximum exceedance probabilities at each grid cell within the basin (using only river pixels with an upstream area $> 1500 \text{ km}^2$). The three different shades of orange–blue indicate the probability: dark ($> 90\%$), medium (75%–90%), and light (50%–75%). Basins that remain white are those in which the probability of unusually high or low flow does not exceed 50% during the 4-month forecast horizon. An example is shown in Fig. 4.

As mentioned in Sect. 2.2.3, the Lisflood river network is based on HydroSHEDS (Lehner et al., 2008). In order to generate the river basins used in GloFAS-Seasonal, the corresponding HydroBASINS (Lehner and Grill, 2013) data were used. HydroBASINS consists of a suite of polygon layers depicting watershed boundaries at the global scale. These watersheds were manually merged using QGIS (QGIS Development Team, 2017) to create a global polygon layer of major river basins based on the river network used in the model.

3.2 River network layer

The second map layer provides similar information at the sub-basin scale by colour-coding the entire model river net-

work according to the maximum exceedance probability during the 4-month forecast horizon. This allows the user to zoom in to their region of interest and view the forecast maximum exceedance probabilities in more detail. Again, only river pixels with an upstream area $> 1500 \text{ km}^2$ are shown. The same colour scheme is used for both the basin overview and river network layers, with blue indicating high flow (exceeding the 80th percentile), orange low flow (falling below the 20th percentile), and darker colours indicating higher probabilities. In the river network layer, additional colours also represent areas where the forecast does not exceed 50% probability of exceeding either the high or low flow threshold (light grey) and where the river pixel lies in a climatologically arid area such that the forecast probability cannot be defined (darker grey–brown). Examples of the river network layer can be seen in both Fig. 4 (globally) and Fig. 5 (zoomed in).

3.3 Reporting points layer

In addition to the two summary map layers, reporting points are provided at both static and dynamic locations throughout the global river network, providing additional forecast information: an ensemble hydrograph and a persistence diagram.

Static points originally consisted of a selection of gauged river stations included in the Global Runoff Data Centre (GRDC; BfG, 2017); this set of points has since been expanded to further include points at locations of particular interest to GloFAS partners. There are now ~ 2200 static reporting points in the GloFAS interface.

Dynamic points are generated to provide the additional forecast information throughout the global river network, including river reaches for which there are no static points. These points are obtained for every new forecast based on a set of selection criteria adapted from the GloFAS flood forecast dynamic point selection criteria (Alfieri et al., 2013).

- The maximum probability of high (low) river flow (exceeding or falling below) the 80th (20th) percentile of the reference climatology) during the 4-month forecast horizon must be $\geq 50\%$ for at least five contiguous pixels of the river network.
- The upstream area of the selected point must be $\geq 4000 \text{ km}^2$.
- Dynamic reporting points are generated starting from the most downstream river pixel complying with the previous two selection criteria. A new reporting point is then generated every 300 km upstream along the river network, unless a static reporting point already exists within a short distance of the new dynamic point or the forecasts further upstream no longer comply with the previous two criteria.

Reporting points are displayed as black circles in the “reporting points” seasonal outlook layer. An example is shown in

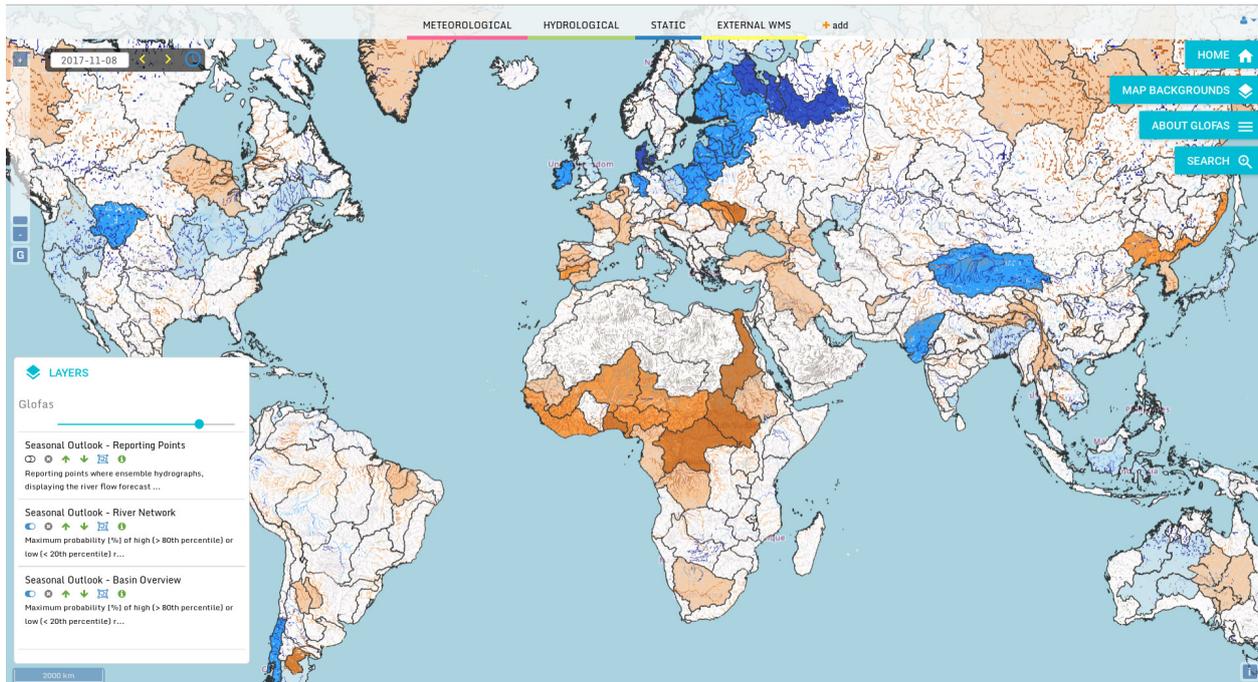


Figure 4. Example screenshot of the seasonal outlook layers in the GloFAS web interface. Shown here are both the “basin overview” layer and “river network” layer, both indicating the maximum probability of unusually high (blue) or low (orange) river flow during the 4-month forecast horizon. The darker the colour, the higher the probability: darkest shading indicates > 90 % probability, medium shading indicates 75 %–90 % probability, and light shading indicates 50 %–75 % probability. A white basin or light grey river pixel indicates that the forecast does not exceed 50 % probability of high or low flow during the forecast horizon. Legends providing this information are available for each layer by clicking on the green “i” next to the layer toggle (shown at the bottom left in this example).

Fig. 5. Clicking on a reporting point brings up a new window containing a hydrograph and persistence diagram alongside some basic information about the location, such as the latitude and longitude, and the upstream area of the point in the model river network. The number of dynamic reporting points can vary from one forecast to the next due to the criteria applied; for example, the March 2018 forecast included ~ 1600 dynamic points in addition to the static points, and thus ~ 3800 reporting points were available globally.

The ensemble hydrographs (also shown in Fig. 5) display a fan plot of the ensemble forecast of weekly averaged river flow out to 4 months, indicating the spread of the forecast and associated probabilities. Also shown are thresholds based on the reference climatology: the median and the 80th and 20th percentiles. These thresholds are displayed as a 3-week moving average of the weekly averaged river flow for the given threshold for the same months of the climatology as that of the forecast (i.e. a forecast for J–F–M–A also displays thresholds based on the reference climatology for J–F–M–A). This allows for a comparison of the forecast to typical and extreme conditions for the time of year.

Persistence diagrams (see Fig. 5) show the weekly probability of exceeding the high and low flow thresholds for the current forecast (bottom row) and previous three forecasts colour coded to match the probabilities indicated in the

map layers. These diagrams are provided in order to highlight the evolution of the forecast, which can indicate whether the forecast is progressing consistently or whether behaviour is variable from month to month.

4 Forecast evaluation

In this section, the GloFAS-Seasonal reforecasts are evaluated using historical river flow observations. Benchmarking a forecasting system is important to evaluate and understand the value of the system and in order to communicate the skill of the forecasts to end users (Pappenberger et al., 2015). This evaluation is designed to measure the ability of the forecasts to predict the correct category of an “event”, i.e. the ability of the forecast to predict that weekly averaged river flow will fall in the upper 80th or lower 20th percentile of climatology using a climatology of historical observations as a benchmark. This can be referred to as the potential usefulness of the forecasts and is of particular importance for decision-making purposes (Arnal et al., 2018). Another key aspect of probabilistic forecasts to consider is their reliability, which indicates the agreement between forecast probabilities and the observed frequency of events.

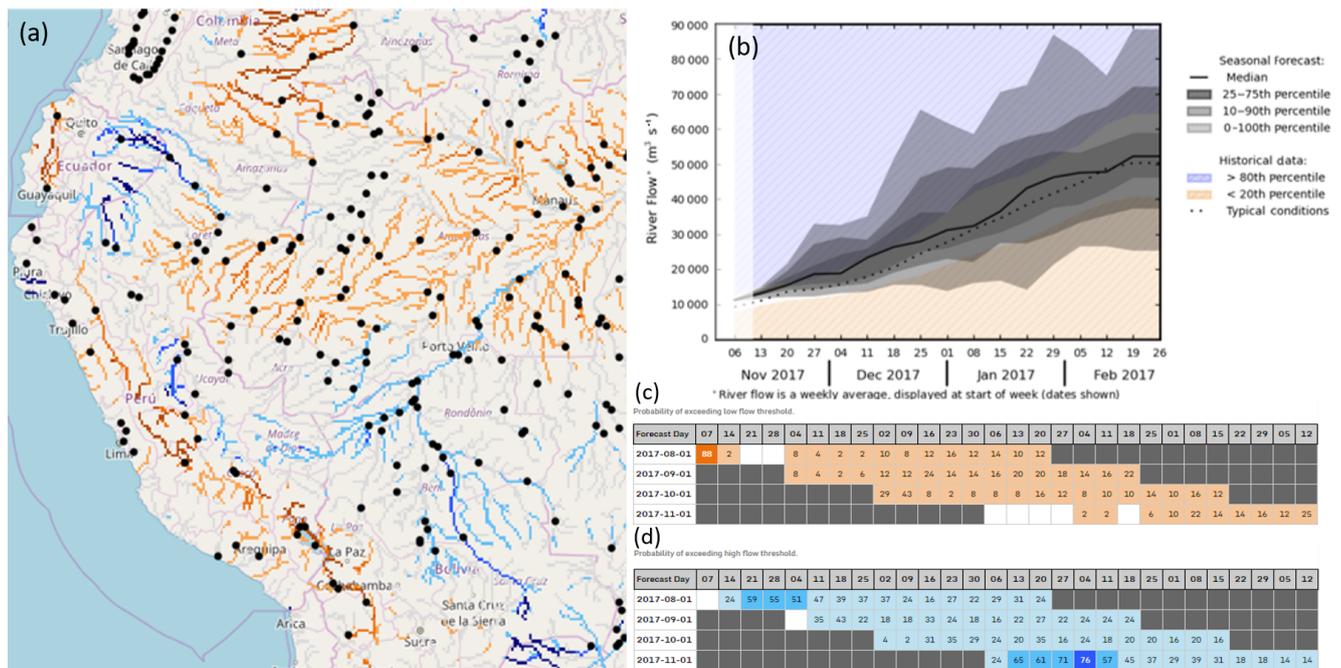


Figure 5. Example of the “reporting points” GloFAS seasonal outlook layer in the web interface (a). Black circles indicate the reporting points, which provide the ensemble hydrograph (b) and persistence diagrams for both low flow (c) and high flow (d). Also shown is an example section of the “river network” seasonal outlook layer indicating the maximum probability of high (blue) or low (orange) river flow during the 4-month forecast horizon. The darker the colour, the higher the probability.

The potential usefulness is assessed using the relative operating characteristic (ROC) curve, which is based on ratios of the proportion of events (the probability of detection, POD) and non-events (the false alarm rate, FAR) for which warnings were provided (Mason and Graham, 1999); in this case warnings are treated as forecasts of river flow exceeding the 80th or falling below the 20th percentile of the reference climatology (see Sect. 2.2.4). These ratios allow for the estimation of the probability that an event will be predicted.

For each week of the forecast (out to 16 weeks, corresponding to the forecasts provided via the interface; for example, the hydrograph shown in Fig. 5), the POD (Eq. 1) and FAR (Eq. 2) are calculated for both the 80th and 20th percentile events at each observation station:

$$\text{POD} = \frac{\text{hits}}{\text{hits} + \text{misses}}, \quad (1)$$

$$\text{FAR} = \frac{\text{false alarms}}{\text{hits} + \text{false alarms}}, \quad (2)$$

where a hit is defined when the forecast correctly exceeded (fell below) the 80th (20th) percentile of the reference climatology during the same week that the observed river flow exceeded (fell below) the 80th (20th) percentile of the observations at that station. It follows that a miss is defined when an event was observed but the forecast did not exceed the threshold, and a false alarm when the forecast exceeded the threshold but no event was observed. From these, the area un-

der the ROC curve (AROC) is calculated, again for both the 80th and 20th percentile events. The AROC ($0 \leq \text{AROC} \leq 1$, where 1 is perfect) indicates the skill of the forecasts compared to the long-term average climatology (which has an AROC of 0.5) and is used here to evaluate the potential usefulness of the forecasts. The maximum lead time at which forecasts are more skilful than climatology ($\text{AROC} > 0.5$) is identified; a forecast with an $\text{AROC} < 0.5$ would be less skilful than climatology and thus not useful.

The reliability of the forecasts is assessed using attributes diagrams, which show the relationship between the forecast probability and the observed frequency of the events. While the ROC measures the ability of a forecasting system to predict the correct category of an event, the reliability assesses how closely the forecast probabilities correspond to the actual chance of observing the event. As such, these evaluation metrics are useful to consider together. As with the ROC calculations, the reliability is assessed for each week of the forecast (out to 16 weeks) and for both the 80th and 20th percentile events. The range of forecast probabilities is divided into 10 bins (0%–10%, 10%–20%, etc.), and the forecast probability is plotted against the frequency at which an event was observed for forecasts in each probability bin. Perfect reliability is exhibited when the forecast probability and the observed frequency are equal; for example, if a forecast predicts that an event will occur with a probability of 60%, then the event should occur on 60% of the occasions that this fore-

cast was made. Attributes diagrams can also be used to assess the sharpness and resolution of the forecasts. Forecasts that do not discriminate between events and non-events are said to have no resolution (a forecast of climatology would have no resolution), and forecasts which are capable of predicting events with probabilities that differ from the observed frequency, such as forecasts of high or 0 probability, are said to have sharpness.

The GloFAS-Seasonal reforecasts (of which there are 216 covering 18 years, as described in Sect. 2.2.4 and Fig. 2) are compared to river flow observations that have been made available to GloFAS, covering 17 years of the study period up to the end of 2015 when the data were collated (see Fig. 2). To ensure a large enough sample size for this analysis, alongside the best possible spatial coverage, the following criteria are applied to the data.

- The weekly river flow data record available for each station must contain no more than 53 % (9 years) missing data. The high and low flow thresholds (the 80th and 20th percentile, respectively) are calculated using the observations for each station and for each week across the 17 years of data, so a sample size of 17 is the maximum possible. A threshold of (up to) 53 % missing data allows for a minimum sample size of eight. Selecting a smaller threshold reduced the number of stations and the spatial coverage across the globe significantly. The percentage of missing data is calculated at each station and for each week of the dataset independently, and as such the number of stations used can vary slightly with time.
- The upstream area of the corresponding grid point in the model river network must be at least 1500 km².

These criteria allow for the use of 1140 ± 14 stations globally. While the dataset contains 6122 stations, just 1664 of these contain data during the 17-year period, and none have the full 17 years of data available. Data from human-influenced rivers have not been removed, as in this study we are interested in identifying the ability of the forecasting system in its current state to predict observed events rather than the ability of the hydrological model to represent natural flow.

4.1 Potential usefulness

In order to gain an overview of the potential usefulness of the GloFAS-Seasonal forecasts across the globe, we map the maximum lead time at which the forecasts are more skilful than climatology (i.e. $\text{AROC} > 0.5$) at each observation station averaged across all forecast months. These results are shown in Fig. 6, and it is clear that forecasts of both high and low flow events are more skilful than climatology across much of the globe, with potentially useful forecasts at many stations out to 4 months ahead. However, there are regions where the forecasts are (on average across all fore-

cast months) not useful (i.e. $\text{AROC} < 0.5$), such as the western USA and Canada (excluding coastlines), much of Africa, and additionally across parts of Europe for low flow events. As forecasts with an AROC larger than but close to 0.5 could be deemed as only marginally more skilful than climatology, we apply a skill buffer, setting the threshold to $\text{AROC} > 0.6$ for a forecast to be deemed as potentially useful. These results are mapped in Fig. 7 and clearly indicate the reduction in the lead time at which forecasts are potentially useful (for both high and low flow events) at many stations, implying that in some locations, forecasts beyond the first 1–2 months are only marginally more skilful than climatology. There are, however, stations in some rivers with an $\text{AROC} > 0.6$ out to 4 months of lead time and many locations across the globe that still indicate that forecasts are potentially useful 1–2 months ahead for both high and low flow events.

These results can be further broken down by season, indicating whether the forecasts are more potentially useful at certain times of the year. Maps showing the maximum lead time at which $\text{AROC} > 0.6$ for each season (for forecasts started during the season; e.g. DJF indicates the average results for forecasts produced on 1 December, 1 January, and 1 February) are provided for high and low flow events in Figs. S1 and S2 in the Supplement, respectively.

The following paragraphs provide an overview of these results for each continent; for further detail please refer to the maps.

South America. For high flow events, forecasts for the Amazon basin in DJF and MAM are potentially useful out to longer lead times (up to 3–4 months) and at more stations than in JJA and SON, with similar results in MAM for low flow events. In contrast, further south, forecasts are most potentially useful JJA and SON up to 4 months ahead. In the more mountainous regions of western South America, forecasts in JJA and SON are generally less skilful than climatology for high and low flow events. In the north-west, however, for some stations, forecasts started in DJF and MAM are potentially useful up to 3 months ahead.

North America. In eastern North America, JJA and SON forecasts are most potentially useful, with more stations indicating an $\text{AROC} > 0.6$ out to 2–3 months ahead. However, during all seasons there are several stations in the east showing skill out to varying lead times. Much of the western half of the continent (excluding coastal areas) sees forecasts that are less skilful than climatology during all seasons, although some stations do indicate skill up to 4 months ahead for high flow, for forecasts started in MAM and JJA, and for low flow in MAM. At many coastal stations in the west, forecasts of high flow events started in DJF, MAM, and JJA indicate skill out to 3–4 months and out to ~ 6 weeks in SON.

Europe. Forecasts for European rivers generally perform best for high flow events in SON and DJF, with the exception of some larger rivers in eastern Europe, for which the forecasts are more potentially useful in JJA and SON. In MAM and JJA, the number of stations indicating no skill is gener-

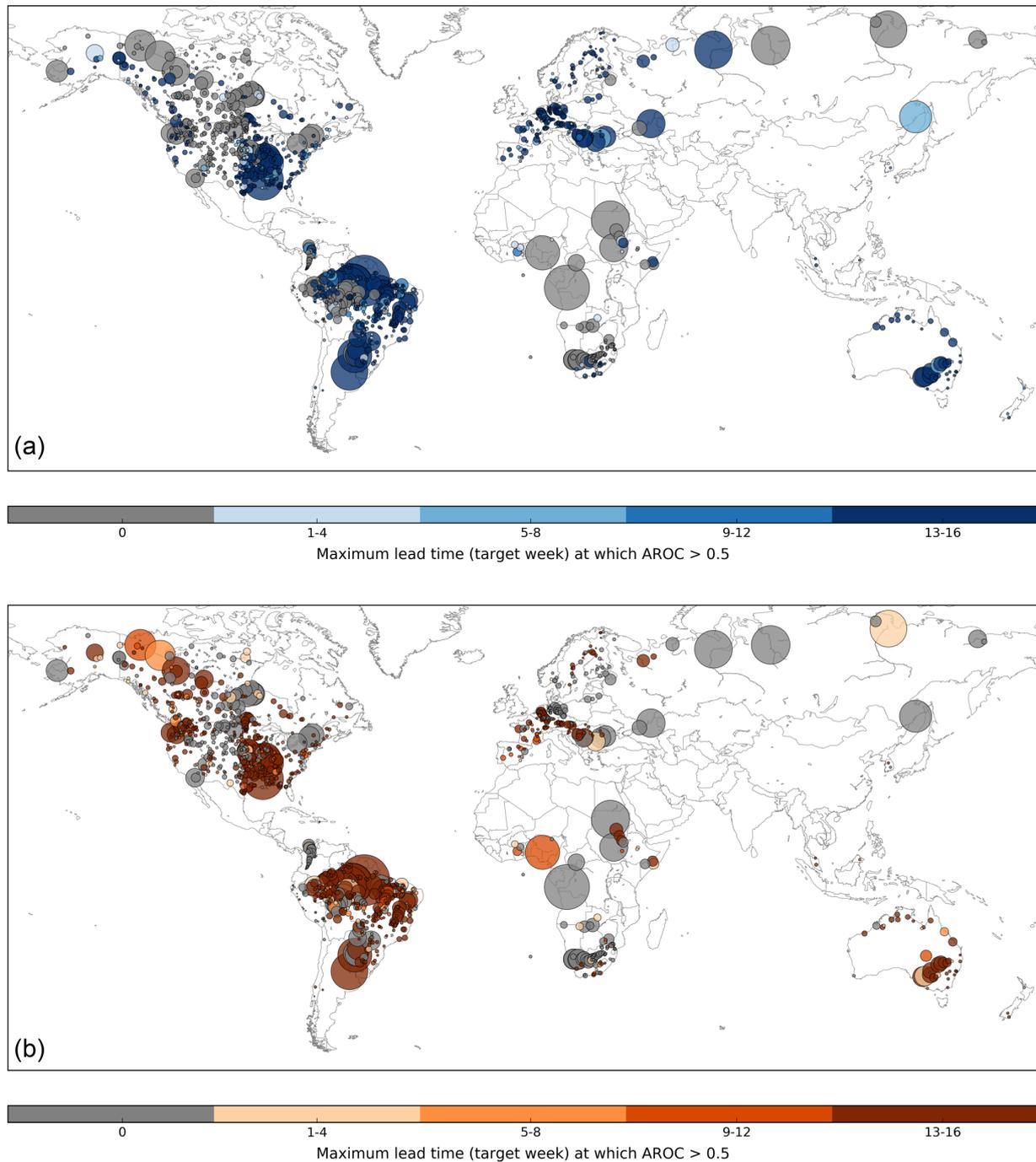


Figure 6. Maximum forecast lead time (target week, averaged across all months) at which the area under the ROC curve (AROC) is greater than 0.5 **(a)** for high flow events (flow exceeding the 80th percentile of climatology) and **(b)** low flow events (flow below the 20th percentile of climatology) at each observation station. This is used to indicate the maximum lead time at which forecasts are more skilful than the long-term average. Dot size corresponds to the upstream area of the location – thus larger dots represent larger rivers and vice versa. Grey dots indicate that (on average, across all months) forecasts are less skilful than climatology at all lead times.

ally higher. In contrast, forecasts for low flow events are less skilful than climatology across much of Europe. Particularly in north-east Europe and Scandinavia, forecasts produced in the summer months of JJA have an $\text{AROC} < 0.6$ at all sta-

tions, with only a few stations indicating any skill in other seasons, whereas in central and south-east Europe forecasts of low flow events are most skilful in JJA and SON out to 3–4 months ahead in the larger rivers. These results are similar

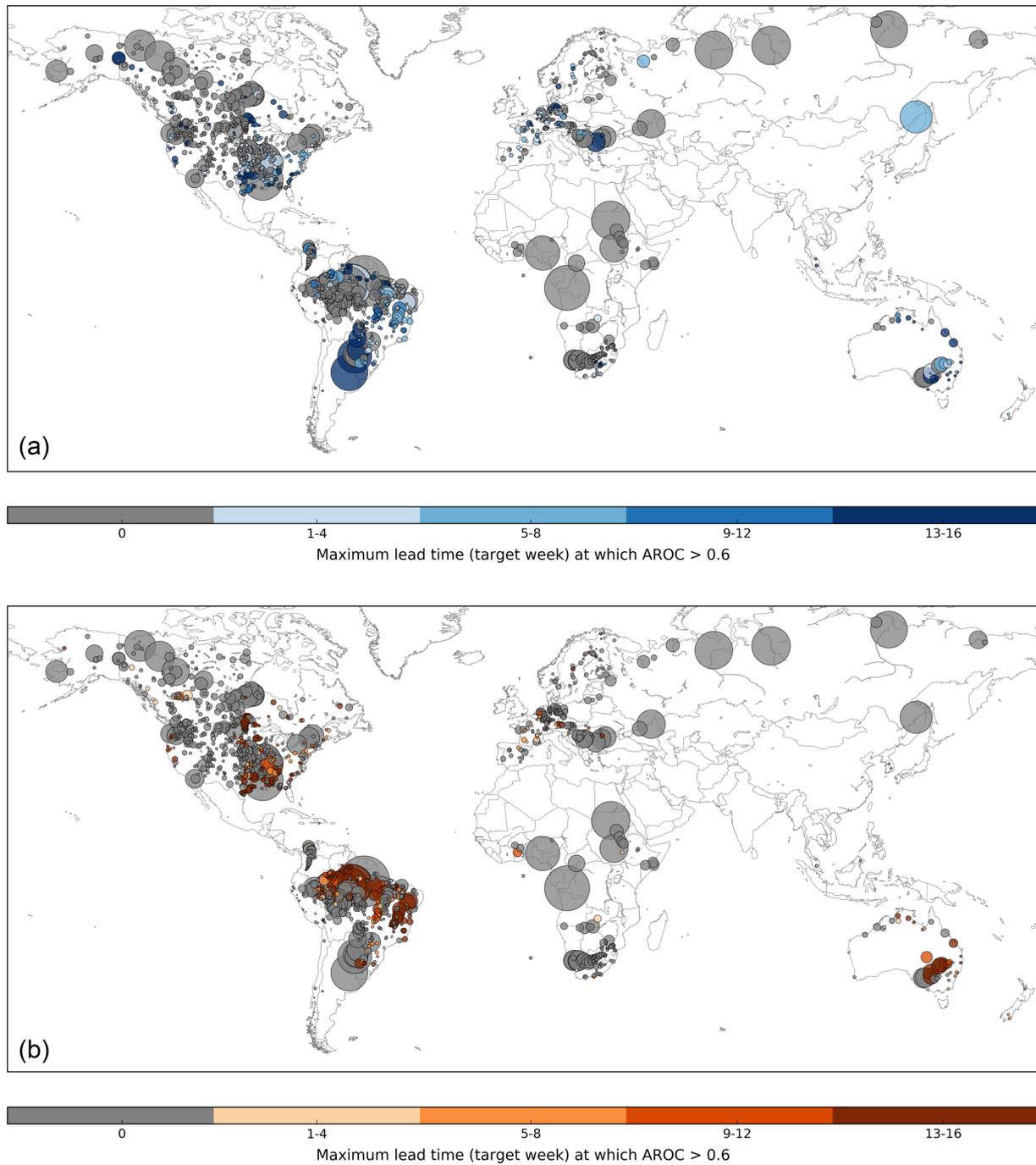


Figure 7. Maximum forecast lead time (target week, averaged across all months) at which the area under the ROC curve (AROC) is greater than 0.6 for **(a)** high flow events (flow exceeding the 80th percentile of climatology) and **(b)** low flow events (flow below the 20th percentile of climatology) at each observation station. This is used to indicate the maximum lead time at which forecasts are deemed skilful. Dot size corresponds to the upstream area of the location – thus larger dots represent larger rivers and vice versa. Grey dots indicate that (on average, across all months) forecasts are less skilful than climatology at all lead times. Maps for each season are provided in the Supplement.

to those of Arnal et al. (2018) for the potential usefulness of the EFAS seasonal outlook.

Asia. Although the number of available stations is very limited, the few stations available in South East Asia indicate

that the forecasts are potentially useful out to 3–4 months ahead, particularly for forecasts started in DJF and MAM preceding the start of the wet season. For low flow events, this skill extends into JJA, whereas forecasts made in SON

towards the end of the wet season tend to be less skilful than climatology.

Australia and New Zealand. Forecasts are most skilful out to longer lead times in the Murray–Darling river basin in the south-east, in particular for forecasts started in JJA and SON during the Southern Hemisphere winter and spring. In northern Australia, forecasts started in DJF and MAM for high flow events and MAM and JJA for low flow events are potentially useful out to 3–4 months ahead. This corresponds with the assessment of the skill of the Bayesian joint probability modelling approach for sub-seasonal to seasonal streamflow forecasting in Australia by Zhao et al. (2016), who found that forecasts in northern Australian catchments tend to be more skilful for the dry season (May to October) than the wet season (December to March). At the three stations in New Zealand, forecasts are only skilful for high flow events during the first month of lead time in DJF and MAM; however, for low flow events forecasts made in SON for the southern stations are potentially useful out to 4 months ahead.

Africa. While the spatial distribution of stations is limited, for high flow events forecasts are seen to be potentially useful at some of the stations in eastern Africa, particularly in SON and to a lesser extent in DJF. In southern Africa, there is skill in DJF and MAM, although the maximum lead time varies significantly from station to station. For low flow, there is little variation between the seasons; forecasts are generally less skilful than climatology across the continent, with some stations in DJF in southern and western Africa indicating skill in the first 1–2 months only.

4.2 Reliability

To provide an overall picture of the reliability of the GloFAS-Seasonal forecasts, attributes diagrams are produced for forecasts aggregated across all observation stations globally for both the 80th and 20th percentile events. In order to assess geographical differences in forecast reliability, attributes diagrams are also produced for forecasts aggregated across the stations within each of the major river basins used in the GloFAS-Seasonal forecast products (see Sect. 3.1). Many of these river basins do not contain a large enough number of stations to produce useful attributes diagrams, and as such the results in this section are presented for one river basin per continent for this initial evaluation. The river basin chosen for each continent is that which contains the largest number of observation stations.

The globally aggregated results (Fig. 8) indicate that, in general, the forecasts have more reliability than a forecast of climatology, though the reliability is less than perfect. It is important to note that the globally aggregated results shown in Fig. 8 mask any variability between river basins. Overall, the reliability appears to be slightly better for forecasts of high flow events than low flow events, and for lower probabilities, indicated by the steeper positive slope showing that as the forecast probability increases, so does the verified

chance of the event. The forecasts for both high and low flow events exhibit sharpness, although more so for high flow events, meaning that they have the ability to forecast probabilities that differ from the climatological average. This is indicated by the histograms inset within the attributes diagrams in Fig. 8; a forecast with sharpness will show a range of forecast probabilities differing from the climatological average (20%), and a forecast with perfect sharpness will show peaks in the forecast frequency at 0% and 100%. Forecasts with no or low sharpness will show a peak in the forecast frequency near the climatological average. A forecast can have sharpness but still be unreliable. Figure 8 also suggests that in general, GloFAS-Seasonal forecasts have a tendency to over-predict the likelihood of an event occurring.

The following paragraphs summarise the forecast reliability for one river basin per continent; for a map of the location of these river basins, please refer to Fig. S3. The attributes diagrams for these river basins for both the 80th and 20th percentile events and for each season are provided in Figs. S4–S8. Each attributes diagram displays the results for forecast weeks 4, 8, 12, and 16, representing the reliability out to 1, 2, 3, and 4 months ahead. There are no river basins in Asia containing enough stations to produce an attributes diagram.

South America, Tocantins River (Fig. S4). For high flow events, forecasts for the Tocantins River indicate good reliability in all seasons, particularly up to 50% probability. Forecasts in the higher-probability bins tend to over-predict, and this over-prediction worsens with lead time. In MAM and JJA, the forecasts tend to slightly under-predict in the lower-probability bins. The forecasts have sharpness, but it is clear that the sample size of high-probability forecasts is limited. There is a tendency to over-predict the likelihood of low flow events in all seasons, but the forecasts show good reliability for the lower-probability bins, particularly in SON and DJF. In JJA, the resolution of the forecasts is low.

North America, Lower Mississippi River (Fig. S5). For high flow events, the sample size of high-probability forecasts is small, and as such it is difficult to evaluate the reliability of these forecasts. The forecasts at lower probabilities have good reliability, particularly out to 2 months ahead in MAM and JJA. In SON and DJF, forecasts are more reliable at longer lead times. There is a tendency to under-predict at low probabilities and over-predict at high probabilities. For low flow events, the forecasts have a tendency to over-predict in all seasons, and the resolution of the forecasts is lower than for high flow events. At higher probabilities, forecasts of low flow events are more reliable than climatology, but the resolution is particularly low for probabilities up to 50–60%. The forecasts for both high and low flow events have sharpness.

Europe, River Rhône (Fig. S6). For the River Rhône, the reliability is better than climatology at all lead times for high flow events, although there is a lack of forecasts of higher probabilities, particularly in MAM and JJA, as may be expected in the summer months. In SON, the reliability of forecasts up to 60–70% is good at all lead times, and in DJF the

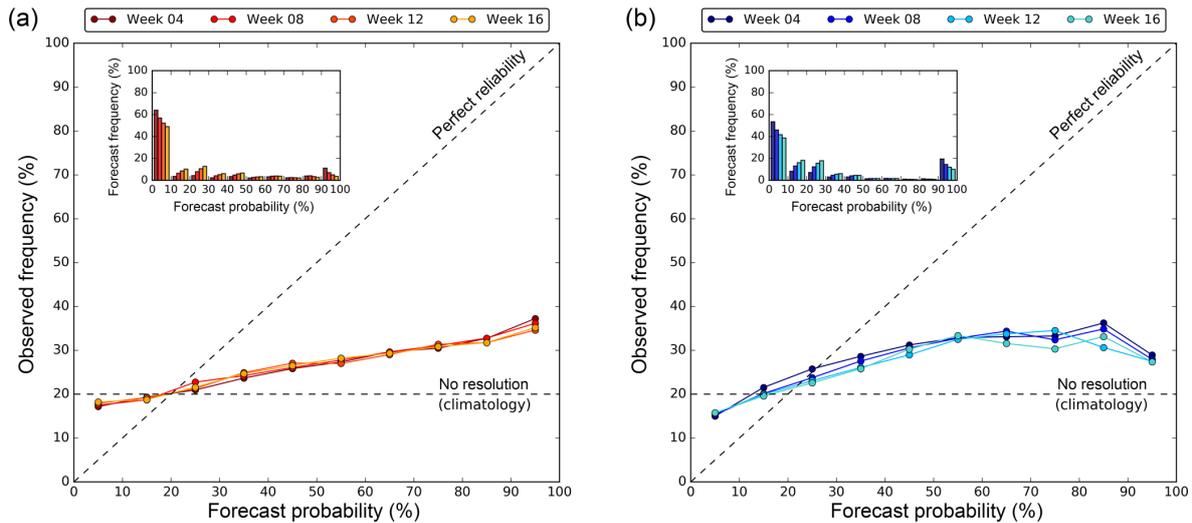


Figure 8. Attributes diagram for forecasts of (a) low flow events (flow below the 20th percentile of climatology) and (b) high flow events (flow exceeding the 80th percentile of climatology) aggregated across all observation stations globally. Results are shown for lead time weeks 4, 8, 12, and 16 and indicate the reliability of the forecasts. The histograms (inset) show the frequency at which forecasts occur in each probability bin and are used to indicate forecast sharpness. Attributes diagrams for selected river basins are provided in the Supplement.

forecasts are more reliable in the first 2 months of lead time for most probability bins. The reliability is less good for low flow events, but is generally better than climatology, particularly in summer (JJA). In winter (DJF), the resolution and reliability of the forecasts is poor. For all seasons and lead times and for both events, the forecasts have sharpness.

Australia, Murray River (Fig. S7). The attributes diagrams for both high and low flow events indicate that forecasts are often over-confident in this river basin, with probabilities of 0%–10% for low flow events and 0%–30% and 90%–100% for high flow events, occurring frequently. As such, the sample size of forecasts in several of the bins is low. For high flow events, forecasts tend to over-predict at high probabilities and under-predict at low probabilities. The reliability is very good up to ~30%, after which the sample size is too small. For low flow events, there is a tendency to under-predict, but based on the forecasts available, the reliability is better than climatology at all lead times. The reliability for low flow events is better in SON and DJF (spring and summer) than MAM and JJA (autumn and winter), and for high flow events there is less differentiation between the seasons.

Africa, Orange River (Fig. S8). For the Orange River, forecasts of high flow events exhibit good reliability for lower probabilities in SON, DJF, and MAM (spring through autumn), particularly at longer lead times in SON and DJF, with a tendency to over-predict at higher probabilities. Resolution and reliability are poor for high flow events in JJA (winter), with probabilities of 90%–100% predicted too frequently. For low flow events, forecasts of 0%–10% are very frequent, and the forecasts under-predict in all seasons, although the reliability is better than climatology at all lead times (based

on a limited sample of forecasts for most probability bins). Reliability for low flow events is best in DJF (summer).

4.3 Discussion

The results presented provide an initial evaluation of the potential usefulness and reliability of GloFAS-Seasonal forecasts. For decision-making purposes, it is important to measure the ability of a forecasting system to predict the correct category of an event. As such, an event-based evaluation of the forecasts is used to assess whether the forecasts were able to correctly predict observed high and low river flow events over a 17-year period and whether it is able to do so with good reliability. The initial results are promising, indicating that the forecasts are, on average, potentially useful up to 1–2 months ahead in many rivers worldwide and up to 3–4 months ahead in some locations. The GloFAS-Seasonal forecasts have sharpness, i.e. they are able to predict forecasts with probabilities that differ from climatology, and overall have better reliability than a forecast of climatology, but with a tendency to over-predict at higher probabilities. It is also clear that there is a frequency bias in the reliability results, as often there is a small sample of high-probability forecasts. Typically, the reliability is seen to be better when there is a higher forecast frequency on which to base the results. As would be expected, the potential usefulness and reliability of the forecasts vary by region, season, and forecast lead time.

Considering the evaluation results by season allows for further analysis of the times of year in which the forecasts are potentially useful and/or reliable. For example, in south-east Australia, forecasts are seen to be potentially useful up to 4 months ahead in JJA and SON, but for forecasts produced

in DJF the skill only extends to 1 month ahead, and forecasts are less skilful than climatology at several of the stations in MAM. In many rivers across the globe, it is the case that forecasts are potentially useful in some seasons, but not in others, and may be more reliable in certain seasons than others. As such, the maps provided in Figs. S1 and S2 are intended to highlight where and when the forecasts are likely to be useful, information that is key in terms of decision-making.

It is clear that there are regions and seasons in which the forecasts are less skilful than climatology and do not have good reliability, and thus in these rivers it would be more useful to use a long-term average climatology than seasonal hydro-meteorological forecasts of river flow. This lack of skill could be due to several factors, such as certain hydrological regimes that may not be well-represented in the hydrological model or may be difficult to forecast at these lead times (for example, snow-dominated catchments or regions where convective storms produce most of the rainfall in some seasons), poor skill of the meteorological forecast input, poor initial conditions from the ERA5-R reanalysis, extensive management of rivers that cannot be represented by the current model, or the lack of model calibration. While this initial evaluation is designed to provide an overview of whether the forecasts are potentially useful and reliable in predicting high and low flow events, more extensive analysis is required to diagnose the sources of predictability in the forecasts and the potential causes of poor skill. Additionally, it is evident that observations of river flow, particularly covering the reforecast period, are both spatially and temporally limited across large areas of the globe. A more extensive analysis should make use of the globally consistent ERA5-R river flow reanalysis as a benchmark in order to fully assess the forecast skill worldwide, including in regions where no observations are available.

The verification metrics used also require that a high or low flow event is predicted with the correct timing in the same week as that in which it occurred. This is asking a lot of a seasonal forecasting system and for many applications, such as water resources and reservoir management, a forecast of the exact week in which an event is expected at a lead time of several months ahead may not be necessary. That such a system shows real skill despite this being a tough test for the model and is able to successfully predict observed high or low river flow in a specific week, several weeks or months ahead, provides optimism for the future of global-scale seasonal hydro-meteorological forecasting. Further evaluation should aim to assess the skill of the forecasts with a more relaxed constraint on the event timing and also make use of alternative skill measures to cover different aspects of the forecast skill, such as the spread and bias of the forecasts. It will also be important to assess whether the use of weekly averaged river flow is the most appropriate way to display the forecasts. While this is commonly used for applications such as drought early awareness and water resources management, there may be other aspects of decision-making,

such as flood forecasting, for which other measures may be more appropriate, for example daily averages or floodiness (Stephens et al., 2015).

Future development of GloFAS-Seasonal will aim to address these evaluation results and improve the skill and reliability of the current forecasts; it will also aim to overcome some of the grand challenges in operational hydrological forecasting, such as seamless forecasting and the use of data assimilation. Seamless forecasting will be key in the future development of GloFAS; the use of two different meteorological forecast inputs for the medium-range and seasonal versions of the model means that discrepancies can occur between the two timescales, thus producing confusing and inconsistent forecast information for users. Additionally, the use of river flow observations could lead to significant improvements in skill through calibration of the model using historical observations and assimilation of real-time data to adjust the forecasts. This remains a grand challenge due to the lack of openly available river flow data, particularly in real time.

5 Conclusions

In this paper, the development and implementation of a global-scale operational seasonal hydro-meteorological forecasting system, GloFAS-Seasonal, was presented, and an event-based forecast evaluation was carried out using two different but complementary verification metrics to assess the capability of the forecasts to predict high and low river flow events.

GloFAS-Seasonal provides forecasts of high or low river flow out to 4 months ahead for the global river network through three new forecast product layers via the openly available GloFAS web interface at <http://www.globalfloods.eu> (last access: 16 August 2018). Initial evaluation results are promising, indicating that in many rivers, forecasts are both potentially useful, i.e. more skilful than a long-term average climatology out to several months ahead in some cases, and overall more reliable than a forecast of climatology. Forecast skill and reliability vary significantly by region and by season.

The initial evaluation, however, also indicates a tendency of the forecasts to over-predict in general, and in some regions forecasts are currently less skilful than climatology; future development of the system will aim to improve the forecast skill and reliability with a view to providing potentially useful forecasts across the globe. Development of GloFAS-Seasonal will continue based on results of the forecast evaluation and on feedback from GloFAS partners and users worldwide in order to provide a forecast product that remains state of the art in hydro-meteorological forecasting and caters to the needs of its users. Future versions are likely to address some of the grand challenges in hydro-meteorological forecasting in order to improve forecast skill, such as data assim-

ilation, and will also include more features, such as flexible percentile thresholds and indication of the forecast skill via the interface. A further grand challenge that is important in terms of global-scale hydro-meteorological forecasting, and indeed for the development of GloFAS, is the need for more observed data (Emerton et al., 2016), which is essential not only for providing initial conditions to force the models, but also for evaluation of the forecasts and continuous improvement of forecast accuracy.

While such a forecasting system requires extensive computing resources, the potential for use in decision-making across a range of water-related sectors, and the promising results of the initial evaluation, suggest that it is a worthwhile use of time and resources to develop such global-scale systems. Recent papers have highlighted the fact that seasonal forecasts of precipitation are not necessarily a good indicator of potential floodiness and called for investment in better forecasts of seasonal flood risk (Coughlan De Perez et al., 2017; Stephens et al., 2015). Coughlan de Perez et al. (2017) state that “ultimately, the most informative forecasts of flood hazard at the seasonal scale could be seasonal streamflow forecasts using hydrological models” and that better seasonal forecasts of flood risk could be hugely beneficial for disaster preparedness.

GloFAS-Seasonal represents a first attempt at overcoming the challenges of producing and providing openly available seasonal hydro-meteorological forecast products, which are key for organisations working at the global scale and for regions where no other forecasting system exists. We provide, for the first time, seasonal forecasts of hydrological variables for the global river network by driving a hydrological model with seasonal meteorological forecasts. GloFAS-Seasonal forecasts could be used in addition to other forecast products, such as seasonal rainfall forecasts and short-range forecasts from national hydro-meteorological centres across the globe, to provide useful added information for many water-related applications from water resources management and agriculture to disaster risk reduction.

Code availability. The ECMWF IFS source code is available subject to a licence agreement, and as such access is available to the ECMWF member-state weather services and other approved partners. The IFS code is also available for educational and academic purposes as part of the OpenIFS project (ECMWF, 2011, 2018a), with full forecast capabilities and including the HTESSEL land surface scheme, but without modules for data assimilation. Similarly, the GloFAS river routing component source code is not openly available; however, the “forecast product” code (prior to implementation in ecFlow) that was newly developed for GloFAS-Seasonal and used for a number of tasks such as computing exceedance probabilities and producing the graphics for the interface is provided in the Supplement.

Data availability. ECMWF’s ERA5 reanalysis and SEAS5 reforecasts are available through the Copernicus Climate Data Store (Copernicus, 2018a). The ERA5-R river flow reanalysis and the GloFAS-Seasonal reforecasts (daily data) are currently available from the authors on request and will be made available through ECMWF’s data repository in due course. The majority of the observed river flow data were provided by the Global Runoff Data Centre (GRDC; BfG, 2017). These data are freely available from <https://www.bafg.de/> (last access: 16 August 2018). Additional data were provided by the Russian State Hydrological Institute (SHI, 2018), the European Flood Awareness System (EFAS, 2017), Somalia Water and Land Information Management (SWALIM, 2018), South Africa Department for Water and Sanitation (DWA, 2018), Colombia Institute of Hydrology, Meteorology and Environmental Studies (IDEAM, 2014), Nicaragua Institute of Earth Studies (INETER, 2016), Dominican Republic National Institute of Hydraulic Resources (INDRHI, 2017), Brazil National Centre for Monitoring and Forecasting of Natural Hazards (Cemaden, 2017), Environment Canada Water Office (Environment Canada, 2014), Nepal Department of Hydrology and Meteorology (DHM, 2017), Red Cross Red Crescent Climate Centre (RCCC, 2018), Chile General Water Directorate (DGA, 2018), and the Historical Database on Floods (BDHI, 2018).

Supplement. The supplement related to this article is available online at: <https://doi.org/10.5194/gmd-11-3327-2018-supplement>.

Author contributions. FP proposed the development of GloFAS-Seasonal, RE wrote the GloFAS-Seasonal forecast product code, and RE and LA designed the forecast products. EZ built the ecFlow suite and produced ERA5-R and the GloFAS-Seasonal reforecasts, and DM provided technical support for the website and operational implementation. RE evaluated the forecasts and wrote the paper, with the exception of Sect. 2.4, written by DM. All authors were involved in discussions throughout development, and all authors commented on the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work has been funded by the Natural Environment Research Council (NERC) as part of the SCENARIO Doctoral Training Partnership under grant NE/L002566/1. Ervin Zsoter, Davide Muraro, Christel Prudhomme, and Peter Salamon were supported by the Copernicus Emergency Management Service – Early Warning Systems (CEMS-EWS; EFAS). Louise Arnal, Hannah L. Cloke, and Florian Pappenberger acknowledge financial support from the Horizon 2020 IMPREX project (grant agreement no. 641811). Elisabeth M. Stephens is thankful for support from NERC and the Department for International Development (grant number NE/P000525/1) under the Science for Humanitarian Emergencies and Resilience (SHEAR) research programme (project FATHUM: Forecasts for Anticipatory HUMANitarian action).

Edited by: Jeffrey Neal

Reviewed by: two anonymous referees

References

- Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., and Pappenberger, F.: GloFAS – global ensemble streamflow forecasting and flood early warning, *Hydrol. Earth Syst. Sci.*, 17, 1161–1175, <https://doi.org/10.5194/hess-17-1161-2013>, 2013.
- Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B., and Pappenberger, F.: Skilful seasonal forecasts of streamflow over Europe?, *Hydrol. Earth Syst. Sci.*, 22, 2057–2072, <https://doi.org/10.5194/hess-22-2057-2018>, 2018.
- Bahra, A.: Managing work flows with ecFlow, *ECMWF Newsl.*, 129, 30–32 available from: <https://www.ecmwf.int/sites/default/files/elibrary/2011/14594-newsletter-no129-autumn-2011.pdf> (last access: 18 April 2018), 2011.
- Balsamo, G., Pappenberger, F., Dutra, E., Viterbo, P., and van den Hurk, B.: A revised land hydrology in the ECMWF model: a step towards daily water flux prediction in a fully-closed water cycle, *Hydrol. Process.*, 25, 1046–1054, <https://doi.org/10.1002/hyp.7808>, 2011.
- BDHI: Base de Donnees Historiques sur les Inondations, available at: <http://bdhi.fr/appli/web/welcome>, last access: 23 April 2018.
- Bell, V. A., Davies, H. N., Kay, A. L., Brookshaw, A., and Scaife, A. A.: A national-scale seasonal hydrological forecast system: development and evaluation over Britain, *Hydrol. Earth Syst. Sci.*, 21, 4681–4691, <https://doi.org/10.5194/hess-21-4681-2017>, 2017.
- Bennett, J. C., Wang, Q. J., Li, M., Robertson, D. E., and Schepen, A.: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model, *Water Resour. Res.*, 52, 8238–8259, <https://doi.org/10.1002/2016WR019193>, 2016.
- Bennett, J. C., Wang, Q. J., Robertson, D. E., Schepen, A., Li, M., and Michael, K.: Assessment of an ensemble seasonal streamflow forecasting system for Australia, *Hydrol. Earth Syst. Sci.*, 21, 6007–6030, <https://doi.org/10.5194/hess-21-6007-2017>, 2017.
- BfG: The GRDC, available at: http://www.bafg.de/GRDC/EN/Home/homepage_node.html (last accessed: 23 April 2018), 2017.
- BoM: Seasonal Streamflow Forecasts: Water Information: Bureau of Meteorology, available at: <http://www.bom.gov.au/water/ssf/about.shtml>, last access: 24 April 2018.
- Candogan Yossef, N., van Beek, R., Weerts, A., Winsemius, H., and Bierkens, M. F. P.: Skill of a global forecasting system in seasonal ensemble streamflow prediction, *Hydrol. Earth Syst. Sci.*, 21, 4103–4114, <https://doi.org/10.5194/hess-21-4103-2017>, 2017.
- Cemaden: Cemaden – Centro Nacional de Monitoramento e Alertas de Desastres Naturais, available at: <http://www.cemaden.gov.br/> (last access: 23 April 2018), 2017.
- Chiew, F. H. S. and McMahon, T. A.: Global ENSO-streamflow teleconnection, streamflow forecasting and interannual variability, *Hydrol. Sci. J.*, 47, 505–522, <https://doi.org/10.1080/02626660209492950>, 2002.
- Chow, V. Te, Maidment, D. R., and Mays, L. W.: Applied hydrology, Tata McGraw-Hill Education, available at: https://books.google.co.uk/books/about/Applied_Hydrology.html?id=RRwidSsBJrEC&redir_esc=y (last access: 17 November 2017), 2010.
- Cloke, H., Pappenberger, F., Thielen, J., and Thiemiig, V.: Operational European Flood Forecasting, in *Environmental Modelling*, John Wiley & Sons, Ltd, Chichester, UK, 415–434, 2013.
- Copernicus: Copernicus Climate Data Store, available at: <https://climate.copernicus.eu/climate-data-store>, last access: 23 April 2018a.
- Copernicus: SWICCA, Service for Water Indicators in Climate Change Adaptation, available at: <http://swicca.climate.copernicus.eu/>, last access: 12 January 2018b.
- Coughlan de Perez, E., Stephens, E., Bischiniotis, K., van Aalst, M., van den Hurk, B., Mason, S., Nissan, H., and Pappenberger, F.: Should seasonal rainfall forecasts be used for flood preparedness?, *Hydrol. Earth Syst. Sci.*, 21, 4517–4524, <https://doi.org/10.5194/hess-21-4517-2017>, 2017.
- Crochmore, L., Ramos, M.-H., and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, *Hydrol. Earth Syst. Sci.*, 20, 3601–3618, <https://doi.org/10.5194/hess-20-3601-2016>, 2016.
- Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J., Zhu, Y., Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J., and Zhu, Y.: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, *Am. Meteorol. Soc.*, 95, 79–98, <https://doi.org/10.1175/BAMS-D-12-00081.1>, 2014.
- DGA: Ministerio de Obras Públicas – Dirección de General de Aguas, available at: <http://www.dga.cl/Paginas/default.aspx>, last access: 23 April 2018.
- DHM: Department of Hydrology and Meteorology, available at: <http://www.dhm.gov.np/> (last access: 23 April 2018), 2017.
- DWA: Department: Water and Sanitation, available at: <http://www.dwa.gov.za/default.aspx>, last access: 23 April 2018.
- ECMWF: OpenIFS, available at: <https://www.ecmwf.int/en/research/projects/openifs> (last access: 16 August 2018), 2011.
- ECMWF: ecFlow Documentation, available at: <https://software.ecmwf.int/wiki/display/ECFLOW/Documentation> (last access: 18 April 2018), 2012.
- ECMWF: SEAS5 user guide, available at: https://www.ecmwf.int/sites/default/files/medialibrary/2017-10/System5_guide.pdf (last access: 18 April 2018), 2017a.
- ECMWF: What are the changes from ERA-Interim to ERA5? – Copernicus Knowledge Base – ECMWF Confluence Wiki, available at: <https://software.ecmwf.int/wiki/pages/viewpage.action?pageId=74764925> (last access: 24 April 2018), 2017b.
- ECMWF: About OpenIFS, available at: <https://software.ecmwf.int/wiki/display/OIFS/About+OpenIFS>, last access: 26 April 2018a.
- ECMWF: ECMWF IFS Documentation CY43R1, available at: https://www.ecmwf.int/search/elibrary/IFS?secondary_title=IFSDocumentationCY43R1, last access: 18 April 2018b.
- EFAS: European Flood Awareness System (EFAS), available at: <https://www.efas.eu/> (last access: 23 April 2018), 2017.

- Emerton, R., Cloke, H. L., Stephens, E. M., Zsoter, E., Woolnough, S. J., and Pappenberger, F.: Complex picture for likelihood of ENSO-driven flood hazard, *Nat. Commun.*, 8, 14796, <https://doi.org/10.1038/ncomms14796>, 2017.
- Emerton, R. E., Stephens, E. M., Pappenberger, F., Pagano, T. C., Weerts, A. H., Wood, A. W., Salamon, P., Brown, J. D., Hjerdt, N., Donnelly, C., Baugh, C. A., and Cloke, H. L.: Continental and global scale flood forecasting systems, *Wiley Interdiscip. Rev. Water*, 3, 391–418, <https://doi.org/10.1002/wat2.1137>, 2016.
- Environment Canada: Water Level and Flow – Environment Canada, available at: <https://wateroffice.ec.gc.ca/> (last access: 23 April 2018), 2014.
- Environmental Systems Research Institute: ArcMap, ArcGIS Desktop, available at: <http://desktop.arcgis.com/en/arcmap/>, last access: 26 April 2018.
- Fekete, B. M., Vörösmarty, C. J., and Lammers, R. B.: Scaling gridded river networks for macroscale hydrology: Development, analysis, and control of error, *Water Resour. Res.*, 37, 1955–1967, <https://doi.org/10.1029/2001WR900024>, 2001.
- GloFAS: GloFAS Web Map Service Time (WMS-T) User Manual, available at: http://www.globalfloods.eu/static/downloads/GloFAS-WMS-T_usermanual.pdf last access: 26 April 2018a.
- GloFAS: GloFAS WMS-T, available at: <http://globalfloods-ows.ecmwf.int/glofas-ows/?service=WMS&request=GetCapabilities>, last access: 16 August 2018b.
- IDEAM: IDEAM, available at: <http://www.ideam.gov.co/> (last access: 23 April 2018), 2014.
- INDRHI: INDRHI – National Institute of Hydraulic Resources, available at: <http://indrhi.gob.do/> (last access: 23 April 2018), 2017.
- INETER: Ineter, Instituto Nicaragüense de Estudios Territoriales, available at: <http://www.ineter.gob.ni/> (last access: 23 April 2018), 2016.
- Lehner, B. and Grill, G.: Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems, *Hydrol. Process.*, 27, 2171–2186, <https://doi.org/10.1002/hyp.9740>, 2013.
- Lehner, B., Verdin, K., and Jarvis, A.: New Global Hydrography Derived From Spaceborne Elevation Data, *Eos, Trans. Am. Geophys. Union*, 89, 93–94, <https://doi.org/10.1029/2008EO100001>, 2008.
- Lorenz, E. N.: Deterministic Nonperiodic Flow, *J. Atmos. Sci.*, 20, 130–141, [https://doi.org/10.1175/1520-0469\(1963\)020<0130:DNF>2.0.CO;2](https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2), 1963.
- Lorenz, E. N.: The essence of chaos, University of Washington Press, 1993.
- Mason, S. J. and Graham, N. E.: Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels, *Weather Forecast.*, 14, 713–725, [https://doi.org/10.1175/1520-0434\(1999\)014<0713:CPROCA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2), 1999.
- McPhaden, M. J., Zebiak, S. E., and Glantz, M. H.: ENSO as an integrating concept in earth science., *Science*, 314, 1740–1745, <https://doi.org/10.1126/science.1132588>, 2006.
- Meißner, D., Klein, B., and Ionita, M.: Development of a monthly to seasonal forecast framework tailored to inland waterway transport in central Europe, *Hydrol. Earth Syst. Sci.*, 21, 6401–6423, <https://doi.org/10.5194/hess-21-6401-2017>, 2017.
- Mo, K. C., Lettenmaier, D. P., Mo, K. C., and Lettenmaier, D. P.: Hydrologic Prediction over the Conterminous United States Using the National Multi-Model Ensemble, *J. Hydrometeorol.*, 15, 1457–1472, <https://doi.org/10.1175/JHM-D-13-0197.1>, 2014.
- OGC: OGC Web Map Service v1.3.0, <https://doi.org/10.3173/air.21.76>, 2015.
- OpenLayers: OpenLayers, available at: <http://openlayers.org/>, last access: 18 April 2018.
- Open Source Geospatial Foundation: MapServer 7.0.1 documentation, available at: <http://mapserver.org/uk/index.html> (last access: 26 April 2018), 2016.
- Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., and Salamon, P.: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction, *J. Hydrol.*, 522, 697–713, <https://doi.org/10.1016/j.jhydrol.2015.01.024>, 2015.
- Prudhomme, C., Hannaford, J., Harrigan, S., Boorman, D., Knight, J., Bell, V., Jackson, C., Svensson, C., Parry, S., Bachiller-Jareno, N., Davies, H., Davis, R., Mackay, J., McKenzie, A., Rudd, A., Smith, K., Bloomfield, J., Ward, R., and Jenkins, A.: Hydrological Outlook UK: an operational streamflow and groundwater level forecasting system at monthly to seasonal time scales, *Hydrolog. Sci. J.*, 62, 2753–2768, <https://doi.org/10.1080/02626667.2017.1395032>, 2017.
- QGIS Development Team: Quantum GIS Geographical Information System, available at: <https://www.qgis.org/>, last access: 4 December 2017.
- RCCC: Home – Red Cross Red Crescent Climate Centre, available at: <http://www.climatecentre.org/>, last access: 23 April 2018.
- SHI: “State Hydrological Institute” (SHI), Russian Federal State Budgetary Organization, available at: <http://www.hydrology.ru/en>, last access: 23 April 2018.
- Stephens, E., Day, J. J., Pappenberger, F., and Cloke, H.: Precipitation and floodiness, *Geophys. Res. Lett.*, 42, 10316–10323, <https://doi.org/10.1002/2015GL066779>, 2015.
- Stockdale, T., Johnson, S., Ferranti, L., Balmaseda, M., and Briceag, S.: ECMWF's new long-range forecasting system SEAS5, *ECMWF Newsl.*, 154, 15–20, available at: <http://www.ecmwf.int/en/about/news-centre/media-resources>, last access: 18 April 2018.
- SWALIM: FAO SWALIM: Somalia Water and Land Information Management, available at: <http://www.faoswalim.org/>, last access: 23 April 2018.
- Van Der Knijff, J. M., Younis, J., and De Roo, A. P. J.: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, *Int. J. Geogr. Inf. Sci.*, 24, 189–212, <https://doi.org/10.1080/13658810802549154>, 2010.
- Ward, P. J., Eisner, S., Flörke, M., Dettinger, M. D., and Kummerow, M.: Annual flood sensitivities to El Niño–Southern Oscillation at the global scale, *Hydrol. Earth Syst. Sci.*, 18, 47–66, <https://doi.org/10.5194/hess-18-47-2014>, 2014a.
- Ward, P. J., Jongman, B., Kummerow, M., Dettinger, M. D., Sperna Weiland, F. C., and Winsemius, H. C.: Strong influence of El Niño Southern Oscillation on flood risk around the world, *P. Natl. Acad. Sci. USA*, 111, 15659–15664, <https://doi.org/10.1073/pnas.1409822111>, 2014b.
- Ward, P. J., Kummerow, M., and Lall, U.: Flood frequencies and durations and their response to El Niño Southern Oscillation: Global analysis, *J. Hydrol.*, 539, 358–378, <https://doi.org/10.1016/j.jhydrol.2016.05.045>, 2016.

- Weisheimer, A. and Palmer, T. N.: On the reliability of seasonal climate forecasts, *J. R. Soc. Interface*, 11, 20131162, <https://doi.org/10.1098/rsif.2013.1162>, 2014.
- Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D. P.: Long-range experimental hydrologic forecasting for the eastern United States, *J. Geophys. Res.*, 107, 4429, <https://doi.org/10.1029/2001JD000659>, 2002.
- Wood, A. W., Kumar, A., and Lettenmaier, D. P.: A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States, *J. Geophys. Res.*, 110, D04105, <https://doi.org/10.1029/2004JD004508>, 2005.
- Yuan, X., Wood, E. F., Chaney, N. W., Sheffield, J., Kam, J., Liang, M., and Guan, K.: Probabilistic Seasonal Forecasting of African Drought by Dynamical Models, *J. Hydrometeorol.*, 14, 1706–1720, <https://doi.org/10.1175/JHM-D-13-054.1>, 2013.
- Yuan, X., Wood, E. F., and Ma, Z.: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, *Wiley Interdiscip. Rev. Water*, 2, 523–536, <https://doi.org/10.1002/wat2.1088>, 2015a.
- Yuan, X., Roundy, J. K., Wood, E. F., Sheffield, J., Yuan, X., Roundy, J. K., Wood, E. F., and Sheffield, J.: Seasonal Forecasting of Global Hydrologic Extremes: System Development and Evaluation over GEWEX Basins, *B. Am. Meteorol. Soc.*, 96, 1895–1912, <https://doi.org/10.1175/BAMS-D-14-00003.1>, 2015b.
- Zhao, T., Schepen, A., and Wang, Q. J.: Ensemble forecasting of sub-seasonal to seasonal streamflow by a Bayesian joint probability modelling approach, *J. Hydrol.*, 541, 839–849, <https://doi.org/10.1016/j.jhydrol.2016.07.040>, 2016.

A5: The 2013/14 Thames basin floods: do improved meteorological forecasts lead to more skilful hydrological forecasts at seasonal time scales?

This paper presents a co-author contribution arising through collaboration during this PhD, summarised in Chapter 3, Sect. 3.4, and has the following reference:

Neumann, J. L., L. Arnal, L. Magnusson and H. Cloke, 2018a: The 2013/14 Thames basin floods: Do improved meteorological forecasts lead to more skillful hydrological forecasts at seasonal time scales?, *J. Hydrometeorol.*, 19, 6, 1059-1075, doi:10.1175/JHM-D-17-0182.1*

* ©2018. The Authors. Journal of Hydrometeorology, a journal of the American Meteorological Society. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.

The 2013/14 Thames Basin Floods: Do Improved Meteorological Forecasts Lead to More Skillful Hydrological Forecasts at Seasonal Time Scales?

JESSICA NEUMANN

Department of Geography and Environmental Science, University of Reading, Reading, United Kingdom

LOUISE ARNAL

Department of Geography and Environmental Science, University of Reading, and European Centre for Medium Range Weather Forecasts, Reading, United Kingdom

LINUS MAGNUSSON

European Centre for Medium Range Weather Forecasts, Reading, United Kingdom

HANNAH CLOKE

Department of Geography and Environmental Science, and Department of Meteorology, University of Reading, Reading, United Kingdom

(Manuscript received 27 September 2017, in final form 7 April 2018)

ABSTRACT

The Thames basin experienced 12 major Atlantic depressions in winter 2013/14, leading to extensive and prolonged fluvial and groundwater flooding. This exceptional weather coincided with highly anomalous meteorological conditions across the globe. Atmospheric relaxation experiments, whereby conditions within specified regions are relaxed toward a reanalysis, have been used to investigate teleconnection patterns. However, no studies have examined whether improvements to seasonal meteorological forecasts translate into more skillful seasonal hydrological forecasts. This study applied relaxation experiments to reforecast the 2013/14 floods for three Thames basin catchments with different hydrogeological characteristics. The tropics played an important role in the development of extreme conditions over the Thames basin. The greatest hydrological forecasting skill was associated with the tropical Atlantic and less with the tropical Pacific, although both captured seasonal meteorological flow anomalies. Relaxation applied over the northeastern Atlantic produced confident ensemble forecasts, but hydrological extremes were underpredicted; this was unexpected with relaxation applied so close to the United Kingdom. Streamflow was most skillfully forecast for the catchment representing a large drainage area with high peak flow. Permeable lithology and antecedent conditions were important for skillfully forecasting groundwater levels. Atmospheric relaxation experiments can improve our understanding of extratropical anomalies and the potential predictability of extreme events such as the Thames 2013/14 floods. Seasonal hydrological forecasts differed from what was expected from the meteorology alone, and thus knowledge is gained by considering both components. In the densely populated Thames basin, considering the local hydrogeological context can provide an effective early alert of potential high-impact events, allowing for better preparedness.

1. Introduction

The prediction of water availability over seasonal time scales is beneficial for many aspects of the water sector, including flood forecasting, water supply, hydropower

 Denotes content that is immediately available upon publication as open access.

Corresponding author: Jessica Neumann, j.l.neumann@reading.ac.uk



This article is licensed under a [Creative Commons Attribution 4.0 license](http://creativecommons.org/licenses/by/4.0/) (<http://creativecommons.org/licenses/by/4.0/>).

DOI: 10.1175/JHM-D-17-0182.1

© 2018 American Meteorological Society

generation, and navigation. For contingency planners, skillful seasonal hydrological forecasts (SHFs) of river and groundwater levels have the potential to provide an indication of possible flood events weeks or months in advance, allowing for more optimal and consistent decisions to be made (Arnal et al. 2017). The operational use of SHFs, however, remains a challenge because of uncertainties posed by the initial hydrologic conditions (e.g., soil moisture, groundwater levels) and seasonal climate forcings (mainly forecasts of precipitation and temperature) that lead to a decrease in skill with increasing lead times (Wood and Lettenmaier 2008; Svensson 2016).

Across the United Kingdom and Europe, seasonal streamflow and groundwater forecast methods are currently being developed for application, for example, the U.K. Met Office Global Seasonal Forecast System (GloSea5), the Hydrological Outlook UK, and the Copernicus European Flood Awareness System (MacLachlan et al. 2015; Mackay et al. 2015; Svensson 2016), supported by Copernicus projects including Service for Water Indicators in Climate Change Adaptation (SWICCA) and End-to-End Demonstrator for Improved Decision-Making in the Water Sector in Europe (EDgE; Copernicus 2017a,b). Recent U.K. developments in SHFs stem from a prolonged period of drought beginning in 2010, which changed rapidly to widespread flooding during the winter of 2013/14. Driven by the consecutive formation of 12 major Atlantic depressions, the period between December 2013 and February 2014 (DJF 2014) was the wettest in the United Kingdom since records began in 1910 (Huntingford et al. 2014; Kendon and McCarthy 2015; Muchan et al. 2015) and the stormiest for at least 143 years when measured by cyclone frequency and intensity (Matthews et al. 2014). Individual storm events did not yield exceptional rainfall, but accumulated levels over the period led to extensive flooding nationwide, the costs of which were estimated at £1.3 billion (Environment Agency 2015a). The Thames River basin (southeast United Kingdom) received more than half a year's typical rainfall during DJF 2014 (Lewis et al. 2015), which led to concurrent fluvial, pluvial (surface water/flash), and groundwater flooding—so-called compound or coincident flood events (Thorne 2014).

To date, seasonal hydrological forecasting studies in the Thames and other lowland catchments have primarily identified initial hydrologic conditions as a dominant source of predictability (see Svensson et al. 2015; Svensson 2016). This is because flow regimes are dominated by slowly released groundwater, and forecast skill can be derived largely from the hydrogeological memory of antecedent conditions. Research conducted elsewhere, however, has found that driving hydrological models with more skillful meteorological inputs can capture observed flood events (Yossef et al. 2013) and improve

hydrological prediction skill for streamflow (Shukla and Lettenmaier 2011; Svensson et al. 2015) and groundwater levels (Almanaseer et al. 2014). The contribution of meteorological forcing to SHF skill has also been found to outweigh that provided from initial hydrologic conditions during times of transition from dry to wet climate conditions (Wood et al. 2016). In the United Kingdom, there is demand to develop and improve the characterization and skill of meteorological inputs (Lewis et al. 2015) to improve hydrological forecasting skill at longer lead times and during winter months (e.g., see Li et al. 2009; Shukla and Lettenmaier 2011; Thober et al. 2014). This is recognized as being particularly important for predicting groundwater levels during extreme events as currently, inadequacies in seasonal rainfall forecasts have resulted in low groundwater forecast skill in all but the most quickly responding U.K. catchments (Mackay et al. 2015). Considering better predictions of meteorological conditions, alongside studies focusing on the role of initial hydrologic conditions, will help ascertain which skill improvements have the greatest potential to benefit hydrological forecasts across the Thames basin.

There has been much discussion regarding the meteorological factors that led to the DJF 2014 floods in southern England. Huntingford et al. (2014) proposed various driving mechanisms for the precipitation anomalies, with the North Atlantic Oscillation (NAO) providing the strongest relationship. A positive NAO, characterized by an atmospheric pressure difference between the Azores and Iceland, is associated with increased delivery of rain-bearing cyclonic weather systems into northern Europe during the winter months (Wilby 2001; Svensson et al. 2015). The importance of the NAO, however, was disputed by van Oldenburg et al. (2015), who stated that a pressure pattern bearing a low to the west of Scotland (as opposed to Iceland) accounted for substantially more of the variance in precipitation during this event and that combinations of major large-scale modes of variability are likely to have caused the stormy conditions (see also Knight et al. 2017).

The exceptional conditions of DJF 2014 have thus also been attributed to a hemispheric pattern of severe weather. Relative to the December 1981–February 2010 ERA-Interim climatology, sea surface temperatures (SSTs) in the tropical Pacific were warmer than usual, which disturbed wind patterns over the northeast Pacific and deflected the Atlantic jet stream northward. This brought cold air to North America while eastern Europe was anomalously warm (Palmer 2014; Watson et al. 2016); this temperature gradient strengthened the jet stream and provided conditions for the continued formation of depressions that affected the United Kingdom (Slingo et al. 2014). These anomalous conditions, however, were not skillfully forecast by the European Centre for

Medium-Range Weather Forecasts (ECMWF) Operational System 4 (S4) seasonal meteorological forecasting system (Molteni et al. 2011). As a result, the ECMWF conducted a set of hindcast atmospheric relaxation experiments (AREs) to better understand the role of tropical sea surface temperatures in forcing the extratropical circulation response. The AREs relaxed the atmosphere toward the ERA-Interim reanalysis state within specified domains highlighted by negative Rossby wave source anomalies (see Rodwell et al. 2015; Magnusson 2017), forcing S4 to more accurately represent the cyclonic weather conditions prevailing in winter 2013/14. The results provided convincing evidence that the temperature and precipitation anomalies in Europe and North America were embedded within a hemispheric regime that was partly forced by tropical and underlying sea surface temperatures via Rossby wave source forcing (associated with its convection and divergent outflow) and that increased precipitation may have acted to reinforce the upstream wave.

This paper will use the ECMWF's AREs from Rodwell et al. (2015) to relate seasonal hydrological forecasting skill to the forecasting skill of meteorological input and its traceability from different atmospheric domains. Seasonal hydrological reforecasts for DJF 2014 were conducted using the European Flood Awareness System (EFAS) with seasonal meteorological input generated from the unforced S4 and three AREs. Specifically, we seek to identify 1) which seasonal meteorological reforecasts perform best, 2) whether increased skill in seasonal meteorological input translates through to more accurate streamflow and groundwater reforecasts for the 2013/14 compound flood event in the Thames River basin, and 3) how hydrological response differs for catchments with different geological and land-use characteristics. We discuss the potential for improvements to seasonal meteorological and hydrological forecasts and the practical value of more skillful seasonal flood forecasts for stakeholders to assist with decision-making in the Thames River basin.

2. Methods

a. Study catchments

The Thames River basin (containing 18 tributary catchments) covers approximately 16 200 km² in the southern United Kingdom. The western side is predominantly rural, comprising agriculture and woodland with rolling hills and wide, flat floodplains. Toward the center and east, the basin becomes increasingly urbanized, encompassing the towns of Reading, Slough, and Greater London. The source of the River Thames is located in the west (elevation up to 350 m MSL) and flows 230 km to Teddington Lock, which is the

official upper tidal limit (elevation 4 m MSL; Fig. 1). The basin encompasses a diverse range of lithologies that greatly influence the flow regime of the Thames and its tributaries, from seasonally spring-fed streams to chalk aquifers with high baseflow and clay-based rivers that are characterized by a flashy response to storm events and high levels of surface runoff (Bloomfield et al. 2009). Anthropogenic channel modifications, abstraction from major aquifers, and discharge points into the river also influence the flow regime; abstraction specifically represents a 5%–12% reduction in typical annual peak flow (Thames Water 2010). Recent estimates identified more than 200 000 properties at risk of flooding from a “100 year” event across the basin (Environment Agency 2009).

For the purposes of forecasting fluvial and groundwater floods, this study focuses on three catchments with contrasting geological and physical characteristics upstream of Teddington Lock that experienced compound flood events during DJF 2014. The Evenlode is a relatively small (429 km²) rural agricultural headwater catchment dominated by a limestone aquifer. The Loddon (682 km²) comprises a rural–urban gradient and variable geology. The area referred to in Fig. 1 as Lower Thames (324 km²) is the farthest point downstream before Teddington Lock and is small and heavily urbanized, with a densely populated floodplain largely overlaying impervious London clay deposits (Fig. 1, Table 1).

b. ECMWF atmospheric relaxation experiments

The rationale behind the AREs was to investigate teleconnection patterns from specified forcing regions. The concept nudges the forecast toward the “true state” in a predefined area during the forecast integration, allowing the downstream impacts from the region to be investigated. The nudging involves adding an extra term to the prognostic equations of the model. Further details about the relaxation technique can be found in section 2.2 in Magnusson (2017). The source regions in this study were selected based on their strong and persistent seasonal-mean forcing on the Rossby waveguide (Rodwell et al. 2015). As these forcing patterns in the source regions potentially had a long predictability, it was expected that the AREs should show impact on the predictability in other parts of the world.

This paper used three ARE model runs (AR_NPAC, AR_WATL, and AR_EATL), each representing a different source of atmospheric relaxation (see Figs. 2d–f and Figs. A1a–c in appendix A). The source regions were chosen where a strong average forcing on the northern midlatitude flow during DJF 2014 was identified, using a Rossby wave source as the diagnostic [see Rodwell et al. (2015) for details]. The AR_NPAC region (centered at 35°N, 150°W) can be physically explained as

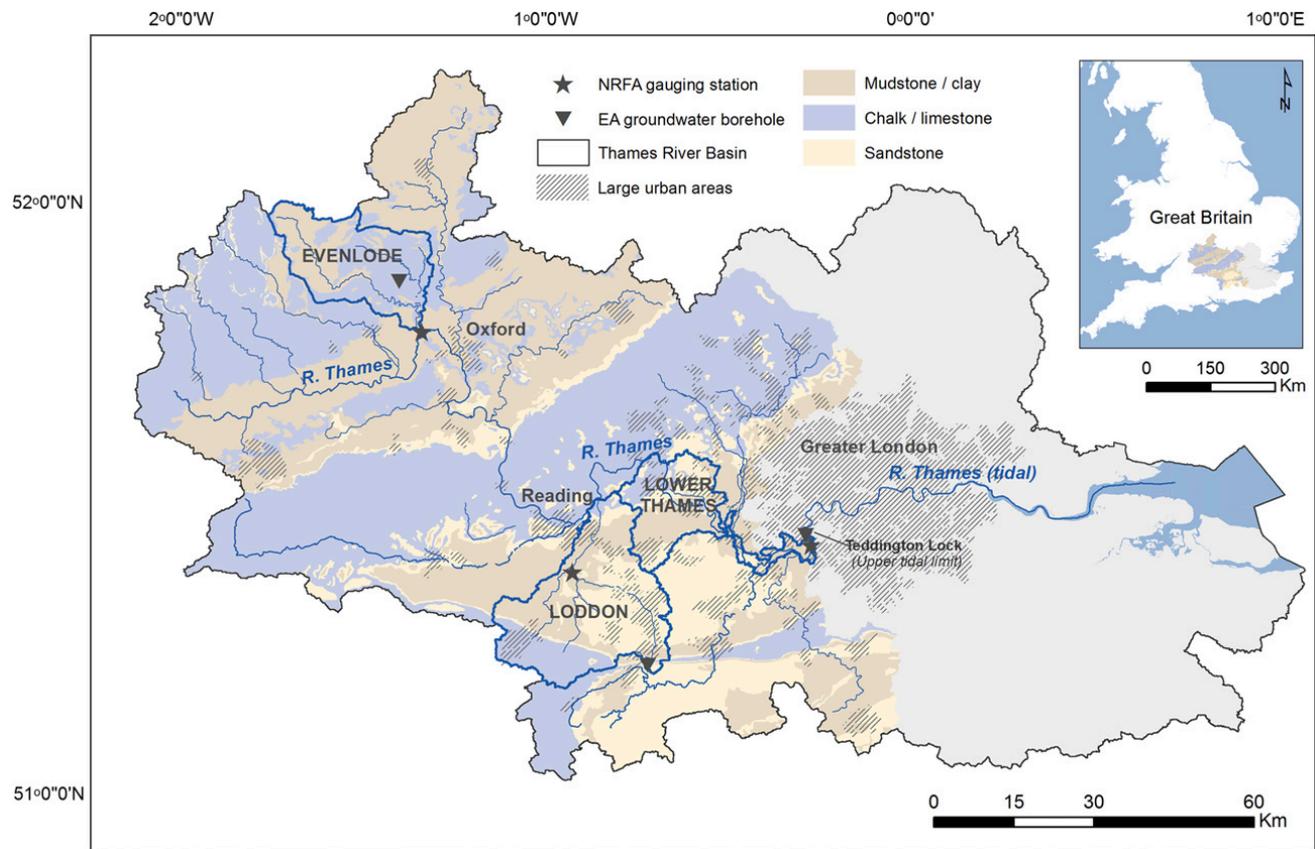


FIG. 1. Thames River basin, showing the geographical context and location of the West Thames National River Flow Archive (NRFA 2017) river gauging stations and EA (EA 2017) groundwater boreholes for the Evenlode, Loddon, and Lower Thames catchments. The West Thames general lithology are from the British Geological Survey (after Bloomfield et al. 2011), catchments and rivers are from the Water Framework Directive (EA 2015b), and the base map and urban areas are from Ordnance Survey (2017).

the region where forcing from the northeast tropical Pacific acted on the midlatitude flow (Fig. 2d). AR_WATL (western Atlantic) was the source region for Atlantic cyclones (35°N, 75°W; Fig. 2e). AR_EATL was over the northeast Atlantic (55°N, 15°W) and was associated with the heavy precipitation experienced during DJF 2014, although not directly linked to any underlying SST anomaly

(Fig. 2f). In each of the regions, the atmosphere was relaxed toward the ERA-Interim reanalysis state to determine the impact of each region.

All seasonal meteorological ensemble forecasts (S4 has 51 ensemble members and AR_NPAC, AR_WATL, and AR_EATL have 28 members) were produced by the ECWMF Integrated Forecasting System (IFS) coupled

TABLE 1. Summary of catchment characteristics and flow regimes. Dominant land use obtained from Centre for Ecology and Hydrology LandCover Map 2007 (NERC 2011) and general lithology (British Geological Survey after Bloomfield et al. 2011). Time to peak calculated according to the Revitalised Flood Hydrograph (ReFH) model (Kjeldsen 2007). Gauged flow regimes under average (Q50) and high extreme (Q10) conditions; values from NRFA (2017).

Catchment	Dominant land use (% of catchment)	General lithology (% of catchment)	Time to peak (Tp)	Gauged flow regimes ($\text{m}^3 \text{s}^{-1}$)
Evenlode	Agriculture = 85% Urban = 2% Seminatural = 13%	Mudstone/clay = 41% Chalk/limestone = 58% Sandstone = 1%	Tp = 14.68 h	Q50 = 2.52, Q10 = 8.93
Loddon	Agriculture = 50% Urban = 20% Seminatural = 30%	Mudstone/clay = 36% Chalk/limestone = 22% Sandstone = 42%	Tp = 9.81 h	Q50 = 2.31, Q10 = 5.92
Lower Thames	Agriculture = 43% Urban = 35% Seminatural = 22%	Mudstone/clay = 53% Chalk/limestone = 16% Sandstone = 31%	Tp = 34.31 h	Q50 = 40.5, Q10 = 161.6

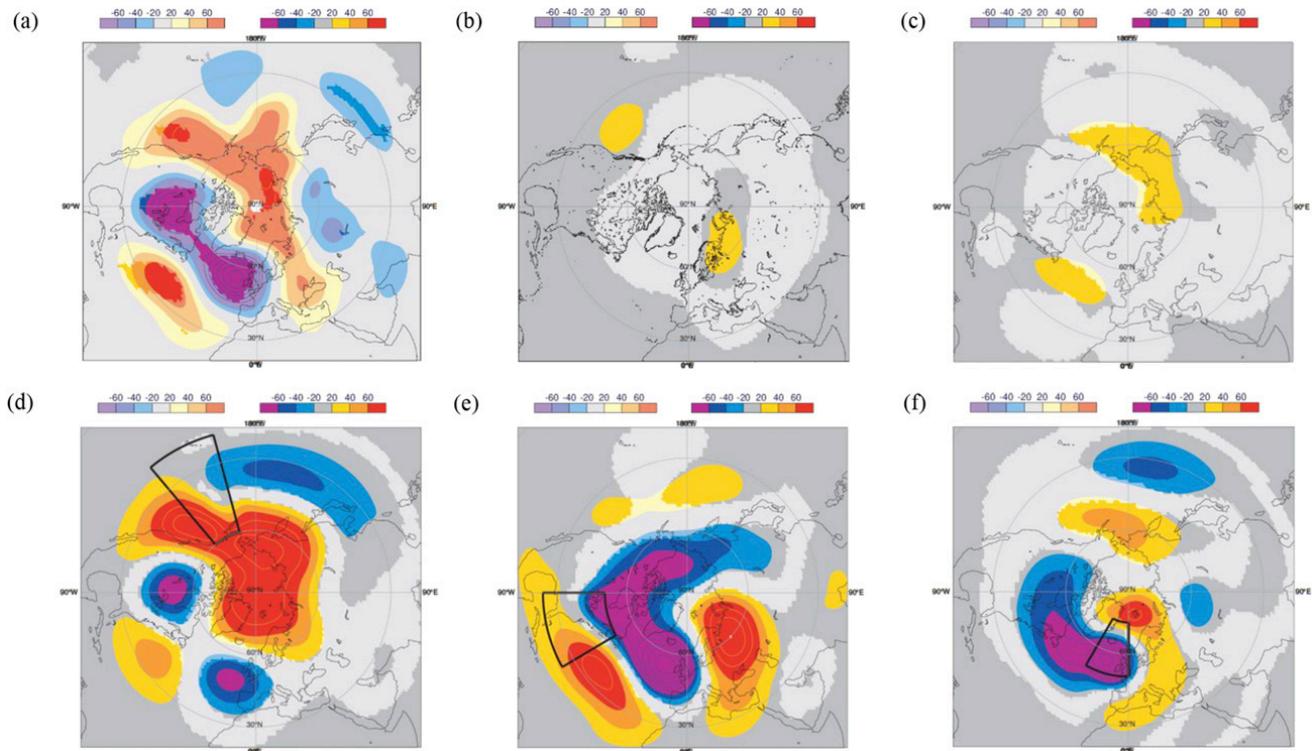


FIG. 2. (top) DJF 2014 anomaly fields of z_{500} : (a) ERA-Interim analysis, (b) S4 (CY36R4), and (c) NO_AR (control) equivalent to the operational S4 forecasts, but with the most recent model cycle (CY40R1). (bottom) Equivalent fields from AREs where the atmosphere was relaxed toward ERA-Interim reanalyses with Rossby wave source centers identified by black boxes: (d) AR_NPAC, centered at 35°N , 150°W ; (e) AR_WATL at 35°N , 75°W ; and (f) AR_EATL at 55°N , 15°W . Model climatology based on three ensemble members, initiated from 1 November for the 30 years of 1981–2010. Statistical significance at the 5% level is estimated from the 28-member distribution and indicated with saturated colors.

atmosphere–ocean–land model. The atmospheric model was run at T255 horizontal resolution ($\sim 80\text{ km}$) with 60 vertical levels (91 for S4), and the NEMO ocean model with 1° horizontal resolution in middle latitudes and higher resolution near the equator. Because of model updates, AREs used a more recent atmospheric model version (CY40R1) than that which is used operationally in S4 (CY36R4). ECMWF produced a 28-member ensemble of unforced control runs (NO_AR; Figs. 2c and A1d); neither cycle was able to predict the observed planetary wave anomaly in DJF 2014.

c. Seasonal hydrological modeling

Hydrological reforecasts were produced using the EFAS seasonal hydrological forecasting suite. EFAS aims to increase preparedness for floods in large European river basins based on operational probabilistic flood forecasts (Bartholmes et al. 2009; Thielen et al. 2009; Smith et al. 2016). The hydrological model used in EFAS is LISFLOOD, a hybrid between a conceptual and a physical rainfall–runoff model combined with a river routing module and run on a $5\text{ km} \times 5\text{ km}$ grid (van der Knijff et al. 2010; Alfieri et al. 2014). LISFLOOD is calibrated using nonnaturalized data. A new seasonal outlook for EFAS was recently developed by the

ECMWF that uses seasonal meteorological ensemble forecasts from S4 as input to LISFLOOD to extend the EFAS flood forecast horizon up to 7 months (Arnal et al. 2018). A reference daily simulation, termed the EFAS water balance (EFAS-WB), which starts from the initial conditions of the previous day and is forced with the most recent observed meteorological fields (interpolated point measurements of precipitation and temperature), is also run. EFAS-WB is used as initial conditions from which the seasonal forecasts are started and provides a best estimate of the hydrological states at a given time for a given grid point, that is, represents the theoretical upper limit of the model performance.

This study used the seasonal meteorological forecasts from ECMWF’s S4 and the three ARE model runs as input to LISFLOOD. All hydrological forecasts were initiated on 1 November 2013 and ran for 4 months with a daily time step to provide ensemble reforecasts for streamflow (routed river flow measured in $\text{m}^3\text{ s}^{-1}$) and groundwater level (storage in upper groundwater zone measured in mm). Catchment-averaged daily cumulative precipitation reforecasts (mm day^{-1}) were also produced.

Raw daily observation data for streamflow ($\text{m}^3\text{ s}^{-1}$) and groundwater level [m above ordnance datum (AOD)] were obtained from National River Flow Archive

(NRFA) gauging stations and Environment Agency (EA) groundwater boreholes. Within each study catchment, one gauging station on the main river and one groundwater borehole were chosen (Fig. 1). All observation points provided a complete daily record over the 4-month reforecasting period, plus data extending back 20 years (or as far back since the start of records) to identify probability exceedance thresholds for that location. Streamflow and groundwater level reforecasts were obtained for the 5-km EFAS grid tile within which the NRFA gauging station and EA groundwater borehole were located to ensure spatial consistency when comparing between forecasts and observations (Fig. 1). Areal precipitation reforecasts were calculated using the arithmetic mean for each catchment.

Continuous ranked probability scores (CRPSs; Hersbach 2000) were used as a measure of streamflow forecast sharpness and accuracy comparing against the simulated water balance (EFAS-WB) and river gauge observation data. To ensure consistency when comparing against different-sized ensembles, the relative percentage difference between the 28- and 51-member CRPS values for S4 were calculated; values ranged from 0% (no difference in the Evenlode) to 0.83% (in the Lower Thames; Fig. B1 in appendix B).

Spearman's rank correlation coefficient p was used to compare the median forecasted groundwater level against the simulated EFAS-WB and against borehole groundwater observations. Spearman's rank is a nonparametric measure of temporal rank correlation, which accounted for groundwater levels being expressed in different units.

Finally the EFAS-WB was compared against gauged daily streamflow observations and borehole groundwater observations as an evaluation of the LISFLOOD performance capability to accurately forecast the events in each catchment—this was achieved using Pearson's correlation coefficient r (to test EFAS-WB streamflow performance) and Spearman's rank p (groundwater performance). A workflow of all the forecasts, models, methods, and analyses used in the paper is shown in Fig. 3.

3. Results

a. Meteorological forcing

Severe weather conditions did not originate from a single event, but from a number of events between late December 2013 and the end of February 2014, as supported by the negative seasonal average anomaly of the 500-hPa geopotential height (z500) over the northeastern Atlantic, with the United Kingdom located at the southeastern edge (Fig. 2a). For a seasonal forecasting system, capturing this structure was key to predicting the wet anomaly over the United Kingdom, but no anomaly was present in the

ensemble mean averaged over the whole season for the S4 forecast (Fig. 2b).

Figures 2d–f show the results from the three AREs. By applying the atmospheric relaxation over the northeastern Pacific (AR_NPAC), the z500 anomalies over the western hemisphere were improved with a negative node over Canada and a positive node over the western Atlantic. There was also a negative anomaly present over the northeastern Atlantic, with a similar position to the analysis but weaker in magnitude (Fig. 2d).

Relative to AR_NPAC, the seasonal anomaly over the eastern Atlantic was better captured both in position and magnitude, with relaxation applied over the eastern part of the United States and the western Atlantic (AR_WATL; Fig. 2e). In the final experiment with the relaxation applied over the eastern Atlantic (AR_EATL), the negative anomaly was inside the relaxation box. However, the magnitude was less than in the analysis and AR_WATL, and the southern extent (outside the box) was not captured (Fig. 2f). The time series of accumulated precipitation shows AR_EATL underpredicted through late December but captured the rainfall better in January and February.

b. Hydrological response to meteorological forcing

1) OVERVIEW

Patterns in simulated EFAS-WB cumulative areal precipitation values (pink line) were consistent across all catchments (Figs. 4–6). There were a few wet days in early November and a dry period into mid-December followed by higher-than-average rainfall conditions, with extreme precipitation events corresponding with Atlantic depressions recorded in mid- to late December, early January, late January, and early February. Over the 4 months, total cumulative areal precipitation (EFAS-WB, pink line) was greatest in the Loddon catchment at 541.2 mm and lower in the Evenlode and Lower Thames at 494.1 and 454.1 mm, respectively (Figs. 4–6).

During early to mid-December, observed gauged daily streamflow (black line) fell below the median (Q50) daily flow record (Table 1) in all three catchments (based on daily flow records from 1994 to 2014; NRFA 2017). Observed streamflow in all catchments then exceeded the Q10 exceedance threshold (percentage of time that streamflow exceeds the 90th percentile) from mid-December through to the end of the study period. Observed borehole daily groundwater levels (black line) exceeded Q50 (EA 2017) in the Loddon toward the end of January but did not reach the Q10 level of 65.9 mm (not shown in Fig. 5). Observed groundwater levels exceeded Q10 in the Evenlode by mid-December and Lower Thames by mid-January (Figs. 4, 6).

Comparing observations against EFAS-WB (model performance; Fig. 7), LISFLOOD was capable of predicting

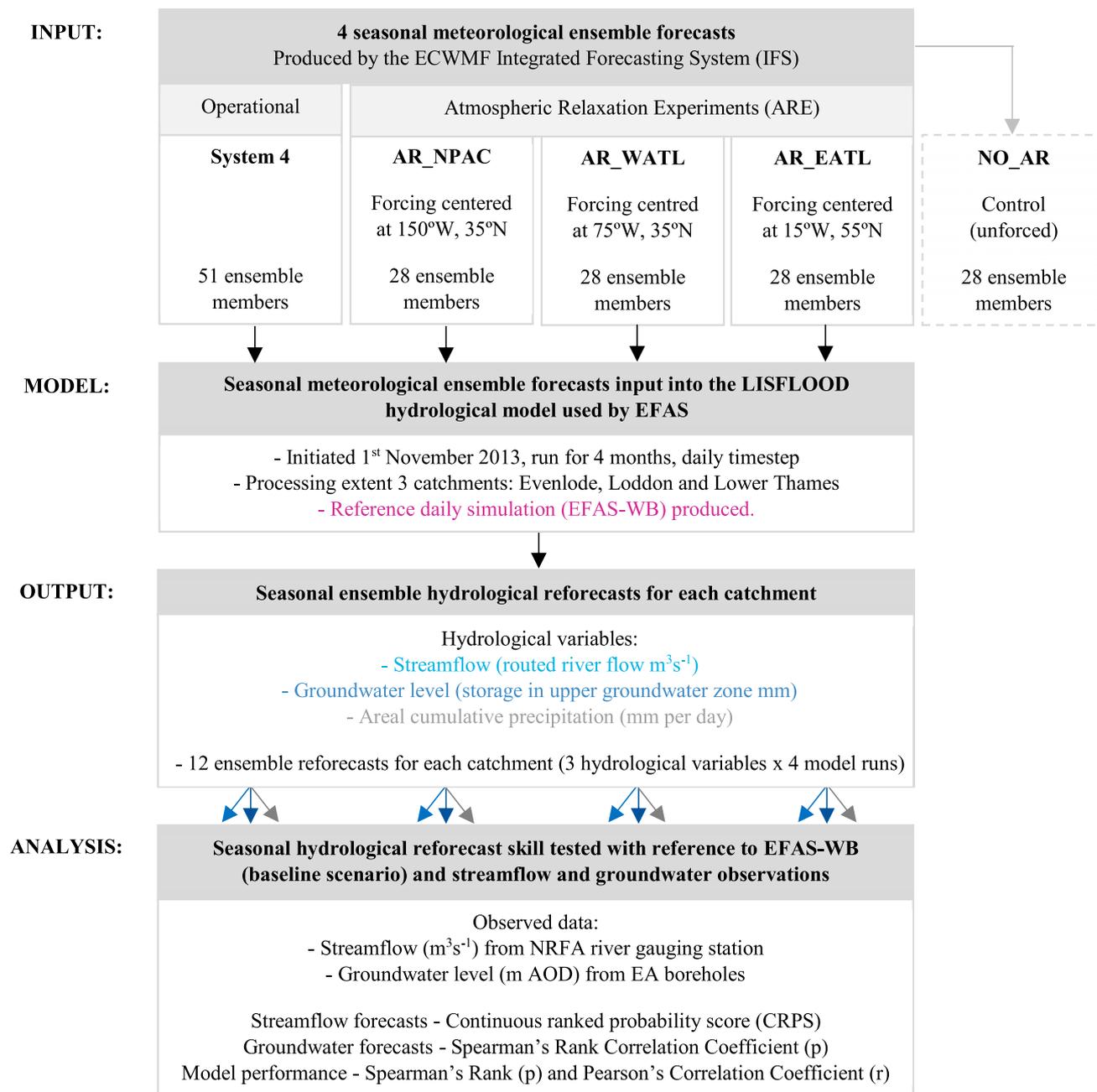


FIG. 3. Workflow detailing seasonal meteorological forecast inputs, hydrological modeling setup, seasonal hydrological reforecast outputs, observational data, and analyses used in the paper.

streamflow and groundwater levels with reasonably high accuracy in all three catchments; correlation coefficients ranged from a positive moderately strong 0.7 for Lower Thames groundwater to a near perfect 0.98 positive correlation for Lower Thames streamflow (Fig. 7).

2) COMPARISON OF SEASONAL HYDROLOGICAL FORECASTS

Visual improvements to areal precipitation forecasts and streamflow forecasts identified by CRPS followed the general pattern: (worst) S4 > AR_NPAC > AR_EATL >

AR_WATL (best) (Figs. 4–6, 8a). This trend was similar for groundwater correlations; all three ARE model runs demonstrated marked improvement compared with S4 forecasts that showed negative correlation with simulated EFAS-WB and borehole groundwater observations in each catchment (Fig. 8b).

S4 forecasted a linear increase in rainfall from 1 November that failed to pick up the low rainfall conditions from the end of November to early December or the extreme precipitation events in mid-December and beyond. S4 also substantially underpredicted the total

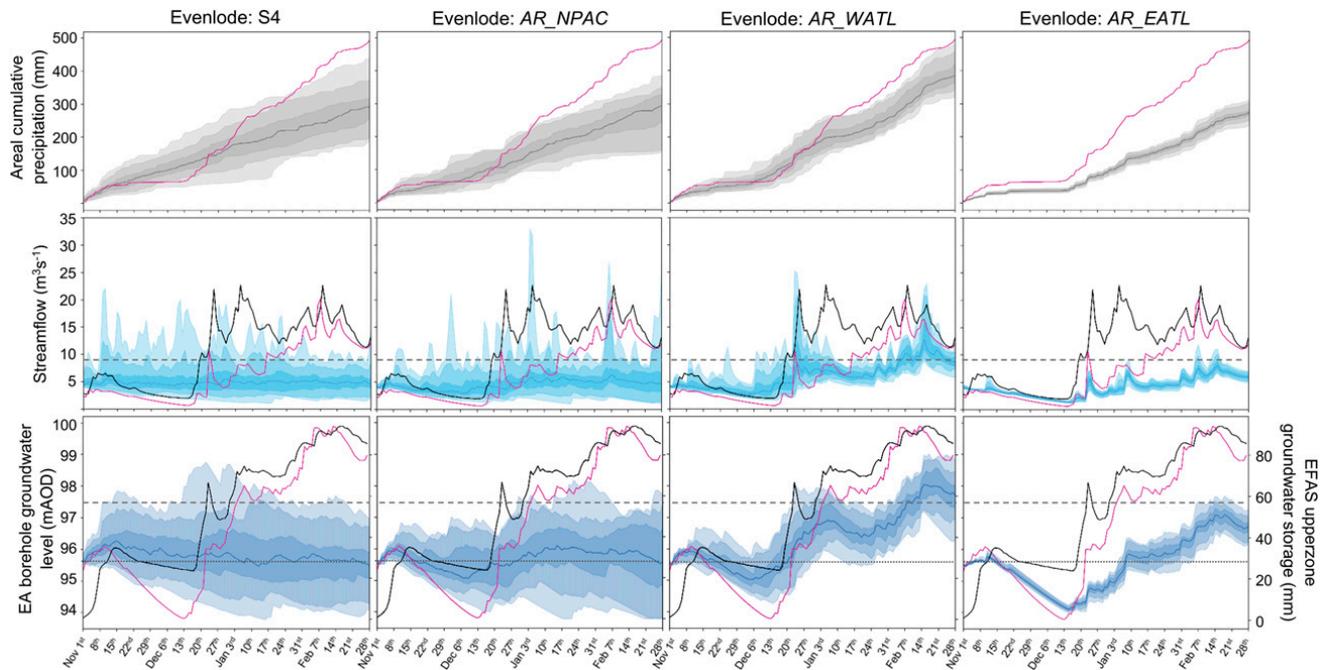


FIG. 4. Precipitation, streamflow, and groundwater levels: a comparison between S4 and the three ARE model runs AR_NPAC, AR_WATL, and AR_EATL for the Evenlode catchment. Forecast shading shows the minimum, 5th, 25th, 75th, 95th, and maximum of the ensemble in all cases. Areal precipitation (mm) = catchment-averaged cumulative daily forecast median values (gray). Streamflow ($\text{m}^3 \text{s}^{-1}$) = daily forecast median (light blue) at the river gauging station. Groundwater level (mm) = daily forecast median (dark blue) at the groundwater borehole. Observations (black) and simulated EFAS-WB (pink) in all cases. Q10 (long dash) and Q50 (short dash) show exceedance thresholds (based on 1994–2014 observation records or the longest available record).

amount of precipitation forecast over the 4-month period. The resulting streamflow forecasts showed minimal forecast skill across all catchments; the median did predict above-average streamflow conditions (up to 90th percentile at times) and low numbers of ensemble members

forecast some extremes, but the timing and magnitude of peak events were largely incorrect, notably during the first 6 weeks (Figs. 4–6). S4 forecasted decreasing groundwater levels over the 4 months, leading to negative correlations with borehole observations; this was

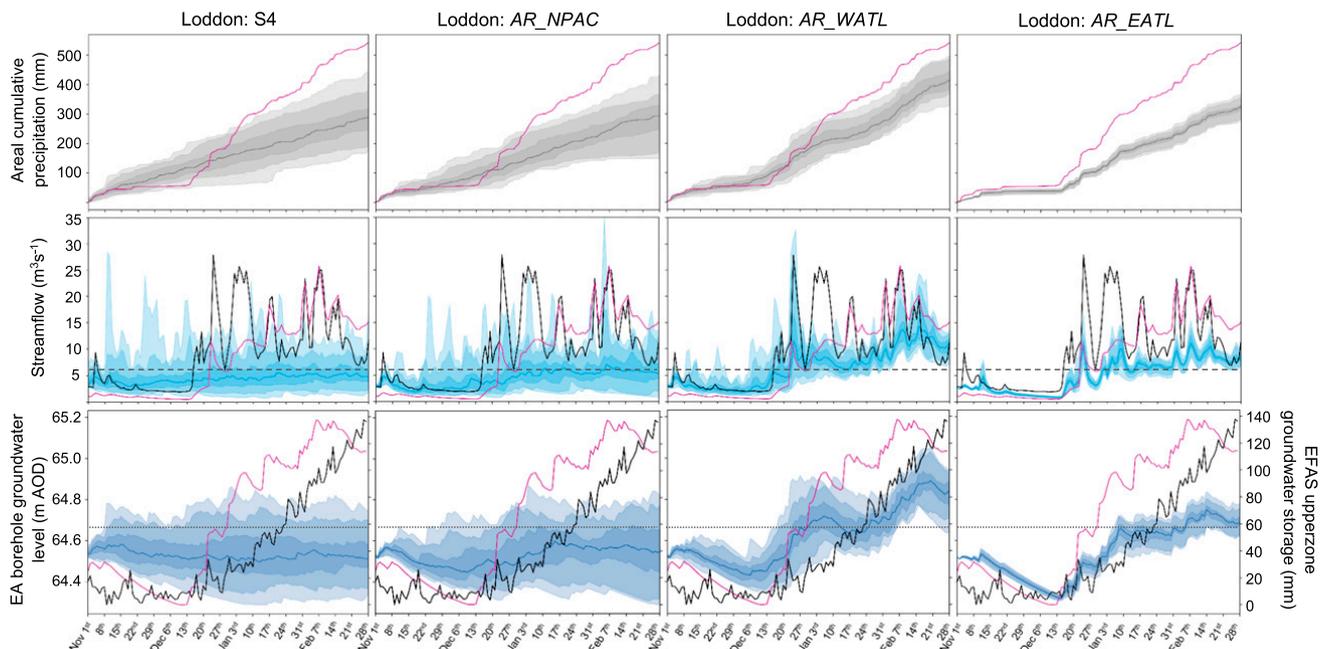


FIG. 5. As in Fig. 4, but for the Loddon catchment.

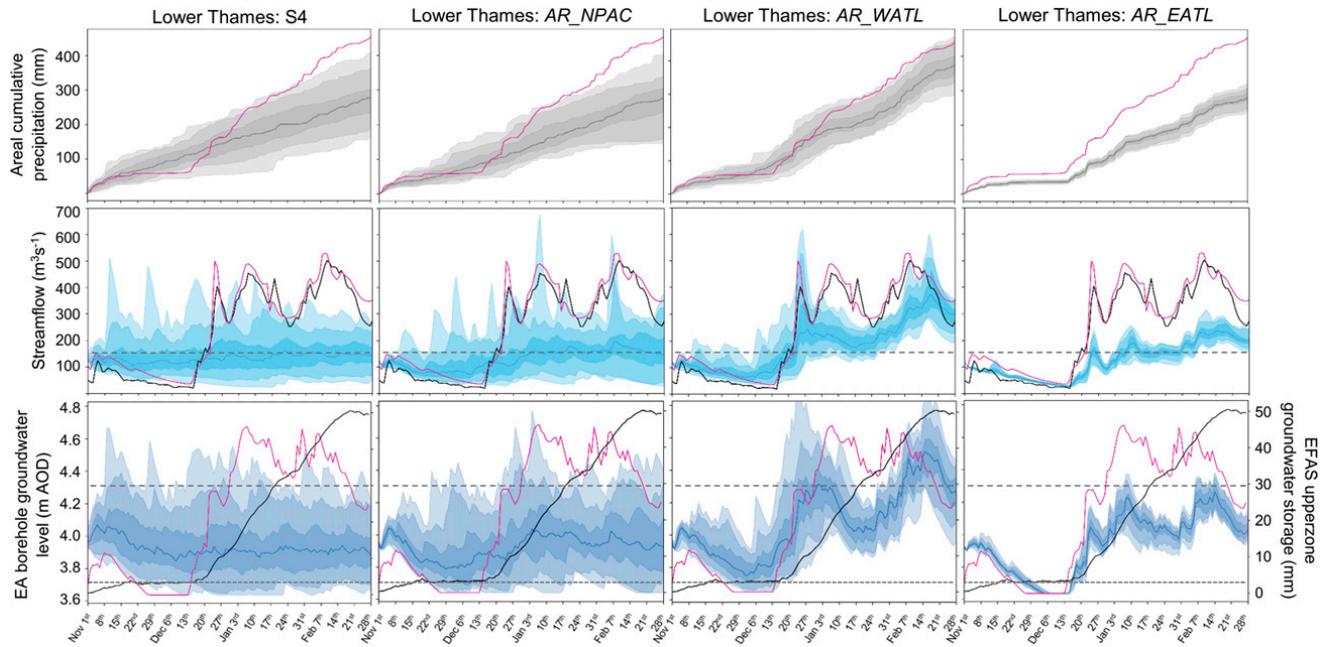


FIG. 6. As in Fig. 4, but for the Lower Thames catchment.

most pronounced in the Evenlode, where observations recorded a 6.10-m increase in groundwater levels in the aquifer (Figs. 4, 8b).

The AR_NPAC precipitation forecast was similar to S4, although sharper with less spread about the median leading to minor improvements in streamflow and groundwater forecasts. The timing of peak streamflow events was more accurately represented, and the magnitude was picked up by the ensemble maximum in many cases. There remained poor forecast quality during the first 6 weeks. Groundwater forecast median showed weak to moderate positive correlation with borehole observations and EFAS-WB (Fig. 8b), although there was a large ensemble spread (Figs. 4–6).

Areal precipitation forecast by the AR_EATL model run was sharp with a good correlation but underprediction in respect to the simulated EFAS-WB values in all catchments. Subsequent streamflow forecasts demonstrated accuracy and sharpness but underprediction and reduced reliability for high extremes. Groundwater forecasts were sharper than S4 and AR_NPAC but also underpredicted against the EFAS-WB (Figs. 4–6).

AR_WATL produced the best areal precipitation forecast in all catchments; the forecast median traced the simulated EFAS-WB cumulative rainfall patterns with relatively high accuracy until mid- to late December when accuracy trailed off. Precipitation forecasts remained sharper than S4 and AR_NPAC, and total rainfall was matched by the forecast maximum in the Evenlode and Lower Thames (Figs. 4–6). Low CRPS and strong positive correlation values indicate a marked improvement for all streamflow and

groundwater forecasts (Figs. 8a,b). Extreme streamflow events were missed from late December to early January in all catchments that correlated with the decreased accuracy in the rainfall forecast. Groundwater forecasts showed regular oscillations in all three catchments (also apparent in AR_EATL and AR_NPAC forecasts; Figs. 4–6).

3) CATCHMENT VARIATION

Observed gauged streamflow patterns (black line), although of different orders of magnitude, were similar for the Evenlode and Lower Thames with consistently high flows from mid-December onward with 5–6 clearly defined peaks (Figs. 4, 6). LISFLOOD successfully modeled the

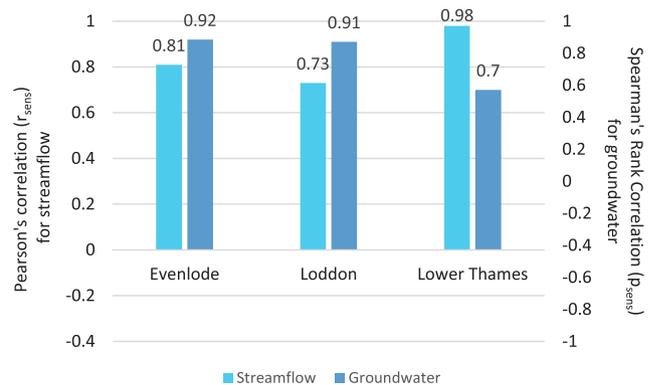


FIG. 7. Comparison between daily simulated EFAS-WB and gauged daily streamflow ($\text{m}^3 \text{s}^{-1}$) and borehole groundwater level (m AOD) observations as a measure of the capability (sensitivity) of the LISFLOOD model performance. Streamflow performance (light blue) tested using Pearson's correlation r_{sens} . Groundwater performance (dark blue) tested using Spearman's rank correlation coefficient ρ_{sens} .

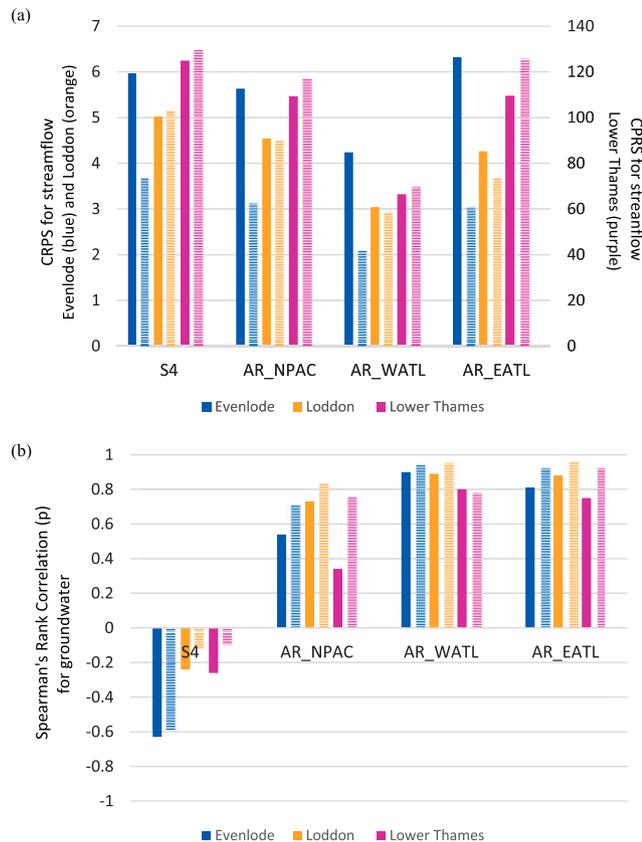


FIG. 8. Comparison of seasonal hydrological forecast skill from the operational S4 and three ARE model runs: AR_NPAC, AR_WATL, and AR_EATL for the Evenlode (blue), Loddon (orange), and Lower Thames (purple) catchments. (a) CRPS comparing daily streamflow ($\text{m}^3 \text{s}^{-1}$) forecast against gauged daily streamflow observations ($\text{m}^3 \text{s}^{-1}$) (CRPS_{obs} shown by solid bars) and against daily simulated EFAS-WB (CRPS_{sim} shown by hashed bars). (b) Spearman's rank correlation coefficient p comparing median daily groundwater level (mm) forecast against borehole daily groundwater observations (m AOD) (p_{obs} , solid bars) and against daily simulated EFAS-WB (p_{sim} , hashed bars).

flow dynamics in the Lower Thames ($r_{\text{sens}} = 0.98$; Fig. 7). The flow pattern was quite accurate in the Evenlode, but overall model performance was lower ($r_{\text{sens}} = 0.81$) as the simulated EFAS-WB did not capture flow pattern between mid-December and the end of January (Figs. 4, 6, and 7). The Loddon had a much flashier response with eight clearly defined peaks (black line), coupled with the shortest time to peak of 9.81 h (Table 1). Model performance was the lowest of the three catchments ($r_{\text{sens}} = 0.73$) as LISFLOOD failed to detect peaks around 17 December and underpredicted the extreme events from late December to mid-January (Figs. 5, 7).

Observed borehole groundwater levels (black line) increased in all three catchments: Evenlode +6.10 m, Loddon +0.90 m, and Lower Thames +1.13 m (Figs. 4–6). The Loddon and Lower Thames recorded average or just below average (Q50) groundwater levels until mid-December, when levels showed a consistent and steady

rise. Groundwater levels recorded in the Evenlode were more responsive following precipitation events and mirrored streamflow dynamics (Fig. 4). LISFLOOD was best able to model groundwater levels in the Evenlode ($p_{\text{sens}} = 0.92$) but was oversensitive in the Lower Thames ($p_{\text{sens}} = 0.70$; Fig. 7).

In respect to streamflow and groundwater level forecasting skill compared against observations in each catchment, CRPS and Spearman's rank p indicated that AR_WATL provided the best forecast skill in all catchments (solid bars in Figs. 8a,b). CRPS_{obs} for the Loddon and Lower Thames followed the aforementioned pattern $S4 > \text{AR_NPAC} > \text{AR_EATL} > \text{AR_WATL}$; however, the AR_EATL model run performed worst in the Evenlode ($\text{CRPS}_{\text{obs}} = 6.32$). Groundwater level forecast skill was consistent across catchments, with S4 performing worst and AR_WATL best (Fig. 8b).

4. Discussion

The winter of 2013/14 was exceptional in regard to the large number of Atlantic depressions that affected the United Kingdom—the Thames basin saw record precipitation levels that led to widespread and prolonged fluvial and groundwater flooding (Kendon and McCarthy 2015), the impacts of which have been well documented (e.g., Slingo et al. 2014; Thorne 2014; Muchan et al. 2015). The drivers of these extreme conditions have also been debated, with papers seeking to identify the atmospheric influences via reviews of multiscale model simulations investigating factors such as atmosphere, ocean, land use and demographics (see Huntingford et al. 2014), correlation analyses (van Oldenburg et al. 2015) and relaxation experiments (Rodwell et al. 2015; Watson et al. 2016; Knight et al. 2017). It is largely accepted that a combination of global meteorological influences were important, but studies that link different meteorological inputs and how these translate through to hydrological forecasting skill have not been conducted. Below, we discuss how identification of skill through the meteorological (ARE) and hydrological (EFAS) seasonal forecasting chain may provide an indication as to the origins of extreme events and the level of predictability that can be gained if the evolution in parts of the system are known. We also highlight the value of more skillful hydrological forecasts during extreme events for stakeholders, taking into account the variation in catchment properties that exist across the Thames basin.

a. Translating meteorological improvements into more skillful hydrological forecasts

From a meteorological perspective, AR_EATL was expected to give the rainfall closest to that observed during DJF 2014 due to the location of southern England at the

edge of the relaxation box. Although this experiment provided the most confident hydrometeorological ensemble forecasts, their value was limited because of underprediction, likely because AR_EATL missed the southward extent of the atmospheric trough and hence did not fully capture the details of the flow anomaly affecting southern England.

Atmospherically speaking, AR_NPAC captured the representation of large-scale flow over the northern Atlantic better than S4, yet this did not translate into an improved precipitation forecast, resulting in low hydrological forecasting skill over the Thames basin. As AR_NPAC gave a stronger anomaly in geopotential height over the eastern Atlantic, one could speculate that systematic model errors affected the Rossby wave train from the Pacific to the Atlantic, leading to misplacement of the anomaly over the northeastern Atlantic. Given the relationship between the tropical Pacific and El Niño–Southern Oscillation (ENSO; [Doblas-Reyes et al. 2013](#)), there was hope that seasonal hydrological predictability could be improved in the future with a better modeled teleconnection from ENSO. Rather, the results point to the importance of the western Atlantic and pose the open question about whether the forcing into this box is linked with the Pacific and/or tropical Atlantic.

The best hydrological forecasts were obtained by the AR_WATL experiment. Climatologically, the eastern United States and Gulf Stream is the most active region for cyclogenesis in the Atlantic ([Hoskins and Hodges 2002](#)), and the representation of the anomaly in this region also captured the downstream anomalies over the northern Atlantic. Whether this is a result of the cold anomaly over North America giving a strong temperature contrast (baroclinicity) over the Gulf Stream or related to the anomaly in the divergent flow from South America as discussed in [Knight et al. \(2017\)](#) has yet to be confirmed. Nonetheless, with future improvements to coupled models, there is scope for an improvement in the teleconnections whereby the results for this study could be revised ([Magnusson et al. 2013](#)).

b. More skillful hydrological forecasts, but missed events, oscillations, and uncertainty

All three AREs led to improvements in meteorological input, which translated through to more skillful streamflow and groundwater level reforecasts compared to S4, with AR_WATL performing best. However, there were consistent trends observed for all ARE model runs across the three catchments. Poor representation of the hydrological variables during the first 6 weeks coincided with the end of the drought period that preceded the extreme wet conditions. [Wood et al. \(2016\)](#) found that this climatological transition period produced the lowest seasonal predictability as initial hydrologic conditions provide minimal contribution, an effect that may have been heightened in the Thames basin, which is largely groundwater driven ([Svensson et al. 2015](#)).

Streamflow forecasts also missed peak streamflow events observed between the end of December and mid-January; timewise, these correlated with the point at which precipitation forecasts diverged from the simulated EFAS-WB, indicating potential meteorological forcing errors. This was likely due to the extreme nature of the rainfall experienced at this time, which was undetected in the meteorological forecast, propagating the error into the hydrological forecast ([Davolio et al. 2008](#)) coupled with the uncertainty prevalent at longer lead times ([Wood and Lettenmaier 2008](#)). Structural issues in LISFLOOD, however, cannot be ruled out, as the EFAS-WB also failed to capture these peak streamflow events in the Evenlode and Loddon catchments. Factors including the variable density of the rain gauge network, lack of horizontal flow (from pixel to pixel) of water in the topsoil and subsoil, and inability to represent finescale geological and morphological characteristics in smaller subbasins, for example, may have limited forecast skill and model performance in these catchments. Nonetheless, recovery of the EFAS-WB toward observations later in the seasonal streamflow forecast (mid-January onward) suggests that these missed events may relate more strongly to meteorological forcing errors.

Groundwater-level forecasts showed oscillations of increasing amplitude as precipitation forecasts improved (most obvious for AR_WATL), with troughs corresponding with the November dry period and missed rainfall events, and peaks shortly following periods of intense precipitation. A rapid response to rainfall has been observed for aquifer recharge rates and groundwater level time series ([Lee et al. 2006](#); [Bloomfield and Marchant 2013](#)), indicating that there can be sensitivity of groundwater forecasts to meteorological forcing data. Here, we investigated the LISFLOOD upper-groundwater-zone response, where processes represent a mix of fast groundwater, including preferential flow rates and subsurface flow through soil macropores ([Thielen et al. 2009](#)), and thus a quicker response to rainfall following a dry period was expected ([Mahmood-ul-Hassan and Gregory 2002](#); [Lee et al. 2006](#)). The cyclical dynamic of the forecast may also represent model processes whereby outflow from the upper zone is released once the amount of water being stored reaches a threshold ([van der Knijff et al. 2010](#)). As such, it is likely that the observed oscillations represent combined effects of the LISFLOOD model setup and sensitivity to rainfall input.

c. Catchment controls on the variation in hydrological skill improvements

There were differences in the observed hydrological response and model performance between catchments, likely explained by the EFAS setup, plus local weather conditions and geographical differences acting at the catchment level. Simulated EFAS-WB streamflow values were low compared

to observations in the largely groundwater-driven Evenlode and flashy-responding Loddon (discussed previously), but were well captured in the Lower Thames, where peak streamflow observations exceeded $\sim 500 \text{ m}^3 \text{ s}^{-1}$. This increased performance may be attributed to the fact that the Thames basin in LISFLOOD is calibrated using gauged daily flow records from the Lower Thames (at Kingston/Teddington Lock). The geographical position of the Lower Thames also represents drainage from the entire upstream catchment, essentially representing a larger basin for which LISFLOOD was designed. The greater coverage of impervious surfaces where LISFLOOD assumes no soil or groundwater storage may also have played a role (Burek et al. 2013). By contrast, groundwater levels were most accurately modeled and forecast in the Evenlode. Despite its small size and position at the headwater, this suggests that LISFLOOD is well set up to capture upper-zone processes in rural land-use catchments dominated by chalk and limestone lithology (see also Mansour et al. 2013). Antecedent dry conditions are also likely to have played an important role, allowing percolation into the aquifer, as explained by the 6.10-m increase in observed groundwater levels (Svensson et al. 2015). By contrast, the relatively small yet linear increase in observed groundwater levels in the Loddon and Lower Thames could be attributed to the locations of boreholes within less permeable lithologies, and in the case of the Lower Thames, a heavily urbanized area (Macdonald et al. 2012). Bloomfield and Marchant (2013) recognized clear effects of fractured chalk and granular sandstone aquifer characteristics on saturated flow and storage during U.K. drought conditions, and it is not unreasonable to expect differences to carry forward into a period of extreme rainfall. The cumulative effects of upstream groundwater abstractions not accounted for in LISFLOOD may also explain the notable difference between simulated EFAS-WB and borehole observations in the Lower Thames—an effect less prevalent in the Loddon and Evenlode, where boreholes are located toward the top of the catchments. Interestingly, the observed groundwater storage in the Loddon and Lower Thames was more consistent with that of the lower (saturated) zone in LISFLOOD (not considered in this study because of data issues), where water is either stored or enters the channel via baseflow, producing a very slow, seasonally linear response to meteorological forcings (van der Knijff et al. 2010; Mackay et al. 2015). Whether the oversensitivity of the simulated upper-zone response in these catchments (notably the Lower Thames, where the EFAS-WB captured groundwater variability at an entirely different frequency) is a result of finer-scale geological and land-use heterogeneity not captured by LISFLOOD (Svensson et al. 2015) or the saturated nature of the impervious deposits that may be better represented by lower-zone processes requires further work.

d. Stakeholder implications and future developments

1) IMPROVE CLIMATE FORCINGS TO DELIVER MORE SKILLFUL HYDROLOGICAL FORECASTS

There is currently a lot of focus on improving operational flood forecasts at seasonal time scales, and extreme events such as those experienced in DJF 2014 raise important questions about whether there are elements of predictability that are being missed by seasonal forecasting systems (Scaife et al. 2014; van Oldenburg et al. 2015; Watson et al. 2016; Knight et al. 2017). Driving hydrological models with inputs from atmospheric relaxation experiments provides a valid indication of what can be achieved from an operational forecasting system if the determinants of prolonged seasonal mean forcing, for example, ENSO, could be captured in the future. While S4 was unable to skillfully capture the seasonal average forcing for DJF 2014, updates such as the fifth-generation ECMWF seasonal forecasting system (SEAS5), which will shortly replace S4, indicate substantial improvements to SST bias in the tropical Pacific, increased model resolution, and a greater ensemble size (Lucas 2017) that may go some way to improving seasonal hydrological predictions.

2) DIFFERENT HYDROLOGICAL MODEL SETUP TO EXTRACT SKILL

While the uncertainty in the forecasts appears to be largest, further analysis might consider adjusting the LISFLOOD model parameters through a process of model calibration (Shi et al. 2008) and/or comparing results with those obtained from a local-scale hydrological model that better captures streamflow and groundwater dynamics in smaller basins. Use of multiple different hydrological models could also help capture a fuller representation of the uncertainty that comes from the hydrology and land surface (e.g., see EDgE; Copernicus 2017b).

3) DJF 2014 FLOOD SIGNALS DETECTED WEEKS IN ADVANCE

Based on numerical weather prediction, the fluvial flood events of DJF 2014 were well forecast at a lead time of 2–3 days and with reasonable accuracy up to 2 weeks ahead of time (Lewis et al. 2015). Groundwater floods, which acted over a longer time scale and were triggered by exceptional aquifer recharge and saturation of permeable deposits, were not well predicted due to the complex dynamics and interactions of the groundwater system with atmosphere and land processes (Mackay et al. 2015). The EA is responsible for managing flood risk in the United Kingdom. Taking the Loddon December floods as an example, a flood alert based on the EA streamflow thresholds at the river gauging station would have been

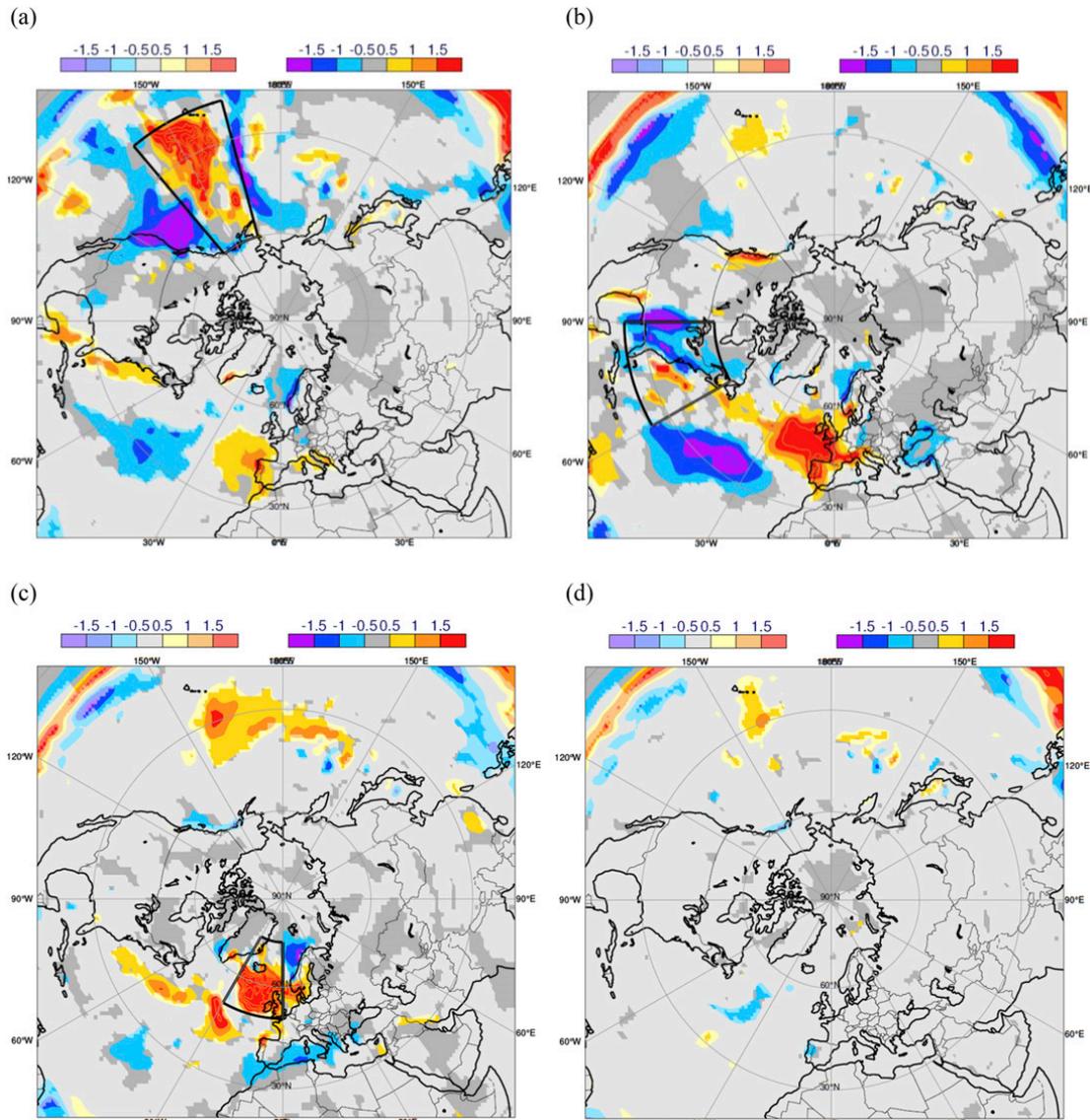
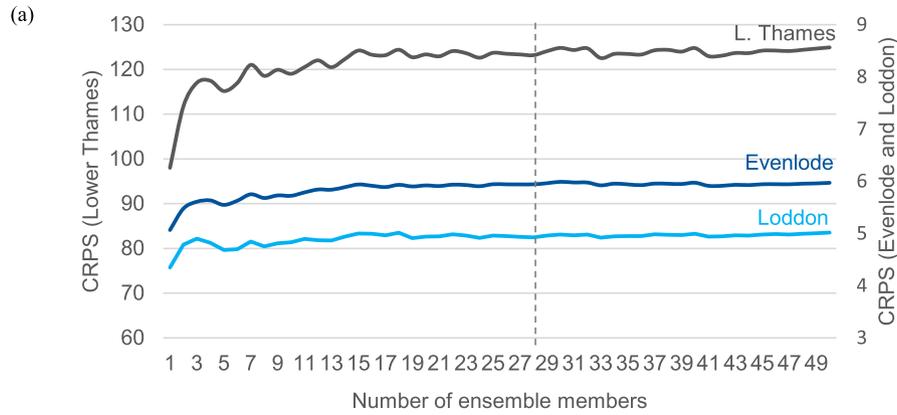


FIG. A1. DJF 2014 anomaly fields of TPO relative to model climatology for 28-member hindcasts made with the coupled model (CY40R1). The equivalent field from AREs where the atmosphere was relaxed toward ERA-Interim reanalyses with Rossby wave source centers identified by black boxes: (a) AR_NPAC, centered at 35°N, 150°W; (b) AR_WATL at 35°N, 75°W; and (c) AR_EATL at 55°N, 15°W. (d) NO_AR (control) equivalent to the operational S4 forecasts (CY36R4), but with the most recent model cycle. Model climatology based on three ensemble members, initiated from 1 November for the 30 years of 1981–2010. Statistical significance at the 5% level is estimated from the 28-member distribution and indicated here with saturated colors.

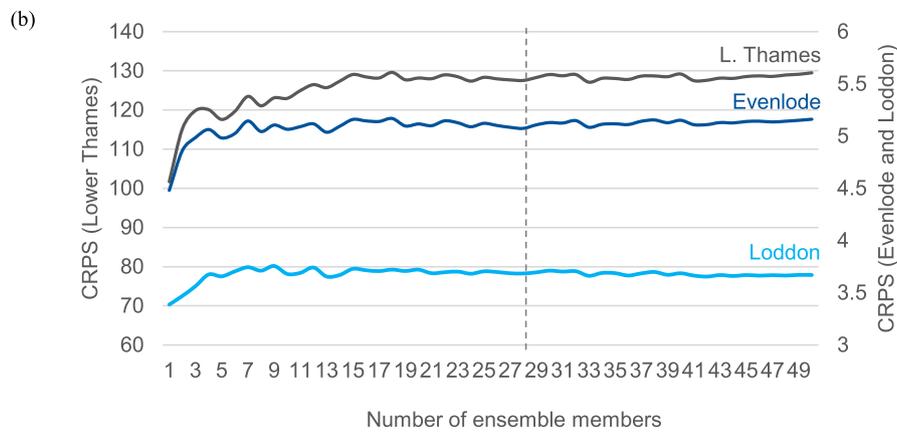
triggered from a value of $11 \text{ m}^3 \text{ s}^{-1}$: in the case of S4 and AR_NPAC, the forecast median did not cross this threshold, although maximum extremes of the ensemble did. For AR_WATL and AR_EATL, a flood alert for the local area would have been observed with 6 weeks lead time (1 November) based on the forecast median. This would have allowed mitigation strategies and low-cost preventative actions to be carried out well in advance while also highlighting an “area to watch” as the season progressed.

The importance of SHF for advance warning should not be underestimated in densely populated areas such

as the Thames basin. Increasing pressures for urban development, intensification of agriculture, and clean water demand a more spatially and temporally integrated approach to management of the water sector (Mansour et al. 2013; Lewis et al. 2015). There is also growing evidence to support an increasing likelihood of Atlantic storms that take a more southerly track akin to DJF 2014 (Slingo et al. 2014), and while the contribution of climate change cannot be definitively related to changes in the U.K. hydrological response (Hannaford 2015), even a small shift in mean climate variability could substantially



CRPS_{obs}	Evenlode	Loddon	Lower Thames
28 mem average	5.066	4.985	123.187
51 mem average	5.066	4.995	124.225
Relative % difference	0.00 %	-0.19 %	-0.83 %



CRPS_{sim}	Evenlode	Loddon	Lower Thames
28 mem average	3.756	4.986	123.346
51 mem average	3.756	4.999	124.271
Relative % difference	0.00 %	-0.25 %	-0.74 %

FIG. B1. CRPS for S4 run using 2–51 ensemble members: (a) forecast against streamflow observations (CRPS_{obs}) and (b) forecast against simulated EFAS-WB (CRPS_{sim}). Tables outline relative percentage differences in CRPS achieved with 28- and 51-member ensembles.

shorten the return periods of such events (Knight et al. 2017). Further studies that trace the meteorological input improvements right through the meteorological–hydrological forecasting chain are therefore strongly advocated.

5. Conclusions

Atmospheric relaxation experiments can improve our understanding of extratropical anomalies and the potential predictability of extreme events such as DJF 2014. Our results highlight that there is meteorological

knowledge to be gained by considering the hydrology, that is, although large-scale seasonal flow anomalies were picked up in the meteorology, these did not always translate through to more skillful hydrological forecasts. Extreme events such as DJF 2014 are difficult to predict with confidence at seasonal time scales, but considering the local hydrogeological context for streamflow and groundwater levels can provide an effective early alert of potentially high impact events, allowing for better preparedness and greater confidence in forecasts as an event approaches.

Acknowledgments. This work was supported and funded by the EU Horizon 2020 IMPREX project (<http://www.imprex.eu/>) (641811). Borehole groundwater level data and flood alert exceedance thresholds were made available by the Environment Agency, and we thank Simon Lewis and Stuart Hyslop at the Environment Agency for useful discussions on seasonal forecasting in the Thames basin. The authors have no conflicts of interest to declare.

APPENDIX A

Atmospheric Relaxation Experiments

AREs for DJF 2014 total precipitation (TPO) anomaly fields (Fig. A1).

APPENDIX B

Continuous Ranked Probability Scores

CRPSs for S4 run using 2–51 ensemble members (Fig. B1).

REFERENCES

- Alfieri, L., F. Pappenberger, F. Wetterhall, T. Haiden, D. Richardson, and P. Salamon, 2014: Evaluation of ensemble streamflow predictions in Europe. *J. Hydrol.*, **517**, 913–922, <https://doi.org/10.1016/j.jhydrol.2014.06.035>.
- Almanaseer, N., A. Sankarasubramanian, and J. Bales, 2014: Improving groundwater predictions utilizing seasonal precipitation forecasts from general circulation models forced with sea surface temperature forecasts. *J. Hydrol. Eng.*, **19**, 87–98, [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0000776](https://doi.org/10.1061/(ASCE)HE.1943-5584.0000776).
- Arnal, L., A. W. Wood, E. Stephens, H. Cloke, and F. Pappenberger, 2017: An efficient approach for estimating streamflow forecast skill elasticity. *J. Hydrometeorol.*, **18**, 1715–1729, <https://doi.org/10.1175/JHM-D-16-0259.1>.
- , H. L. Cloke, E. Stephens, F. Wetterhall, C. Prudhomme, J. Neumann, B. Krzeminski, and F. Pappenberger, 2018: Skillful seasonal forecasts of streamflow over Europe? *Hydrol. Earth Syst. Sci.*, **22**, 2057–2072, <https://doi.org/10.5194/hess-22-2057-2018>.
- Bartholmes, J. C., J. Thielen, M.-H. Ramos, and S. Gentilini, 2009: The European Flood Alert System EFAS—Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrol. Earth Syst. Sci.*, **13**, 141–153, <https://doi.org/10.5194/hess-13-141-2009>.
- Bloomfield, J. P. and B. P. Marchant, 2013: Analysis of groundwater drought building on the standardised precipitation index approach. *Hydrol. Earth Syst. Sci.*, **17**, 4769–4787, <https://doi.org/10.5194/hess-17-4769-2013>.
- , D. J. Allen, and K. J. Griffiths, 2009: Examining geological controls on baseflow index (BFI) using regression analysis: An illustration from the Thames Basin, UK. *J. Hydrol.*, **373**, 164–176, <https://doi.org/10.1016/j.jhydrol.2009.04.025>.
- , S. H. Bricker, and A. J. Newell, 2011: Some relationships between lithology, basin form and hydrology: A case study from the Thames basin, UK. *Hydrol. Processes*, **25**, 2518–2530, <https://doi.org/10.1002/hyp.8024>.
- Burek, P. A., J. van der Knijff, and A. De Roo, 2013: LISFLOOD distributed water balance and flood simulation model: Revised user manual. JRC Tech. Rep. 78917, 139 pp., <https://doi.org/10.2788/24719>.
- Copernicus, 2017a: SWICCA: Service for Water Indicators in Climate Change Adaptation. SMHI, accessed 10 August 2017, <http://swicca.climate.copernicus.eu/>.
- , 2017b: EDgE. Climate Change Service, accessed 10 August 2017, <http://edge.climate.copernicus.eu/>.
- Davolio, S., M. M. Miglietta, T. Diomede, C. Marsigli, A. Morgillo, and A. Moscatello, 2008: A meteo-hydrological prediction system based on a multi-model approach for precipitation forecasting. *Nat. Hazards Earth Syst. Sci.*, **8**, 143–159, <https://doi.org/10.5194/nhess-8-143-2008>.
- Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. L. Rodrigues, 2013: Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdiscip. Rev.: Climate Change*, **4**, 245–268, <https://doi.org/10.1002/wcc.217>.
- Environment Agency, 2009: Thames Catchment Flood Management Plan—Managing flood risk. Summary Rep., 28 pp., <https://www.gov.uk/government/publications/thames-catchment-flood-management-plan>.
- , 2015a: The costs and impacts of the winter 2013 to 2014 floods. Project Summary SC140025. 2 pp., <https://www.gov.uk/government/publications/the-costs-and-impacts-of-the-winter-2013-to-2014-floods>.
- , 2015b: WFD River Basin Districts Cycle 2. Accessed 10 August 2016, <https://data.gov.uk/dataset/wfd-river-basin-districts-cycle-2>.
- , 2017: Groundwater level measurements (AFA075).
- Hannaford, J., 2015: Climate-driven changes in UK river flows: A review of the evidence. *Prog. Phys. Geogr.*, **39**, 29–48, <https://doi.org/10.1177/0309133314536755>.
- Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2).
- Hoskins, B. J., and K. I. Hodges, 2002: New perspectives on the Northern Hemisphere winter storm tracks. *J. Atmos. Sci.*, **59**, 1041–1061, [https://doi.org/10.1175/1520-0469\(2002\)059<1041:NPOTNH>2.0.CO;2](https://doi.org/10.1175/1520-0469(2002)059<1041:NPOTNH>2.0.CO;2).
- Huntingford, C., and Coauthors, 2014: Potential influences on the United Kingdom's floods of winter 2013/14. *Nat. Climate Change*, **4**, 769–777, <https://doi.org/10.1038/nclimate2314>.
- Kendon, M., and M. McCarthy, 2015: The UK's wet and stormy winter of 2013/2014. *Weather*, **70**, 40–47, <https://doi.org/10.1002/wea.2465>.

- Kjeldsen, T. R., 2007: Flood estimation handbook. Supplementary Rep. 1, Centre for Ecology and Hydrology, 68 pp., <https://www.ceh.ac.uk/sites/default/files/FEH%20Supplementary%20Report%20hi-res.pdf>.
- Knight, J. R., and Coauthors, 2017: Global meteorological influences on the record UK rainfall of winter 2013–14. *Environ. Res. Lett.*, **12**, 074001, <https://doi.org/10.1088/1748-9326/aa693c>.
- Lee, L. J. E., D. S. L. Lawrence, and M. Price, 2006: Analysis of water-level response to rainfall and implications for recharge pathways in the Chalk aquifer, SE England. *J. Hydrol.*, **330**, 604–620, <https://doi.org/10.1016/j.jhydrol.2006.04.025>.
- Lewis, H., and Coauthors, 2015: From months to minutes—Exploring the value of high-resolution rainfall observation and prediction during the UK winter storms of 2013/2014. *Met. Apps.*, **22**, 90–104, <https://doi.org/10.1002/met.1493>.
- Li, H., L. Luo, E. F. Wood, and J. Schaake, 2009: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. *J. Geophys. Res.*, **114**, D04114, <https://doi.org/10.1029/2008JD010969>.
- Lucas, D., 2017: Implementation of seasonal forecast SEAS5. ECMWF, accessed 19 September 2017, <https://software.ecmwf.int/wiki/display/FCST/Implementation+of+Seasonal+Forecast+SEAS5#ImplementationofSeasonalForecastSEAS5-References>.
- Macdonald, D., A. Dixon, A. Newell, and A. Hallways, 2012: Groundwater flooding within an urbanised flood plain. *J. Flood Risk Manage.*, **5**, 68–80, <https://doi.org/10.1111/j.1753-318X.2011.01127.x>.
- Mackay, J. D., C. R. Jackson, A. Brookshaw, A. A. Scaife, J. Cook, and R. S. Ward, 2015: Seasonal forecasting of groundwater levels in principal aquifers of the United Kingdom. *J. Hydrol.*, **530**, 815–828, <https://doi.org/10.1016/j.jhydrol.2015.10.018>.
- MacLachlan, C., and Coauthors, 2015: Global Seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system. *Quart. J. Roy. Meteor. Soc.*, **141**, 1072–1084, <https://doi.org/10.1002/qj.2396>.
- Magnusson, L., 2017: Diagnostic methods for understanding the origin of forecast errors. *Quart. J. Roy. Meteor. Soc.*, **143**, 2129–2142, <https://doi.org/10.1002/qj.3072>.
- , M. Alonso-Balmaseda, S. Corti, F. Molteni, and T. Stockdale, 2013: Evaluation of forecast strategies for seasonal and decadal forecasts in presence of systematic model errors. *Climate Dyn.*, **41**, 2393–2409, <https://doi.org/10.1007/s00382-012-1599-2>.
- Mahmood-ul-Hassan, M., and P. J. Gregory, 2002: Dynamics of water movement on Chalkland. *J. Hydrol.*, **257**, 27–41, [https://doi.org/10.1016/S0022-1694\(01\)00530-3](https://doi.org/10.1016/S0022-1694(01)00530-3).
- Mansour, M., J. Mackay, C. Abesser, A. Williams, L. Wang, S. Bricker, and C. Jackson, 2013: Integrated Environmental Modeling applied at the basin scale: Linking different types of models using the OpenMI standard to improve simulation of groundwater processes in the Thames Basin, UK. *MODFLOW and More: Translating Science into Practice*, Golden, CO, Colorado School of Mines, 5 pp., <http://nora.nerc.ac.uk/501789/>.
- Matthews, T., C. Murphy, R. Wilby, and S. Harrigan, 2014: Stormiest winter on record for Ireland and UK. *Nat. Climate Change*, **4**, 738–740, <https://doi.org/10.1038/nclimate2336>.
- Molteni, F., and Coauthors, 2011: The new ECMWF seasonal forecast system (System 4). ECMWF Tech. Memo. 656, 49 pp., <https://www.ecmwf.int/sites/default/files/elibrary/2011/11209-new-ecmwf-seasonal-forecast-system-system-4.pdf>.
- Muchan, K., M. Lewis, J. Hannaford, and S. Parry, 2015: The winter storms of 2013/2014 in the UK: Hydrological responses and impacts. *Weather*, **70**, 55–61, <https://doi.org/10.1002/wea.2469>.
- National River Flow Archive, 2017: Search for gauging stations. Accessed 10 July 2017, <http://nrfa.ceh.ac.uk/data/search>.
- NERC, 2011: Land Cover Map 2007 dataset documentation, version 1.0. Centre for Ecology and Hydrology, 19 pp., <https://www.ceh.ac.uk/sites/default/files/LCM2007%20dataset%20documentation.pdf>.
- Ordnance Survey, 2017: OS open data. Subsets used: Strategi and Miniscale, accessed 16 August 2016, <https://www.ordnancesurvey.co.uk/opendatadownload/products.html>.
- Palmer, T., 2014: Record-breaking winters and global climate change. *Science*, **344**, 803–804, <https://doi.org/10.1126/science.1255147>.
- Rodwell, M. J., L. Ferranti, L. Magnusson, A. Weisheimer, F. Rabier, and D. Richardson, 2015: Diagnosis of northern hemispheric regime behaviour during winter 2013/14. ECMWF Tech. Memo. 769, 12 pp., <https://www.ecmwf.int/sites/default/files/elibrary/2015/15305-diagnosis-northern-hemispheric-regime-behaviour-during-winter-201314.pdf>.
- Scaife, A. A., and Coauthors, 2014: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2514–2519, <https://doi.org/10.1002/2014GL059637>.
- Shi, X., A. J. Wood, and D. P. Lettenmaier, 2008: How essential is hydrologic model calibration to seasonal streamflow forecasting? *J. Hydrometeor.*, **9**, 1350–1363, <https://doi.org/10.1175/2008JHM1001.1>.
- Shukla, S., and D. Lettenmaier, 2011: Seasonal hydrologic prediction in the United States: Understanding the role of initial hydrologic conditions and seasonal climate forecast skill. *Hydrol. Earth Syst. Sci.*, **15**, 3529–3538, <https://doi.org/10.5194/hess-15-3529-2011>.
- Slingo, J., and Coauthors, 2014: The recent storms and floods in the UK. Met. Office Rep., 27 pp., <http://nora.nerc.ac.uk/id/eprint/505192/1/N505192CR.pdf>.
- Smith, P., and Coauthors, 2016: On the operational implementation of the European Flood Awareness System (EFAS). ECMWF Tech. Memo. 778, 34 pp., <https://www.ecmwf.int/sites/default/files/elibrary/2016/16337-operational-implementation-european-flood-awareness-system-efas.pdf>.
- Svensson, C., 2016: Seasonal river flow forecasts for the United Kingdom using persistence and historical analogues. *Hydrol. Sci. J.*, **61**, 19–35, <https://doi.org/10.1080/02626667.2014.992788>.
- , and Coauthors, 2015: Long-range forecasts of UK winter hydrology. *Environ. Res. Lett.*, **10**, 064006, <https://doi.org/10.1088/1748-9326/10/6/064006>.
- Thames Water, 2010: Hydrological context for water quality and ecology preliminary impact assessments, technical appendix B. Thames Water Utilities Ltd 2W0H Lower Thames Operating Agreement, 66 pp.
- Thielen, J., J. Bartholmes, M.-H. Ramos, and A. P. J. De Roo, 2009: The European Flood Alert System—Part 1: Concept and development. *Hydrol. Earth Syst. Sci.*, **13**, 125–140, <https://doi.org/10.5194/hess-13-125-2009>.
- Thober, S., A. Wood, L. Samaniego, M. Clark, R. Kumar, and M. Zink, 2014: The elasticity of hydrological forecast skill with respect to initial conditions and meteorological forcing for two major flood events in Germany. *Geophysical Research Abstracts*, Vol. 16, Abstract EGU2014-3089-1, <http://meetingorganizer.copernicus.org/EGU2014/EGU2014-3089-1.pdf>.
- Thorne, C., 2014: Geographies of UK flooding in 2013/14. *Geogr. J.*, **180**, 297–309, <https://doi.org/10.1111/geoj.12122>.
- van der Knijff, J. M., J. Younis, and A. P. J. De Roo, 2010: LISFLOOD: A GIS-based distributed model for river basin

- scale water balance and flood simulation. *Int. J. Geogr. Inf. Sci.*, **24**, 189–212, <https://doi.org/10.1080/13658810802549154>.
- van Oldenburg, G. J., D. B. Stephenson, A. Sterl, R. Vautard, P. Yiou, S. S. Drijfhout, H. von Storch, and H. van den Dool, 2015: Drivers of the 2013/14 winter floods in the UK. *Nat. Climate Change*, **5**, 490–491, <https://doi.org/10.1038/nclimate2612>.
- Watson, P. A. G., A. Weisheimer, J. R. Knight, and T. N. Palmer, 2016: The role of the tropical West Pacific in the extreme Northern Hemisphere winter of 2013/2014. *J. Geophys. Res. Atmos.*, **121**, 1698–1714, <https://doi.org/10.1002/2015JD024048>.
- Wilby, R. L., 2001: Seasonal forecasting of river flows in the British Isles using North Atlantic pressure patterns. *Water Environ. J.*, **15**, 56–63, <https://doi.org/10.1111/j.1747-6593.2001.tb00305.x>.
- Wood, A. W., and D. P. Lettenmaier, 2008: An ensemble approach for attribution of hydrologic prediction uncertainty. *Geophys. Res. Lett.*, **35**, L14401, <https://doi.org/10.1029/2008GL034648>.
- , T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark, 2016: Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill. *J. Hydrometeor.*, **17**, 651–667, <https://doi.org/10.1175/JHM-D-14-0213.1>.
- Yossef, N. C., H. Winsemius, A. Weerts, R. van Beek, and M. F. P. Bierkens, 2013: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing. *Water Resour. Res.*, **49**, 4687–4699, <https://doi.org/10.1002/wrcr.20350>.

A6: An efficient approach for estimating streamflow forecast skill elasticity

This paper presents the published version of Chapter 4, Sect. 4.2 of this thesis, with the following reference:

Arnal, L., A. W. Wood, E. Stephens, H. L. Cloke and F. Pappenberger, 2017b: An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity, *J. Hydrometeorol.*, 18, 1715–1729, doi:10.1175/JHM-D-16-0259.1*

* ©2017. The Authors. Journal of Hydrometeorology, a journal of the American Meteorological Society. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.

An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity

LOUISE ARNAL

*Department of Geography and Environmental Science, University of Reading, and European
Centre for Medium-Range Weather Forecasts, Shinfield Park,
Reading, United Kingdom*

ANDREW W. WOOD

Research Applications Laboratory, NCAR, Boulder, Colorado

ELISABETH STEPHENS

*Department of Geography and Environmental Science, University of Reading, Reading,
United Kingdom*

HANNAH L. CLOKE

*Department of Geography and Environmental Science, and Department of Meteorology, University
of Reading, Reading, United Kingdom*

FLORIAN PAPPENBERGER

*European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, and School of
Geographical Sciences, University of Bristol, Bristol, United Kingdom*

(Manuscript received 1 November 2016, in final form 10 March 2017)

ABSTRACT

Seasonal streamflow prediction skill can derive from catchment initial hydrological conditions (IHCs) and from the future seasonal climate forecasts (SCFs) used to produce the hydrological forecasts. Although much effort has gone into producing state-of-the-art seasonal streamflow forecasts from improving IHCs and SCFs, these developments are expensive and time consuming and the forecasting skill is still limited in most parts of the world. Hence, sensitivity analyses are crucial to funnel the resources into useful modeling and forecasting developments. It is in this context that a sensitivity analysis technique, the variational ensemble streamflow prediction assessment (VESPA) approach, was recently introduced. VESPA can be used to quantify the expected improvements in seasonal streamflow forecast skill as a result of realistic improvements in its predictability sources (i.e., the IHCs and the SCFs)—termed “skill elasticity”—and to indicate where efforts should be targeted. The VESPA approach is, however, computationally expensive, relying on multiple hindcasts having varying levels of skill in IHCs and SCFs. This paper presents two approximations of the approach that are computationally inexpensive alternatives. These new methods were tested against the original VESPA results using 30 years of ensemble hindcasts for 18 catchments of the contiguous United States. The results suggest that one of the methods, end point blending, is an effective alternative for estimating the forecast skill elasticities yielded by the VESPA approach. The results also highlight the importance of the choice of verification score for a goal-oriented sensitivity analysis.

 Denotes content that is immediately available upon publication as open access.

 Supplemental information related to this paper is available at the Journals Online website: <http://dx.doi.org/10.1175/JHM-D-16-0259.s1>.

Corresponding author: Louise Arnal, l.l.s.arnal@pgr.reading.ac.uk; louise.arnal@ecmwf.int



This article is licensed under a [Creative Commons Attribution 4.0 license](http://creativecommons.org/licenses/by/4.0/) (<http://creativecommons.org/licenses/by/4.0/>).

DOI: 10.1175/JHM-D-16-0259.1

1. Introduction

Unprecedented increases in computer capabilities have shaped the last several decades' advances in numerical weather prediction (NWP), and with them, the development of environmental forecasting and modeling systems. This has led to a shift in the strategy of operational forecasting centers toward more integrated modeling and forecasting approaches, such as coupled systems and Earth system models (ESMs), with the final aim to extend the limits of predictability (i.e., from subseasonal to seasonal forecasting). These developments are supported by the assimilation of more and better-quality observation data as well as the increase in model resolutions and complexity. However, such advances can be very expensive and data hungry and may not yield proportional improvements.

Seasonal hydrological forecasts are predictions of the future states of the land surface hydrology (e.g., streamflow), up to a few months ahead. They are valuable for applications such as reservoir management for hydropower, agriculture and urban water supply, spring flood and drought prediction, and navigation, among others (Clark et al. 2001; Hamlet et al. 2002; Chiew et al. 2003; Wood and Lettenmaier 2006; Regonda et al. 2006; Luo and Wood 2007; Kwon et al. 2009; Cherry et al. 2005; Viel et al. 2016). They have the potential to provide early warning for increased preparedness (Yuan et al. 2015). Traditionally, seasonal streamflow forecasts have relied upon land surface memory, the persistence in the land surface (e.g., catchment) initial hydrological conditions (IHCs; of soil moisture, groundwater, snowpack, and the current streamflow). IHCs are one of the most important predictability sources of seasonal streamflow forecasts and were thus the starting point for the development of the ensemble streamflow prediction (ESP) approach in the 1970s (Wood et al. 2016b). The ESP was first developed and used for reservoir management purposes. It is produced by running a hydrological model with observed meteorological inputs to produce current observed IHCs, from which the forecast is started, and the forcing over the forecast period is undertaken using an ensemble of historical meteorological observations (Day 1985). The ESP method assumes that the model states to initialize a forecast are perfectly estimated, while the future climate is completely unknown. However, the skill of the ESP decreases significantly after one to a few months of lead time over most parts of the world because of a decrease in the land surface memory with time. The achievable predictability from the ESP thus depends on the persistence of the IHCs, which can

vary as a function of the season (i.e., the transition between dry and wet seasons can, for example, be hard to forecast) and the location and size of the catchment (i.e., the streamflow in a large catchment with a slow response time and/or situated in a region with negligible precipitation inputs during the forecast period will for example be easier to forecast; Wood and Lettenmaier 2008; Shukla et al. 2013; van Dijk et al. 2013; Yuan et al. 2015).

More recently, seasonal climate predictability derived from large-scale climate precursors [e.g., El Niño–Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO)] has been used to enhance seasonal streamflow forecasting (e.g., Wood et al. 2002; Yuan et al. 2013; Demargne et al. 2014; Mendoza et al. 2017). Such systems produce streamflow forecasts by initializing a hydrological model to estimate IHCs and forcing the model with inputs based on seasonal climate forecasts (SCFs; of temperature and precipitation) instead of historical observations. Their skill is also still limited because of the rapid decrease in precipitation forecasting skill beyond two weeks of lead time, and the skill is variable in both space and time (Yuan et al. 2011; van Dijk et al. 2013; Slater et al. 2017). In Europe, for instance, the skill is higher in winter in regions where the winter precipitation is highly correlated with the NAO. Regions with high skill include the Iberian Peninsula, Scandinavia, and regions around the Black Sea (Bierkens and van Beek 2009). In the contiguous United States (CONUS), the skill is on average higher over (semi)arid western catchments, due to the persistence of the IHCs influence up to three months of lead time. The skill can be higher in some regions of the western CONUS (i.e., California, the Pacific Northwest, and Great Basin) in the winter and fall due to higher precipitation forecasting skill in strong ENSO phases (Wood et al. 2005).

Increasing the seasonal streamflow forecast skill remains a challenge: one that is being tackled by improving IHCs and SCFs using a variety of techniques. Techniques include model developments and data assimilation and can vary in computational expense. However, over the past several decades, it has been shown that operational streamflow forecast quality has not significantly improved (Pagano et al. 2004; Welles et al. 2007). This is the motivation for the use of sensitivity analysis techniques to guide future forecasting developments for seasonal streamflow forecasting and is the basis for this paper.

It is in this context that the attribution of seasonal streamflow forecast uncertainty to the IHC and SCF errors has been researched extensively. Wood and Lettenmaier (2008) introduced a method based on two hindcasting end points: the ESP and the reverse ESP. In contrast to the ESP, which only represents the

uncertainty in the future climate, the reverse ESP only represents the uncertainty in IHCs by using an ensemble of initial model states taken from historical simulations to initialize a prediction forced by a single set of observed meteorological inputs. Typically, the input uncertainty attenuates over a period of months as the influence of the perfect future climate input increasingly determines model states.

Comparing the skill of the ESP versus reverse-ESP seasonal streamflow forecasts allows one to identify the dominant predictability source (and conversely uncertainty source) of seasonal streamflow forecasting (i.e., the IHCs or the SCFs), and its evolution in both space and time. It was successfully used to disentangle the relative importance of initial conditions and boundary forcing errors on seasonal streamflow forecast uncertainties by several authors: for example, for catchments in the United States (Wood and Lettenmaier 2008; Li et al. 2009; Shukla and Lettenmaier 2011), in France (Singla et al. 2012), in Switzerland (Staudinger and Seibert 2014), in China (Yuan et al. 2016; Yuan 2016), and in the Amazon (Paiva et al. 2012), as well as for the entire globe (Shukla et al. 2013; Yossef et al. 2013; MacLeod et al. 2016). This work is instructive as it enables the dominant predictability source to be identified (i.e., where efforts and resources should be targeted) to focus improvement, which could potentially lead to more skillful seasonal streamflow predictions.

This method was extended by Wood et al. (2016a, hereafter W16) via a method called variational ensemble streamflow prediction assessment (VESPA), which involves assessing intermediate IHC and SCF uncertainty points between the perfect and climatological points applied in ESP and reverse ESP. The approach allows the calculation of a metric called “skill elasticity,” that is, the sensitivity of streamflow forecast skill to IHC and SCF skill changes. A key drawback of the VESPA approach, however, is that it is computationally intensive. For each catchment and initialization month of a forecast, the response surface was defined through the use of dozens of multidecadal variable-skill ensemble hindcasts, ultimately amounting to millions of simulations. In contrast, the ESP and reverse-ESP skill can be estimated from a single set of ensemble hindcasts spanning a historical period. The IHC and SCF skill variation method was also highly specific to the particular model state configuration and involved a relatively simplistic linear blending procedure. The elasticity calculations were furthermore based only on a single verification score of forecast skill (i.e., coefficient of determination R^2) for the analysis. An ensemble forecast has many attributes, for example, the skill, the reliability, the resolution, and the uncertainty of the forecast, among others. To obtain a complete picture of

the forecast quality, the scores should encompass many of these attributes, as each verification score will give us different information about the forecast quality.

The drawbacks of VESPA motivate us to assess two computationally inexpensive methods of estimating the forecast skill elasticities, using only the original ESP and reverse-ESP results that depend on the single hindcast series as mentioned above. The two methods are termed end point interpolation (EPI) and end point blending (EPB). In the first part of this paper, we compare results from the two methods tested on 18 catchments of the CONUS to the original results from the VESPA, using a single verification score. The objective of this part is to investigate whether the new methods can discriminate the influence of IHC and SCF errors on seasonal streamflow forecasting uncertainties and to assess the ability of those new methods to correctly estimate the forecast skill elasticities. In the second part, additional verification scores are applied for streamflow forecast verification, supporting the second objective of the paper, which is to explore the sensitivity of the results obtained from the two new methods and the VESPA approach to the choice of the verification score.

2. Methods, data, and evaluation strategy

a. The VESPA approach

In this work, as in W16, the term “perfect” refers to current observed meteorological data and the term climatological refers to the whole distribution of historical observed data. Figure 1 presents the ESP (Fig. 1a), the reverse ESP (Fig. 1b), the climatology (Fig. 1c), and the VESPA forecast (Fig. 1d), as generated in W16. The ESP, the reverse ESP, the perfect forecast, and the climatology are all end points of the uncertainty in the sense that the uncertainty in those forecasts is either perfect or climatological. They are the end points of the VESPA approach.

VESPA aims to produce streamflow forecasts from IHCs and SCFs with an uncertainty situated between the perfect and the climatological uncertainty (Fig. 1d). Forecasts were generated by linearly blending the climatological and perfect IHCs (i.e., model moisture states) and the climatological and perfect SCFs (i.e., meteorological forcings of precipitation, evapotranspiration, and temperature), subsequently used to run the hydrological model. The weights used for blending the data were ($w = 0, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 1.0$), applied so that a weight of zero is the perfect knowledge and unity is the climatological knowledge, with w_{IHC} and w_{SCF} denoting the weights used to blend the IHCs and the SCFs, respectively (W16). An ESP forecast results from the weights $w_{\text{IHC}} = 0$ and $w_{\text{SCF}} = 1$, the reverse ESP from $w_{\text{IHC}} = 1$ and $w_{\text{SCF}} = 0$, the perfect

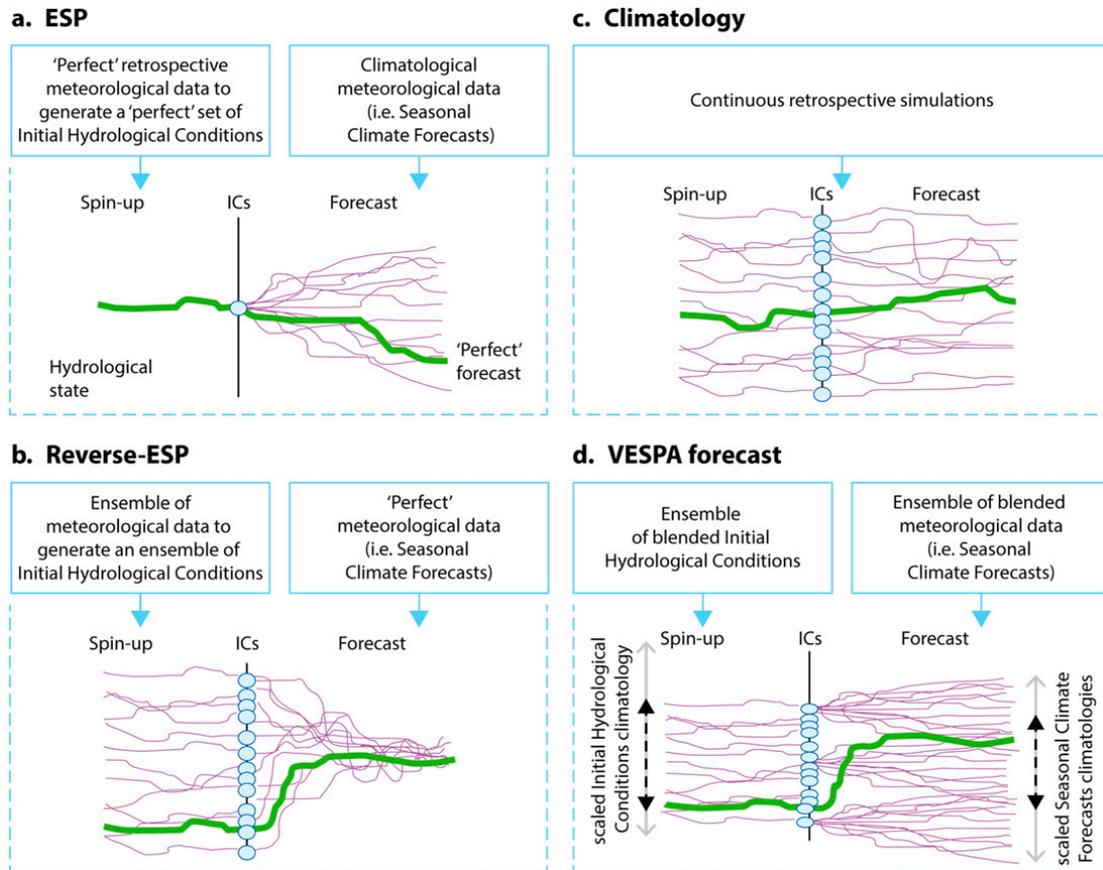


FIG. 1. Schematic of (a) the ESP, (b) the reverse ESP, (c) the climatology, and (d) the VESPA (this figure is adapted from Fig. 3 in W16).

forecast from $w_{IHC} = 0$ and $w_{SCF} = 0$, and the climatology from $w_{IHC} = 1$ and $w_{SCF} = 1$.

To plot the skill of the VESPA forecasts as a function of the IHC and SCF skill, W16 used skill surface plots (Fig. 2), interpolating forecast skill results from different IHC and SCF weighting combinations. The axes represent the SCF and IHC skill, derived respectively from the blending weights w_{SCF} and w_{IHC} using the following equations (W16):

$$\text{SCF skill} = 100 \times (1 - w_{SCF}^2) \quad \text{and} \quad (1)$$

$$\text{IHC skill} = 100 \times (1 - w_{IHC}^2). \quad (2)$$

The SCF and the IHC skill values obtained from these equations are the percentage of climatological variance explained in the respective predictability source (i.e., SCF and IHC; W16). Each SCF skill-IHC skill combination corresponds to a specific VESPA forecast, the skill of which can be plotted on the skill surface plot (black plus signs in Fig. 2). The blue circles are the end points of the VESPA forecasts: the reverse ESP (revESP in Fig. 2), the perfect forecasts, the ESP, and the climatology (climo in Fig. 2). The skill surface plots are hence a graphical representation of the

response surface obtained from the VESPA sensitivity analysis.

The VESPA seasonal streamflow forecasts were generated by W16 using lumped Sacramento Soil Moisture Accounting (SAC-SMA) and SNOW-17 catchment models for unimpaired catchments. The models were forced with daily inputs in precipitation, temperature, and potential evapotranspiration and were calibrated and validated against observed daily streamflow from the U.S. Geological Survey (USGS). Eighty-one skill variations of a 30-yr hindcast (from 1981 to 2010) were produced for 424 catchments in the CONUS, starting at the beginning of each month (i.e., forecast initialization dates), with lead times up to 6 months.

b. Alternative methods to the VESPA approach

In this paper we present two alternative methods of the VESPA approach, the EPI and the EPB. These methods aim to reproduce the response surface obtained from the VESPA approach by using the same 30-yr hindcast ensembles produced by W16, aggregated over the first three months with zero lead time for each initialization date (referred to as 3-month streamflow forecast hereafter) and corresponding exclusively to the end points (i.e., the ESP, the reverse ESP, the perfect forecast, and the climatology).

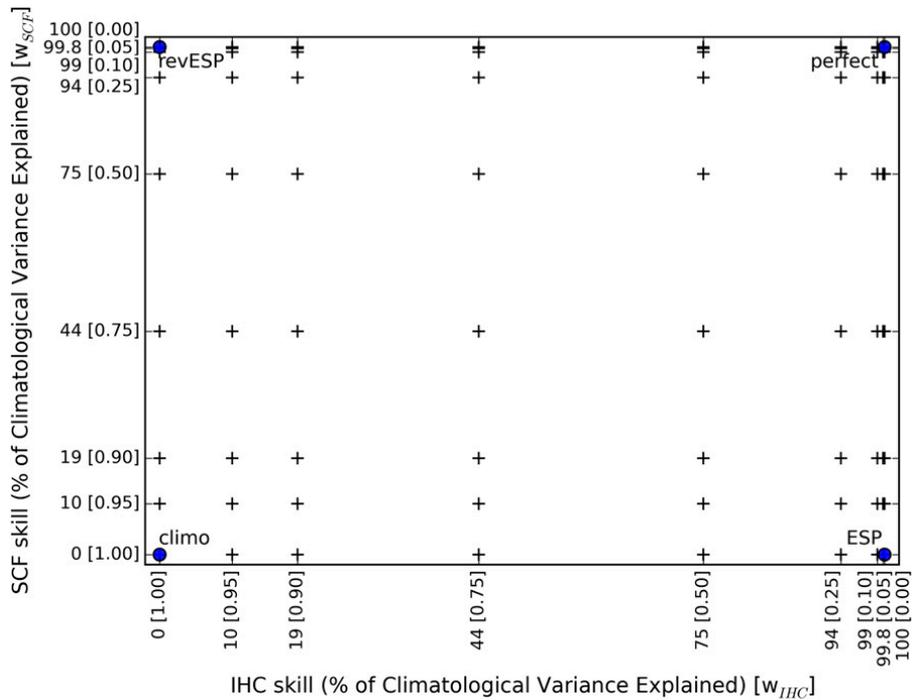


FIG. 2. Schematic of a skill surface plot. The y and x axes display the SCF and the IHC skill, respectively. They are expressed as a percentage of the climatological variance explained by the respective predictability source. The blending weights, w_{SCF} and w_{IHC} , from which the skill values are derived are shown in square brackets in the figure.

The two new methods were tested for a subset of the CONUS-wide catchment dataset presented in W16 (Fig. 3), comprising 18 catchments from the large USGS Hydro-Climatic Data Network (HCDN; Lins 2012). The 18 selected catchments cover a large range of hydrometeorological conditions, including the maritime climate regime of the U.S. West Coast catchments; the humid regime of the eastern United States (south of the Great

Lakes) with rainfall-driven runoff and variable winter snow in the most northern catchments; and the Intermountain West and northern Great Plains regions, where streamflow is greatly influenced by the snow cycle.

1) END POINT INTERPOLATION

The EPI produces a response surface by interpolating the forecast skill of the end points throughout the skill

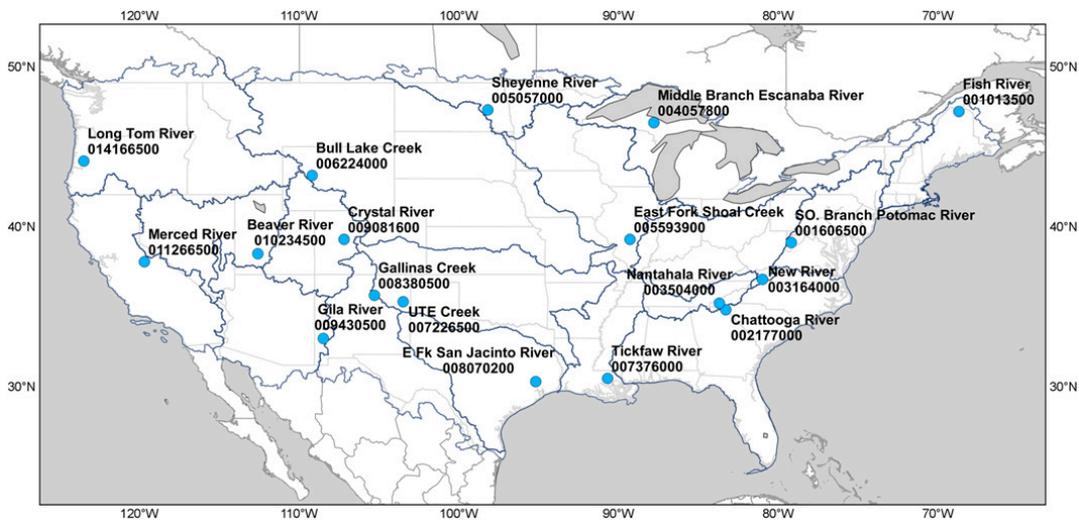


FIG. 3. Map of the 18 catchments of the CONUS selected for the analysis and the HCDN regions (dark blue outlines).

surface plot. Both linear (i.e., linear barycentric interpolation) and cubic interpolation techniques were tested. However, results will be shown for the linear interpolation only as the cubic interpolation did not provide noticeable improvements to the linear interpolation given that the interpolation is based on only four points situated at the corners of the response surface. The linear EPI was performed for each forecast initialization date and for each catchment.

2) END POINT BLENDING

The EPB generates hindcasts for each $w_{\text{SCF}}-w_{\text{IHC}}$ combination (i.e., each plus sign in Fig. 2; w_{SCF} and w_{IHC} are selected to be the same blending weights used by W16, for the purpose of comparison). For each combination point, a new ensemble of 100 members was generated by blending the four end points, given a specific weighted average. The percentage of each end point used [EP(%); i.e., the number of members randomly selected from each end point], was calculated for each combination point using the following equation:

$$\text{EP}(\%) = (1 - |x_{\text{EP}} - w_{\text{IHC}}|) \times (1 - |y_{\text{EP}} - w_{\text{SCF}}|), \quad (3)$$

where x_{EP} and y_{EP} are the w_{IHC} and w_{SCF} values of the end point for which the percentage is calculated, respectively. For example, if the w_{IHC} and w_{SCF} match the end point values, 100% of the EPB hindcast members are resampled from that end point (i.e., the end point skill is reproduced). This was done for each forecast initialization date and for each catchment.

To produce the skill surface plots for the EPB method, the SCF and IHC skill was calculated using the same equations as in W16 [i.e., Eqs. (1) and (2), respectively].

c. The evaluation strategy

The aim of this paper is to compare two computationally inexpensive alternative methods to the VESPA approach, the EPI and the EPB. To this end, the paper unfolds into two distinct objectives. First, we want to

investigate whether the EPI and/or the EPB can discriminate the influence of IHC and SCF errors on seasonal streamflow forecasting uncertainties and reproduce VESPA skill elasticity estimates. This will validate the use of one or both methods as alternative to the VESPA approach. Second, we want to explore the sensitivity of the results obtained from the EPI, the EPB, and the VESPA methods to the choice of the verification score. This will be an attempt to demonstrate the importance of the choice of the verification score for forecast verification and communication.

1) CAN EPI AND EPB DISCRIMINATE THE INFLUENCE OF IHC AND SCF ERRORS ON SEASONAL STREAMFLOW FORECAST UNCERTAINTIES?

To explore the first objective of this paper, skill surface plots were produced for the EPI, the EPB, and the VESPA methods. As in W16, the seasonal streamflow forecast skill depicted in the skill surface plots was calculated from the R^2 of forecast ensemble means with the observations, where perfect forecasts (model simulations driven by the observed meteorology) were treated as observations to calculate the R^2 . As discussed at length in W16, this choice deliberately excludes the model errors as a source of forecast uncertainty.

The skill surface plots obtained from the EPI and the EPB methods were subsequently compared qualitatively and quantitatively to the skill surface plots obtained for the VESPA approach. The qualitative analysis consisted in visually inspecting the patterns contained in the skill surface plots, giving an indication of the dominant predictability source on the streamflow forecast skill. The quantitative analysis of the results was based on the calculation of the skill elasticities for the IHCs and the SCFs (E_{IHC} and E_{SCF} , respectively), for the EPI, the EPB, and the VESPA methods, averaged across three transects of a quadrant situated in the center of the response surface, according to the following equations:

$$E_{\text{IHC}} = 100 \times \left\{ \frac{S(F[75, 19]) - S(F[19, 19])}{75\% - 19\%} + \frac{S(F[75, 44]) - S(F[19, 44])}{75\% - 19\%} + \frac{S(F[75, 75]) - S(F[19, 75])}{75\% - 19\%} \right\} / 3 \quad (4)$$

and

$$E_{\text{SCF}} = 100 \times \left\{ \frac{S(F[19, 75]) - S(F[19, 19])}{75\% - 19\%} + \frac{S(F[44, 75]) - S(F[44, 19])}{75\% - 19\%} + \frac{S(F[75, 75]) - S(F[75, 19])}{75\% - 19\%} \right\} / 3. \quad (5)$$

The numerators, expressed as $S(F[\cdot]) - S(F[\cdot])$, contain the gradients in the streamflow forecast skill

between IHC skill (or SCF skill) values of 75% and 19% (the denominator). The values in between the

square brackets of the numerator are the IHC skill followed by the SCF skill values, which indicates a certain $w_{\text{SCF}}-w_{\text{IHC}}$ combination point in the example skill surface plot in Fig. 2. In the denominator, the IHC and SCF skill gradients are gradients in the percentage of the climatological variance explained in the respective predictability source. The skill elasticities (E_{IHC} and E_{SCF}) are positively oriented, where a skill elasticity of zero is obtained when the predictability source has no influence on the skill of the streamflow forecast, while positive (negative) elasticities mean that an improvement in the predictability source will lead to higher (lower) streamflow forecast skill. The skill elasticities were calculated for all three methods and for the 3-month streamflow forecasts produced for each catchment and forecast initialization date.

The only difference between Eqs. (4) and (5) and the skill elasticities calculated in W16 is that they chose to calculate skill elasticities around the ESP point in the skill surface plots. Here, we choose to calculate skill elasticities across a quadrant within the skill surface plot (between 75% and 19% of the climatological variance explained in the IHC and the SCF) in order for the skill elasticity values calculated in this paper to reflect the forecast skill gradients within the response surface. This is done differently to W16 because the aim of this paper is to compare (qualitatively and quantitatively) the skill surface plots obtained from the EPI and the EPB methods to the VESPA approach.

2) WHAT IS THE SENSITIVITY OF THE RESPONSE SURFACE TO THE CHOICE OF THE VERIFICATION SCORE?

To investigate the second objective of this paper, several verification scores were calculated for each method (i.e., the EPI, the EPB, and the VESPA approach). These scores were selected in order to cover key attributes of the forecasts verified, and they include

- the mean absolute error (MAE) of forecast ensemble means, relative to the perfect forecasts and
- the continuous rank probability score (CRPS) and its decomposition:
 - the potential CRPS (CRPS_{spot}), where CRPS_{spot} = resolution – uncertainty, and
 - the reliability part of the CRPS (CRPS_{reli}).

The potential CRPS is the CRPS value that a forecast with perfect reliability would have. The uncertainty is the variability of the observations and the resolution is the ability of the forecast to distinguish situations with distinctly different frequencies of occurrence. The CRPS reliability is a measure of the bias and the spread of the system.

The CRPS was chosen as it is a widely used score to assess the overall quality of an ensemble hydrometeorological

forecast. The CRPS moreover has the advantage that it can be decomposed into different scores in order to look at the many different attributes of an ensemble forecast. The CRPS for a single forecast is equivalent to the MAE, which is why the latter was chosen.

For all of the above verification scores, the corresponding skill scores were calculated for each point of the skill surface plots from

$$\text{skill score}_{\text{forecast}} = 1 - \frac{\text{score}_{\text{forecast}}}{\text{score}_{\text{reference}}}, \quad (6)$$

where the $\text{score}_{\text{reference}}$ is the score of the climatology point, for each method, each initialization date, and each catchment. A perfect forecast results in a forecast skill score of unity and a forecast with equal quality as the reference forecast corresponds to a skill score of zero. Any forecasts of lower quality than the reference forecast produce negative skill score values. Skill scores were calculated in order to have a similar score range as the R^2 (i.e., a climatological score of zero and a perfect score of one), for comparative purposes.

Skill elasticities were subsequently calculated for all the skill scores, using Eqs. (4) and (5), for all three methods and for the 3-month streamflow forecasts produced for each catchment and forecast initialization date. From these skill elasticity values, the influence of improvements in the IHCs and SCFs on the seasonal streamflow forecast skill can be assessed, in terms of the forecasts' overall performance (considering the mean of the ensemble or the full ensemble spread, from the MAE and the CRPS, respectively), their resolution and uncertainty (CRPS_{spot}), and their reliability (CRPS_{reli}).

3. Results

a. Can EPI and EPB discriminate the influence of IHC and SCF errors on seasonal streamflow forecast uncertainties?

For the first part of this study, the Crystal River (Colorado; USGS gauge 009081600), a snowmelt-driven catchment, will be used as a test case to illustrate the skill surface plots obtained from the EPI and the EPB methods, compared to the VESPA approach. Precipitation is the highest in winter and spring in this catchment and falls as snow between November and April. In April, the snow starts melting and consequently the soil moisture and streamflow both increase.

Figure 4 displays the skill surface plots obtained for the VESPA (Fig. 4a), the linear EPI (Fig. 4b), and the EPB methods (Fig. 4c), from R^2 for the 3-month streamflow forecast for the Crystal River, for initializations on the first of each month (each row in Fig. 4).

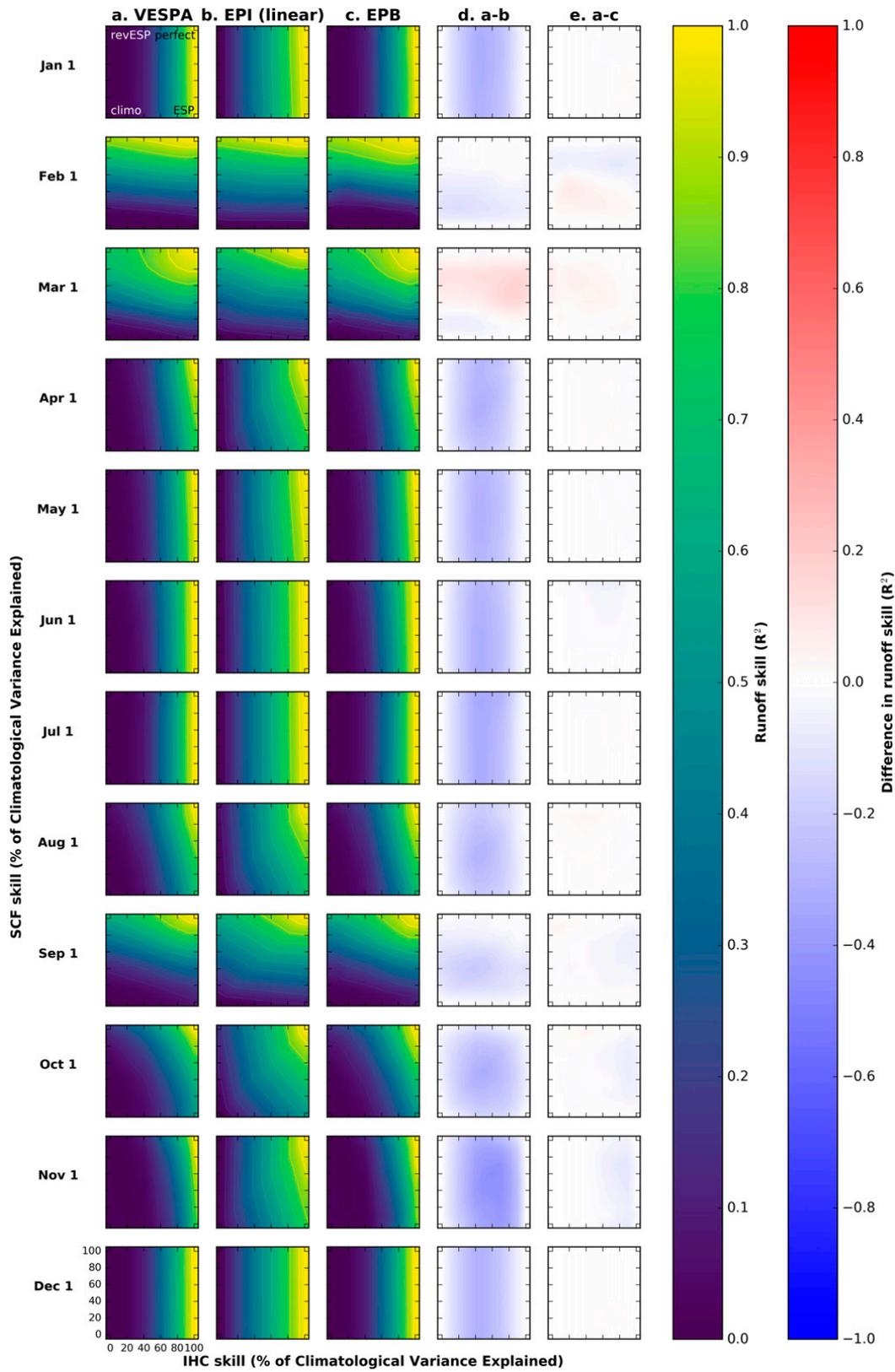


FIG. 4. Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R^2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for the Crystal River (USGS gauge 009081600), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.

Figures 4d and 4e show the differences between the skill surface plots obtained for the VESPA and EPI methods and the VESPA and EPB methods, respectively. A first visual comparison of the skill surface plots obtained from the linear EPI method (Fig. 4b) and the EPB method (Fig. 4c) with those obtained from the VESPA approach (Fig. 4a) for the Crystal River tells us that the skill surface plots obtained from all three methods are very similar. For each initialization date, the orientation of the gradients in streamflow forecast skill appears identical. The EPI and the EPB methods seem to correctly indicate the dominant predictability source on the 3-month streamflow forecast skill, for each initialization date for this catchment. Similar results were obtained for the other 17 catchments (see Figs. S1–S17 in the supplemental material). Forecasts made on the first of February, March, and September show a sensitivity to the SCF skill (i.e., horizontal or near to horizontal orientation of the streamflow forecast skill gradients), while all other forecasts are dominantly sensitive to the IHC skill (i.e., vertical or near to vertical orientation of the streamflow forecast skill gradients).

The gradients in streamflow forecast skill contained in the EPI skill surface plots (Fig. 4b) differ moderately from the gradients obtained from the VESPA approach (Fig. 4a). This can be observed in Fig. 4d, showing the differences between the skill surface plots obtained for both methods. The VESPA approach gives very strong gradients, causing a rapid decrease in streamflow forecast skill with a decrease in one of the predictability sources' skill, depending on the initialization date. In comparison, the EPI method indicates a gradual decrease in streamflow forecast skill with a decrease in one of the two predictability sources, depending on the initialization date. The streamflow forecast skill gradients produced by the EPI method are a reflection of the interpolation method used (i.e., here linear), and because the corner points lack information about describing curvature of the surface at interior points, they cannot fully capture nonlinearities in the skill gradients across the skill surface. For some interior points, this limitation of the EPI method could estimate very different skill elasticities than those obtained from the VESPA approach.

The skill surface plots produced by the EPB method (Fig. 4c) show minor differences in the streamflow forecast skill gradients when compared to the skill surface plots generated by the VESPA approach (Fig. 4a). This can be seen in Fig. 4e, which shows the differences between the skill surface plots obtained for both methods. To further inspect those differences, they will be explored quantitatively (i.e., by comparing the skill elasticities) below.

To quantify the accuracy of the patterns contained in the EPI and the EPB skill surface plots compared to the patterns of the VESPA skill surface plots, IHC and SCF skill

elasticities (i.e., E_{IHC} and E_{SCF} , respectively) were calculated across a quadrant situated within the response surface for all three methods, for the 18 catchments and each forecast initialization date, from Eqs. (4) and (5), respectively. Figure 5 presents the skill elasticities for nine of the 18 catchments (the plots for the other nine catchments are shown in Fig. S18). Each plot corresponds to a catchment and shows the skill elasticities obtained from the VESPA, the linear EPI, and the EPB methods as a function of the forecast initialization date. From the nine different plots, the skill elasticities given by the EPB method appear almost identical to the VESPA approach, whereas the skill elasticities obtained from the EPI method differ in some places. This confirms that the patterns of the EPB method are very similar to the patterns of the VESPA approach, with it being the closest out of the two tested methods.

The value of the SCF skill elasticity (i.e., E_{SCF}) in relation to the value of the IHC skill elasticity (i.e., E_{IHC}), for a given method, indicates the dominant predictability source on the 3-month streamflow forecast skill (here calculated from the R^2). For a selected method, equal SCF and IHC skill elasticity values signifies that equal improvements in both the SCFs and the IHCs will lead to equal improvements in the streamflow forecast skill. If E_{SCF} is superior (inferior) to E_{IHC} , it reflects a larger potential increase in streamflow forecast skill by improving the SCFs (IHCs). Although the EPI method almost always indicates the same dominant predictability source as the two other methods, the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill (i.e., the exact values of the skill elasticities) often differs. For many catchments and forecast initialization dates, the EPI appears to underestimate the skill elasticities produced by the VESPA method.

The nine different catchments for which the skill elasticities are presented in Fig. 5 display three different types of behavior, best captured by the VESPA approach and the EPB method. For the three catchments in Fig. 5 (left), improvements in the IHCs would yield the highest improvements in the 3-month streamflow forecast skill for spring to summer initializations (April–August for the Crystal River, March–July for the Fish River, and March–June for the Middle Branch Escanaba River) and in the winter (October–January for the Crystal River, November–December for the Fish River, and in December for the Middle Branch Escanaba River). SCF improvements would lead to better 3-month streamflow forecast skill for forecasts initialized in the late winter and summer to fall (February–March and September for the Crystal River, February and August–October for the Fish River, and January–February and July–September for the Middle Branch Escanaba River). For the three catchments in Fig. 5 (middle),

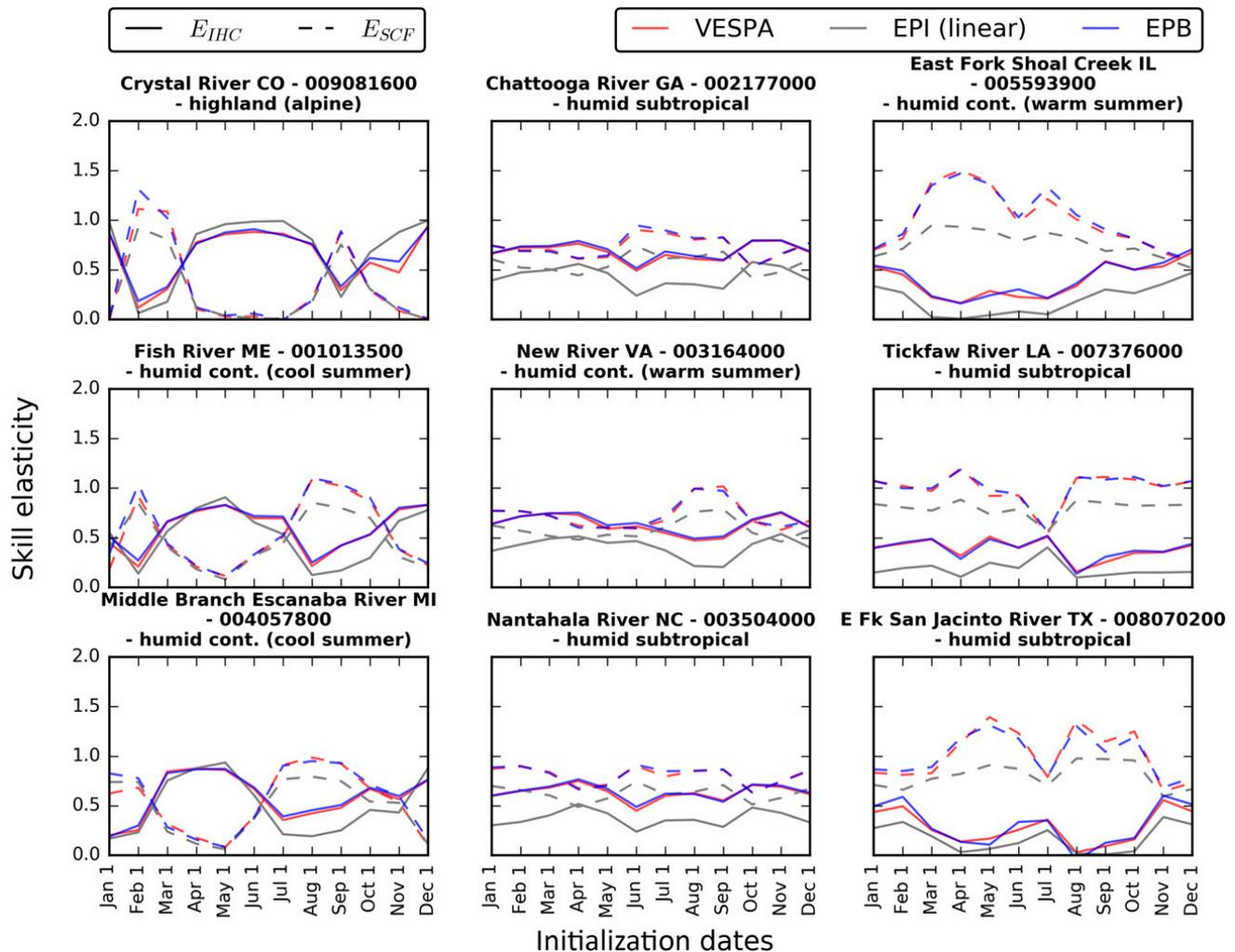


FIG. 5. Streamflow forecast skill elasticities for the IHCs (i.e., E_{IHC} , solid line) and the SCFs (i.e., E_{SCF} , dashed line), calculated across a quadrant situated within the 3-month streamflow forecast skill surface plots for the VESPA (red), the linear EPI method (gray), and the EPB method [blue; using Eqs. (4) and (5)]. Each plot shows the evolution of the IHC and SCF skill elasticities with the initialization date for a given catchment. The climatological regions of the catchments are indicated in the plots' headings. The skill surface plots from which these skill elasticities were calculated are presented in Fig. 4 and Figs. S1–S17.

a notable feature is that the 3-month streamflow forecast skill would benefit from SCF improvements for summer initializations (June–September for the Chattooga and the Nantahala Rivers and July–September for the New River). Finally, for the three catchments in Fig. 5 (right), the 3-month streamflow forecast skill would benefit from improvements in the SCFs for all initialization dates. This is true with the exception of forecasts initialized in December for East Fork Shoal Creek. It is important to note that there is uncertainty around these estimates. However, this is a good first indication of the sensitivity of 3-month streamflow forecast skill (measured from the R^2) to IHC and SCF errors for each forecast initialization date and each catchment.

The skill elasticities produced by the EPB method appear to be almost identical to the skill elasticities

obtained from the VESPA approach, with occasional marginal differences. This suggests that the EPB method captures nearly exactly the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill, obtained from the VESPA approach. Both methods additionally indicate the same dominant predictability source: the predictability source which, once improved, could lead to the largest increase in 3-month streamflow forecast skill. The EPB method will therefore be used as an alternative to the VESPA approach to investigate the second objective of this paper.

b. What is the sensitivity of the response surface to the choice of the verification score?

To investigate the sensitivity of the response surface to the choice of the verification score, and therefore to

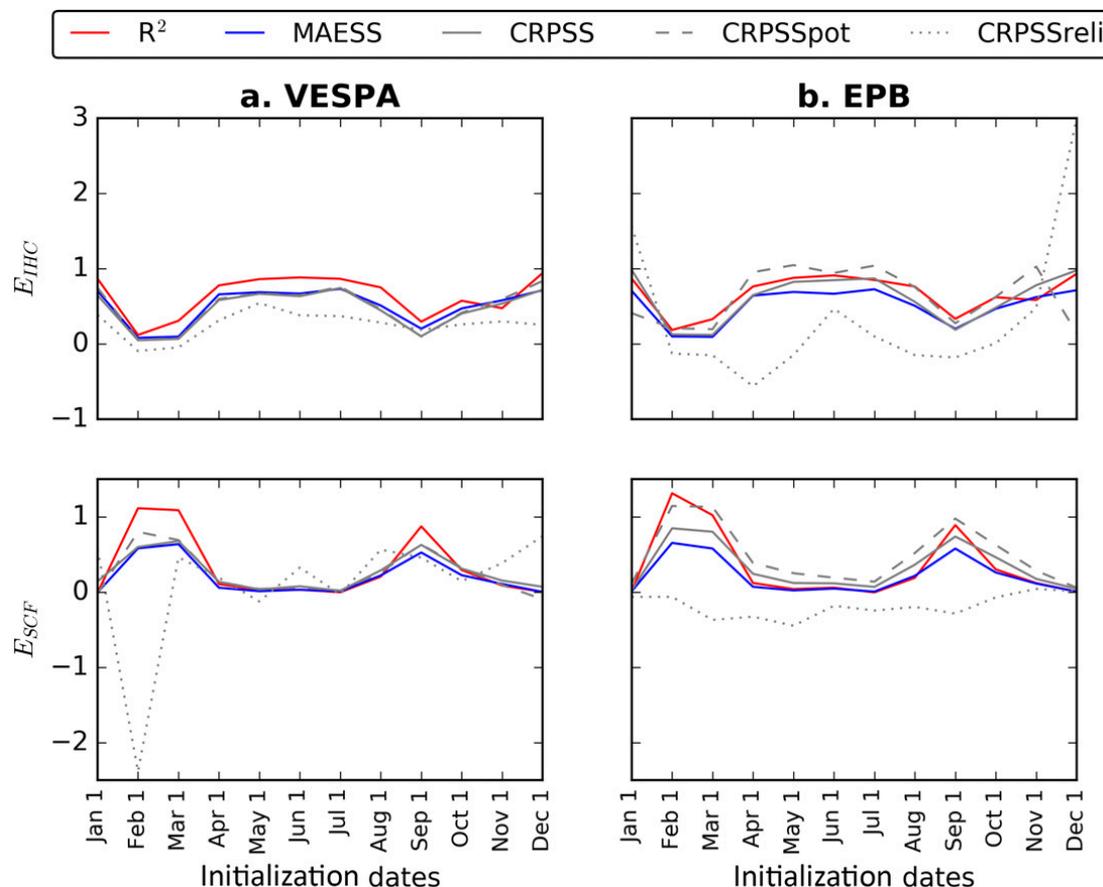


FIG. 6. Streamflow forecast skill elasticities for the (top) IHCs (i.e., E_{IHC}) and (bottom) SCFs (i.e., E_{SCF}) as a function of forecast initialization dates, for hindcasts produced from 1981 to 2010 for the Crystal River (USGS gauge 009081600). These skill elasticities were calculated across a quadrant situated within the 3-month streamflow forecast skill surface plots [from Eqs. (4) and (5)] for several verification scores (R^2 in red, MAESS in blue, CRPSS in gray solid line, CRPSSpot in gray dashed line, and CRPSSreli in gray dotted line). The results are shown for (a) the VESPA approach and (b) the EPB method.

the attribute of the forecast, several scores were computed to evaluate the streamflow forecast quality. The R^2 , the mean absolute error skill score (MAESS), and the continuous rank probability skill score (CRPSS) were calculated to evaluate the forecasts' overall performance in terms of the ensemble mean and the entire ensemble. The potential CRPSS (CRPSSpot) was computed to look at the forecasts' resolution and uncertainty, and the CRPSS reliability (CRPSSreli) was computed to look at the forecasts' reliability. The Crystal River (USGS gauge 009081600) will here again be used as a test case to illustrate this part of the results.

Figure 6 presents the IHC and SCF skill elasticities [i.e., E_{IHC} and E_{SCF} ; in Fig. 6 (top) and Fig. 6 (bottom), respectively] as a function of forecast initialization date for the Crystal River catchment. These are calculated from Eqs. (4) and (5), for all the mentioned verification scores, for the VESPA approach (Fig. 6a) and the EPB method (Fig. 6b). If we compare the skill elasticities obtained from the VESPA approach with

the skill elasticities obtained from the EPB method, it appears that both methods produce very similar elasticities for the R^2 , the MAESS, and the CRPSS. This further confirms the results of the first part of the analysis, which highlighted the similarity of the EPB results to the VESPA results and extends it to multiple attributes of the seasonal streamflow forecasts. However, slight differences between the skill elasticities produced by the two methods can be observed for the CRPSSpot, and significant differences exist for the CRPSSreli. These dissimilarities are discussed further below.

If we now compare the skill elasticities obtained for the various verification scores for both methods, it is clear that the R^2 , the MAESS, the CRPSS, and the CRPSSpot give very similar skill elasticities. This hints that those verification scores overall agree on the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill. However, a few dissimilarities can be observed for some of the forecast initialization dates. This is, for example, the case for forecasts made in

the spring and in summer, where the E_{IHC} appears lower for the MAESS and the CRPSS (and the CRPSSpot for the VESPA approach) compared to the E_{IHC} obtained for the R^2 for both methods. It is also apparent for forecasts made on the first of February, March, and September, where the E_{SCF} calculated for the MAESS and the CRPSS (and the CRPSSpot for the VESPA approach) is lower than the E_{SCF} obtained for the R^2 for both methods. For both examples, it infers that improvements in the IHC and the SCF skill could lead to larger improvements in the streamflow forecast skill in terms of the R^2 rather than in terms of the MAESS and the CRPSS (and the CRPSSpot for the VESPA approach). Overall, this indicates that the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill differs relative to the choice of the verification score.

While the R^2 , the MAESS, the CRPSS, and the CRPSSpot give a very similar picture, the skill elasticities obtained for the CRPSSreli appear very different, occasionally reaching negative values. These negative values indicate a loss in streamflow forecast skill (in terms of the forecast reliability) as a result of improvements in one of the two predictability sources, while all the other verification scores suggest a gain in streamflow forecast skill (in terms of the forecast ensemble mean and the ensemble overall performance, its resolution, and uncertainty) with improvements in one of the two predictability sources.

The substantial differences in skill elasticities obtained for the CRPSSreli from the VESPA versus EPB method suggest that there are limitations to the ability of EPB to reconstruct the full ensemble information present in VESPA, and of VESPA (applied with relatively small ensembles at the end points) to estimate sensitivities for complex verification scores such as reliability. The reliability verification score is influenced by the combination of bias, spread, and other ensemble properties and exhibits more noisy outcomes here than were obtained for other verification scores. A negative elasticity may occur because the ensemble spread has narrowed without sufficient improvements in bias, for instance. The behavior of the elasticity of reliabilities is even more difficult to diagnose, but we suspect that the presence of noise (erroneous local minima or maxima) or curvature in the associated VESPA skill surface greatly undermines the linear blending techniques.

Overall, these results suggest that improvements in the skill of either of the two predictability sources will impact streamflow forecast skill differently depending on the attribute (i.e., verification score) of the forecast skill that is considered and whether the ensemble mean or the full ensemble is used.

4. Discussion

a. Implications and limitations of the results

W16 introduced the VESPA approach, a sensitivity analysis technique used to pinpoint the dominant predictability source of seasonal streamflow forecasting (i.e., the IHCs and the SCFs), as well as quantifying improvements that can be expected in seasonal streamflow forecast skill as a result of realistic improvements in those key predictability sources. Despite being a powerful sensitivity analysis approach, VESPA presents two key limitations.

- 1) It is computationally intensive, requiring multiple ensemble hindcasts to define the skill response surface (81 were used in the VESPA paper vs one for the EPB and the EPI techniques).
- 2) It requires a complex state and forcing blending procedure that may introduce additional uncertainties, biases, or interactions between the predictability sources (Saltelli et al. 2004; Baroni and Tarantola 2014) that are not accounted for or difficult to quantify. This is not necessary in any of the end points required in the two approaches presented here, which rely instead on analyzing a single conventional hindcast dataset that is more likely to be feasible for forecasting centers.

The central aim of this paper was to address the first limitation of the VESPA approach by presenting two computationally inexpensive alternative methods: the EPI and the EPB methods. Both methods successfully identified the dominant predictability source of 3-month streamflow forecasts for a given catchment and forecast initialization date (i.e., given by the orientation of the streamflow forecast skill gradients in the skill surface plots). However, the EPB was more successful in reproducing the VESPA skill elasticities—the exact streamflow forecast skill gradients situated within the skill surface plots (for skill and accuracy verification scores including the R^2 , the MAESS, the CRPSS, and the potential CRPSS to a certain extent). These skill elasticities indicate the influence of changes in IHC and SCF skill on streamflow forecast skill.

The new methods, by differing in their setup from the VESPA approach, do not inherit the drawbacks specific to this approach and mentioned above. The EPI and the EPB methods nevertheless have their own limitations.

The EPI (both for the linear and cubic interpolation methods; the latter was not shown) did not fully capture the VESPA skill elasticities because of the nature of the method that produces predefined gradients within the skill surface plots (i.e., defined by the interpolation method used). Additionally, curvature or local minima

or maxima (if any) of the response surface cannot be represented by the EPI method. The EPB, on the other hand, performs better at reflecting curvature in the skill response surface, hence local elasticities between the end points. The EPB method aimed at reproducing VESPA elasticities only by manipulating the output of a single hindcast dataset (interpreted as ESP, reverse ESP, the perfect forecast, and climatology). The EPB method cannot match exactly the forecasts created by the VESPA approach, as it does not account for the idiosyncrasies in model forecast behavior, such as interactions between the predictability sources. Furthermore, it is likely that the more the model investigated is nonlinear or exhibits skill response thresholds, the more the results obtained from the EPB method will differ from the ones obtained from the VESPA approach. These results overall allow that the EPB method can be used as an inexpensive alternative method to the VESPA approach, yet with the potential limitations of the method stated above.

For the first part of the analysis, the streamflow forecast quality was evaluated in terms of the forecasts' skill from the R^2 . The use of multiple verification scores is, however, essential to obtain a more complete perspective of forecast quality. Thus, we explored the performance of the two new methods and the VESPA approach for a range of additional verification scores. The results, presented for the EPB method and the VESPA approach, showed differences in the response surfaces obtained for the various verification scores (i.e., the R^2 , the MAESS, the CRPSS, and its decomposition). This suggests distinct sensitivities of the seasonal streamflow forecast attributes (i.e., overall performance of the forecast ensemble mean and its full ensemble, forecast resolution, uncertainty, and reliability) to changes in the IHC and SCF skill. Ideally, a sensitivity analysis should be goal oriented, that is, it should be performed with prior knowledge of the intended use of the results (Saltelli et al. 2004; Pappenberger et al. 2010; Baroni and Tarantola 2014), which may favor using one verification score over another.

This paper covered selected limitations of the work presented by W16. However, many areas were left unexplored and could be interesting topics in which to focus future research. First, a major area inherent to model-based sensitivity analyses is that their results are model dependent (Saltelli et al. 2000); thus, the extent to which they can be transferred to reality depends on the model fidelity. The results presented in this paper are specific to the forecasting system and similar systems on which this analysis was based and should be used as an indicator of catchment sensitivities. As noted in W16, an extension of the elasticity analysis to include

observations and a model error component would provide valuable insights. Another possible approach could be to use the results from various forecasting systems as input to the sensitivity analysis, in order to achieve a multimodel consensus view of the skill. As shown in Cloke et al. (2017), a multimodel forcing framework can be highly beneficial for streamflow forecasting compared to a single model forecasting approach, provided the models are chosen judiciously so as to provide a rational characterization of forecasting uncertainty. Second, the dependence of blending technique performance versus VESPA on the characteristics of the skill surface (e.g., linear or nonlinear) bears further investigation. Finally, this sensitivity analysis leaves generic the concept of improvements in either of the predictability sources, although the space–time nature of improvements may be consequential. This work could therefore be extended by studying the effect of degradations in the temporal and spatial accuracy of the input data, thereby indicating the relative value of improvements in the spatial or temporal predictability for a specific catchment and a specific time of the year.

b. The wider context

The new strategy of operational forecasting centers is to move toward more integrated operational modeling and forecasting approaches, such as land surface–atmosphere coupled systems, and beyond that, Earth system models. These advances are enabled by the continuous growth of computing capabilities, a better understanding of physical processes and their interactions throughout all compartments of the Earth system, and the availability and use of more and better observation data (i.e., satellite data). Despite all these advances, most forecasts still reflect substantial uncertainty that grows with time and limits the predictability of observed events beyond a few weeks of lead time. The rapid progress has led our systems to be ever more data hungry as increases in model complexity and resolution are sought. These computationally expensive developments are not always feasible; hence, model developers must be creative and constantly weigh the costs and benefits of improving one aspect over another, such as increasing the resolution or complexity of the models (Flato 2011).

In this context, sensitivity analyses appear more than ever as a natural tool to establish priorities in improving predictions based on Earth system modeling. Such analyses are a powerful and valuable tool to support the examination of uncertainty and predictability across spatial and temporal scales and for various applications. They can be used for a large range of activities, including examining model structure,

identifying minimum data standards, establishing priorities for updating forecasting systems, designing field campaigns, and providing realistic insights into the potential benefits of efforts to improve a forecasting system to managers with prior knowledge of their costs (Cloke et al. 2008; Lilburne and Tarantola 2009; W16).

However, sensitivity analyses must be easily reproducible to be effective in supporting each new model or forecast system update, and the results should easily be applied in order to constitute a “continuous learning process” (Baroni and Tarantola 2014). In other words, a sensitivity analysis should be a simple, tractable tool for addressing a multifaceted challenge.

5. Conclusions

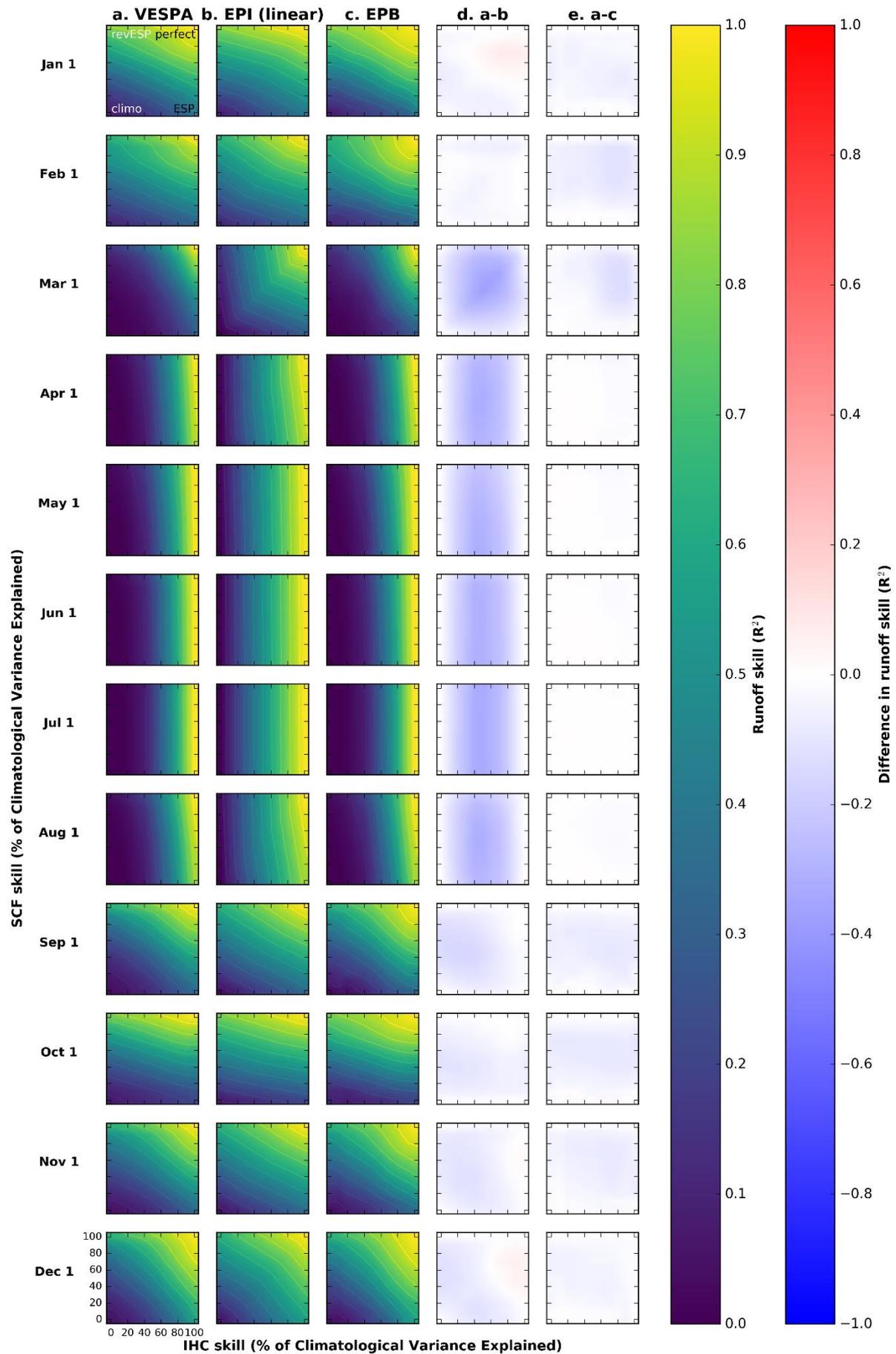
This paper presents two computationally inexpensive alternative methods to the VESPA approach for estimating forecast skill sensitivities and elasticities. Of these, the end point blending (EPB) method provides a useful substitute to the VESPA approach. Despite the existence of some differences between the EPB and the VESPA outcomes, the EPB successfully identifies the dominant predictability source (i.e., IHCs and SCFs) of seasonal streamflow forecast skill, for a given catchment and forecast initialization date. The EPB method can additionally reproduce the VESPA forecast skill elasticities, indicating the degree of influence of changes in IHC and SCF skill on the streamflow forecast skill. The paper also draws attention to how the choice of verification score impacts the forecast’s sensitivity to improvements made to the predictability sources. With a good understanding of the limitations of the methods, such a sensitivity analysis approach can be a valuable tool to guide future forecasting and modeling developments.

Acknowledgments. L. Arnal, A. W. Wood, H. L. Cloke, and F. Pappenberger gratefully acknowledge financial support from the Horizon 2020 IMPREX project (Grant Agreement 641811) (project IMPREX: www.imprex.eu). E. Stephens’ time was funded by the Leverhulme Early Career Fellowship ECF-2013-492. We also acknowledge high-performance computing support from Yellowstone (ark:/85065/d7wd3xhc) provided by NCAR’s Computational and Information Systems Laboratory, sponsored by the National Science Foundation. Last, A. W. Wood is thankful for support from the U.S. Bureau of Reclamation under Cooperative Agreement R11AC80816 and from the U.S. Army Corps of Engineers (USACE) Climate Preparedness and Resilience Program (Award Number 1254557).

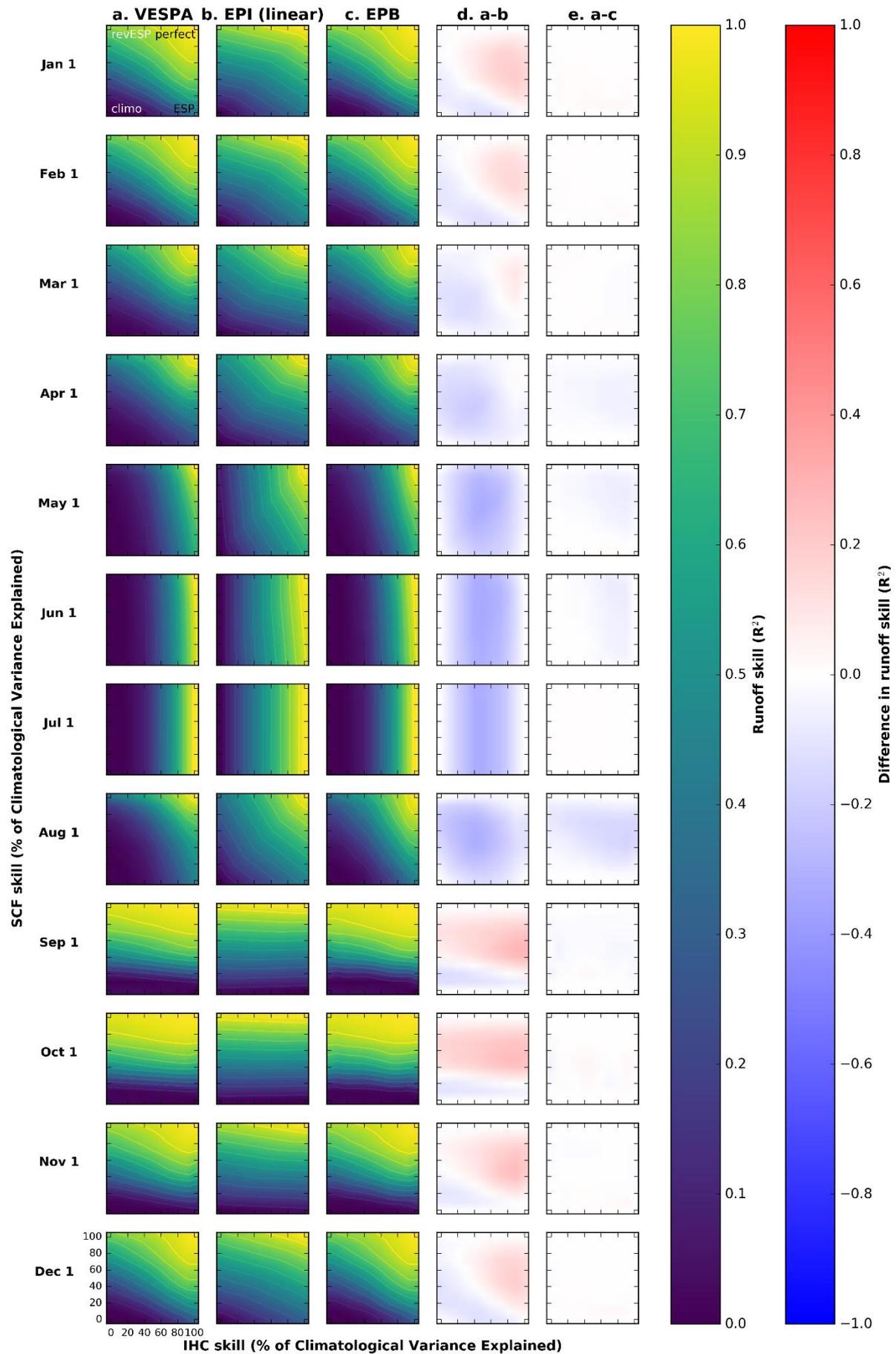
REFERENCES

- Baroni, G., and S. Tarantola, 2014: A general probabilistic approach for uncertainty and global sensitivity analysis of deterministic models: A hydrological case study. *Environ. Modell. Software*, **51**, 26–34, doi:10.1016/j.envsoft.2013.09.022.
- Bierkens, M. F. P., and L. P. H. van Beek, 2009: Seasonal predictability of European discharge: NAO and hydrological response time. *J. Hydrometeorol.*, **10**, 953–968, doi:10.1175/2009JHM1034.1.
- Cherry, J., H. Cullen, M. Visbeck, A. Small, and C. Uvo, 2005: Impacts of the North Atlantic Oscillation on Scandinavian hydropower production and energy markets. *Water Resour. Manage.*, **19**, 673–691, doi:10.1007/s11269-005-3279-z.
- Chiew, F. H. S., S. L. Zhou, and T. A. McMahon, 2003: Use of seasonal streamflow forecasts in water resources management. *J. Hydrol.*, **270**, 135–144, doi:10.1016/S0022-1694(02)00292-5.
- Clark, M. P., M. C. Serreze, and G. J. McCabe, 2001: Historical effects of El Niño and La Niña events on the seasonal evolution of the montane snowpack in the Columbia and Colorado River basins. *Water Resour. Res.*, **37**, 741–757, doi:10.1029/2000WR900305.
- Cloke, H. L., F. Pappenberger, and J.-P. Renaud, 2008: Multi-method global sensitivity analysis (MMGSA) for modelling floodplain hydrological processes. *Hydrol. Processes*, **22**, 1660–1674, doi:10.1002/hyp.6734.
- , —, P. Smith, and F. Wetterhall, 2017: How do I know if I’ve improved my continental scale flood early warning system? *Environ. Res. Lett.*, **12**, 044006, doi:10.1088/1748-9326/aa625a.
- Day, G. N., 1985: Extended streamflow forecasting using NWSRFS. *J. Water Resour. Plann. Manage.*, **111**, 157–170, doi:10.1061/(ASCE)0733-9496(1985)111:2(157).
- Demargne, J., and Coauthors, 2014: The science of NOAA’s operational Hydrologic Ensemble Forecast Service. *Bull. Amer. Meteor. Soc.*, **95**, 79–98, doi:10.1175/BAMS-D-12-00081.1.
- Flato, G. M., 2011: Earth system models: An overview. *Wiley Interdiscip. Rev.: Climate Change*, **2**, 783–800, doi:10.1002/wcc.148.
- Hamlet, A. F., D. Huppert, and D. P. Lettenmaier, 2002: Economic value of long-lead streamflow forecasts for Columbia River hydropower. *J. Water Resour. Plann. Manage.*, **128**, 91–101, doi:10.1061/(ASCE)0733-9496(2002)128:2(91).
- Kwon, H.-H., C. Brown, K. Xu, and U. Lall, 2009: Seasonal and annual maximum streamflow forecasting using climate information: Application to the Three Gorges Dam in the Yangtze River basin, China. *Hydrol. Sci. J.*, **54**, 582–595, doi:10.1623/hysj.54.3.582.
- Li, H., L. Luo, E. F. Wood, and J. Schaake, 2009: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. *J. Geophys. Res.*, **114**, D04114, doi:10.1029/2008JD010969.
- Lilburne, L., and S. Tarantola, 2009: Sensitivity analysis of spatial models. *Int. J. Geogr. Inf. Sci.*, **23**, 151–168, doi:10.1080/13658810802094995.
- Lins, H. F., 2012: USGS Hydro-Climatic Data Network 2009 (HCDN-2009). USGS Fact Sheet 2012-3047, 4 pp. [Available online at <http://pubs.usgs.gov/fs/2012/3047/>.]
- Luo, L., and E. F. Wood, 2007: Monitoring and predicting the 2007 U.S. drought. *Geophys. Res. Lett.*, **34**, L22702, doi:10.1029/2007GL031673.
- MacLeod, D., H. Cloke, F. Pappenberger, and A. Weisheimer, 2016: Evaluating uncertainty in estimates of soil moisture memory with a reverse ensemble approach. *Hydrol. Earth Syst. Sci.*, **20**, 2737–2743, doi:10.5194/hess-20-2737-2016.

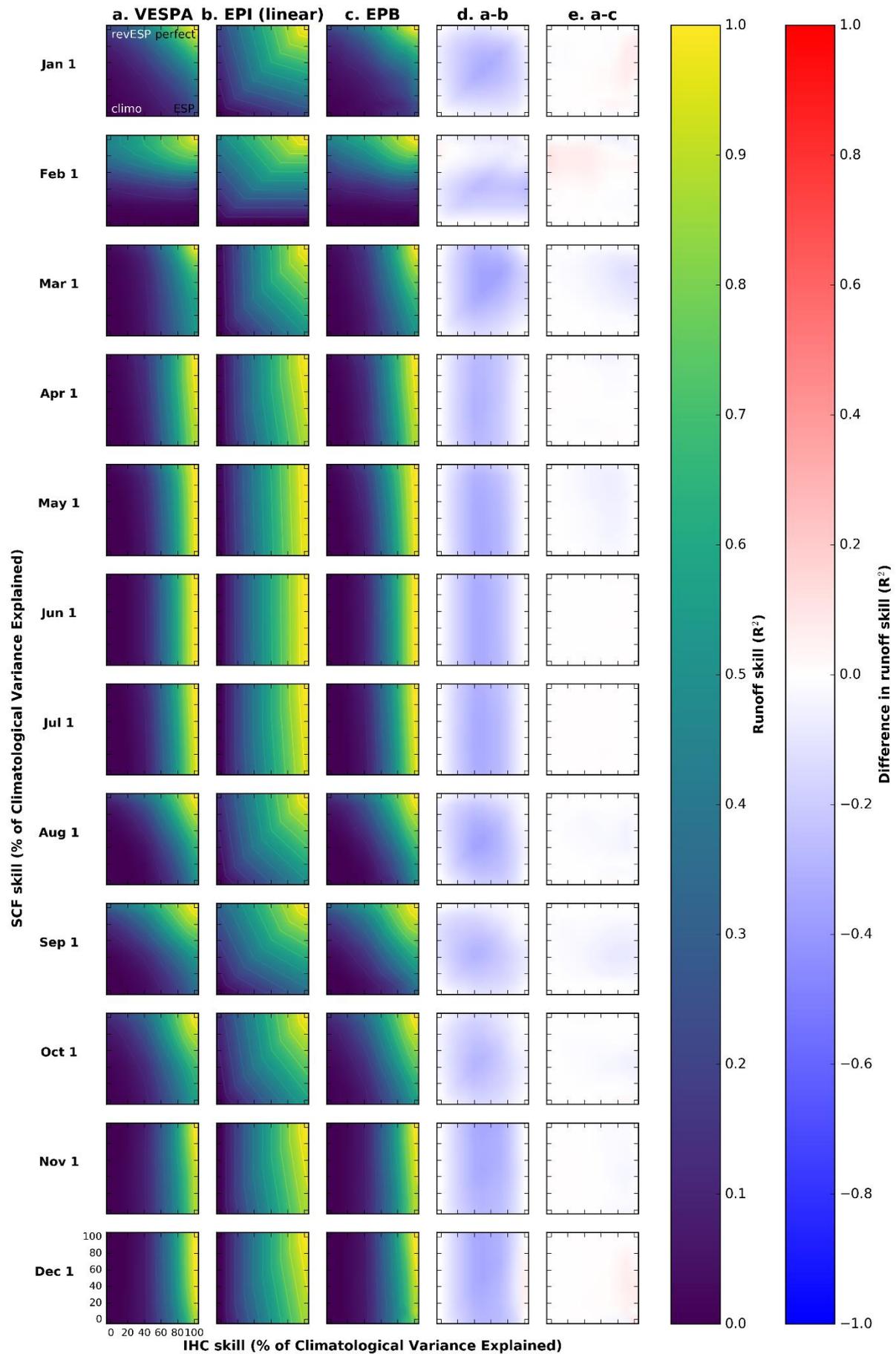
- Mendoza, P. A., A. W. Wood, E. A. Clark, E. Rothwell, M. P. Clark, B. Nijssen, L. D. Brekke, and J. R. Arnold, 2017: An intercomparison of approaches for improving predictability in operational seasonal streamflow forecasting. *Hydrol. Earth Syst. Sci. Discuss.*, doi:10.5194/hess-2017-60.
- Pagano, T., D. Garen, and S. Sorooshian, 2004: Evaluation of official western U.S. seasonal water supply outlooks, 1922–2002. *J. Hydrometeor.*, **5**, 896–909, doi:10.1175/1525-7541(2004)005<0896:EOOWUS>2.0.CO;2.
- Paiva, R. C. D., W. Collischonn, M. P. Bonnet, and L. G. G. de Gonçalves, 2012: On the sources of hydrological prediction uncertainty in the Amazon. *Hydrol. Earth Syst. Sci.*, **16**, 3127–3137, doi:10.5194/hess-16-3127-2012.
- Pappenberger, F., M. Ratto, and V. Vandenberghe, 2010: Review of sensitivity analysis methods. *Modelling Aspects of Water Approach Directive Implementation*, P. A. Vanrolleghem, Ed., IWA Publishing, 191–265.
- Regonda, S. K., B. Rajagopalan, M. Clark, and E. Zagana, 2006: A multimodel ensemble forecast approach: Application to spring seasonal flows in the Gunnison River basin. *Water Resour. Res.*, **42**, W09404, doi:10.1029/2005WR004653.
- Saltelli, A., S. Tarantola, and F. Campolongo, 2000: Sensitivity analysis as an ingredient of modeling. *Stat. Sci.*, **15**, 377–395, doi:10.1214/ss/1009213004.
- , ———, ———, and M. Ratto, 2004: *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. John Wiley & Sons, 218 pp.
- Shukla, S., and D. P. Lettenmaier, 2011: Seasonal hydrologic prediction in the United States: Understanding the role of initial hydrologic conditions and seasonal climate forecast skill. *Hydrol. Earth Syst. Sci.*, **15**, 3529–3538, doi:10.5194/hess-15-3529-2011.
- , J. Sheffield, E. F. Wood, and D. P. Lettenmaier, 2013: On the sources of global land surface hydrologic predictability. *Hydrol. Earth Syst. Sci.*, **17**, 2781–2796, doi:10.5194/hess-17-2781-2013.
- Singla, S., J. P. Céron, E. Martin, F. Regimbeau, M. Déqué, F. Habets, and J. P. Vidal, 2012: Predictability of soil moisture and river flows over France for the spring season. *Hydrol. Earth Syst. Sci.*, **16**, 201–216, doi:10.5194/hess-16-201-2012.
- Slater, L. J., G. Villarini, and A. A. Bradley, 2017: Evaluation of the skill of North-American Multi-Model Ensemble (NMME) global climate models in predicting average and extreme precipitation and temperature over the continental USA. *Climate Dyn.*, doi:10.1007/s00382-016-3286-1, in press.
- Staudinger, M., and J. Seibert, 2014: Predictability of low flow—An assessment with simulation experiments. *J. Hydrol.*, **519**, 1383–1393, doi:10.1016/j.jhydrol.2014.08.061.
- van Dijk, A. I. J. M., J. L. Peña-Arancibia, E. F. Wood, J. Sheffield, and H. E. Beck, 2013: Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide. *Water Resour. Res.*, **49**, 2729–2746, doi:10.1002/wrcr.20251.
- Viel, C., A.-L. Beaulant, J.-M. Soubeyroux, and J.-P. Céron, 2016: How seasonal forecast could help a decision maker: An example of climate service for water resource management. *Adv. Sci. Res.*, **13**, 51–55, doi:10.5194/asr-13-51-2016.
- Welles, E., S. Sorooshian, G. Carter, and B. Olsen, 2007: Hydrologic verification: A call for action and collaboration. *Bull. Amer. Meteor. Soc.*, **88**, 503–511, doi:10.1175/BAMS-88-4-503.
- Wood, A. W., and D. P. Lettenmaier, 2006: A test bed for new seasonal hydrologic forecasting approaches in the western United States. *Bull. Amer. Meteor. Soc.*, **87**, 1699–1712, doi:10.1175/BAMS-87-12-1699.
- , and ———, 2008: An ensemble approach for attribution of hydrologic prediction uncertainty. *Geophys. Res. Lett.*, **35**, L14401, doi:10.1029/2008GL034648.
- , E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.*, **107**, 4429, doi:10.1029/2001JD000659.
- , A. Kumar, and D. P. Lettenmaier, 2005: A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States. *J. Geophys. Res.*, **110**, D04105, doi:10.1029/2004JD004508.
- , T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark, 2016a: Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill. *J. Hydrometeor.*, **17**, 651–668, doi:10.1175/JHM-D-14-0213.1.
- , T. Pagano, and M. Roos, 2016b: Tracing the origins of ESP. HEPEX, accessed 24 October 2016. [Available online at <https://hepex.irstea.fr/tracing-the-origins-of-esp/>.]
- Yossef, N. C., H. Winsemius, A. Weerts, R. van Beek, and M. F. P. Bierkens, 2013: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing. *Water Resour. Res.*, **49**, 4687–4699, doi:10.1002/wrcr.20350.
- Yuan, X., 2016: An experimental seasonal hydrological forecasting system over the Yellow River basin—Part 2: The added value from climate forecast models. *Hydrol. Earth Syst. Sci.*, **20**, 2453–2466, doi:10.5194/hess-20-2453-2016.
- , E. F. Wood, L. Luo, and M. Pan, 2011: A first look at Climate Forecast System version 2 (CFSv2) for hydrological seasonal prediction. *Geophys. Res. Lett.*, **38**, L13402, doi:10.1029/2011GL047792.
- , ———, J. K. Roundy, and M. Pan, 2013: CFSv2-based seasonal hydroclimatic forecasts over the conterminous United States. *J. Climate*, **26**, 4828–4847, doi:10.1175/JCLI-D-12-00683.1.
- , ———, and Z. Ma, 2015: A review on climate-model-based seasonal hydrologic forecasting: Physical understanding and system development. *Wiley Interdiscip. Rev.: Water*, **2**, 523–536, doi:10.1002/wat2.1088.
- , F. Ma, L. Wang, Z. Zheng, Z. Ma, A. Ye, and S. Peng, 2016: An experimental seasonal hydrological forecasting system over the Yellow River basin—Part 1: Understanding the role of initial hydrological conditions. *Hydrol. Earth Syst. Sci.*, **20**, 2437–2451, doi:10.5194/hess-20-2437-2016.



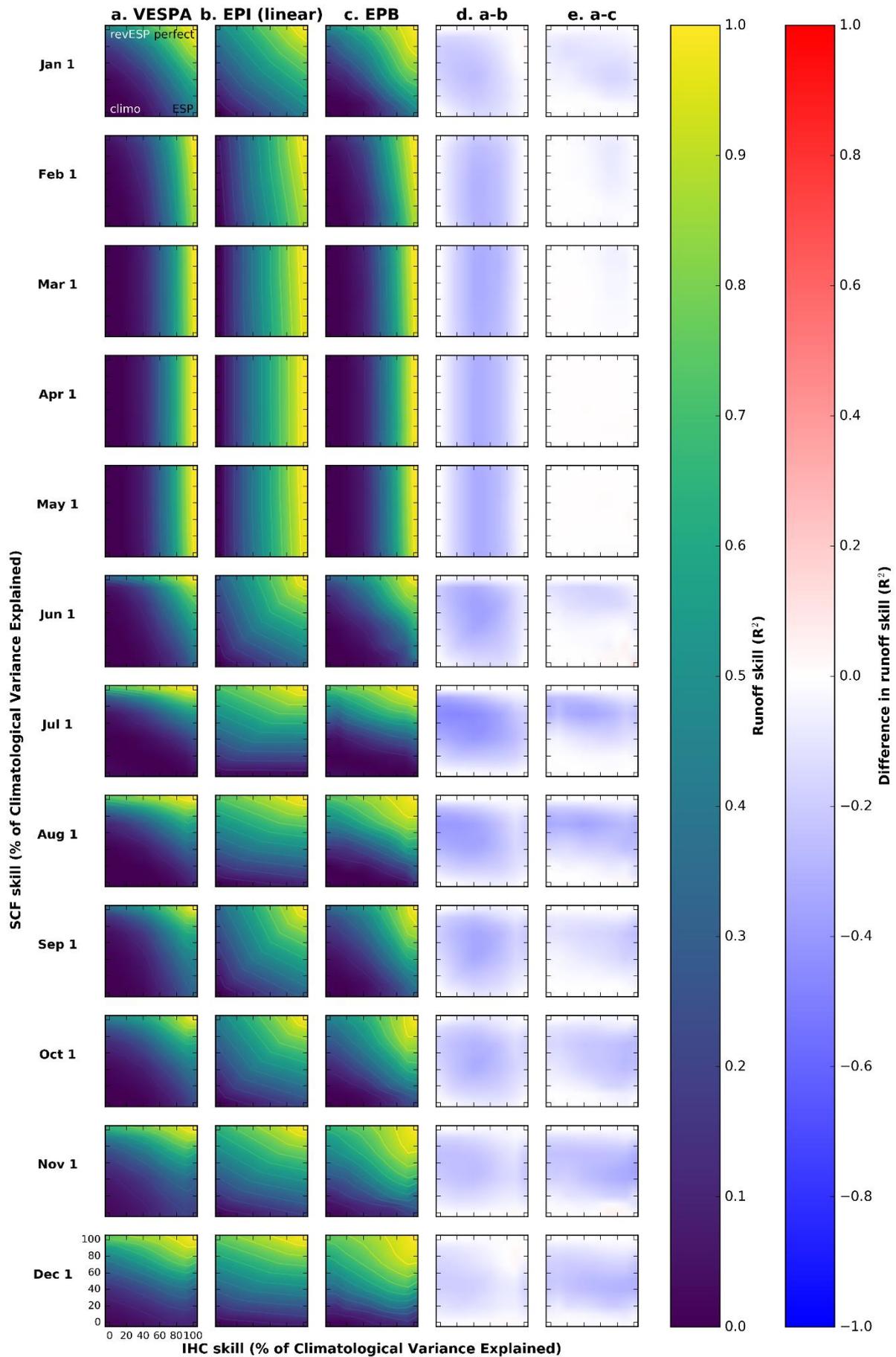
Supplementary Figure 1 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for Merced River (CA; USGS gauge 011266500), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



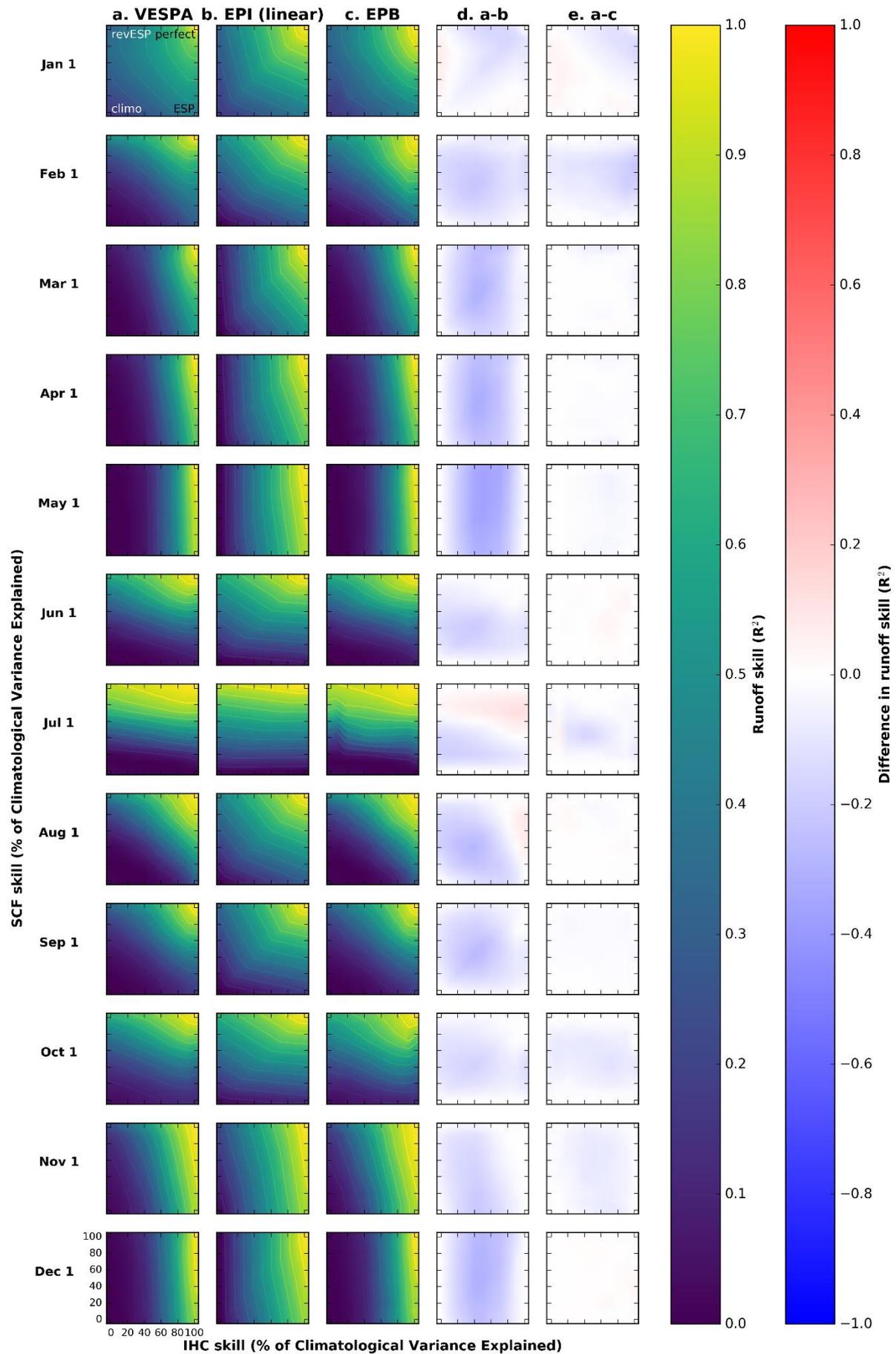
Supplementary Figure 2 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R² of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for Long Tom River (OR; USGS gauge 014166500), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



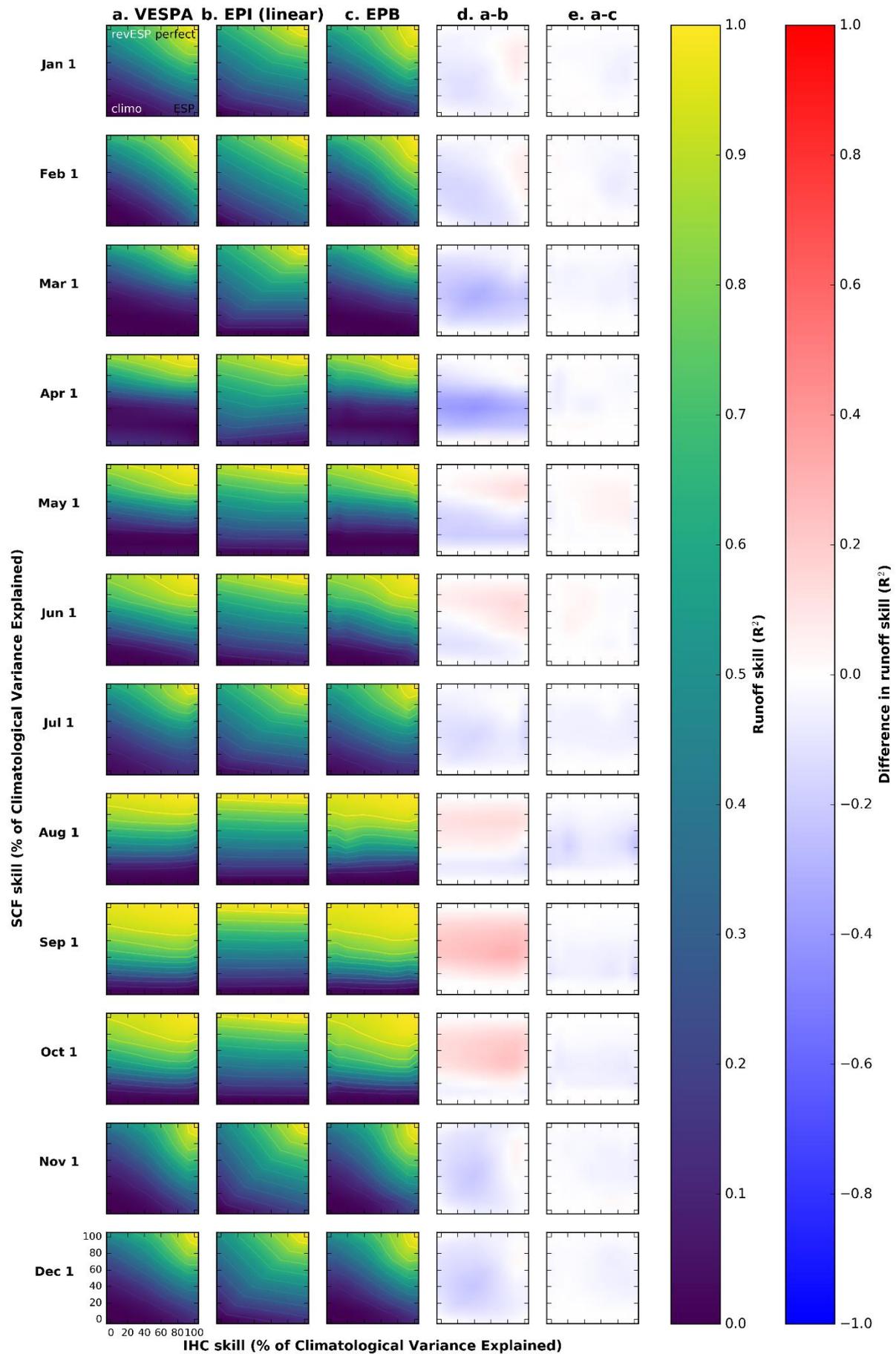
Supplementary Figure 3 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R² of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for Beaver River (UT; USGS gauge 010234500), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



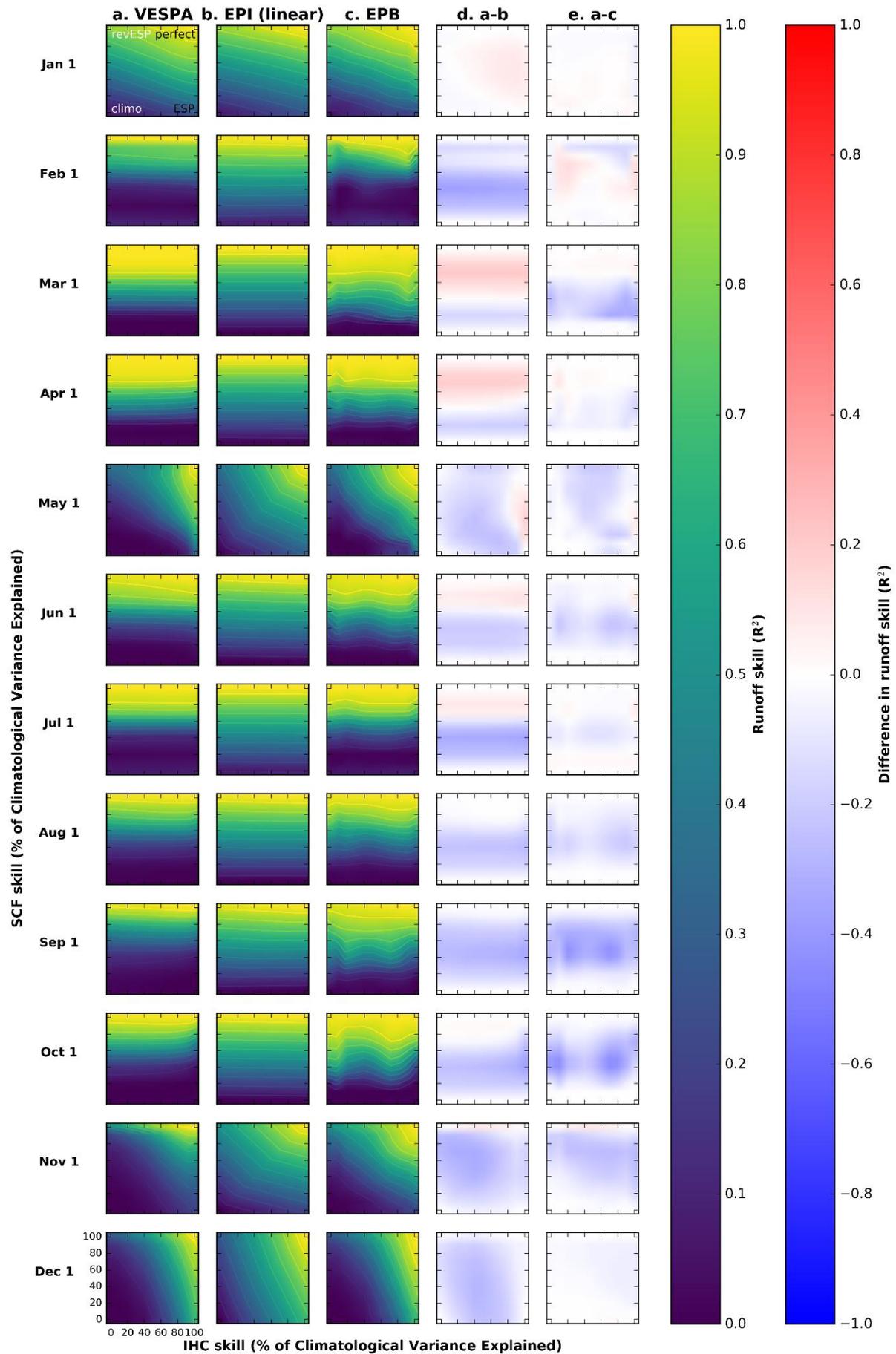
Supplementary Figure 4 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for Gila River (NM; USGS gauge 009430500), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



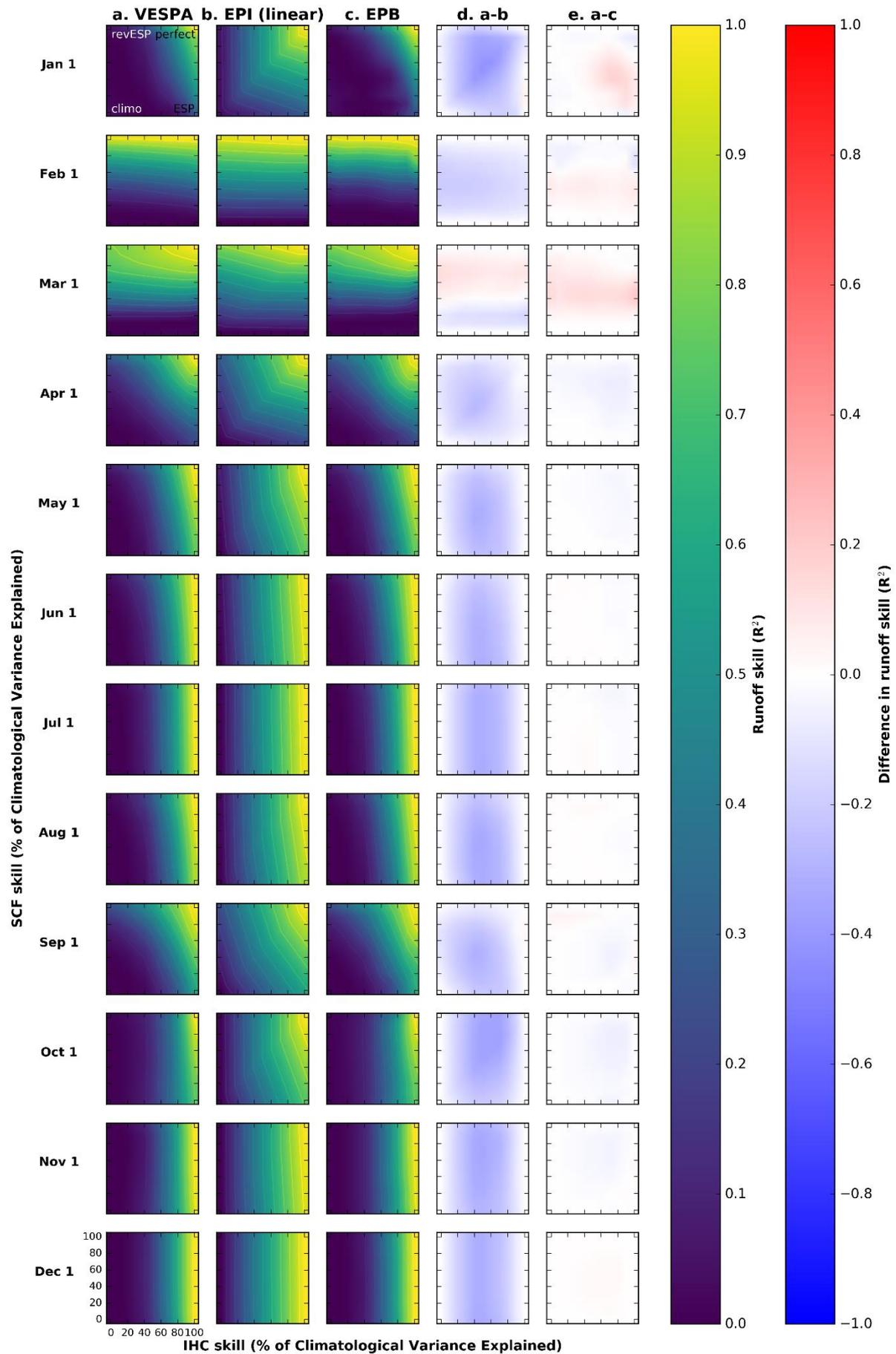
Supplementary Figure 5 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for Gallinas Creek (NM; USGS gauge 008380500), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



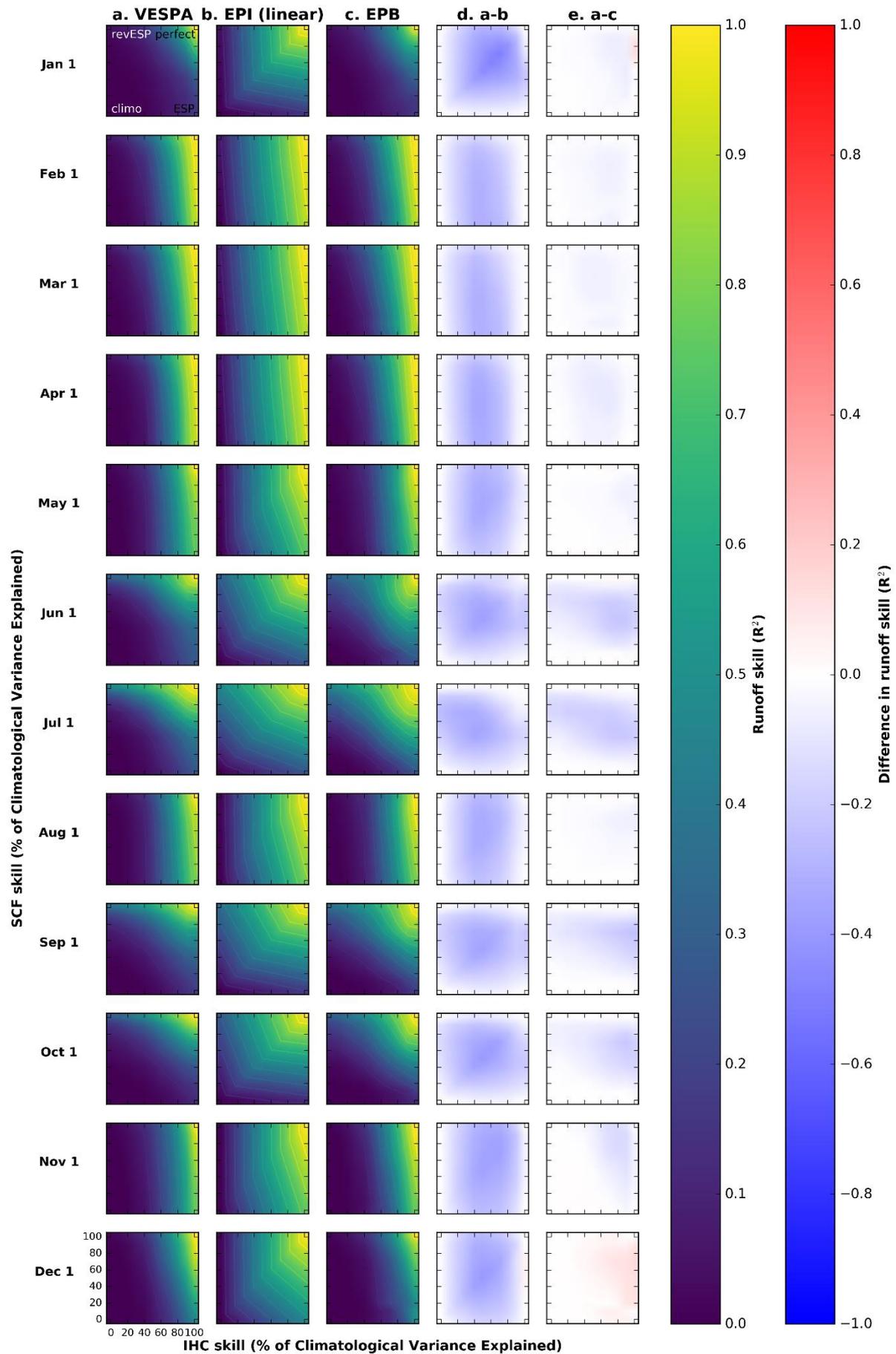
Supplementary Figure 6 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for E Fk San Jacinto River (TX; USGS gauge 008070200), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



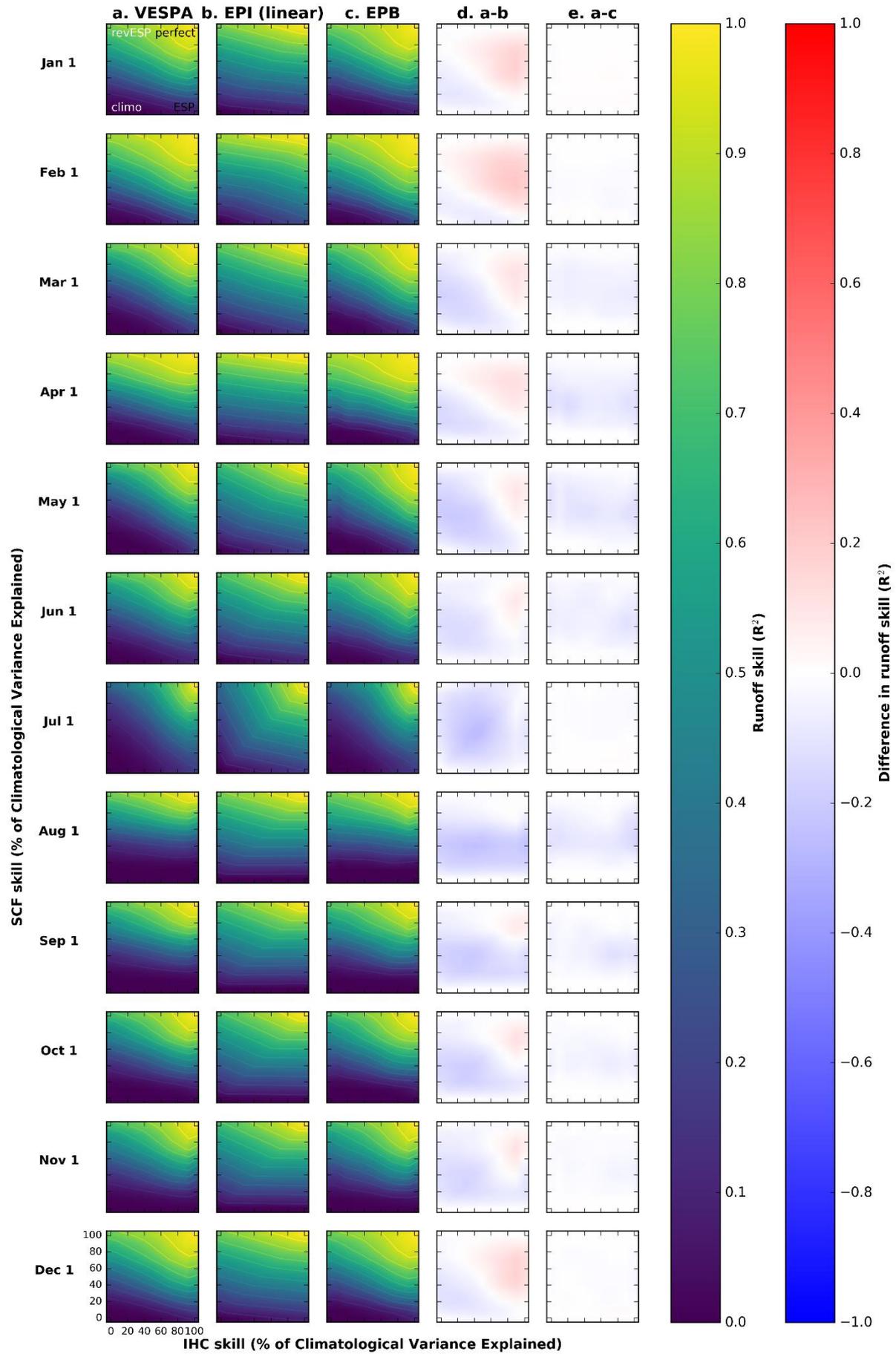
Supplementary Figure 7 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for UTE Creek (NM; USGS gauge 007226500), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



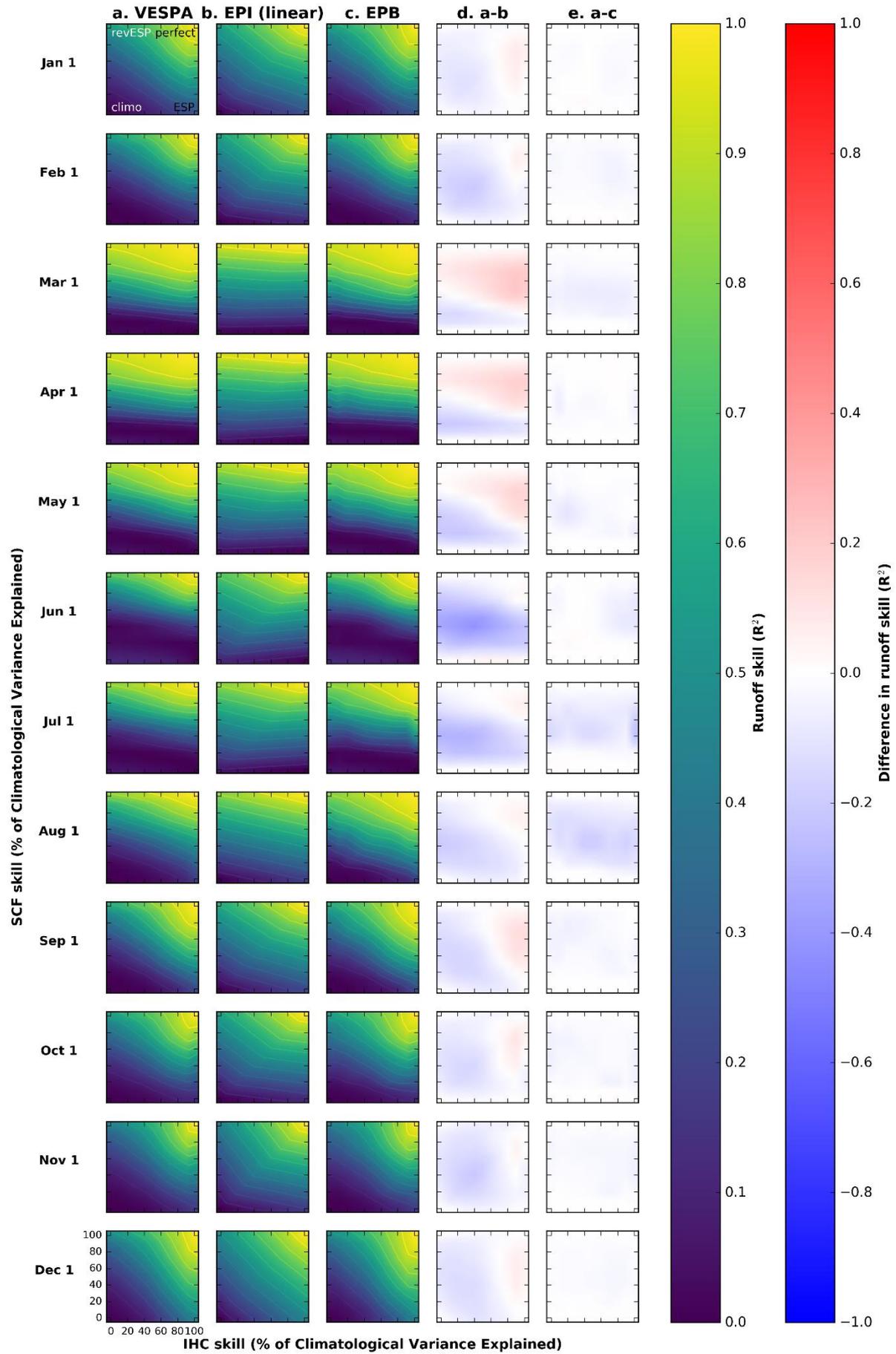
Supplementary Figure 8 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for Bull Lake Creek (WY; USGS gauge 006224000), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



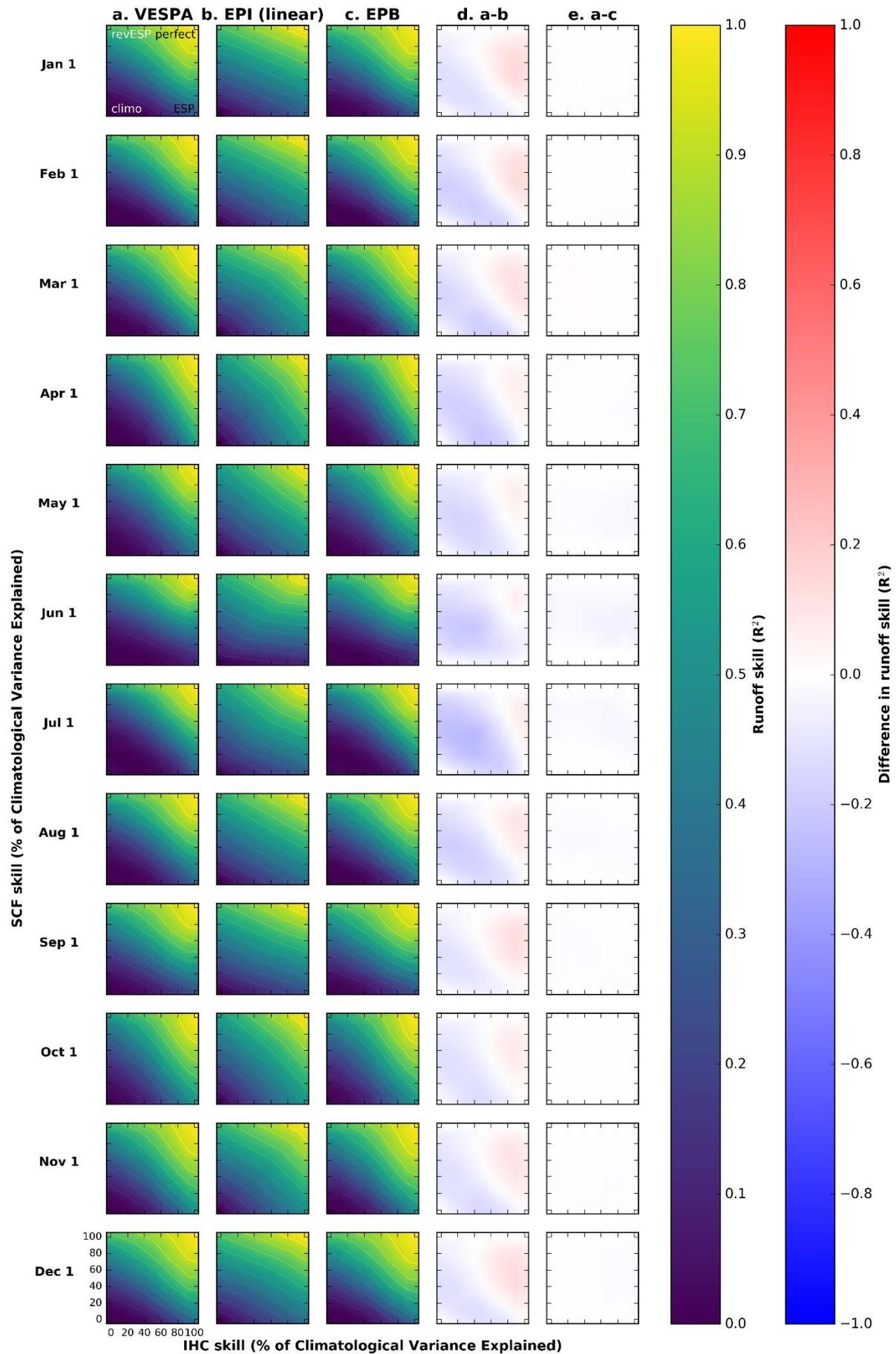
Supplementary Figure 9 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for Sheyenne River (ND; USGS gauge 005057000), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



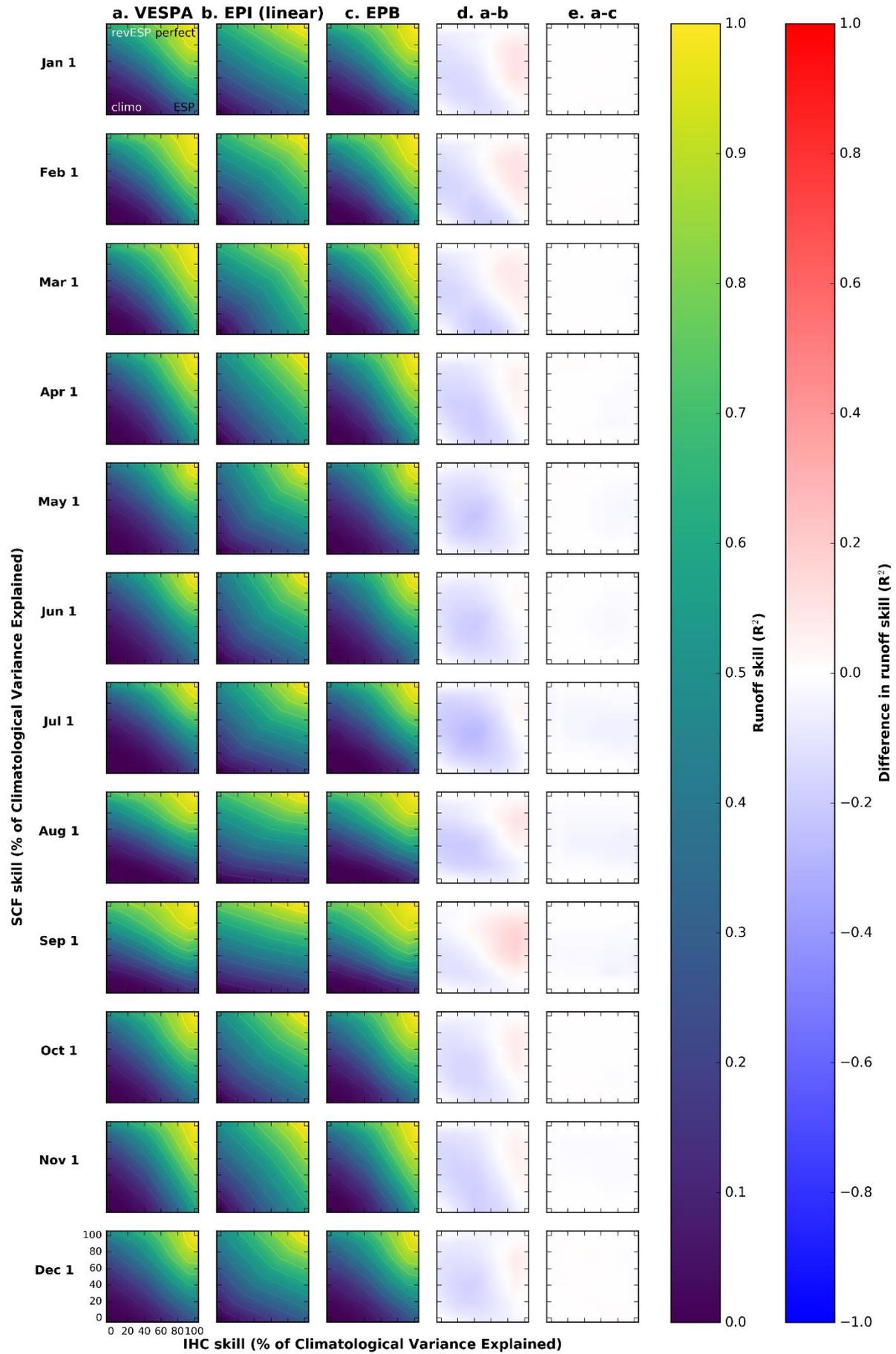
Supplementary Figure 10 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for Tickfaw River (LA; USGS gauge 007376000), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



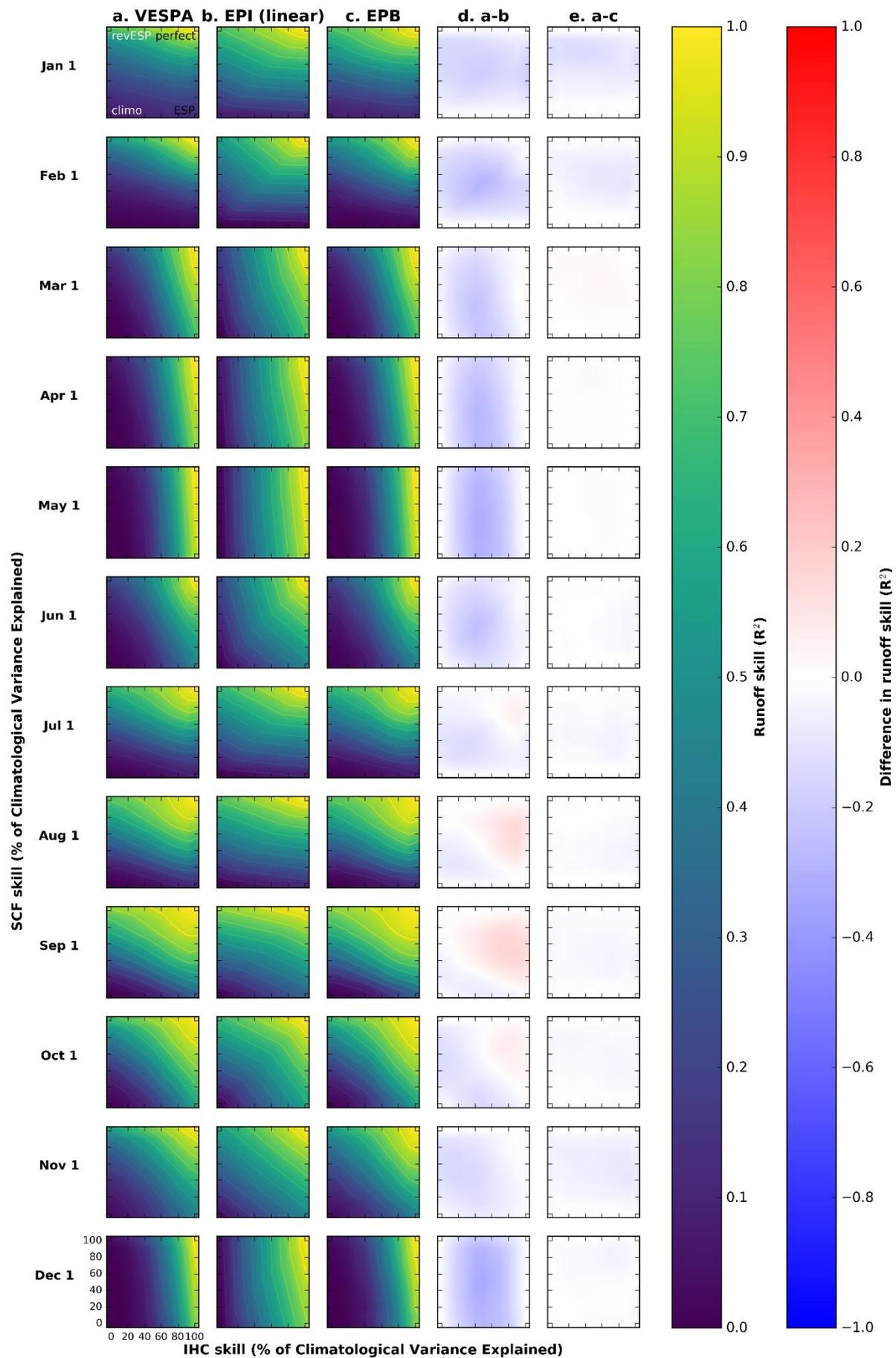
Supplementary Figure 11 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R² of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for East Fork Shoal Creek (IL; USGS gauge 005593900), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



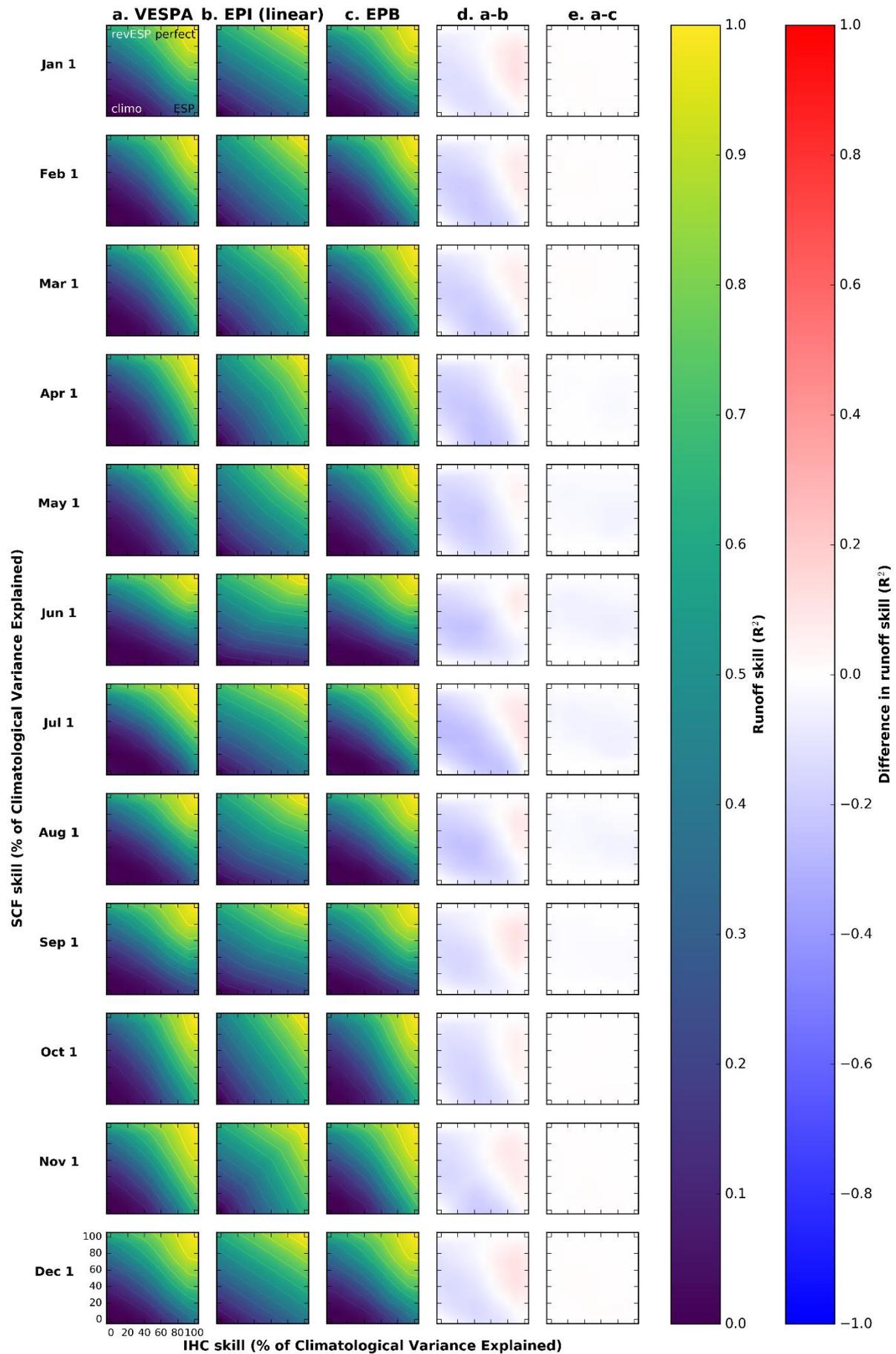
Supplementary Figure 12 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for Nantahala River (NC; USGS gauge 003504000), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



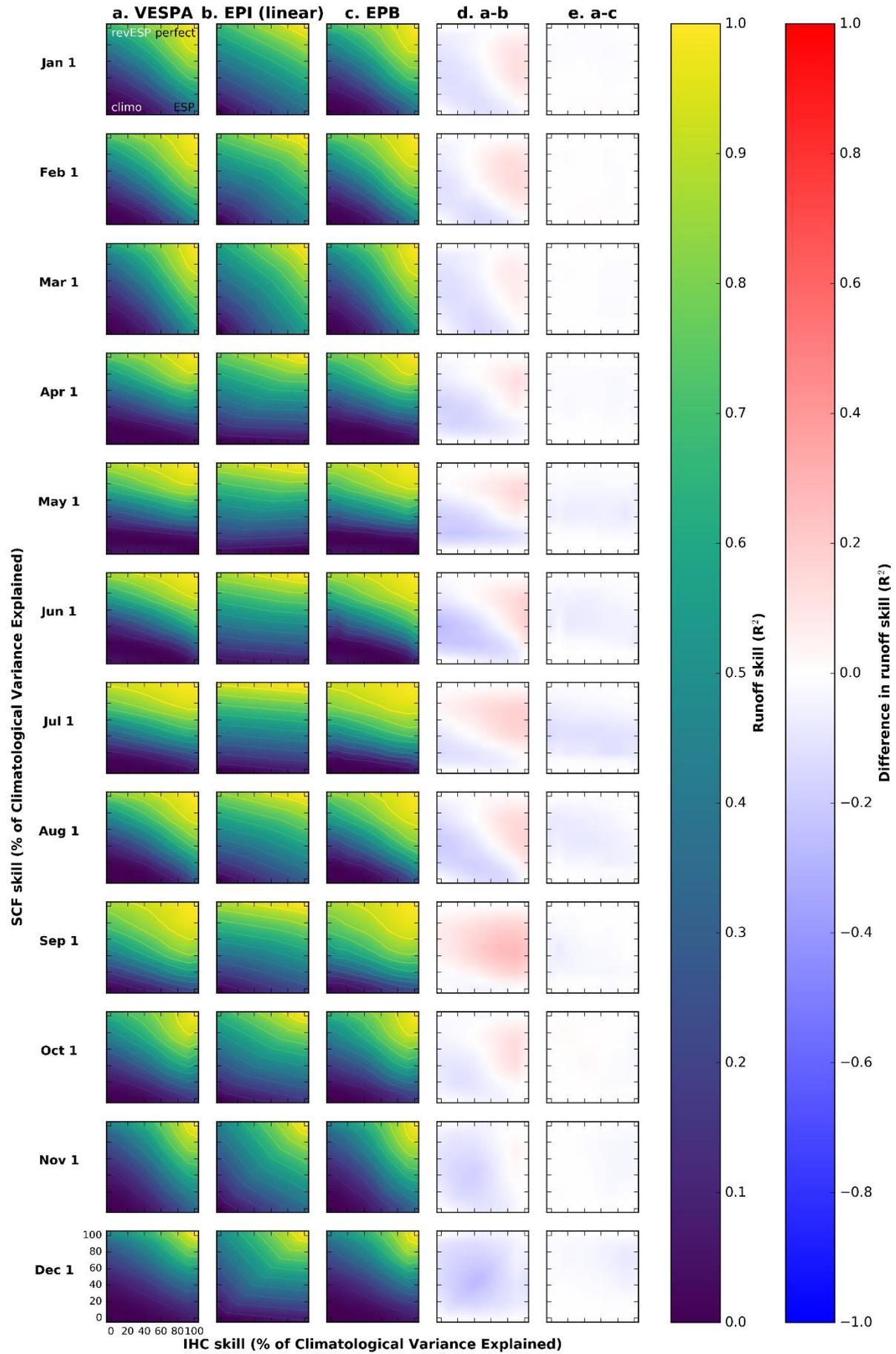
Supplementary Figure 13 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for New River (VA; USGS gauge 003164000), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



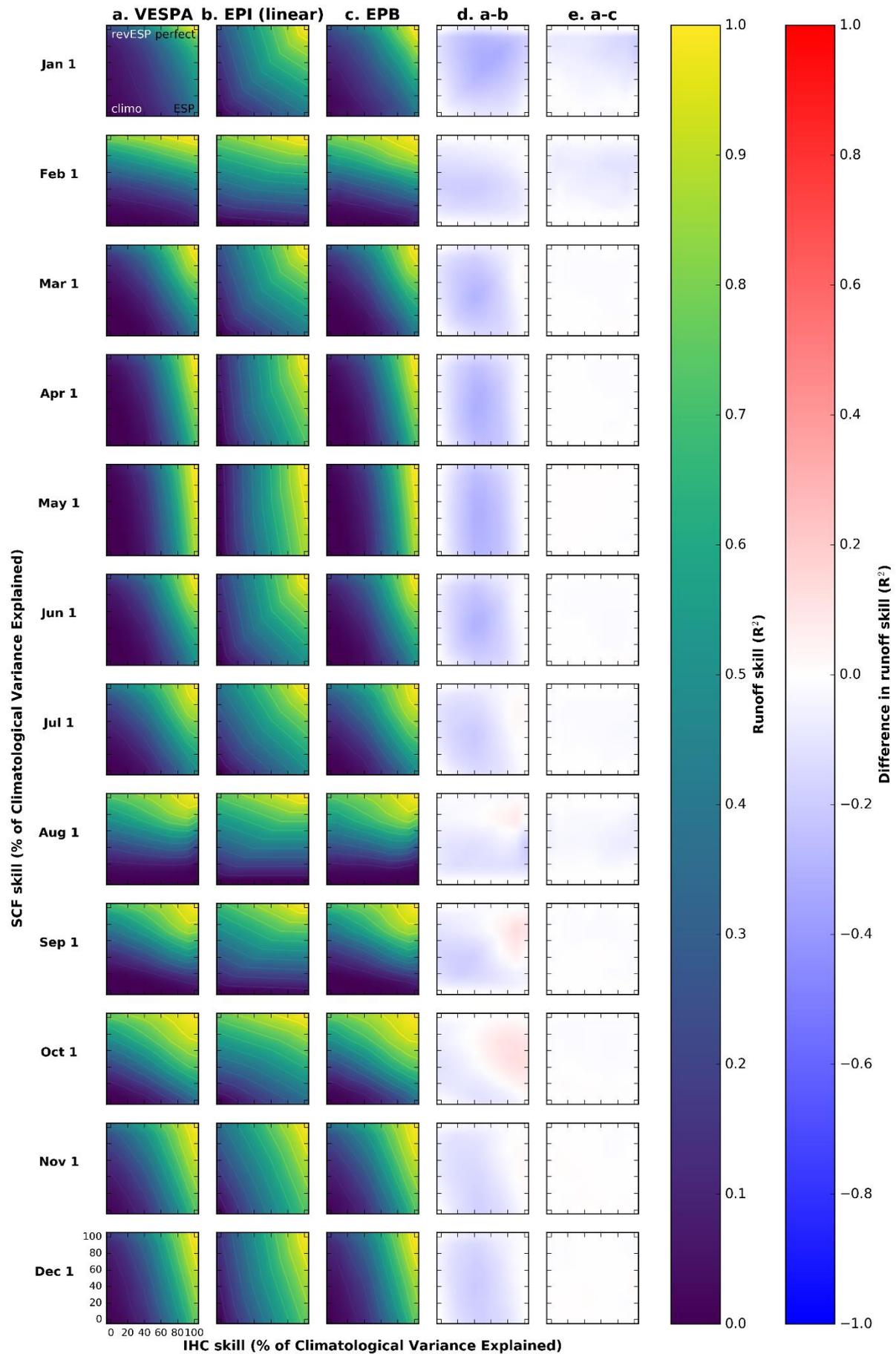
Supplementary Figure 14 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for Middle Branch Escanaba River (MI; USGS gauge 004057800), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



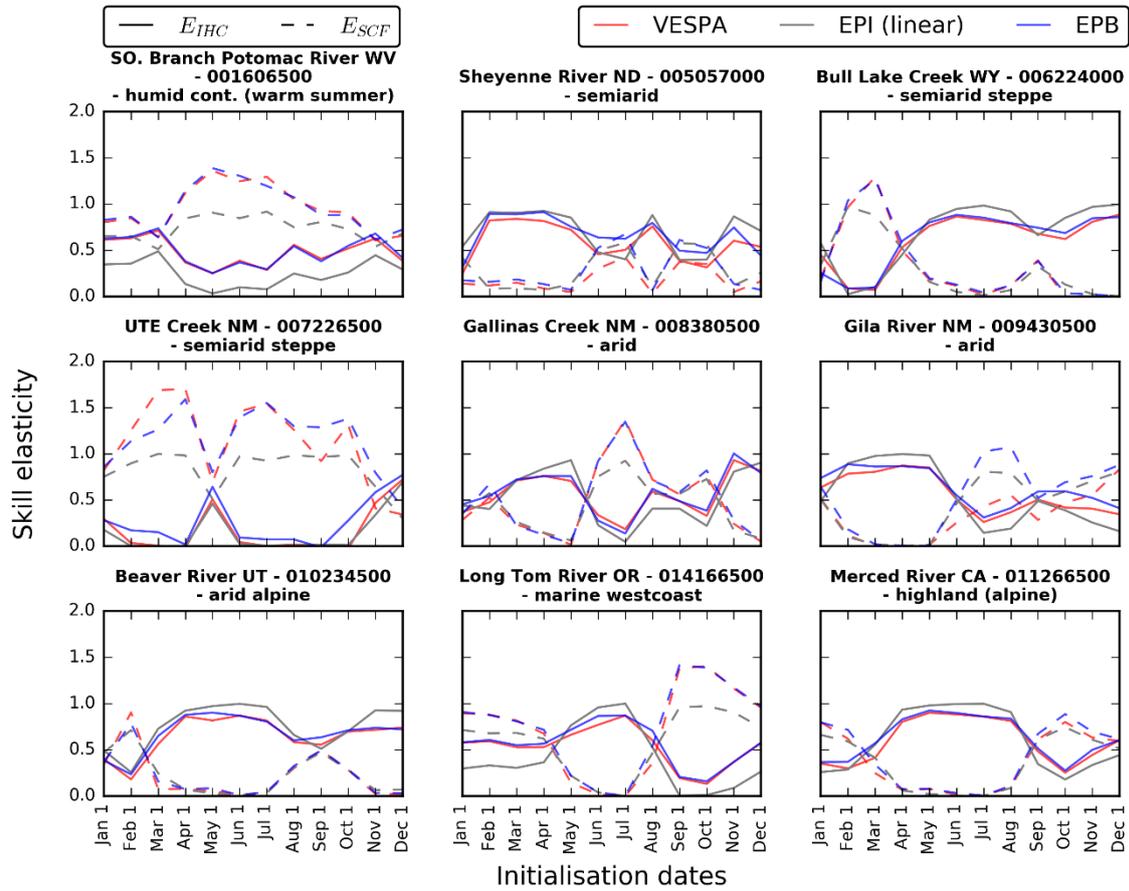
Supplementary Figure 15 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for Chattooga River (GA; USGS gauge 002177000), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



Supplementary Figure 16 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R² of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for SO. Branch Potomac River (WV; USGS gauge 001606500), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



Supplementary Figure 17 Skill surface plots obtained for (a) the VESPA, (b) the linear EPI, and (c) the EPB methods. The skill is calculated from the R2 of the 3-month streamflow forecast ensemble means against the perfect forecasts, for hindcasts produced from 1981 to 2010 for Fish River (ME; USGS gauge 001013500), with forecast initializations on the first day of each month. Differences between the skill surface plots obtained for (d) the VESPA and linear EPI methods and (e) the VESPA and EPB methods are also shown.



Supplementary Figure 18 Streamflow forecast skill elasticities for the IHCs (i.e., E_{IHC} , solid line) and the SCFs (i.e., E_{SCF} , dashed line), calculated across a quadrant situated within the 3-month streamflow forecast skill surface plots for the VESPA (red), the linear EPI method (gray), and the EPB method [blue; using Eqs. (4) and (5)]. Each plot shows the evolution of the IHC and SCF skill elasticities with the initialization date for a given catchment. The climatological regions of the catchments are indicated in the plots' headings. The skill surface plots from which these skill elasticities were calculated are presented in Fig. 4 and Figs. S1–S17.

A7: The sensitivity of sub-seasonal to seasonal streamflow forecasts to meteorological forcing quality, modelled hydrology and the initial hydrological conditions

This paper presents a co-author contribution arising through collaboration during this PhD, summarised in Chapter 4, Sect. 4.3, and has the following reference:

Arnal, L. et al.: “IMPRES D4.2 - The sensitivity of sub-seasonal to seasonal streamflow forecasts to meteorological forcing quality, modelled hydrology and the initial hydrological conditions”. Deliverable of EU H2020 project “IMPRES – Improving predictions and management of hydrological extremes” (contract n° 641811), 2017a*

* This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 641811.



IMPROVING PREDICTIONS AND MANAGEMENT OF HYDROLOGICAL EXTREMES



 @imprex_eu



Funded by
the Horizon 2020
Framework Programme
of the European Union

Grant agreement No. 641811



Deliverable	The sensitivity of sub-seasonal to seasonal streamflow forecasts to meteorological forcing quality, modelled hydrology and the initial hydrological conditions
Related Work Package:	WP4: Improved predictability of hydrological extremes
Deliverable lead:	University of Reading
Author(s):	Louise Arnal, Hannah L. Cloke, Linus Magnusson, Bastian Klein, Dennis Meissner, Alberto de Tomas, Johannes Hunink, Ilias Pechlivanidis, Louise Crochemore, Sara Suarez, Abel Solera, Joaquin Andreu, Jeff Knight, Felicity Liggins, Albrecht Weerts, Maria-Helena Ramos, Guillaume Thirel
Contact for queries	Louise.arnal@ecmwf.int
Grant Agreement Number:	n° 641811
Instrument:	HORIZON 2020
Start date of the project:	01.10.2015
Duration of the project:	48 months
Website:	www.IMPRES.eu
Abstract	This report presents the sensitivity of seasonal discharge forecasting to meteorological forcing quality, modelled hydrology and initial hydrological conditions and the dependency of the skill to the basin characteristics. The analyses are focused on key sectoral applications of IMPRES.

Dissemination level of this document



X	PU	Public
	PP	Restricted to other programme participants (including the Commission Services)
	RE	Restricted to a group specified by the consortium (including the European Commission Services)
	CO	Confidential, only for members of the consortium (including the European Commission Services)

Versioning and Contribution History

Version	Date	Modified by	Modification reasons
v.01	02/02/2017	L. Arnal	First draft submission
	08/02/2017		Review by Bart vd Hurk
v.02	31/03/2017	L. Arnal	Revised version submission
	04/04/2017		Review by Bart vd Hurk
v.03	07/04/2017	L. Arnal	Modification of the introduction



Table of Contents

Executive summary	6
Glossary.....	10
1 Introduction	11
2 Sub-seasonal to seasonal streamflow forecasting: background, applications and limitations.....	13
2.1 An overview of sub-seasonal to seasonal streamflow forecasting.....	13
2.2 Sectoral applications.....	14
2.2.1 Flood forecasting - University of Reading.....	14
2.2.2 Navigation - BfG.....	15
2.2.3 Agriculture - FutureWater	17
2.2.4 Hydropower - SMHI.....	18
2.2.5 Reservoir management - UPV.....	18
2.2.6 Flood & low flow forecasting - Deltares.....	20
2.3 Sensitivity analyses as a tool to diagnose seasonal streamflow forecasting uncertainties.....	21
2.4 Comparative analysis in large sample hydrology.....	22
3 Data and methods.....	24
3.1 The forecasting systems.....	24
3.2 The forecasting systems intercomparison	29
3.3 The EPB sensitivity analysis.....	38
3.4 Seasonal hydrological forecasts - Clustering of the skill	41
4 Results	44
4.1 The forecasting systems intercomparison	44
4.1.1 Central European Rivers.....	44
4.1.2 The Thames River Basin	56
4.1.3 The Segura and Tagus River Basins.....	63
4.1.4 The Jucar River Basin.....	68
4.1.5 Swedish Rivers.....	77
4.2 The EPB sensitivity analysis.....	81



4.3	Comparative analysis.....	87
5	Lessons learnt	91
6	References	95
7	Annex A – Tabulated overview on hydrological model features	98



Executive summary

Information about streamflow for the coming months (sub-seasonal time-scale) to seasons is needed for decision-making in many sectors of society. Examples are in a reservoir management context, for applications such as hydropower generation, water allocation for drinking water and agriculture, navigation, flood and drought mitigation. Here, sub-seasonal and seasonal forecasts can be a valuable tool. Compared to short range forecasts, these forecasts allow for an increased operational margin for early warning and maximised benefits. However, the potential skill on the longer time-scales are limited due to a low inherent predictability (of the atmosphere and hydrosphere) and limited quality of models and observations.

In order to meet these needs and tackle those challenges, IMPREX will (1) analyse the current skill of state-of-the-art sub-seasonal to seasonal streamflow forecasts over Europe and (2) improve their capabilities, with a focus on extreme events (i.e., high and low flows) and variables, aggregation periods, seasons and lead times of interest to the users of the forecasts involved in IMPREX (which cover the water sectors mentioned above).

This deliverable consists of three parts. The first part is a technical intercomparison of the performance of five different sub-seasonal to seasonal streamflow forecasting systems, operated by partners of IMPREX: ECMWF, SMHI, FW, BfG, UPV and Deltares. This is to be done for key locations in Europe, selected based on the case studies of the project. They include Central European River and Swedish River stations and the Thames, Segura, Tagus, and Jucar River basins. The forecasting systems investigated in this deliverable all use the same meteorological forecasting system, ECMWF's System 4 (with or without applying a bias correction method to the latter) and a variety of hydrological models. This intercomparison therefore enables us to identify the contribution of hydrological model structure and the presence of a bias correction of the meteorological forecasts to the streamflow forecasting skill on sub-seasonal to seasonal time scales and for an array of diverse locations, seasons and extreme events in Europe. This first part has revealed several major differences between the seasonal streamflow forecasting systems and their



impacts on the relevant water sectors. Notably, the BfG seasonal streamflow forecasts tend to underestimate the observed streamflow for the Central European River stations, which could be a problem for the navigation sector, dependent on accurate low flow forecasts in the summer. The ECMWF seasonal streamflow forecasts appear to systematically overestimate the spring flow and to underestimate the winter flow for the Central European River stations and the Thames River basin. The latter could be an issue for the flood protection sector as the forecasts would be prone to missing or underestimating the magnitude of flood events. The ECMWF forecasts are however very accurate for summer flow forecasting in the Segura and Tagus River Basins. This could be highly beneficial for the agricultural sector in this region. Both the ECMWF and the SMHI seasonal streamflow forecasts underestimate the May flow for Swedish River stations, which could be a challenge for the hydropower generation sector, relying on accurate spring flow predictions. The SMHI forecasts seem to generally overestimate the winter flow for the Central European River stations and the Thames River basin, which could once again be a problem for the flood protection sector as the forecasts would potentially lead to false alarms. The FW seasonal streamflow forecasts tend to overestimate largely the early spring-spring flow for the Tagus River Basin and to underestimate slightly the flow during all year for the Segura River Basin. Both biases could be challenging for the agricultural sector.

This first technical intercomparison forms a benchmark, to which improved systems from other IMPREX tasks can be compared. The intercomparison can moreover be enriched during the course of IMPREX, with more stations and scores.

The second part of the deliverable consists in a sensitivity analysis, specifically designed to diagnose the relative contributions of errors in the initial hydrological conditions (IHC) and in the meteorological forecasts (MF, sometimes called seasonal climate forecasts [SCF]) on sub-seasonal to seasonal streamflow forecasting uncertainty. This sensitivity analysis was carried out using the ECMWF and the BfG seasonal streamflow forecasts and highlighted several significant results. The analysis indicates that improving the IHC would yield a higher improvement of the



seasonal streamflow forecasts for the first month of lead time, after which the SCF become rapidly more influential on the skill of the streamflow forecasts. This signal is however contrasted in space and time, highlighting geographical and seasonal variations of the flow generating mechanisms in Europe. For example for streamflow forecasts made in the summer (May-July) and with one month of lead time, there appears to be a larger number of regions in Europe where the IHC dominate the quality of the streamflow forecasts, compared to forecasts made in the winter with the same lead time. This is probably due to the lower rainfall over Europe during the summer months, leading to groundwater dominated streamflow. For most leeward regions in Scandinavia, the IHC dominate the quality of the forecasts made in the winter, with one to three months of lead time. This is potentially due to precipitation falling as snow in the winter, leading to groundwater dominated streamflow in the winter and snowmelt driven flow in the following spring. For most regions of the Iberian Peninsula, the IHC seem relatively more important for streamflow forecasts made in the summer (June-September), with one to three months of lead time. This is probably due to groundwater dominated streamflow in the summer in those regions and a land surface memory spanning several months. Over the eastern part of central Europe, streamflow forecasts made in the spring seem to be more sensitive to the IHC, which might translate snowmelt driven spring flow.

The third part of this deliverable is devoted to the identification of the key drivers (beyond IHC and SCF) that control and influence the hydrological forecasting skill. For this, an alternative sensitivity analysis was designed based on the results from about 35000 European basins, which allows linking the seasonal hydrological forecasting skill (from the SMHI forecasts) to the regional physiographic-hydro-climatic characteristics. This analysis showed that seasonal hydrological forecasting skill is mainly dependent on the basin's hydrological regime. Other factors, such as the elevation and the remaining bias in temperature, were also identified to be important aspects (i.e., dependence of response of mountainous basins to temperature). Another significant result is that seasonal hydrological forecasting skill



seems to be limited for relatively flashy basins, experiencing strong flow dynamics over the year (i.e., less memory in the system).

The results of this deliverable will guide future research in IMPREX, indicating where improvements should be made in the forecasting chain (improvements to the IHC, the SCF) in order to improve the seasonal streamflow forecasts over Europe.



Glossary

Bias correction: Process aiming at removing systematic errors in the output of a model. Methods include: linear scaling, distribution-based Scaling, quantile mapping, to cite a few.

Lead time: The time between the initiation and completion of a forecast.

Discharge: River discharge is the volume of water flowing through a river channel at any given point and is measured in cubic metres per second (m³/s).

Target month or target season: The season or month for which the forecast is made.

Forecast quality: How well a forecast compares against a corresponding observation of what actually occurred, or some good estimate of the true outcome.

Sensitivity analysis: The study of how the uncertainty in the output of a model or system can be apportioned to different sources of uncertainty in its inputs.

Skill elasticity: A measure of the sensitivity of the seasonal discharge forecasting skill to changes in the skill of its two main predictability sources: the initial hydrological conditions or the seasonal climate forcing.

Initial hydrological conditions (IHC): The hydrological states (soil moisture, snow cover, water already in the river, among others) at or close to the start of the forecast run.

Seasonal climate forcing (SCF): The seasonal meteorological forecast used as input to a hydrological model.



1 Introduction

Information about streamflow during the coming month (sub-seasonal time-scale) and season is needed for decision-making in many sectors. Examples are in a reservoir management context, for applications such as hydropower generation, water allocation for drinking water and agriculture, navigation, flood and drought mitigation. Here, sub-seasonal and seasonal forecast can be a valuable tool. Compared to short range forecasts, these forecasts allow for an increased operational margin for early warning and maximised benefits. However, the potential skill on the longer time-scales are limited due to a low inherent predictability and limited quality of models and observations.

In order to meet these needs, IMPREX will (1) analyse the current skill of state-of-the-art sub-seasonal to seasonal streamflow forecasts over Europe and (2) improve their capabilities, with a focus on extreme events (i.e., high and low flows) and variables, aggregation periods, seasons and lead times of interest to the users of the forecasts involved in IMPREX (which cover the water sectors mentioned above).

This deliverable consists of three parts. The first part will use the verification scoreboard designed in WP4 (deliverable 4.1) to analyse and compare the skill of multiple sub-seasonal to seasonal streamflow forecasting systems, operated by partners of IMPREX. This will be done for key locations selected based on the case studies of the project. Since the forecasting systems investigated here all use the same meteorological forecasting system (with or without applying a bias correction method) and a variety of hydrological models, this work will enable us to identify the contribution of (hydrological) model structure and the presence of a bias correction of the seasonal meteorological forecasts to the streamflow forecasting skill on sub-seasonal to seasonal time scales and for an array of diverse locations, seasons and extreme events in Europe. This first technical intercomparison will form a benchmark, to which more stations, scores and improved systems from other IMPREX tasks can be added and compared. The main aim of this part is to highlight major differences and similarities between the performance of the seasonal



streamflow forecasting systems and the potential impacts of those performances on the sectoral applications at stake in the case study areas. This technical part will also inform later tasks of the IMPREX project, such as the multi-modelling and the data assimilation exercises.

The second part of this deliverable will inform future IMPREX work through a sensitivity analysis, specifically designed to diagnose the relative contributions of initial hydrological conditions (IHC) and errors of the meteorological forecast (MF, sometimes called seasonal climate forecast [SCF]) on sub-seasonal to seasonal streamflow forecasting quality. This will indicate the potential achievable improvements in sub-seasonal to seasonal streamflow forecasting skill through the improvement in either one of the two error (or predictability) sources.

The third part of this deliverable is devoted to the identification of the key drivers (beyond IHC and SCF) that control and influence the hydrological forecasting skill. For this an alternative sensitivity analysis was designed based on the results from about 35000 European basins, which allows linking the skill to the regional physiographic-hydro-climatic characteristics.

The aim of this work is to produce a 'hydrological sensitivity chart', providing information about the state-of-the-art in terms of sub-seasonal to seasonal streamflow forecasting, as well as about potential targeted improvements on which IMPREX should focus. An overview of sub-seasonal to seasonal streamflow forecasting and the use of sensitivity analyses to diagnose its uncertainties are given in Section 2. Section 3 introduces the methodology, with an overview of the forecasting systems, the data and the methods used for the analyses. The results are subsequently presented in Section 4 and finally discussed in Section 5. Section 5 additionally states the lessons learnt and recommendations for future work.



2 Sub-seasonal to seasonal streamflow forecasting: background, applications and limitations

2.1 An overview of sub-seasonal to seasonal streamflow forecasting

The first seasonal streamflow forecasting methods were statistical methods, regression-based, using antecedent hydrological conditions (i.e., snowpack measurements, soil moisture, among others) to give an indication of the streamflow for the following months (Church, 1935; Wood and Lettenmaier, 2006). With the understanding of hydrological processes and the advances in computer technologies, the first numerical hydrological models were created (Helms et al. 2008). In the 1970s, one of the first dynamical forecasting system was constructed using a hydrological model, initialising it with observed hydrological conditions (IHC) and forcing it with historical time series of observed precipitation and temperature from all the previous years of recorded meteorological observations. This method was introduced by the National Weather Service (NWS) in the United States and was termed the Extended Streamflow Prediction (ESP) system (Twedt et al., 1974; Day, 1985). The ESP nowadays stands for Ensemble Streamflow Prediction and describes the same forecasting process.

Despite its strength, the ESP is limited by the fact that it is based on the assumption that the historical weather can give an accurate indication of the future weather. In the 1950s, the use of seasonal meteorological forecasts for seasonal streamflow forecasting for water management was first investigated but its skill was judged too poor for operational purposes (Pagano and Garen, 2005). The 1970s were a milestone for seasonal meteorological forecasting, due to the understanding of atmosphere-ocean-land interactions and the importance of teleconnections forecasting on seasonal time scales (such as the ENSO, NAO, etc; Pagano and Garen, 2005). It is however not until the late 1990s that seasonal meteorological forecasts were used for operational purposes, as a result of the very strong El-Niño of 1997-98 (Pagano and Garen, 2005).



Statistical forecasting techniques are still widely used, sometimes based on complex regression methods, harnessing the teleconnection indicators (Wang et al, 2011). It is only recently that dynamical seasonal streamflow forecasting (based on forcing a hydrological model with meteorological seasonal forecasts to obtain seasonal hydrological forecasts) has become a real potential to surpass statistical seasonal streamflow forecast skill (Easy et al. 2006). Statistical-dynamical hybrid systems also exist, for instance the use of teleconnection indicators to resample the historical observed meteorological years, removing anti-analogues, to force a hydrological model (Schaake, 1978; Pagano and Garen, 2005; Bierkens and van Beek, 2009).

2.2 Sectoral applications

Sub-seasonal to seasonal streamflow forecasts are valuable for many applications of the water sector, including reservoir management for hydropower generation and water allocation for drinking water and agriculture, navigation and flood and drought mitigation. These applications are diverse in terms of their needs and operational use of the forecasts. For example, the flood protection sector is more vulnerable to high river flow, while the navigation, agriculture, hydropower and reservoir management sectors are more vulnerable to low flows. Additionally, the flood protection sector requires accuracy in the timing and the intensity of an event, while the hydropower sector requires information on the flow accumulations for the spring. The various sectors and their individual needs and current operational practices are described below.

2.2.1 Flood forecasting - University of Reading

Flood forecasting is currently done successfully at short to medium time scales (up to a month ahead). Beyond this lead time, the capacity of the forecasts to indicate the potential for an extreme event to happen is still limited, let alone the exact day or even week when this event might happen and the exact location of this event. This is the main reason for which the Environment Agency (EA) does not currently use any sub-seasonal to seasonal forecasts for their decision-making. The main need for decision-making in a flood context is the probability of an event happening, an indication of how extreme the event will be and the estimate date of



the event. The EA bases their decisions on a very low threshold (i.e., allowing a high false alarm ratio), as the loss for not taking any action is much larger than the cost of taking action. Their strategies could be categorised as risk-averse, as the consequences of a false alarm are lower than for a miss. There is nonetheless the potential to integrate sub-seasonal to seasonal information in their current system. Information for the longer time scale could give an indication of the trend in discharge for the following months and flag areas to watch for these coming months, following a "ready-set-go" approach (Goddard et al. 2014). The EA has expressed interest in this kind of information.

2.2.2 Navigation - BfG

Monthly to seasonal forecasts are required for Inland Waterway Transport (IWT) for the the medium- to long-term planning and enhancement of the water bound logistic chain (stock management, adjustment of the industrial production chain, modal split planning). Information about the future evolution of flow and water levels in the large rivers is especially required by the stakeholders before and within the typical low flow seasons when transport capacity on rivers is limited. The required forecast lead time depends on the specific waterway user and the decisions to be taken. It ranges from weeks, for example to shift cargo from shipping to another means of transportation, to months, to adapt the fleet / usable transport capacity (see Klein and Meissner, 2016). Despite the great demand and interest of the IWT sector, no operational forecasts with lead times exceeding 8 days are available at the moment for the Rivers Rhine (max. lead time 4 days), Elbe (max. published lead time 2- 8 days depending on the gauge) and Upper Danube (max. published lead time 2-4 days depending on the gauge), mainly due to the large uncertainties and the limited skill on monthly and seasonal time scales.

In order to provide stakeholders with monthly to seasonal forecast information, a prototype is being developed in the context of IMPREX. To model the water balance and the flow in rivers the hydrological model LARSIM-ME is applied. The hydrological model was set up for the large rivers in Germany including their



international parts (model acronym LARSIM-M(iddle)E(urope)) and covers the catchments of the River Rhine, River Elbe, River Weser/Ems, River Odra and River Danube up to gauge Nagymaros in Hungary. The total catchment size simulated by the model is approx. 800 000 km². The spatial resolution is 5km x 5 km. As meteorological forcings resampled observed climatology (ESP) and seasonal forecasts from ECMWF Seasonal Forecast System 4 are used. 2-m temperature of the past 24 hours and daily total precipitation of System4 are interpolated to a common 50km x 50km grid (multiple of the 5km x 5km raster). Both variables were bias corrected on the 50km x 50km grid using linear scaling with the meteorological observation dataset set used for the baseline simulation (also aggregated to the 50km x 50km grid as reference data). As seasonal forecasts tend to drift towards their own model climate with increasing lead time, giving rise to model bias, separate bias correction factors have been estimated for each forecast initialisation date (starting on the first day of each calendar month) and monthly lead times (first month, second month, etc, to sixth month). In total 12 x 6 = 72 scaling factors for precipitation and 72 additive terms for temperature were calculated for each 50km x 50km raster to correct the model drift of ECMWF's System 4. In the next step temperature and precipitation are downscaled to the 5km x 5km model grid. In future versions of the navigation related seasonal forecasting prototype NavSEAS-ME seasonal forecasts from GloSea5 from UK Metoffice will be included in addition.

To analyse the potential skill of ECMWF-System4 for navigation related seasonal forecasting, the reforecast data set 01.01.1981 - 01.04.2011 as well as the pre-operational and operational forecasts of the period 01.04.2011 - 31.12.2015 are applied. In the reforecast, the number of ensemble members is limited to 15 for the initialisation months January, March, June, July, September, October, and December. The number of ensemble members is extended to 51 for the initialisation months February, May, August and November. From April 2011 onwards, the (operational) ensemble size is 51 for all initialisation months. For verification the ensemble size of the operational forecasts was reduced to the ensemble size of the reforecast (15 members) for the initialisation months January, March, June, July, September,



October, and December. The hydrological re-forecasts with LARSIM-ME are evaluated for relevant low- and medium flow indicators.

2.2.3 Agriculture - FutureWater

Irrigated agriculture is the main economic activity of Campo de Cartagena in the Segura River basin, Spain. However, water scarcity compromises such activity, which is mainly dependant on the water input it receives from the connected Tagus River basin. Mitigation measures of droughts in Spain are based on a number of drought indicators that are derived from the available water in the storage reservoirs. In order to anticipate drought episodes, decision-makers need to forecast the corresponding reservoir inflows. In the case of the Segura River, forecasts are currently estimated from simple regressions of river discharges from the preceding 6 months, leading to updated management plans twice a year.

In order to provide stakeholders in the basin with a more robust forecasting system that would allow them to better anticipate drought episodes and put into practice more effective allocation and mitigation practices, a prototype of a hydrological seasonal forecasting system is presented. The prototype uses the Spatial Processes in HYdrology model (SPHY) forced with the ECMWF's System 4 (15 ensembles) seasonal meteorological forecasts to predict monthly river inflows at the reservoirs of the upper basins of the Segura and Tagus Rivers. The model was first calibrated for the 1980-2000 period (using 1979 as a warm-up year) against discharge observations at three stations located at the major storage reservoirs: Entrepeñas and Buendía in the Tagus basin and Fuensanta in the Segura basin (see Figure 1). The Cenajo station in the Segura basin was not included in the calibration due to data availability on water transfers between catchments in the Segura basin, but it was included in the simulation runs.

The system focuses on four major periods (initialisation months) relevant for the regional climatology (January, April, July and October), with a forecasting lead time of three months, aiming at finding the most suitable period(s) to take decisions.



2.2.4 Hydropower - SMHI

The regulated mountainous basins are commonly highly influenced by snowmelt runoff and volumes in hydropower production, particularly when a multi-reservoir system is present. In the case study of the Umeälven River (Sweden), seasonal forecasts of snowmelt runoff volumes, together with ground based and remote sensing snow cover monitoring, are key inputs to the decision models of the hydropower companies when planning the production for the current and next winter seasons. It is very common that the operational seasonal forecasts are based on an ESP. Reservoir operators are interested in accumulated forecasts of inflows over the spring flood period (April to July). Forecasts for the April-July accumulated runoff are issued once a month from January until the start of the melt season in April. An important driver is the reservoir level at the end of summer, where a trade-off between water usage for power production during the spring period and the desire to have high water levels at the end of the summer is present. Unnecessary release of water that cannot be used for production is recognised as spill and loss of potential production which can be translated into an economical value. Spill of water may happen when the remaining spring flood volumes were underestimated and reservoirs filled up too early. Therefore, score metrics that are based on volume errors are appropriate measures to describe the improvements in forecasting skill.

2.2.5 Reservoir management - UPV

In the Júcar River Basin, an important characteristic is the semi-aridity of the climate that leads to high hydrological variability, resulting in recurrent periods of drought lasting several years (more than 4 years in some instances). In order to decrease the vulnerability of the water resources system, large reservoirs were built and conjunctive use of surface and groundwater is a regular practice, also integrating wastewater reclamation and reuse. Therefore, integrated and improved management of the water resources system is essential. In addition, proactive drought management requires continuous monitoring and assessment of risk in order to anticipate measures. For this purpose, reliable seasonal forecasts of climate



variables (i.e., precipitation and temperature) and hydrological forecasts (river flows) are needed for the management of the system, which in this case is based on the risk of failure in the supply for all uses, mainly water allocation for drinking water and agriculture.

The analysis must be performed in an integrated way for all elements of the water resources system of the entire basin. Otherwise, physical connections between elements (rivers, aquifers, returns from irrigation and urban uses, etc.) and implications of any decisions in the rest of the system (even from downstream to upstream), would be ignored and results would not be realistic.

For reservoir management, the key is to be able to use the decision support system (DSS) to estimate the risk of failure in the supply of water to all users, as well the risk of failure in the compliance with the established ecological flows, during the next 12 to 24 months (anticipation period). If the risk is considered to be too high, then measures must be proposed and their efficacy be tested with the DSS. A key result is also the forecast of the volume of water remaining in the reservoirs system at the end of the irrigation season. Deterministic and probabilistic forecasts will allow management measures that optimize farmer yields, maintaining high reliability of supplies to the cities, and with an adequate degree of environmental protection

Currently, the analysis is performed using flow forecasts in several places of the basin (we will focus on 5) obtained by multivariate synthetic flow forecasts generation conditioned to the present time state of the system and to past flows. The objective would be to improve the flow forecasts by incorporating short term and seasonal meteorological forecasts as forcing inputs.

For this case study, at UPV we have compared river flow data, obtained from the hydrological model E-HYPE, with regional river flow observations. Hydrological model data were provided by SMHI, corresponding to the continentally calibrated E-HYPE model for the period 1980-2010 and for five sub-basins of the Júcar River basin. Regional observations are naturalized river flows (NRF) for these sub-basins. This comparison has the purpose of testing the reliability of the E-HYPE model and to evaluate the need for a bias correction.



2.2.6 Flood & low flow forecasting - Deltares

In the Netherlands, salt intrusion occurs when the river flows of the Rhine and Meuse are low and coincidentally, wind storms push sea water into the river mouth. As a result, water boards cannot take in water to flush their polders, as these could suffer from saline seepage. This problem is a prerequisite for accurate and reliable forecasts of river flows, water levels, tide and surge, water demand and availability in the polder areas, salt concentrations and intrusion. Rijkswaterstaat (the Ministry of Infrastructure and the Environment) is currently predicting river flows up to 10-15 days for the main rivers Rhine and Meuse.

For drought forecasting in the Netherlands, the National Hydrological Model (LHM) was operationalised to support water management (e.g., lakes, surface water, etc) between April and November (Berendrecht et al., 2011). This system is forced with measured and forecasted river flows at the boundary and areal precipitation and potential evaporation (250x250m). Timely information about low flow conditions at the monthly to seasonal scale (1-3 months) can help to take measures such as raising the level of Lake IJssel. This might also have consequences for the flood risk (e.g., due to windstorms causing surge on Lake IJssel) and flow forecasts for the Rhine and Meuse should therefore be accurate and reliable.



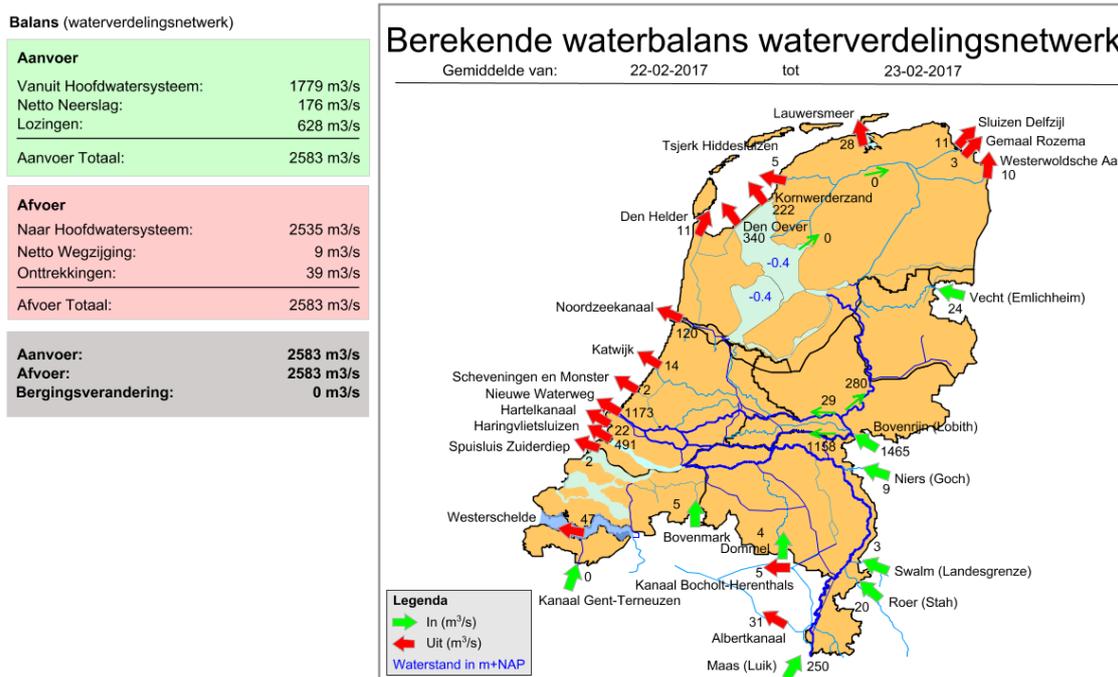


Figure 1 Example display of the operational water management system for the Netherlands, showing a daily computed or forecasted water balance for the surface water network for the Netherlands.

2.3 Sensitivity analyses as a tool to diagnose seasonal streamflow forecasting uncertainties

Despite great advances in sub-seasonal to seasonal streamflow forecasting in the last decade, the forecasting skill in Europe is still limited. This is due to a combination of errors, such as the poor seasonal meteorological forecasting skill in the extra-tropics (Arribas et al., 2010), errors in the IHC, hydrological model and downscaling errors.

Sensitivity analyses are a useful tool to diagnose the sensitivity of the model output (here hydrological variables such as discharge) to the model inputs (SCF, IHC and model parameters; Saltelli et al., 2004; 2008). They can be used for a variety of motivations, ranging from research prioritisation (improving solely certain aspects of the forecasting chain) to model simplification (Saltelli et al., 2008). To this end,



forecasting systems intercomparison can help disentangle the sources of uncertainty and/or corroborate skill both in time and space.

Another sensitivity analysis method widely used in seasonal streamflow forecasting is based on the ESP and the reverse-ESP and was first introduced by Wood and Lettenmaier (2003; 2008). The reverse-ESP can only be run in hindcast and is produced by forcing the hydrological model with a single meteorological trace (the meteorological observations for that specific time of the year). The hydrological model is initialised with an ensemble of historical IHCs (resampled from that same initialisation month for all the previous years). Contrary, the ESP is started from a single set of current IHC and forced with an ensemble of historical meteorological observations (resampled from the past meteorological observations available for all previous years and for the same time of the year as the one for which the ESP forecast is run). The ESP can be run as a forecast or in hindcast.. Whereas the uncertainty in the ESP is given by the SCF, the uncertainty in the reverse-ESP is given by the IHC. By comparing the ESP and reverse-ESP skill for a catchment-season-lead time combination, it is possible to tell which component of the forecast mainly leads the uncertainty (i.e., the SCF or the IHC). Recently, this method was extended by Wood et al. (2016) to a method called VESPA (Variational Ensemble Streamflow Prediction Assessment). The VESPA method aims at assessing intermediate uncertainty points between the climatological and 'perfect' (i.e., current observed meteorological data) skill present in the reverse-ESP and the ESP. This method allows the calculation of a metric called 'skill elasticity', a measure of the potential to increase the seasonal streamflow forecasting skill as a result of increasing the SCF or the IHC skill. In this deliverable we will use an alternative method to the VESPA method, a description of which is given in Section 3.3.

2.4 Comparative analysis in large sample hydrology

Large-scale (i.e. continental) multi-basin modelling can complement the "deep" knowledge from basin-based modelling and enhance process understanding, increase robustness of generalisations, facilitate classification of basin behaviour and



prediction, support better understanding of prediction uncertainty, and go beyond sensitivities related to IHC and SCF (Pechlivanidis and Arheimer, 2015). This type of modelling has the potential to cross regional and international boundaries whilst the analysis over a number of basins allows the consideration of different geophysical and climatic zones (Gupta et al., 2014); hence it can provide a deeper understanding of the underlying sensitivities in the forecasting skill. Such modelling type can also advance hydrological science since it finds a numerical background for comparative hydrology (Blöschl et al., 2013). The use of a large sample of stations, particularly when analyses are conducted at the continental scale (i.e., as in Europe), can also allow for exploration of emerging patterns and facilitate comparative hydrology, allowing to test sensitivities for many catchments with a wide range of environmental conditions (Blöschl et al., 2013).

However, understanding processes in large systems is challenging, given that physical properties (e.g., vegetation and soil type) generally exhibit high spatial variability, which consequently results in significant differences in system behaviour and predictability. As expected, this spatial heterogeneity introduces further high uncertainty on the categorisation of important drivers that influence the predictive hydrological skill. In addition, large river basins are often strongly influenced by human activities (e.g., irrigation, hydropower production, groundwater use) for which information is rarely available and therefore rarely described in hydrological model processes; hence introducing additional uncertainty regarding process understanding and description. Although such modelling type has limitations which vary in space, in here we make the step forward to gain insights in spatial patterns of hydrological skill at the large scale, and link this to the characteristics of the basin system.



3 Data and methods

3.1 The forecasting systems

All the partners of this deliverable use dynamical ensemble seasonal forecasting systems. These systems all use the same seasonal meteorological forecasts but are diverse in terms of the hydrological models and the presence or not of a bias correction method for the seasonal meteorological forcing. This was done in order to obtain insights into the seasonal discharge forecast sensitivity to the hydrological model type. An overview of the various systems and their characteristics is given in Table 1. For more details on the hydrological models used, see Annex A.



Table 1 Dynamical ensemble seasonal hydrological forecasting systems.

Partner	Meteorological forecasts	Interpolation method	Bias correction of the meteorological forecasts	Hydrological model	# of ensemble members	Hindcasts period for which scores calculated	Forecast starting dates	Lead time and time step	Spatial domain
ECMWF	System 4	Inverse distance weighting. Temperature was first corrected using the elevation	None	LISFLOOD (5x5km)	15, extended to 51 every three months	1990 - 2010	On the first of every month	Up to 7 months, daily values	Europe
BfG	System 4	Precipitation Voronoi	Linear scaling, Separate	LARSIM-ME (5x5km)	15, extended to	1990 - 2010	On the first of	Up to 6 months	Catchments of the Rivers



		tessellation Temperature constant lapse rate and inverse distance weighting	scaling factors are derived in dependence of initialization month and monthly lead time		51 every three months		every month		Rhine, Elbe, Weser/Ems, Odra and Danube up to gauge Nagymaros in Hungary
SMHI	System 4		Distribution- Based Scaling (DBS) approach	E-HYPE (215 km ²)	15	1990 - 2010	On the first of every month	Up to 7 months	Europe
FW	System 4		Bias correction using Spain02 observation data (Herrera et al. 2016)	SPHY (5x5km - Tagus; 2x2km Segura)	15	1990 - 2010	On the first of: January, April, July and October	Up to 3 months	Tagus and Segura River basins



UPV	System 4		None	EVALHID (semi-distributed application at sub-basin scale)	15	1990 - 2015	On the first of every month	Up to 7 months, daily values	Jucar River basin
DELTA S*	System 4	Precipitation HYRAS data set extended with emulated HYRAS Temperature HYRAS data set extended with a constant	None	wflow_hbv (1.44 km ²)	15, extended to 51 every three months	1980 - 2015	On the first of every month	Up to 3 months, daily values	Rhine



		lapse rate based on DEM and inverse distance weighting							
	System 4	EFAS forcing dataset (See ECMWF)	None	W3RA (0.5 km ² and 0.05 km ²)	15, extended to 51 every three months	1990 - 2014	On the first of every month	Up to 3 months, daily values	Europe

*As the seasonal forecast runs from Deltares were not ready at the time of this deliverable, results of discharge simulations from the two models shown above for Deltares, as well as discharge simulations produced from the HBV96 model will be shown in the results. The W3RA was run with EFAS historical forcing data from 1991-2014, while the lumped HBV96 and distributed wflow_hbv models were run using HYRAS data from 1991-2006.



3.2 The forecasting systems intercomparison

The first part of this deliverable compares the performance of the dynamical sub-seasonal to seasonal hydrological forecasting systems listed above (see Table 1). For the intercomparison, a common set of stations was selected, based on data available to the partners of this deliverable (see Figure 2 and Table 2). Observed discharge data was distributed for the corresponding stations by a few partners to all partners involved in WP4, in order to have a consistent verification across partners.

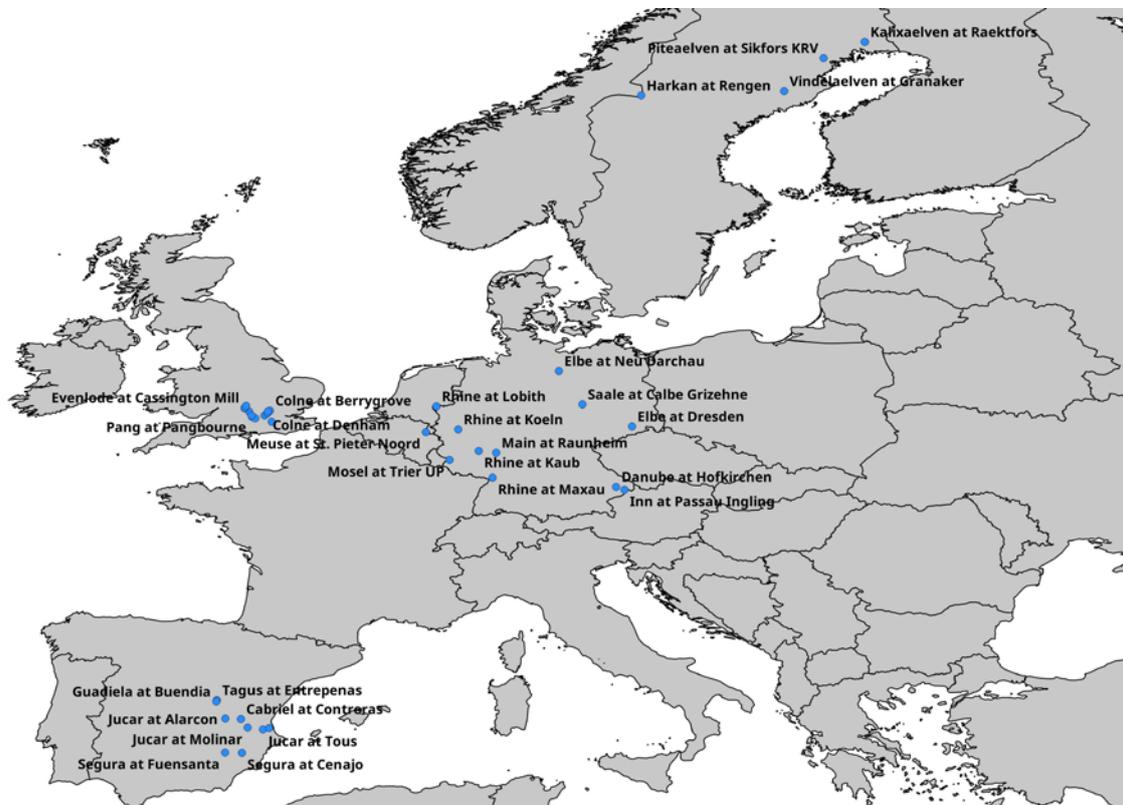


Figure 2 Map of the stations used for the analysis.



Table 2 Observed discharge data for the selected stations.

Case study	Data source	Station code	Station name	Station coordinates (lat, lon)	Drainage area (km ²)	Elevation (m)	River (country)
The Thames River Basin	Observed discharge from the NRFA (National River Flow Archive)	UOR39088	Rickmansworth	51.64199826, -0.461235789	105	47.1	Chess (UK)
		UOR39072	Royal Windsor Park	51.48562525, -0.589407615	7046	13.5	Thames (UK)
		UOR39068	Castle Mill	51.23842724, -0.31118153	316	39.2	Mole (UK)
		UOR39034	Cassington Mill	51.786274, -1.351677333	430	60.2	Evenlode (UK)
		UOR39027	Pangbourne	51.48462376, -1.08759943	170.9	39.6	Pang (UK)
		UOR39021	Enslow Mill	51.86135121, -1.301356378	551.7	65	Cherwell (UK)
		UOR39016	Theale	51.43274011, -	1033.4	43.4	Kennet (UK)



				1.066925886			
		UOR39013	Berrygrove	51.67066472, - 0.380218207	352.2	54.7	Colne (UK)
		UOR39010	Denham	51.56631768, - 0.483890807	743	34.1	Colne (UK)
		UOR39008	Eynsham	51.77524074, - 1.356429965	1616.2	59.7	Thames (UK)
		UOR39007	Swallowfield	51.37737889, - 0.951223862	354.8	42.3	Blackwater (UK)
		UOR39002	Days Weir	51.63852061, - 1.179455188	3444.7	45.8	Thames (UK)
Central European Rivers	The Global Runoff Data centre, 56068 Koblenz, Germany	GRDC634030 0	Calbe Grizehne	51.916608,11.8099 82	23719	49.36	Saale (Germany)
		GRDC634012 0	Dresden	51.054456,13.7388 29	53096	102.68	Elbe (Germany)
		GRDC634280	Hofkirchen	48.67657,13.11427	47496	299.6	Danube



		0		9			(Germany)
		GRDC633510	Kaub	50.085613,7.76500	103488	67.66	Rhine (Germany)
		0		8			
		GRDC633506	Koeln	50.937359,6.96322	144232	34.97	Rhine (Germany)
		0		5			
		GRDC633520	Maxau	49.038933,8.30553	50196	97.76	Rhine (Germany)
		0		5			
		GRDC634011	Neu Darchau	53.232337,10.8887	131950	5.68	Elbe (Germany)
		0		73			
		GRDC634390	Passau Ingling	48.5629,13.443071	26063	289.19	Inn (Germany)
		0					
		BFG2409530	Raunheim	50.016067,8.44824	27142	82.9	Main (Germany)
		2		8			
		GRDC633650	Trier UP	49.732655,6.624	23857	121	Mosel (Germany)
		0					
	Observed discharge from	GRDC643506	Lobith	51.84, 6.11	160800	8.53	Rhine (The Netherlands)
		0					



	RWS	GRDC6421101	St Pieter Noord	50.83, 5.71	21300	44	Meuse (The Netherlands)
The Segura and Tagus River Basins	Reservoir inflows	FWUT_1	Entrepenas	40.2938,-2.4456	3825.5	636	Tagus (Spain)
		FWUT_2	Buendia	40.236,-2.4657	3355.7	636	Guadiela (Spain)
		FWSG_1	Fuensanta	38.2333,-2.1224	1210.1	524	Segura (Spain)
		FWSG_2	Cenajo	38.22,-1.4629	1394.4	335	Segura (Cenajo)
The Jucar River Basin	Reservoir inflows and Spain02	UPV8001	Alarcon	39.564597,-2.112084	2937	831	Jucar (Spain)
		UPV8009	Contreras	39.543559,-1.502849	3266	1030	Cabriel (Spain)
		UPV8026	El Molinar	39.207931,-1.239957	7912	690	Jucar (Spain)



		UPV8030	Tous	39.132927,- 0.650729	17821	64	Jucar (Spain)
		UPV8089	Sueca	38.939532,- 0.478048	21497	18	Jucar (Spain)
Swedish Rivers	The Global Runoff Data centre, 56068 Koblenz, Germany	GRDC6233510	Granaker	64.239979,19.66624	11850.5	NA	Vindaelven (Sweden)
		GRDC6233710	Sikfors KRV	65.532833,21.20877 8	10816.1	NA	Piteaelven (Sweden)
		GRDC6233850	Raectfors	66.170645,22.81577 5	23102.9	NA	Kalixaelven (Sweden)
		GRDC6233470	Rengen	64.069902,14.09551 9	1110.1	NA	Harkan (Sweden)



In order to compare the performance of the different ensemble seasonal hydrological forecasting systems, several scores were chosen including both deterministic and probabilistic scores and covering the main attributes of ensemble forecasting relevant for sectoral applications (see Section 3b). These scores include:

- Deterministic scores:
 - o **The Mean Absolute Error (MAE)** (cawcr, 2015):

$$MAE = \frac{1}{N} \sum_{i=1}^N |F_i - O_i|$$

The MAE ranges from 0 to an upper boundary defined by the system's variability, with a perfect score of 0, and indicates the average magnitude of the forecast errors. Where F_i is the ensemble mean and O_i is the observed discharge for the same time. N is the sample size, it is the total number of forecasts made for the same target month and with the same lead time and temporal aggregation type. This score does not indicate the direction of the forecast deviations, which will be calculated using the Mean Error (ME).

- o **The Mean Error (ME)** (cawcr, 2015):

$$ME = \frac{1}{N} \sum_{i=1}^N (F_i - O_i)$$

The ME ranges from $-\infty$ to $+\infty$, with a perfect score of 0, and is a measure of the average forecast error, considering the ensemble mean. It indicates the forecast average additive bias (i.e., its tendency to underestimate or overestimate observed discharge). Note, a good ME score does not guarantee that the forecast is perfect as overestimations and underestimations made by the latter can compensate each other.

- o The **normalised volumetric term of the Kling-Gupta Efficiency** (*beta*, Gupta et al., 2009):

$$beta = 1 - \sqrt{(\beta - 1)^2}$$

β is defined as the ratio of the monthly mean of the forecasts (the output of the model forced by meteorological forecasts) over the monthly mean of the





perfect forecasts (the output of the model forced by the reference forcing dataset); note that the range of the values for each term varies between $-\infty$ and 1 with 1 being the optimum.

- Probabilistic scores:
 - o **The Continuous Ranked Probability Score (CRPS)** (Hersbach, 2000):

$$CRPS(P, x_a) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 dx$$

Where P is the ensemble forecast cumulative distribution function (cdf) and P_a is the observation cdf and is defined by:

$$P_a(x) = H(x - x_a)$$

For the observed discharge x_a , with $H(x)$ the Heaviside function:

$$H(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$$

The CRPS ranges from 0 to $+\infty$, with a perfect score of 0, and is a measure of the difference between the forecast and the observation cdfs. A perfect score of 0 is achieved in the case of a perfect deterministic forecast. The CRPS is a measure of the forecast accuracy and sharpness. It can be further decomposed into reliability, resolution and uncertainty components, according to:

$$CRPS = reliability + CRSPotential$$

Where the potential CRPS is the CRPS value that a forecast with perfect reliability (reliability=0) would have, expressed as:

$$CRSPotential = uncertainty - resolution$$

The reliability is a measure of the bias and the spread of the system. The uncertainty is the variability of the observations and the resolution is the



ability of the forecast to distinguish situations with distinctly different frequencies of occurrence. The components all range from 0 to $+\infty$, with a perfect score of 0. The CRPS and its components were averaged over all the forecasts made for the same forecast initialisation date and with the same lead time and temporal aggregation (monthly averages here).

o **The Brier score (BS)** (cawcr, 2015):

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2 = \frac{1}{N} \sum_{k=1}^K n_k (p_k - \bar{o}_k)^2 - \frac{1}{N} \sum_{k=1}^K n_k (\bar{o}_k - \bar{o})^2 + \bar{o}(1 - \bar{o})$$

Where N is the sample size, the total number of forecasts made for the same target season and with the same lead time, temporal aggregation type and for the same event. o_i is a binary observation, it is 1 if a predefined event happened and 0 if it did not. p_i is the forecast probability of the event happening. The BS ranges from 0 to 1, with a perfect score of 0 and is a measure of the mean squared error of the probability forecasts over the verification sample. The events selected to calculate the Brier score are the upper and the lower terciles of the observed discharge for the specific season for which the score is calculated. These thresholds were chosen in order to have a large enough sample as this score is sensitive to the climatological frequency of the event: the rarer an event is, the easier it will be to obtain a good BS without necessarily having any real skill. The BS can be further decomposed into a (1) reliability, (2) resolution and (3) uncertainty part. The BS and its three parts were averaged over all the forecasts made for the same target season and with the same lead time, temporal aggregation (monthly averages here) and for the same event (upper or lower terciles).

o **The skill scores:** the forecast skill was also calculated for the CRPS and the BS using the following equation:

$$Skillscore = 1 - \frac{score_{forecast}}{score_{reference}}$$





For the reference, two benchmarks were selected. The first benchmark is the climatology of observed discharge and the corresponding skill scores of the CRPS and the BS are called the CRPSS_CLI and the BSS_CLI, respectively. The climatology covers the same period as is covered by each forecasting system (excluding the year analysed) and is the climatology of a given target month (or season for the Brier Score). The second benchmark is the ESP corresponding to each system, for the same forecast initialisation date, lead time and temporal aggregation (monthly averages here). The corresponding skill scores of the CRPS and the BS are called the CRPSS_ESP and the BSS_ESP, respectively

The analysis will present scores measured for the discharge forecasted from various forecast starting dates or target seasons (for the BS), lead times, monthly aggregations and several stations in Europe. This intercomparison will provide a spatio-temporal overview of the performance of the seasonal hydrological forecasting systems overall as well as for extreme events (high and low flows).

3.3 The EPB sensitivity analysis

The VESPA method is a sensitivity analysis method in the sense that it measures the response of the model output (discharge in our case) to a known variation in the model input(s) (here the SCF and the IHC). It was designed and tested on 424 catchments in the contiguous United States (CONUS), for which it successfully exposed the relative contributions of the two sources of errors (SCF and IHC) on seasonal streamflow forecasting uncertainty. Moreover, the 'skill elasticity' produced by the VESPA method indicates the potential to improve the seasonal streamflow forecasting skill by improving the SCF and/or the IHC skill. This information is valuable for guiding resources in seasonal forecasting system development towards useful improvements. One drawback of the VESPA method however is that it is computationally expensive to run as it is based on a very large number of simulations. Recently, an alternative and cheaper method called EPB (End Point Blending) was designed and tested on 18 catchments of the CONUS for which it gave almost identical results to the VESPA method (Arnal et al. 2017). Because the EPB sensitivity analysis is a



reliable and computationally cheap method which can give insightful results in the context of seasonal streamflow forecasting improvements, it will be used in this deliverable.

The EPB is constructed by combining four sources of data (also called end points): the ESP, the reverse-ESP, the climatology and the 'perfect' forecast. The term 'perfect' refers to current observed meteorological data and the term climatological refers to the whole distribution of historical meteorological observed data. Each end point corresponds to a combination of IHC and SCF weights (w_{IHC} and w_{SCF} respectively; the axes on Figure 3). A weight of 0 is the 'perfect' knowledge (upper right corner on Figure 3) whereas a weight of 1 is the climatological knowledge of either of the two predictability sources (bottom left corner on Figure 3). A 'perfect' forecast (forecast generated by starting a hydrological model with the current IHC and forcing it with the current observed meteorological data) has a w_{IHC} and a w_{SCF} of 0. The climatological forecast ('climo' on Figure 3; forecast generated by starting a hydrological model with all historical IHC and forcing it with all historical observed meteorological data) has a w_{IHC} and a w_{SCF} of 1 by definition. The reverse-ESP is forced with a single meteorological trace, the meteorological observations for that specific time of the year (w_{SCF} of 0) and the model is initialised with a range of historical IHC (w_{IHC} of 1). The ESP is forced with historical observed meteorological data (w_{SCF} of 1) and the current IHC (w_{IHC} of 0).

The EPB combines these four end points for each intermediate SCF and IHC weights ($w = 0, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 1.0$), as shown on Figure 3 below. Those intermediate weights were chosen in order to coincide with the VESPA method (Wood et al. 2016). For each w_{SCF} - w_{IHC} combination (each cross on Figure 3 below), a new 100-member hindcast is generated by a weighted averaging of the forecasts carried out for the four end points. The percentage of each end point used, EP [%] (i.e., the number of members randomly selected from each end point), is given for each combination point by the following equation:

$$EP[\%] = (1 - |x_{EP} - w_{IHC}|) \times (1 - |y_{EP} - w_{SCF}|)$$





Where x_{EP} and y_{EP} are the w_{IHC} and w_{SCF} values of the end point for which the percentage is calculated, respectively. For example, if the w_{IHC} and w_{SCF} match the end point values, 100 percent of the EPB hindcast members are resampled from that end point (i.e., the end point skill is reproduced). This was done for each forecast initialisation date for a given location.

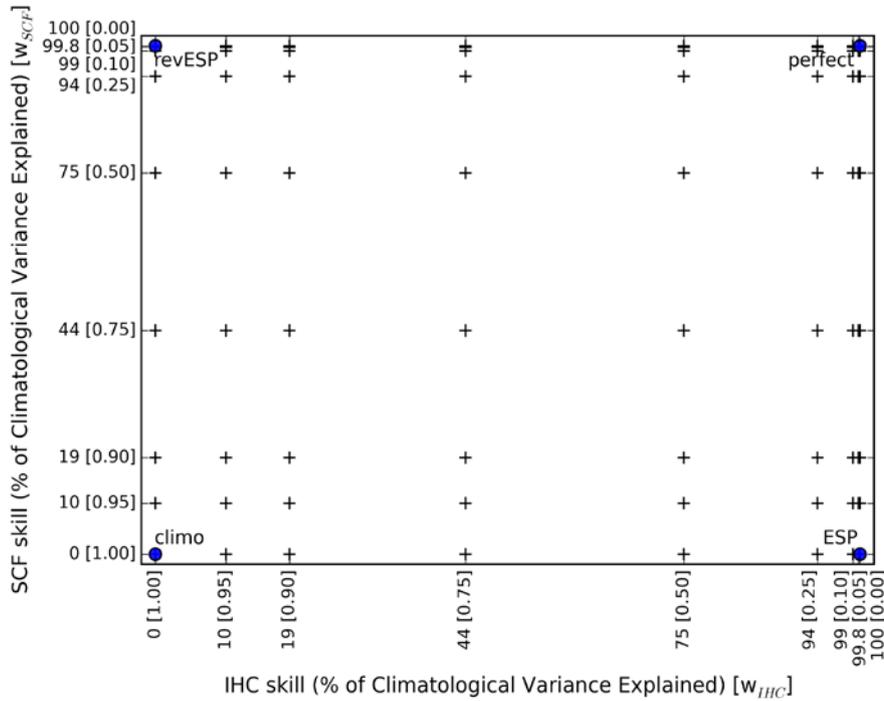


Figure 3 Resampling surface for the EPB sensitivity analysis method (taken from Arnal et al. 2017).

Once the new EPB hindcasts have been generated, their quality can be calculated for each combination point. A plot of the forecast quality as a function of IHC and SCF skill can then be drawn and is called skill surface plot in Wood et al. (2016). Finally, for each response surface (i.e., skill elasticity plot) skill elasticities for the IHC and the SCF (E_{IHC} and E_{SCF} respectively) can be measured from the scores from the following equations:

$$E_{IHC} = 100 \times \left\{ \frac{S(F[75,19]) - S(F[19,19])}{75\% - 19\%} + \frac{S(F[75,44]) - S(F[19,44])}{75\% - 19\%} + \frac{S(F[75,75]) - S(F[19,75])}{75\% - 19\%} \right\} / 3$$

$$E_{SCF} = 100 \times \left\{ \frac{S(F[19,75]) - S(F[19,19])}{75\% - 19\%} + \frac{S(F[44,75]) - S(F[44,19])}{75\% - 19\%} + \frac{S(F[75,75]) - S(F[75,19])}{75\% - 19\%} \right\} / 3$$



The numerators, expressed as $S(F[-]) - S(F[-])$, are the streamflow forecast skill gradients between IHC skill (or SCF skill) values of 75% and 19% (the denominator). The values in the square brackets of the numerator are the IHC skill followed by the SCF skill values, indicating a wSCF - wIHC combination point in the example skill surface plot (i.e., Figure 3). In the denominator, the IHC and SCF skill gradients are gradients in the percentage of the climatological variance explained in the respective predictability source. The skill elasticities (E_{IHC} and E_{SCF}) are positively oriented; where a skill elasticity of zero is obtained when the predictability source has no influence on the skill of the streamflow forecast, while positive (negative) elasticities mean that an improvement in the predictability source will lead to higher (lower) streamflow forecast skill.

For this deliverable, we calculated the IHC and SCF skill elasticities for the ECMWF seasonal discharge forecasts described in Table 1, for each initialisation date (the first of each month), monthly forecast aggregations from 1 to 7 months of lead time and over 74 geoclimatological regions in Europe. These 74 regions were selected as they are the same regions for which the ECMWF seasonal streamflow forecast is currently operational in EFAS (European Flood Awareness System). The skill elasticities are based on the CRPSS, calculated against the climatological forecast. This analysis assumes that the model is perfect as the CRPS is calculated against the 'perfect' forecast (i.e., discharge simulation) and not actual discharge observations. Additionally, skill elasticities were calculated for the BfG seasonal discharge forecasts described in Table 1, for each initialisation date (the first of each month), monthly forecast aggregations from 1 to 6 months of lead time and for the stations shared by the BfG and presented in Table 2.

3.4 Seasonal hydrological forecasts - Clustering of the skill

To better understand the potential factors influencing the skill of a model and to identify regions of similarity, we apply classification and regression trees (CART). Here, we explored the spatial runoff patterns across the entire subcontinent by analysing the skill in all 35408 catchments modelled by the E-HYPE model. CART is a recursive-partitioning algorithm that





classifies the space defined by the input variables/descriptors (i.e. physiographic-hydrologic-climatic characteristics, and remaining climatic biases) based on the output variable (i.e. beta skill for lead month 2 and month March). The tree consists of a series of nodes, where each node is a logical expression based on a similarity metric in the input space (physiographic-hydro-climatic characteristics etc.). CART also provides information on the probabilities of different output groups at each leaf node. In this case, beta (see section 3.2) is divided into five groups - bad ($\beta < 0.2$), poor ($0.2 < \beta < 0.4$), medium ($0.4 < \beta < 0.6$), good ($0.6 < \beta < 0.8$) and very good ($\beta > 0.8$), which are termed C0, C1, C2, C3 and C4 respectively. A terminal leaf exists at the end of each branch of the tree, where the probability of belonging to any of the three output groups can be inspected. Here we summarised the basin characteristics into climatic, topographic, human impacts, biases in forcing input and hydrologic bias (Table 3). We next calculate the predictors' importance (and rank them) by summing changes in the risk due to splits on every predictor and dividing the sum by the number of branch nodes.

It is important to note that in order to avoid the high dimensionality in the CART analysis, the hydrologic signatures were firstly clustered into 11 groups with each group receiving an ID (named FlowID). We applied a k-means clustering approach within the 12-dimensional space (consisting of the 12 calculated flow signatures in Table 3) to categorise the subbasins based on their combined similarity in flow signatures. Through the mapping of the spatial pattern we gained insight into the similarities of catchment functioning and could identify the dominant flow generating processes for specific regions.

Table 3 Basin characteristics used in the clustering analysis.

Climatology characteristics	(7)	Topography	Human impact	Forcing	biases	Hydrologic signatures (12)
		(4)	(1)	(2)		
Precipitation (mm/month); Prec.		Area (km ²); Area	Degree of regulation (%); DoR	Precipitation (%); BiasPrec.		Mean annual specific runoff; Qm
Temperature (°C); Temp.		Elevation (m); Elev.		Temperature (%); BiasTemp.		Normalised high flow; q05
Snow depth (cm/month); Snow		Relief ratio (-); Relief				Normalised low flow; q95
Actual evaporation (mm/month); AET		Slope (%); Slope				Normalised relatively low flow; q70



Potential evaporation (mm/month); P				Slope of flow duration curve; mFDC
Dryness index (-); P/Prec.				Range of Parde coefficient; DPar
Evaporative index (-); AET/P				Coefficient of variation; CV
				Flashiness; Flash
				Normalised peak distribution; PD
				Rising limb density; RLD
				Declining limb density; DLD
				Baseflow index; BFI





4 Results

4.1 The forecasting systems intercomparison

A set of scores was added from each forecasting system (from the ECMWF, SMHI, BfG and FW) into the scoreboard for the selected stations shown in Table 2. This allows to have a first view of the similarities as well as differences between the forecasting systems' performances and highlights common forecasts' behaviours across river basins.

The seasonal discharge forecasts' quality depends on the target month, the lead time and the station for which the forecast is made. We will split the results according to the geographical location of the river basins, as there are some noticeable similar characteristics in terms of forecast performance for stations in a given area of Europe.

4.1.1 Central European Rivers

For stations of the Central European Rivers case study, scores were calculated from the SMHI, the BfG and the ECMWF forecasting systems. From this set of stations, there appears to be two types of forecast performance behaviours. For the most western Central European Rivers stations included in this deliverable (the Main at Raunheim, the Rhine at Koeln, Kaub and Maxau and the Mosel at Trier UP), all forecasts show similar CRPS values, with larger errors from November-April. Figure 4 is an example of the CRPS for the three systems for the Rhine at Koeln.

For the most eastern Central European Rivers stations included in this deliverable (the Elbe at Neu Darchau and Dresden and the Saale at Calbe Grizehne), the SMHI forecasts display larger CRPS values than the two other systems, especially from December-April. Figure 5 is an example of the CRPS for the three systems for the Elbe at Dresden.

In general however, the BfG and the ECMWF seasonal discharge forecasts have a lower CRPS than the SMHI forecasts for the first forecast month.



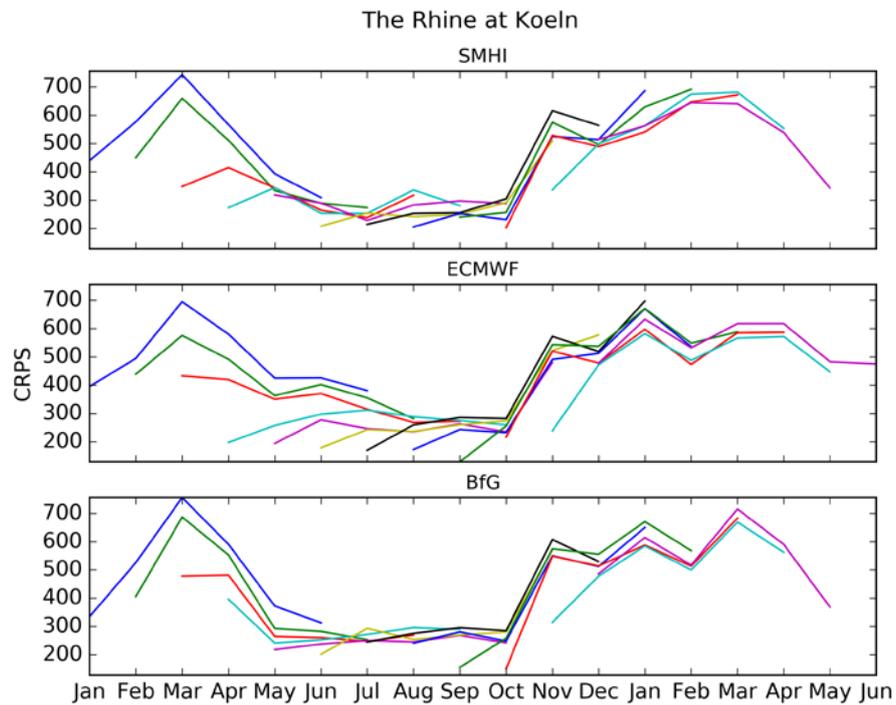


Figure 4 Continuous Ranked Probability Score (CRPS) for the Rhine at Koeln for (top) the SMHI forecasts, (middle) the ECMWF forecasts and (bottom) the BfG forecasts. CRPS = 0 denotes a perfect forecast. The CRPS is given for each forecast initialisation date (on the first of each month, different colours) and for 6 months of lead time (for the SMHI and the BfG forecasts) or 7 months of lead time (for the ECMWF forecasts).



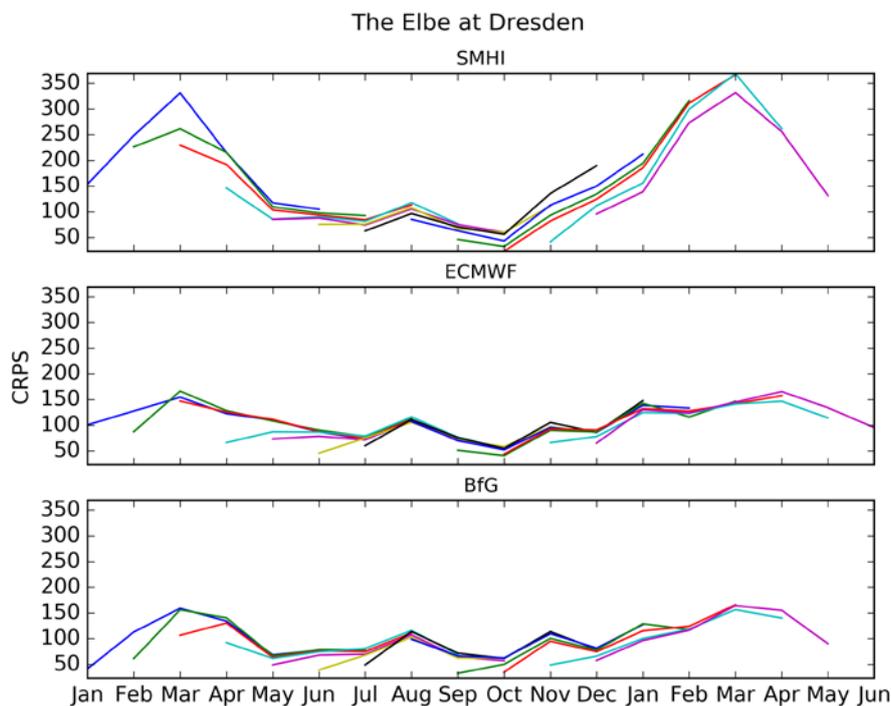


Figure 5 Same as Figure 4 but for the Elbe at Dresden.

While the seasonal forecasts with the W3RA model are not available at this time, results are shown from the simulation run (spanning 1990-2014, with one year of spin up) obtained from the W3RA model from Deltares for the Elbe at Dresden. The results, presented on Figure 6, are plotted as monthly values of the mean absolute error (MAE; comparable with the CRPS as shown above for the SMHI, ECMWF and BfG models). From these results, it appears that the W3RA model has the largest errors from October-May. This is similar to the pattern of the CRPS observed on Figure 5 for the SMHI forecasts for the same station.



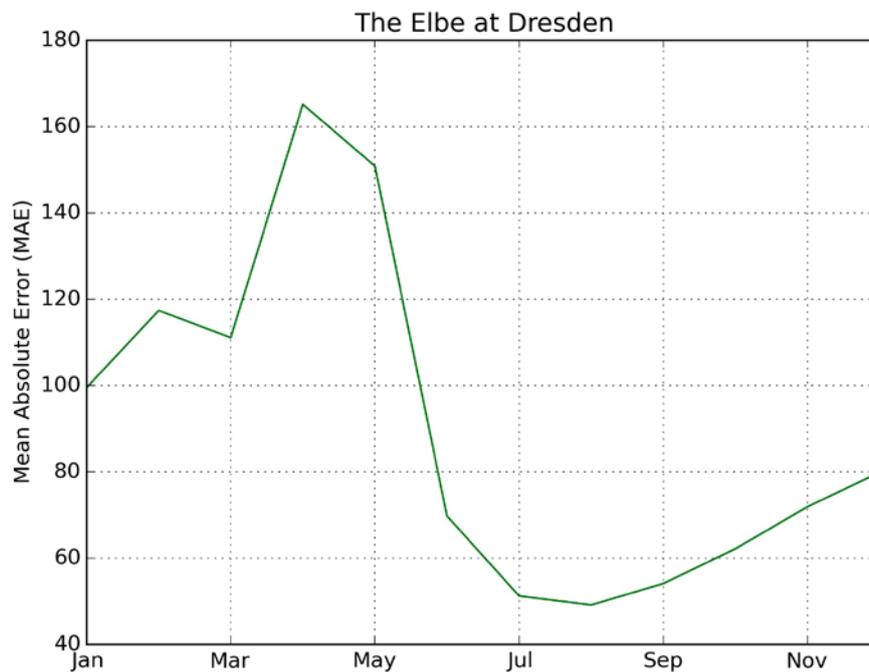


Figure 6 Mean Absolute Error (MAE) for the Elbe at Dresden (at lead time 0) for the W3RA model for the period 1991-2014.

For the Rhine at Lobith, the results are also plotted as monthly values of the mean absolute error for the W3RA model, the lumped HBV96 model and the distributed wflow_hbv model (Figure 7). Note that W3RA was run with EFAS historical forcing data, while the lumped HBV96 and distributed wflow_hbv models were run using HYRAS data (details given in Table 1). From Figure 7, it appears that the W3RA model displays larger errors than the two other models almost all year long, especially in summer. This could be an indication that the EFAS historical forcing data has large uncertainties for this station. It could furthermore be due to a misrepresentation of essential discharge generating mechanisms in this region by the W3RA model.



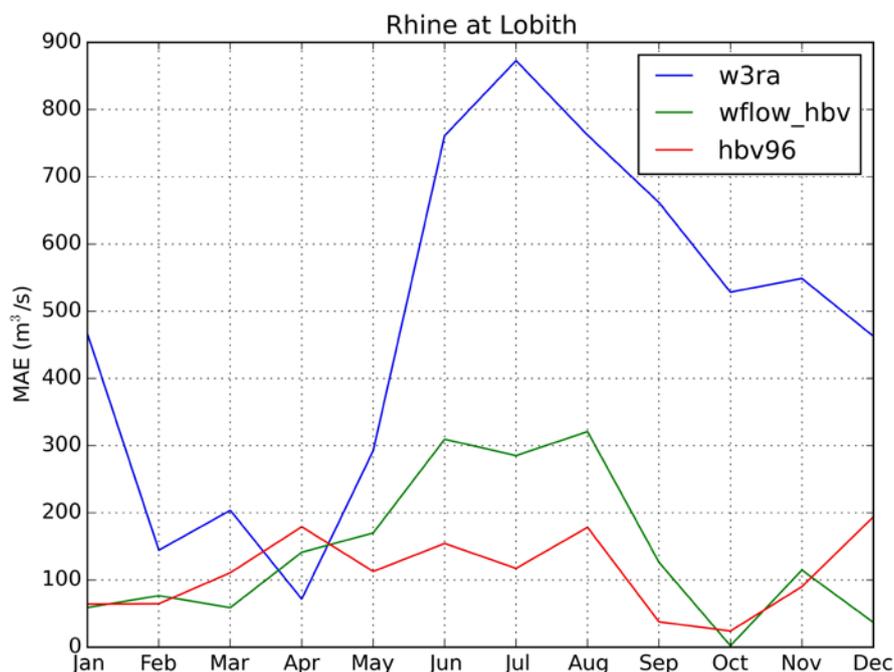


Figure 7 MAE for the Rhine at Lobith (at lead time 0) for the W3RA (1991-2014), the lumped HBV96 and the distributed wflow_hbv models (both for the period 1991-2006).

Figures 8 and 9 show the bias (i.e., the ME) for all forecast initialisation dates and all lead times for the Rhine at Koeln and the Elbe at Dresden, respectively. For most Central European Rivers stations included in this analysis, the SMHI forecasts overestimate the observed discharge in the winter to spring months. This is both the case for the Rhine at Koeln (Figure 8) and the Elbe at Dresden (Figure 9). This positive bias could be due to a hydrological model error, where the model releases more water as river flow than is observed because it cannot store enough water as groundwater. For the Rhine at Koeln (and a few other stations of the most western Central European Rivers stations, not shown), the SMHI forecasts additionally present a negative bias for the rest of the year.

The ECMWF forecasts overall overestimate the observed discharge during the spring months (more largely at longer lead times) while underestimating the winter discharge. This is both true for the Rhine at Koeln and the Elbe at Dresden (see Figures 8 and 9). This positive bias extends into the early summer months for some stations. These biases could be due to meteorological forecast error as the input meteorological forecasts used to produce the ECMWF seasonal discharge forecasts was not bias corrected, contrary to the BfG and the



SMHI forecasts. It seems that ECMWF generates too much of the precipitation falling as snow in winter, leading to underestimated discharge in those months and a snowmelt compensation in spring.

The BfG forecasts underestimate the observed discharge for winter and early spring months or all target months, depending on the station (see Figures 8 and 9). This behaviour could either be due to the bias correction of the meteorological forecasts input to the hydrological model, which produces too dry conditions compared to the observed amount, or to the hydrological model which stores too much incoming water as groundwater.

These are general characteristics of the SMHI, BfG and ECMWF forecasts and the magnitude of the bias depend on the station, target month and lead time for which the forecasts were made.

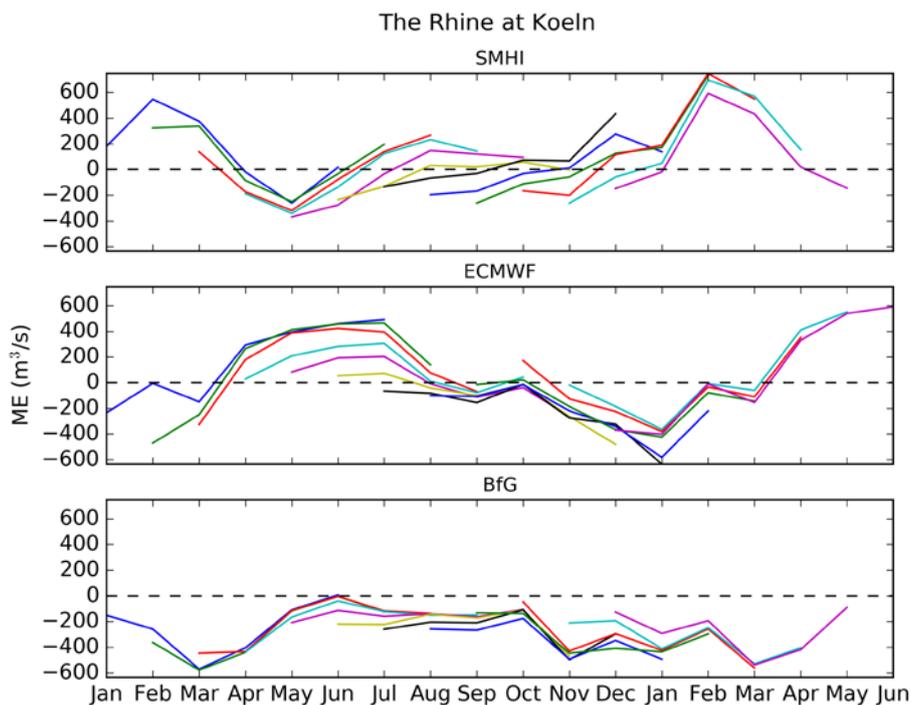


Figure 8 Mean Error (ME) for the Rhine at Koeln for (top) the SMHI forecasts, (middle) the ECMWF forecasts and (bottom) the BfG forecasts. ME = 0 denotes no bias, while ME > 0 denotes a positive forecast bias and the ME < 0 a negative forecast bias. The ME is given for each forecast initialisation date (on the first of each month, different





colours) and for 6 months of lead time (for the SMHI and the BfG forecasts) or 7 months of lead time (for the ECMWF forecasts).

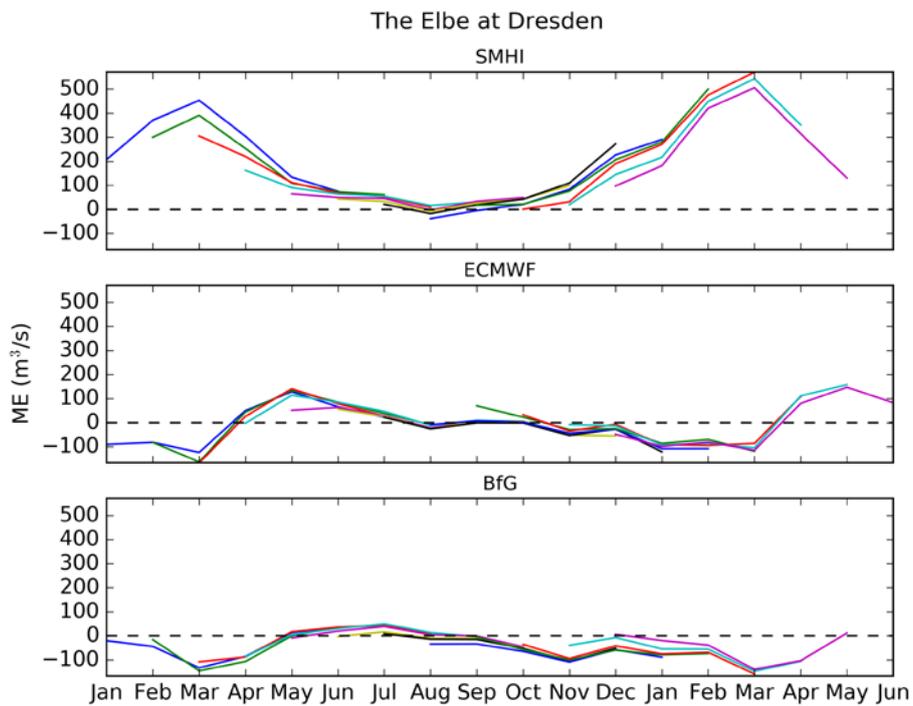


Figure 9 Same as Figure 8 but for the Elbe at Dresden.

Figure 10 shows the bias (i.e., the ME) for the W3RA model simulation for the Elbe at Dresden. From this figure, it can be seen that the W3RA underestimates the discharge for all months of the year, especially in April.



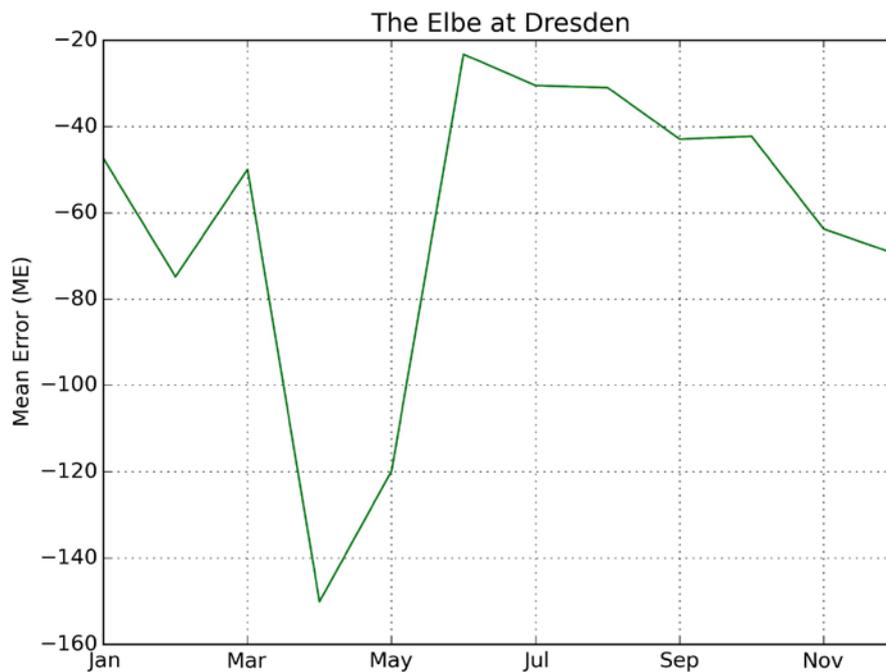


Figure 10 ME for the Elbe at Dresden (at lead time 0) for the W3RA model for the period 1991-2014.

Figure 11 displays the Mean Error (ME) for the Rhine at Lobith for the W3RA, the lumped HBV96 and the distributed wflow_HBV models. For this station, the W3RA model largely overestimates the discharge for all months. The wflow_hbv model overestimates the observed discharge mostly in the summer, while it underestimates it slightly in November and January-February. The HBV96 model underestimates the observed discharge for October-November and overestimates it for the rest of the year.



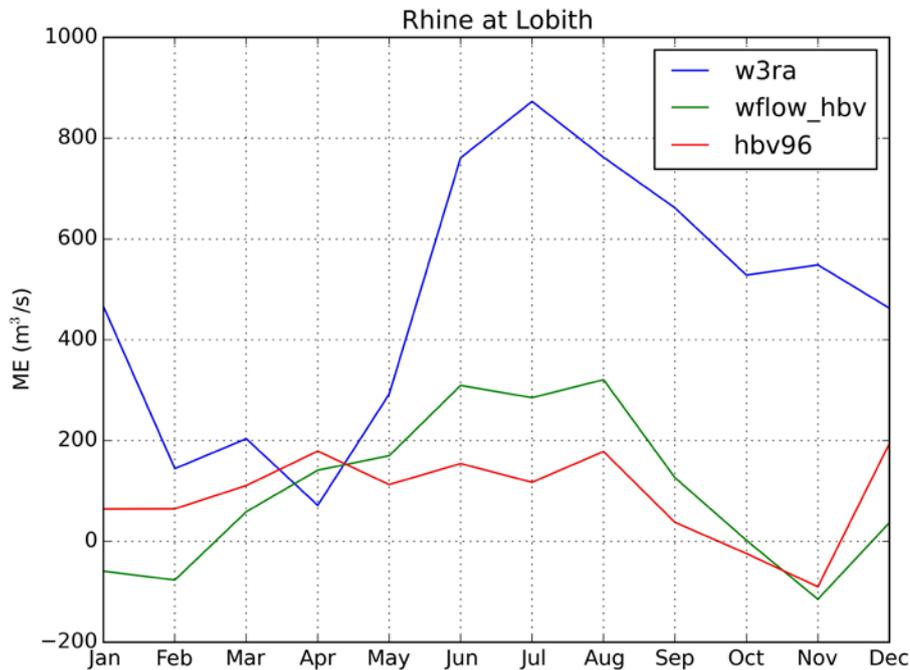


Figure 11 ME for the Rhine at Lobith (at lead time 0) for the W3RA (1991-2014), the lumped HBV96 and the distributed wflow_hbv models (both for the period 1991-2006).

In terms of the reliability of the forecasts (i.e., CRPS reliability), the results are contrasted and vary from station to station. For the most western Central European Rivers stations, the CRPS reliability appears highly influenced by the forecast lead time as well as the event which is being forecasted. Figure 12 is an example of the CRPS reliability for the three systems for the Rhine at Koeln. For this station, the SMHI forecasts are less reliable from February-March and May at 1 month lead time. The ECMWF forecasts are less reliable from May-July and January. The BfG forecasts are less reliable from March-April.

For the most eastern Central European Rivers stations, the CRPS reliability, the ECMWF and the BfG forecasts display a better reliability than the SMHI forecasts, especially for December-April. Figure 13 is an example of the CRPS reliability for the three systems for the Elbe at Dresden.



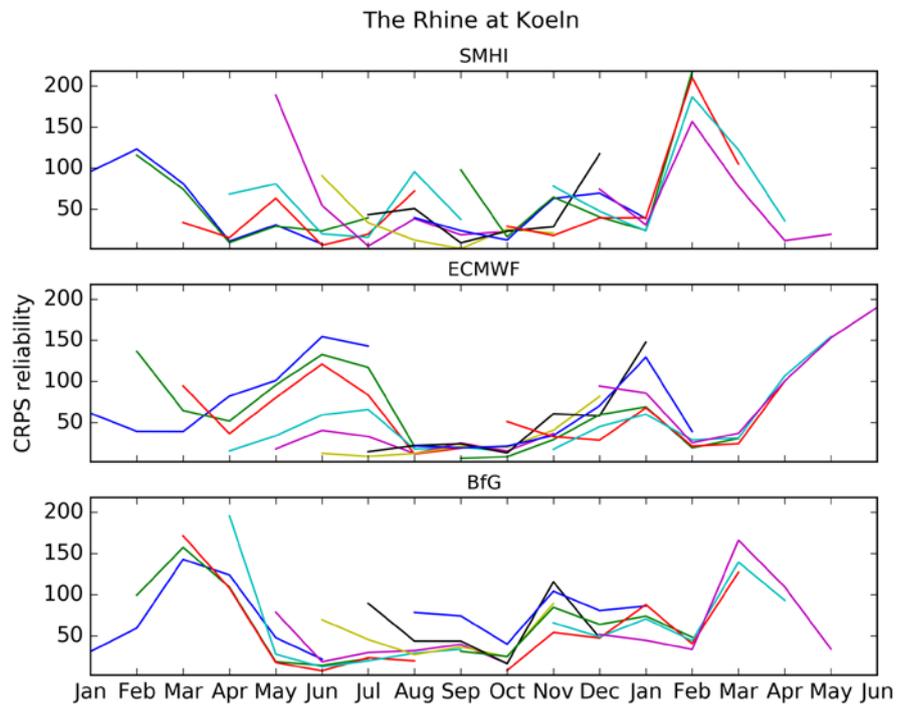


Figure 12 CRPS reliability for the Rhine at Koeln for (top) the SMHI forecasts, (middle) the ECMWF forecasts and (bottom) the BfG forecasts. CRPS reliability = 0 denotes a perfect forecast reliability. The CRPS reliability is given for each forecast initialisation date (on the first of each month, different colours) and for 6 months of lead time (for the SMHI and the BfG forecasts) or 7 months of lead time (for the ECMWF forecasts).



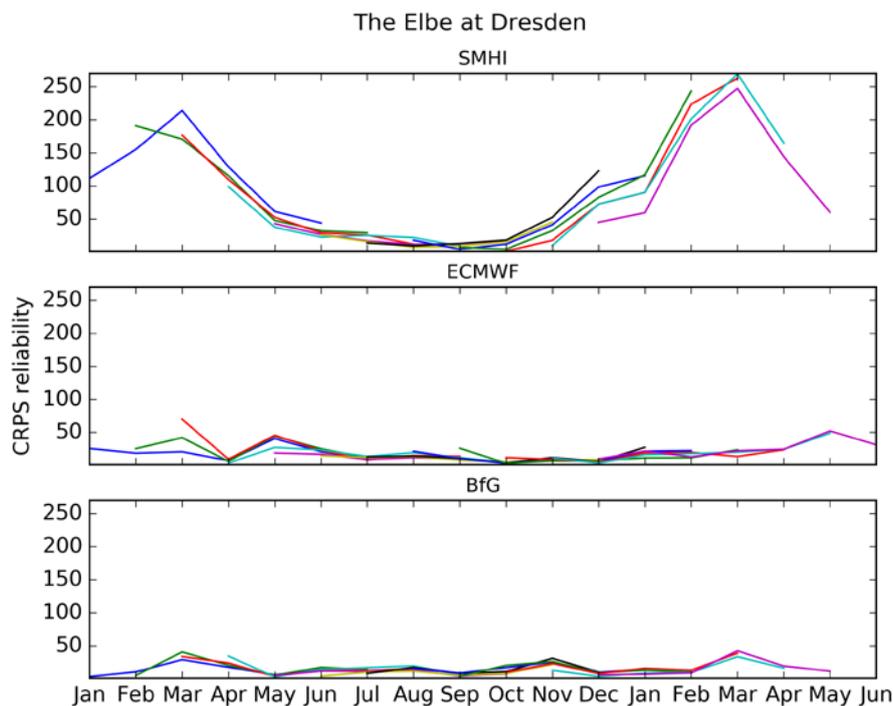


Figure 13 Same as Figure 12 but for the Elbe at Dresden.

If we look at the seasonal discharge forecasts skill, when compared to the observed discharge climatology (CRPSS_CLI), it appears that the seasonal discharge forecasts produced by the three systems are more accurate and sharp than the observed discharge climatology for the first month to two months of lead time, depending on the station and the event forecasted. In some cases however, the seasonal discharge forecasts show a lower performance than the observed discharge climatology, for all lead times. In general, the ECMWF and the BfG forecasts are more skilful than the SMHI forecasts for the first month of lead time. Figures 14 and 15 are examples of the CRPSS_CLI for the three systems for the Rhine at Koeln and the Elbe at Dresden, respectively.



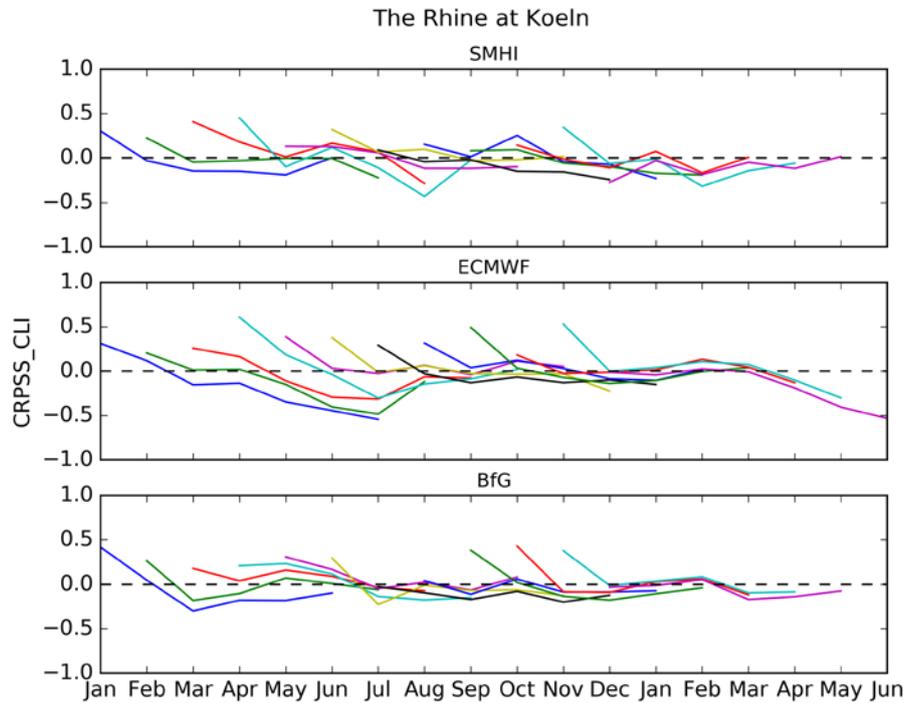


Figure 14 Continuous Ranked Probability Skill Score (CRPSS) of the seasonal discharge forecast against the observed discharge climatology for the Rhine at Koeln for (top) the SMHI forecasts, (middle) the ECMWF forecasts and (bottom) the BfG forecasts. CRPSS = 1 denotes a perfect forecast skill. The CRPSS is given for each forecast initialisation date (on the first of each month, different colours) and for 6 months of lead time (for the SMHI and the BfG forecasts) or 7 months of lead time (for the ECMWF forecasts).



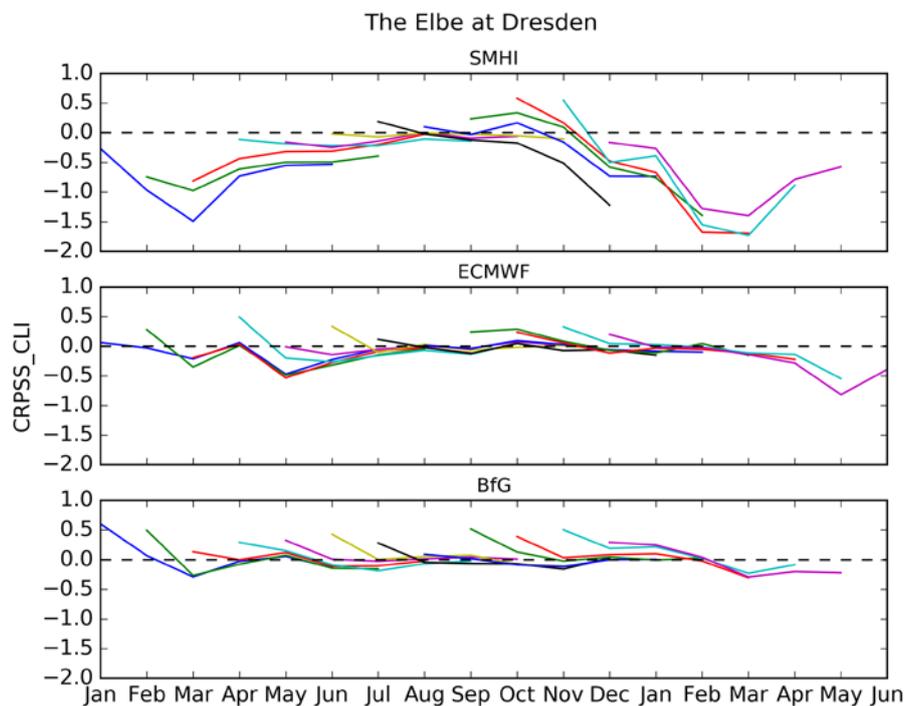


Figure 15 Same as Figure 14 but for the Elbe at Dresden.

For the upper and the lower terciles of the observed discharge, all forecasts show a very similar accuracy (Brier score of 0.2-0.3 on average over all lead times and all target seasons). There are however slight differences for single stations, for which the SMHI forecasts have a lower performance than the other forecasts for the summer target season for the lower tercile (BS33) and for the summer and the winter target season for the upper tercile (BS66; not shown).

4.1.2 The Thames River Basin

For stations for the Thames River Basin case study, scores were calculated from the SMHI and the ECMWF forecasting systems. From this set of stations, there are several observable forecast performance behaviours. For several stations, both the SMHI and the ECMWF forecasts appear overall less accurate and sharp (in terms of the CRPS) from October-April. Figure 16 is an example of such a behaviour and shows the CRPS for the Thames at Royal Windsor Park for both systems.



For several other stations, the SMHI forecasts are less accurate and sharp than the ECMWF forecasts throughout the year, especially from November-April. This can be seen on Figure 17 of the CRPS for the Pang at Pangbourne for both systems.



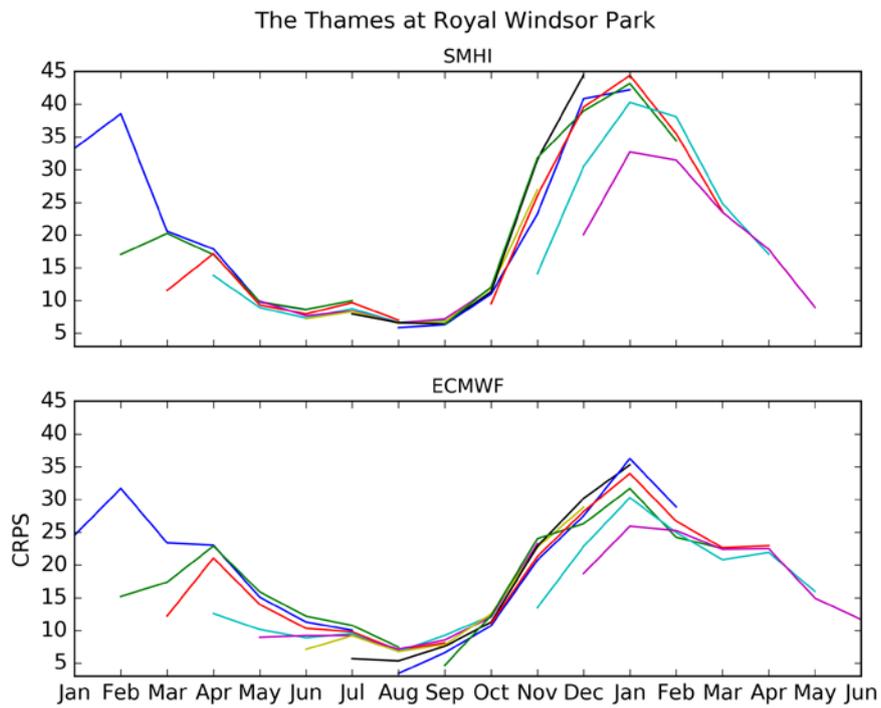


Figure 16 Same as Figure 4 but for the SMHI and the ECMWF forecasts only for the Thames at Royal Windsor Park.

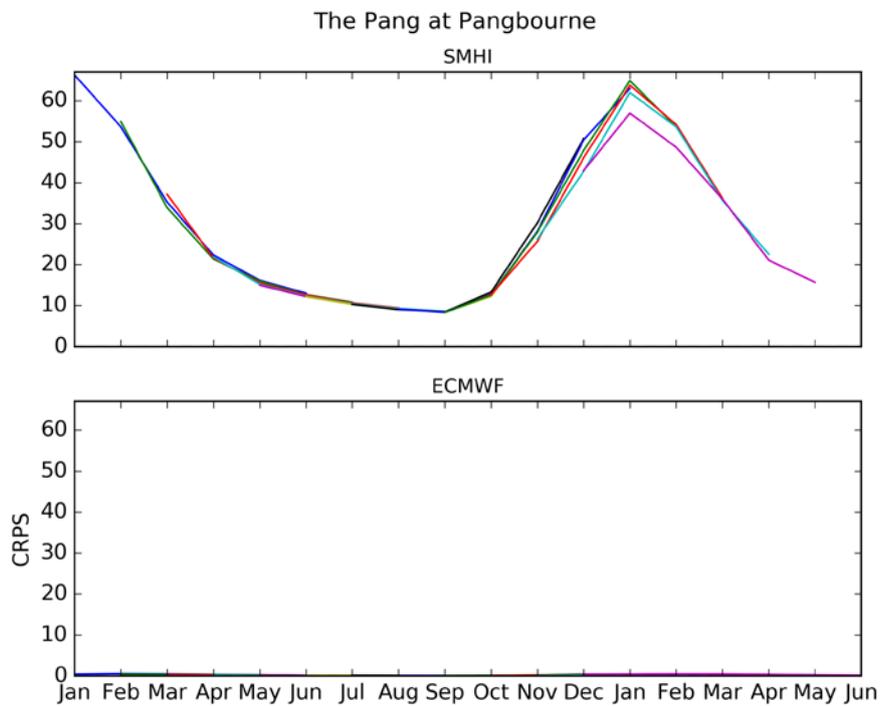


Figure 17 Same as Figure 4 but for the SMHI and the ECMWF forecasts only for the Pang at Panbourne.



In terms of the forecast bias (i.e., the ME) for both stations, the SMHI forecasts overestimate the November-March discharge (as seen for the Thames at Royal Windsor Park and the Pang at Pangbourne; Figures 18 and 19 respectively) and sometimes underestimate the April-May discharge (as seen for the Thames at Royal Windsor Park on Figure 18). A similar positive bias was described for stations of the Central European Rivers (Section 4.1.1) and could be here again due to a hydrological model error, where the model releases more water as river flow than is observed because it cannot store enough water as groundwater.

While there is almost no bias for the Pang at Pangbourne for the ECMWF forecasts (Figure 19), they underestimate the discharge from November-January and overestimate it more largely from February-June (increasingly with lead time) for the Thames at Royal Windsor Park (Figure 18). A similar bias pattern was observed for stations of the Central European Rivers and could be here again due to an overestimated percentage of precipitation falling as snow in the winter, leading to underestimated discharge in those months and a snowmelt compensation in spring.



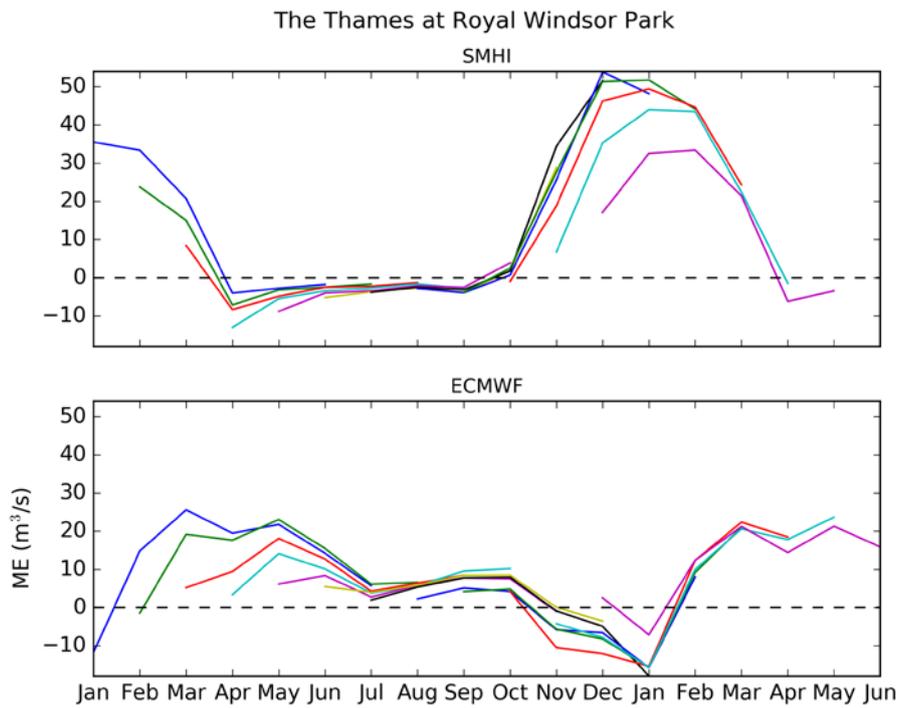


Figure 18 Same as Figure 8 but for the SMHI and the ECMWF forecasts only for the Thames at Royal Windsor Park.

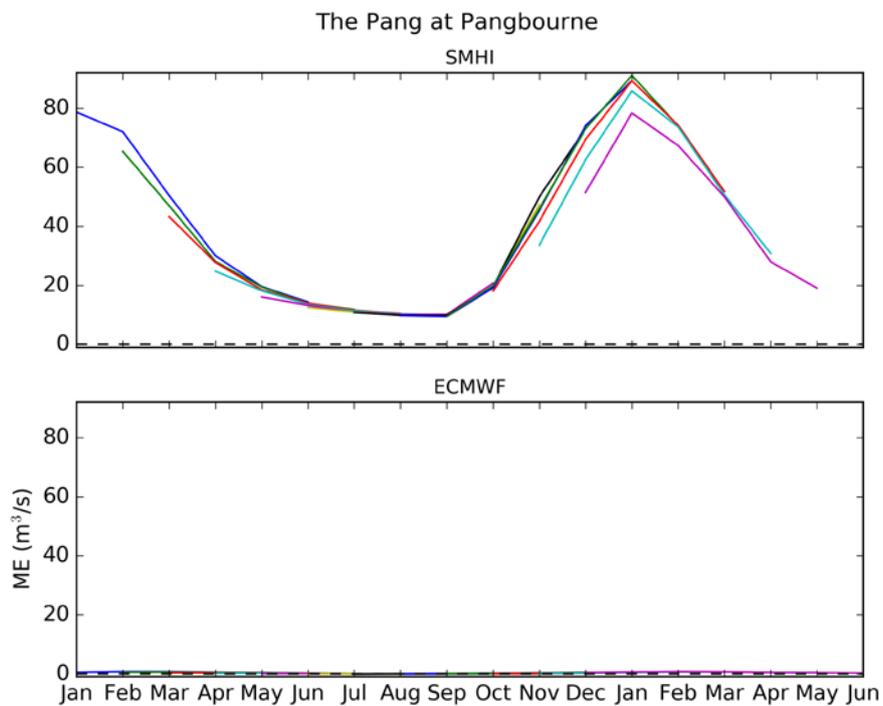


Figure 19 Same as Figure 8 but for the SMHI and the ECMWF forecasts only for the Pang at Panbourne.

In terms of reliability (i.e., CRPS reliability; see Figure 20 for an example at the Thames at Royal Windsor Park), the SMHI forecasts are overall less reliable than the ECMWF forecasts, especially for forecasts for the winter and the spring target months. The ECMWF forecasts are sometimes less reliable than the SMHI forecasts for the late summer or for the winter target months for a few stations. For both systems, the forecasts sometimes become more reliable with lead time, for a forecast made for the winter target months, which is a counter-intuitive behaviour.

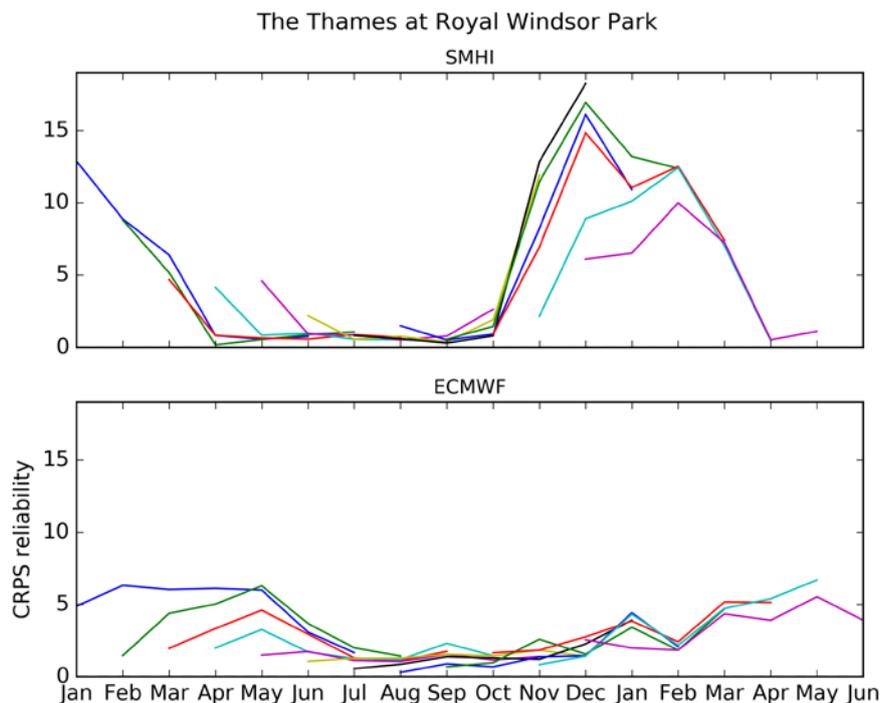


Figure 20 Same as Figure 12 but for the SMHI and the ECMWF forecasts only for the Thames at Royal Windsor Park.

In terms of the seasonal discharge forecasts skill, when compared to the observed discharge climatology (i.e., CRPSS_CLI), for several stations the SMHI and the ECMWF seasonal discharge forecasts exhibit a positive skill for the first month of lead time, which then





decreases with increasing lead time. This can be observed on Figure 21 of the CRPSS_CLI for the Thames at Royal Windsor Park for the SMHI and the ECMWF forecasts. The speed of decrease of the skill for both systems depends on the station and the event forecasted and can be negative for one month of lead time in some cases (not shown). For the Thames at Royal Windsor Park, the skill of the SMHI forecasts decreases the most from November-February, while the skill of the ECMWF forecasts decreases the most from March-July.

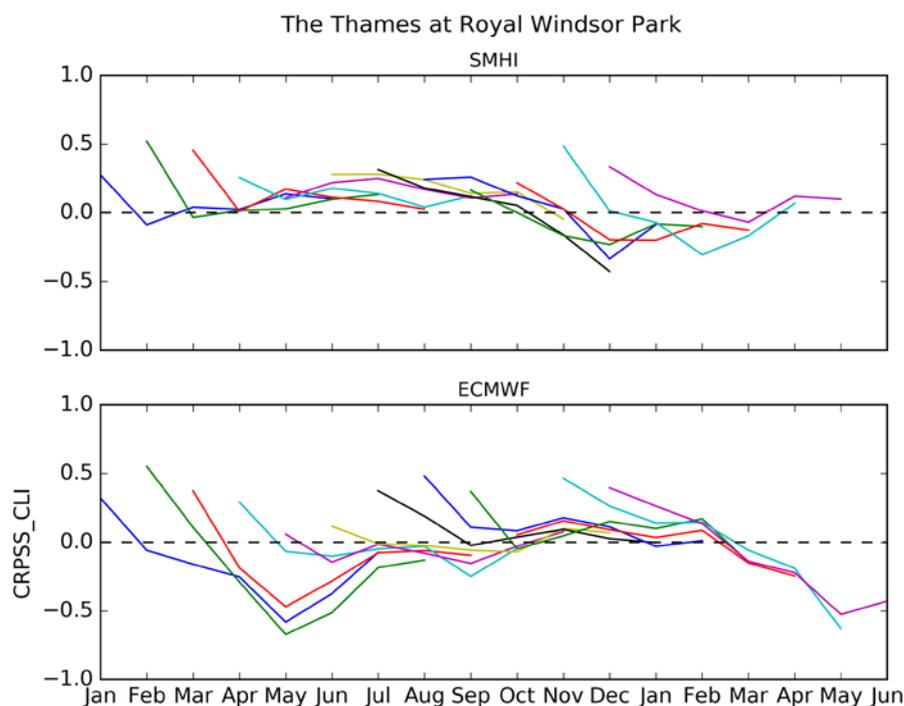


Figure 21 Same as Figure 14 but for the SMHI and the ECMWF forecasts only for the Thames at Royal Windsor Park.

The forecasts' accuracy for the lower and the upper terciles of the observed discharge is variable depending on the station and the target season. In general however, both the SMHI and the ECWMF forecasts show a similar accuracy (Brier score of 0.2 on average over all lead times and all target seasons). For several stations, the SMHI forecasts have a worst BS33 and BS66 for all target seasons and all lead times. However, the ECMWF forecasts are less accurate than the SMHI forecasts for the upper tercile for JJA and MAM target months for a few stations (not shown).



4.1.3 The Segura and Tagus River Basins

For stations of the Segura and Tagus River Basins, scores were calculated from the SMHI, the ECMWF and FW forecasting systems. From this set of stations, the SMHI forecasts are on average less accurate and sharp than the two other forecasting systems, especially for forecasts made for the winter and the spring. The ECMWF forecasts appear more accurate and sharp for forecasts made for the summer for all stations shared for this river basin. See Figure 22 as an example for the Tagus at Entrepenas.



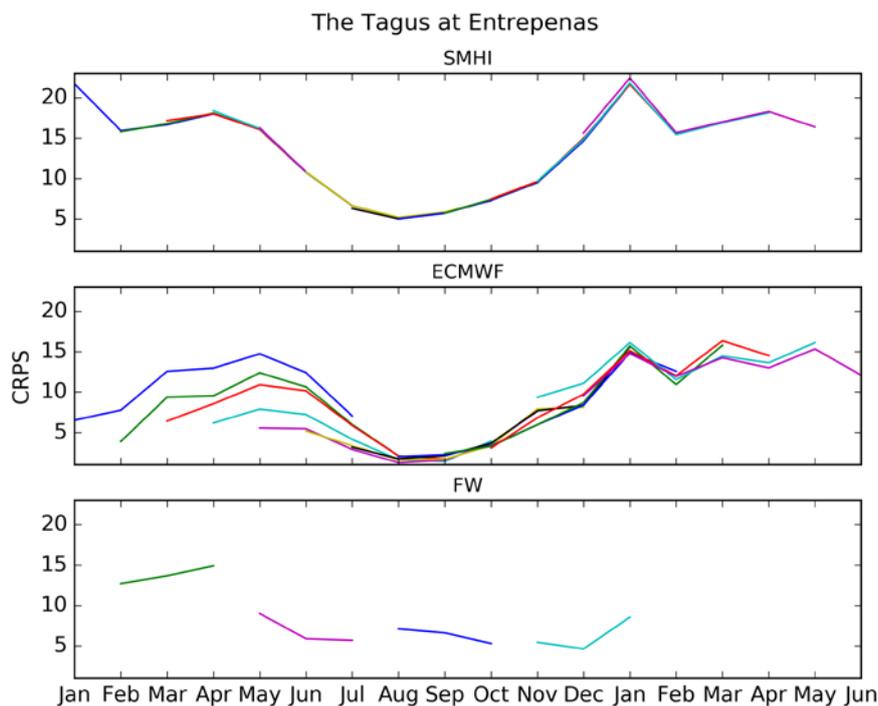


Figure 22 CRPS for the Tagus at Entrepenas for (top) the SMHI forecasts, (middle) the ECMWF forecasts and (bottom) the FW forecasts. The CRPS is given for each forecast initialisation date (on the first of each month for the SMHI and ECMWF forecasts and on the first of January, April, July and October for the FW forecasts; different colours) and for 6 months of lead time (for the SMHI forecasts), 7 months of lead time (for the ECMWF forecasts) or 3 months of lead time (for the FW forecasts).

Figures 23 and 24 show the forecast biases (i.e., the ME) for the Tagus at Entrepenas and the Segura at Cenajo, respectively. For these stations, the SMHI forecasts either underestimate or overestimate the observed discharge, more largely for the spring and the winter. The ECMWF forecasts have large biases for the winter and the spring as well, either positive or negative depending on the station. The summer biases are however the closest to zero for the ECMWF forecasts. The FW forecasts display large positive biases from February-April for the two stations on the Tagus River (see Figure 23 as an example for the Tagus at Entrepenas) and small negative biases throughout the whole year for the two stations on the Segura River (see Figure 24 as an example for the Segura at Cenajo).



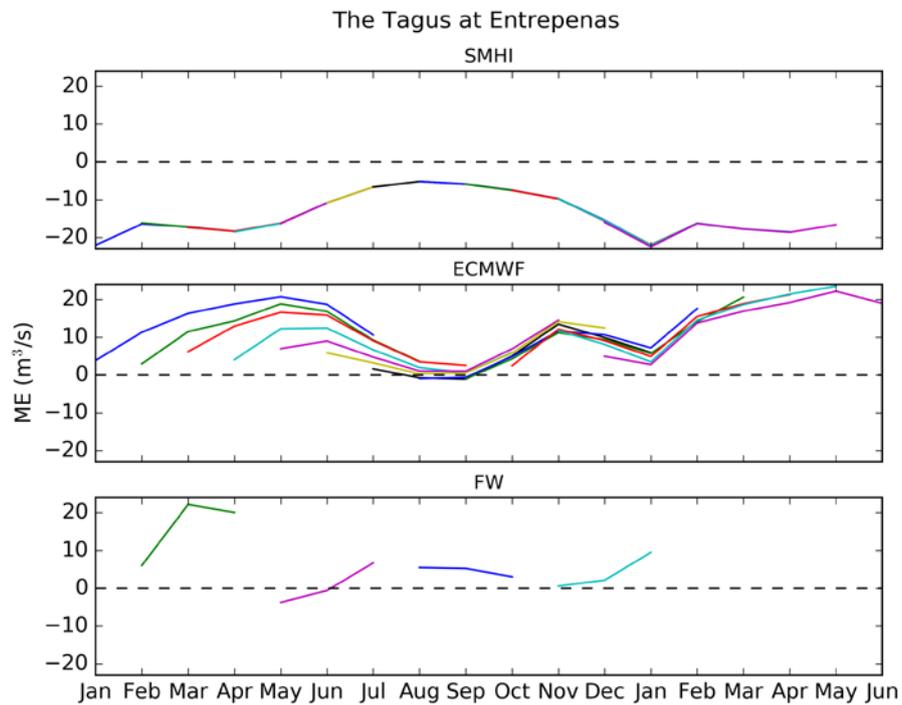


Figure 23 ME for the Tagus at Entrepenas for (top) the SMHI forecasts, (middle) the ECMWF forecasts and (bottom) the FW forecasts. The ME is given for each forecast initialisation date (on the first of each month for the SMHI and ECMWF forecasts and on the first of January, April, July and October for the FW forecasts; different colours) and for 6 months of lead time (for the SMHI forecasts), 7 months of lead time (for the ECMWF forecasts) or 3 months of lead time (for the FW forecasts).



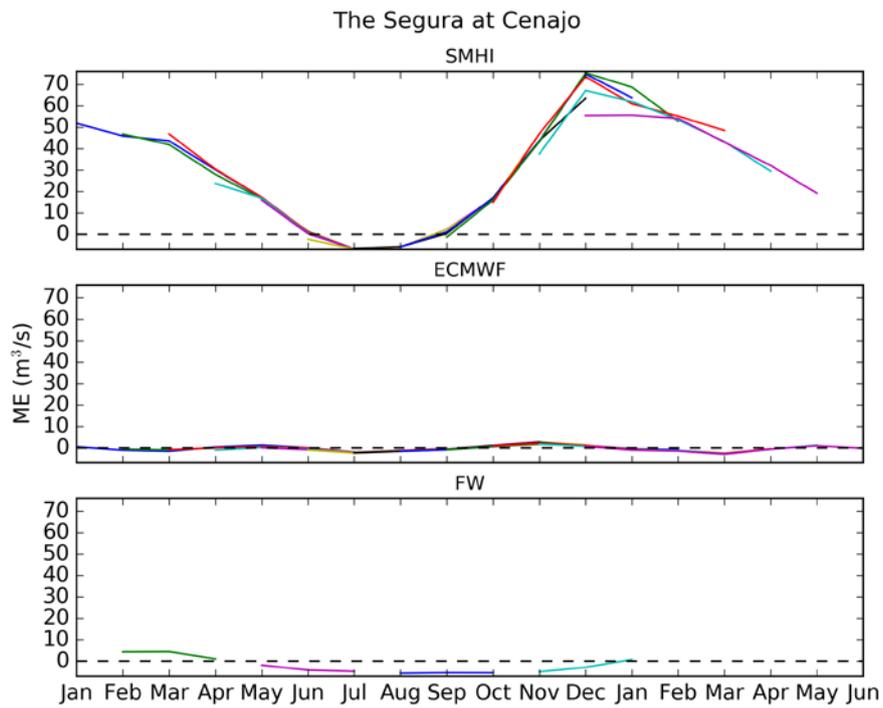


Figure 24 Same as Figure 23 but for the Segura at Cenajo.

The SMHI appears less reliable than the two other forecasts for all stations of the Segura and Tagus River Basins, especially for the winter and the spring (see Figure 25 as an example for the Tagus at Entrepenas).



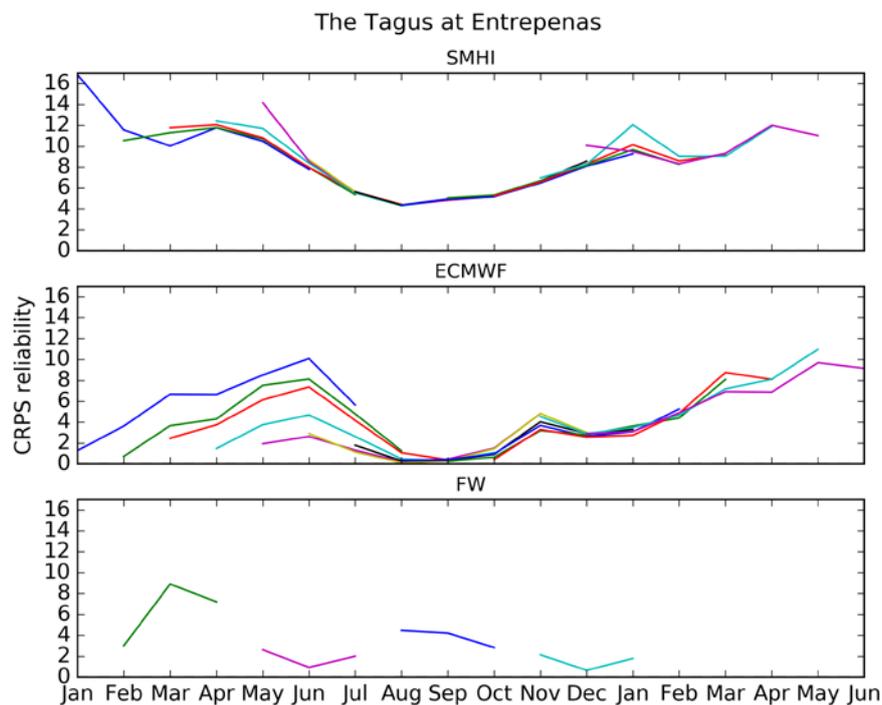


Figure 25 CRPS reliability for the Tagus at Entrepenas for (top) the SMHI forecasts, (middle) the ECMWF forecasts and (bottom) the FW forecasts. The CRPS reliability is given for each forecast initialisation date (on the first of each month for the SMHI and ECMWF forecasts and on the first of January, April, July and October for the FW forecasts; different colours) and for 6 months of lead time (for the SMHI forecasts), 7 months of lead time (for the ECMWF forecasts) or 3 months of lead time (for the FW forecasts).

In terms of the seasonal discharge forecasts skill, when compared to the observed discharge climatology (i.e., the CRPSS_CLI), the SMHI, ECMWF and FW forecasts appear less skilful for the summer. The SMHI forecasts are also less skilful than the ECMWF and the FW forecasts. Both of these results can be seen on Figure 26 of the CRPSS_CLI for the Tagus at Entrepenas.



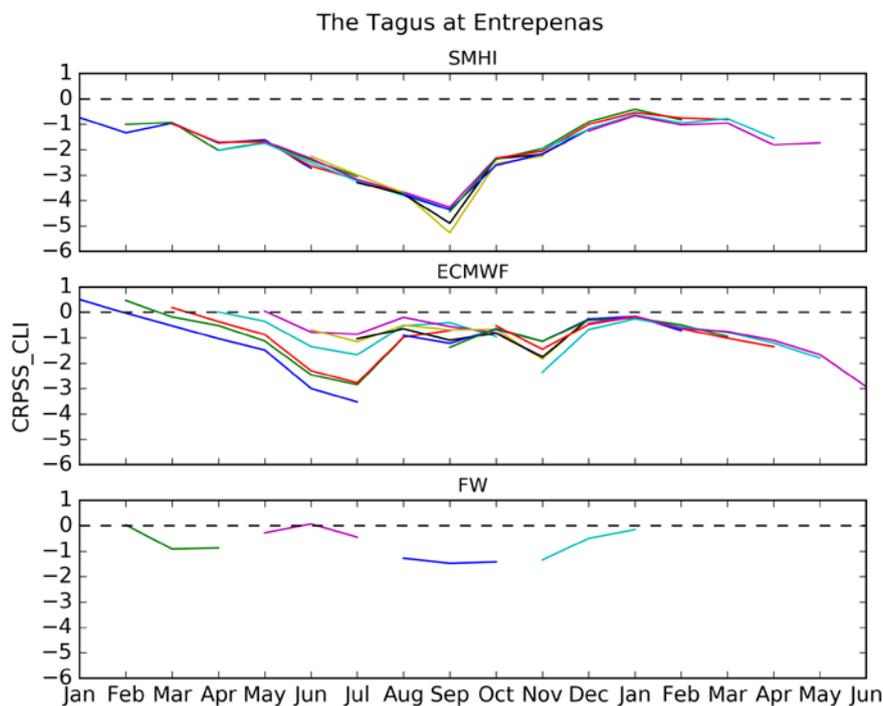


Figure 26 CRPSS of the seasonal discharge forecast against the observed discharge climatology for the Tagus at Entrepenas for (top) the SMHI forecasts, (middle) the ECMWF forecasts and (bottom) the FW forecasts. The CRPSS is given for each forecast initialisation date (on the first of each month for the SMHI and ECMWF forecasts and on the first of January, April, July and October for the FW forecasts; different colours) and for 6 months of lead time (for the SMHI forecasts), 7 months of lead time (for the ECMWF forecasts) or 3 months of lead time (for the FW forecasts).

4.1.4 The Jucar River Basin

The analysis that will be performed in following sectoral work packages (i.e., WP8-Hydroelectricity and WP11-Agriculture), as well as in WP13-Sectoral integration, must be performed using the entire river basin domain, and integrating all the relevant elements of the water resources system. This requires producing forecasts for streamflows in several sites. The reliability of the forecasts and the statistical consistency for time correlations and cross correlation between sites are crucial factors to foster the use of the forecasts in real management of the water resources system. Otherwise, the impacts of droughts would be underestimated.



Since the Jucar River Basin is strongly anthropized, all forecasts and comparisons must be done in terms of natural flows (i.e., flows that would happen if man would not produce changes due to storage and releases from reservoirs, pumping from aquifers, and diversion and return flows from consumptive uses). Natural flows provide a consistent baseline in order to compare the performance of different programmes of measures in planning and management of the basin.

Therefore, we performed a comparison between E-HYPE results and the historical data in five different points or sub-basins of the Jucar River Basin. Four of them are inflows to the main reservoirs (Alarcon, Contreras, Molinar, Tous) and the fifth is located in Sueca, at the lower part of the basin. All of these stations are crucial from the point of view of water management.

In Figure 27, the comparison between the average monthly flows in the five mentioned locations produced by the SMHI-E-HYPE model and the historical re-naturalized flows at the same locations is depicted. As it can be seen, flows produced by E-HYPE in all sites are almost zero in the summer, while the historical values are much higher. This can be explained by the important natural regulation due to aquifers upstream and in the middle section of the basin. It seems that the E-HYPE model is not able to capture this important characteristic of the basin.

On the same figure, the average monthly flows produced by the hydrological model EVALHID, which is currently used by the UPV for the Jucar River Basin, are also depicted. Flows in the summer produced by the EVALHID model are much closer to the historical values for all stations.



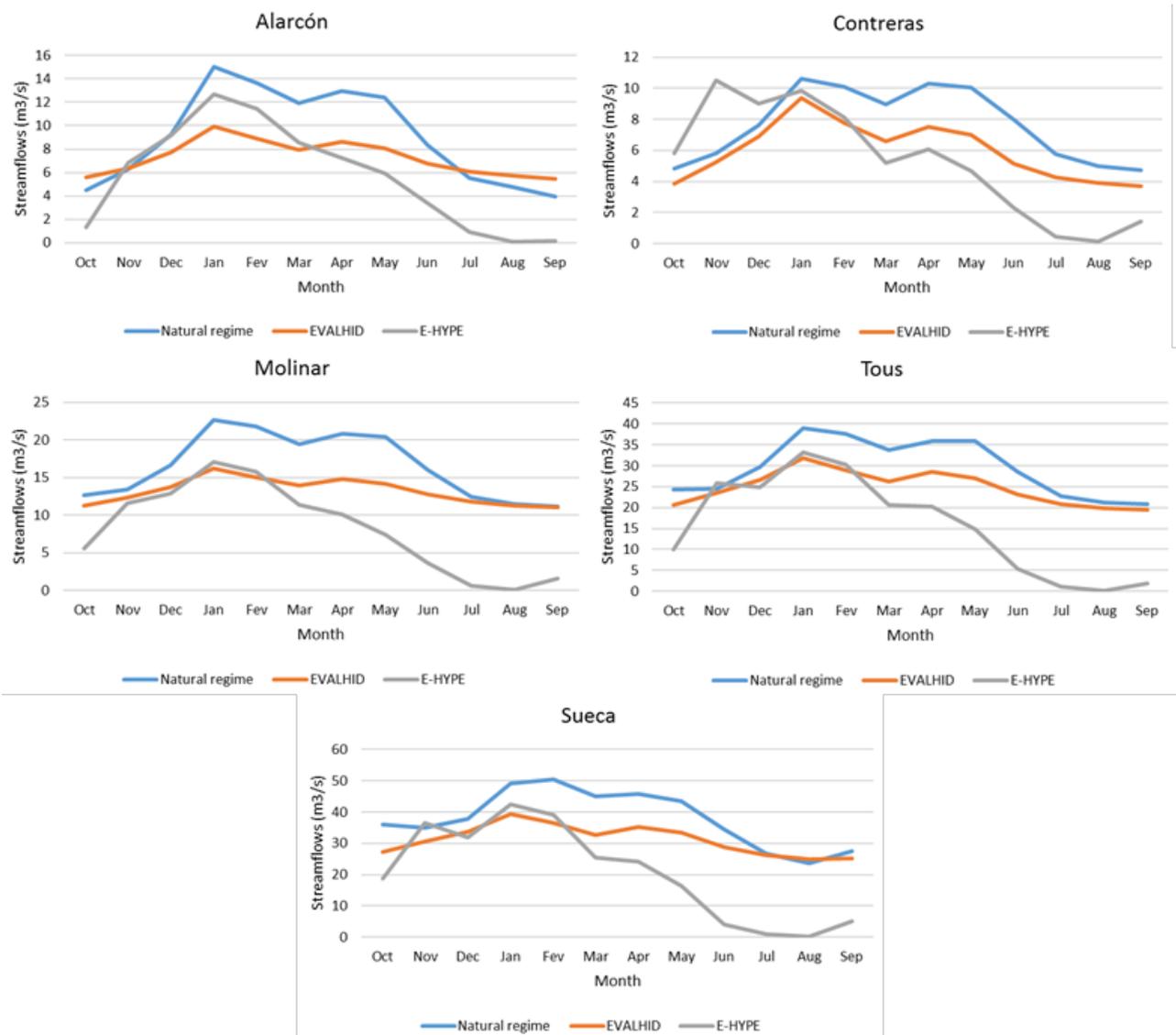


Figure 27 Comparison between streamflows from SMHI Pan-European data (E-HYPE), historical data as naturalized river flows and data from our hydrological model (EVALHID) in the five main sub-basins of Júcar River Basin: Alarcón, Contreras, Molinar, Tous and Sueca.

This essential mismatch between the E-HYPE results and observed values cannot be overcome by any bias correction, since in the summer months rainfall is almost negligible. This is only feasible by means of a conceptual modification of the model and recalibration in order to capture the real behaviour of the basin by the model. This will be discussed in the following months of the IMPREX project.



For this deliverable, forecasts produced by the EVALHID hydrological model forced with ECMWF meteorological data as input (i.e., precipitation and mean temperature) is compared to the ECMWF forecasts.

Figure 28 displays the CRPS obtained from the UPV forecasts for all five stations of the Jucar River Basin. From this figure, it appears that the forecast performance varies depending on the selected station, but the forecasts are on average more accurate and sharp in the spring and summer. The largest errors can be found in the winter months.

Figure 29 displays the CRPS for the ECMWF forecasts for the Jucar at Alarcon and at Tous. Results for these stations show a different behaviour of the ECMWF forecasts. For the Jucar at Alarcon, the largest errors are observed for the winter, similarly to what was observed for the UPV forecasts. For the Jucar at Tous however, the largest errors are situated in the summer.



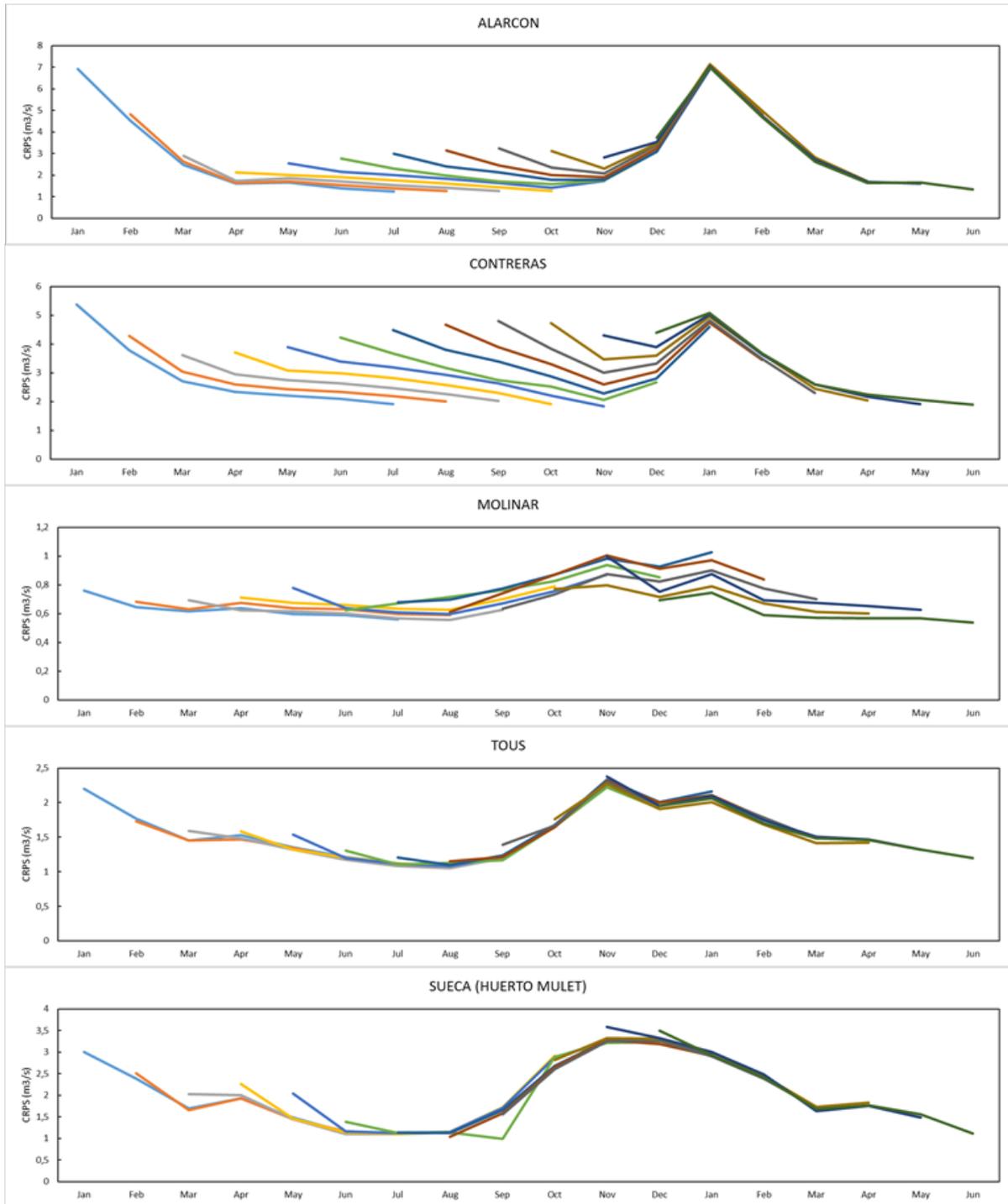
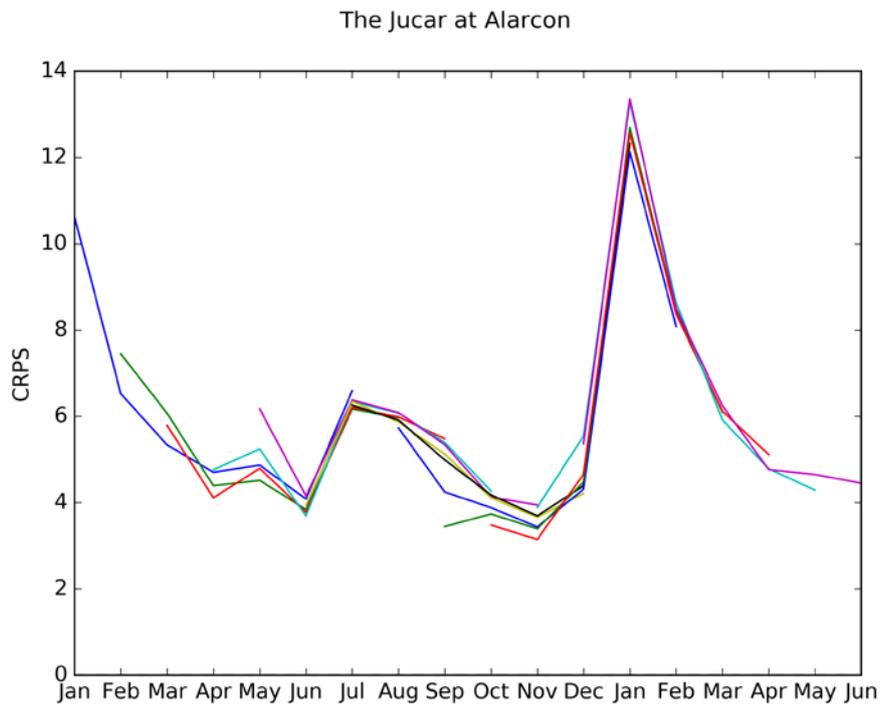


Figure 28 CRPS for the Jucar River at Alarcon, Contreras, Molinar, Tous and Sueca for the UPV forecasts. The CRPS is given for each forecast initialisation date (on the first of each month, different colours) and for 7 months of lead time.





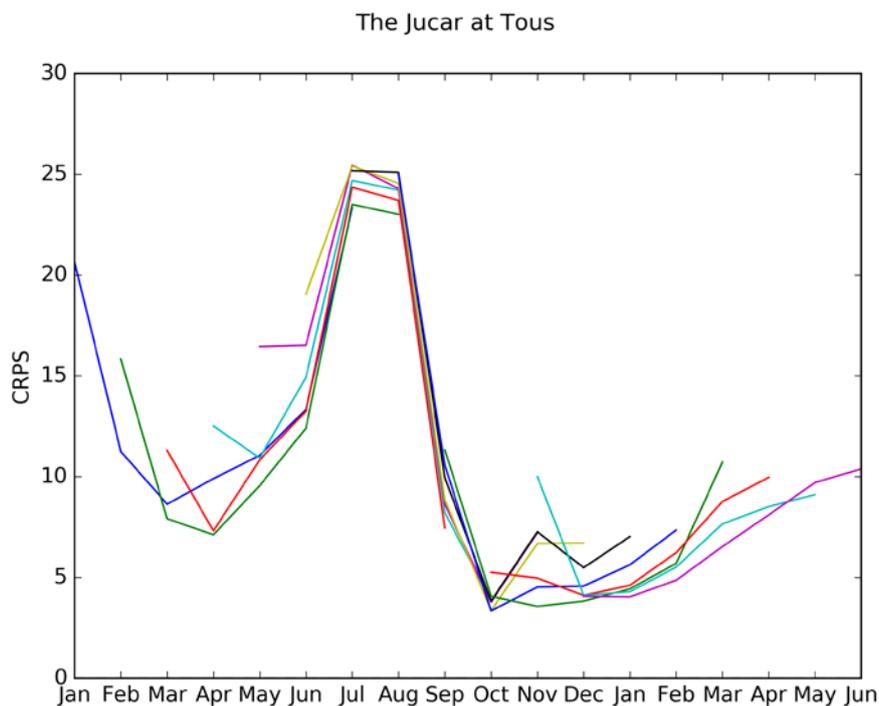
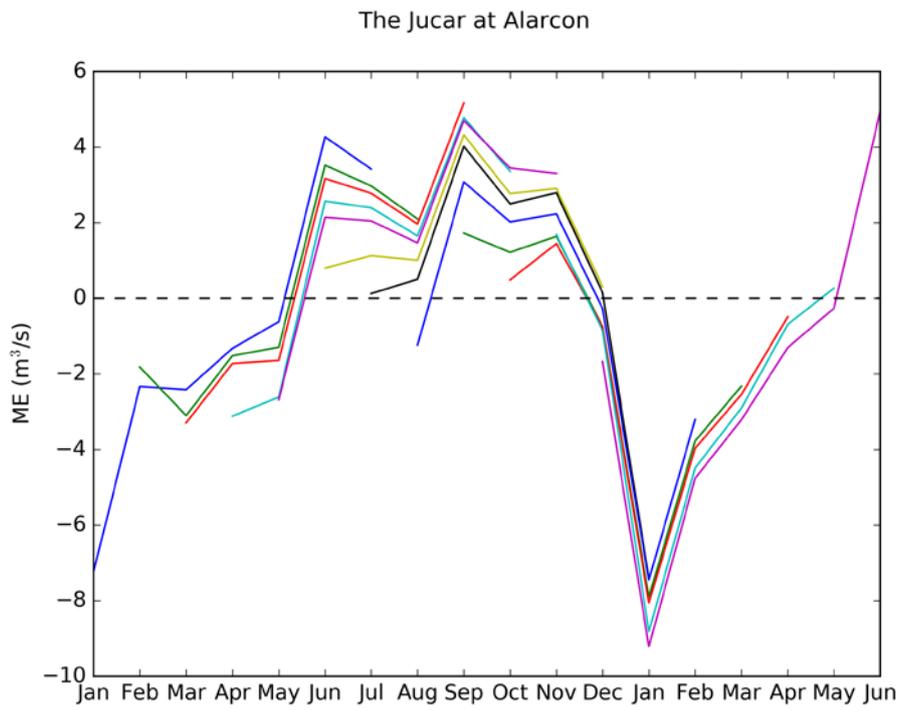


Figure 29 CRPS for (top) the Jucar at Alarcon and (bottom) the Jucar at Tous for the ECMWF forecasts. The CRPS is given for each forecast initialisation date (on the first of each month, different colours) and for 7 months of lead time.

Regarding to the UPV forecasts biases, different behaviours can be seen depending on the selected station (see Figure 30). For the Alarcon and Contreras stations, bias is positive all year except for January when it reaches negative values. This could be due to the fact that those two stations are situated in the mountainous headwaters of the river, where the faster discharge generating processes are likely misrepresented by the model and/or the precipitation is underestimated for this time of the year at those stations. The Molinar and Sueca stations display a slight positive bias all year long, while the Tous presents a negative bias all year.

For the ECMWF forecasts (see Figure 31), the biases are similar to the UPV forecasts biases for the Jucar at Alarcon, which could be due to the same model misrepresentation and/or precipitation underestimation. For the Jucar at Tous, the ECMWF forecasts present slightly different and much larger biases than the UPV forecasts, with negative biases in the summer and positive biases the rest of the year. This hints towards an underestimation of the groundwater discharge by the ECMWF forecasts for this station.





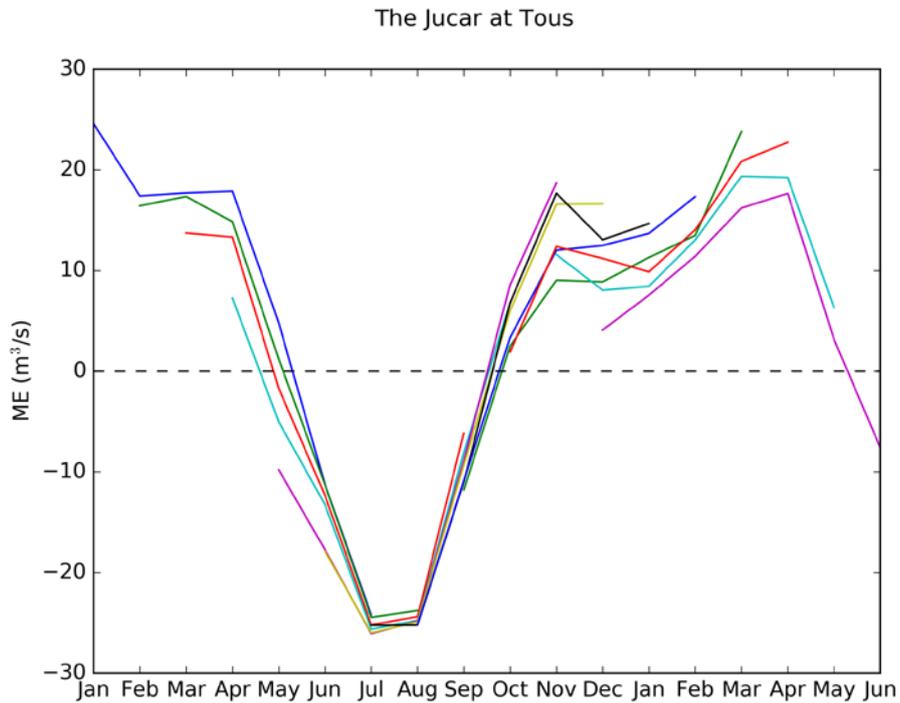


Figure 31 ME for the Jucar River at Alarcon and Tous for the ECMWF forecasts. The ME is given for each forecast initialisation date (on the first of each month, different colours) and for 7 months of lead time.

4.1.5 Swedish Rivers

For stations on the Swedish Rivers, scores were calculated from the SMHI and the ECMWF forecasting systems. From this set of stations, the SMHI and the ECMWF forecasts appear equally accurate and sharp, with larger errors from May-September for all four stations shared. This can be seen on Figure 32, which displays the CRPS for the Vindelaelven at Granaker for both systems.



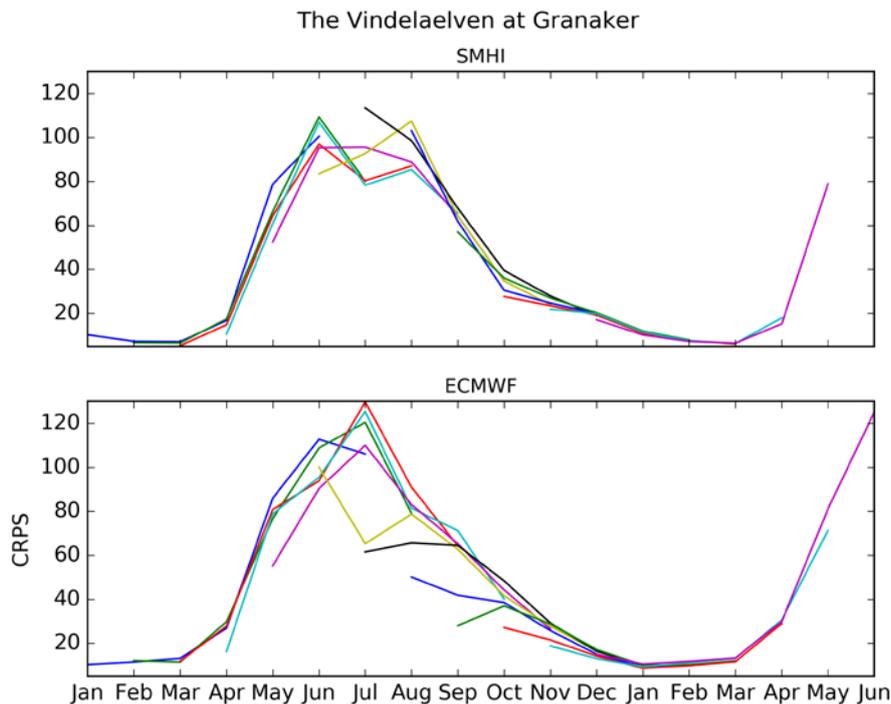


Figure 32 Same as Figure 16 but for the Vindelaelven at Granaker.

Figure 33 displays the bias (i.e., the ME) for the Vindelaelven at Granaker for both systems. On this figure, the SMHI forecasts appear to underestimate both May and July-September observed discharges, while they overestimate the June observed discharge. The ECMWF forecasts underestimate the May observed discharge and overestimate largely the June-August observed discharge. This last characteristic of the ECMWF forecasts is however only seen for forecasts made in May or earlier. For forecasts made in June or after, the June-August observed discharge is underestimated.

These behaviours of the SMHI and the ECMWF forecasts were observed for the other shared stations for Swedish Rivers. For the SMHI forecasts, this could be due to a large underestimation of groundwater storage and recharge in the winter, subsequently leading to underestimated flows in the summer. For the ECMWF forecasts, the biases could be due to a delayed snowmelt process in the model, either due to model errors or to biased seasonal temperature forecasts input into the model.



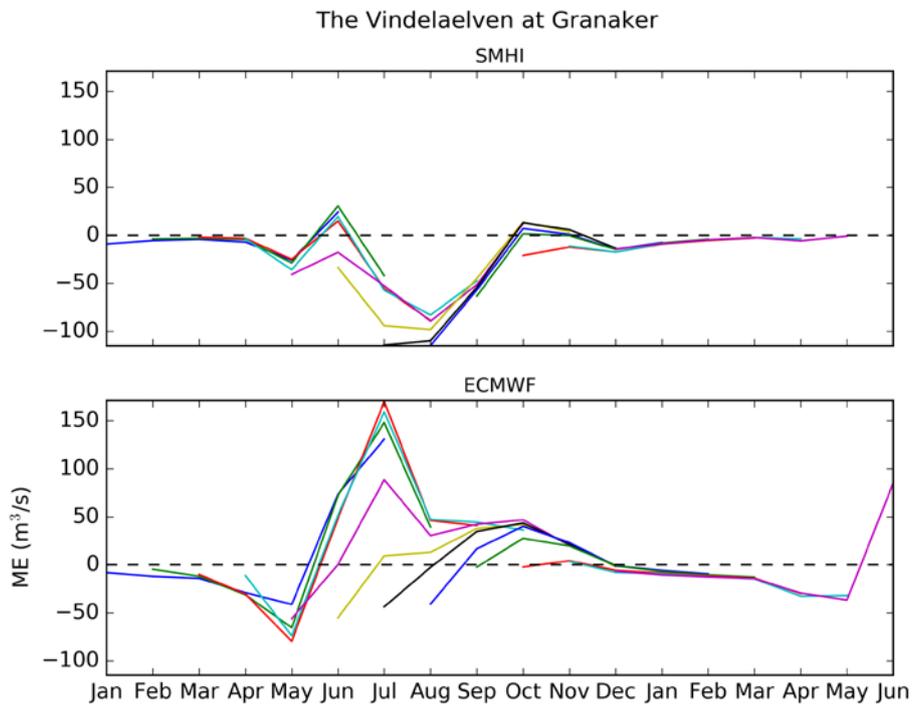


Figure 33 Same as Figure 18 but for the Vindelaelven at Granaker.

For both systems, the reliability is the worst during from June-August (see Figure 34), approximately when the largest CRPS errors were observed (see Figure 32).



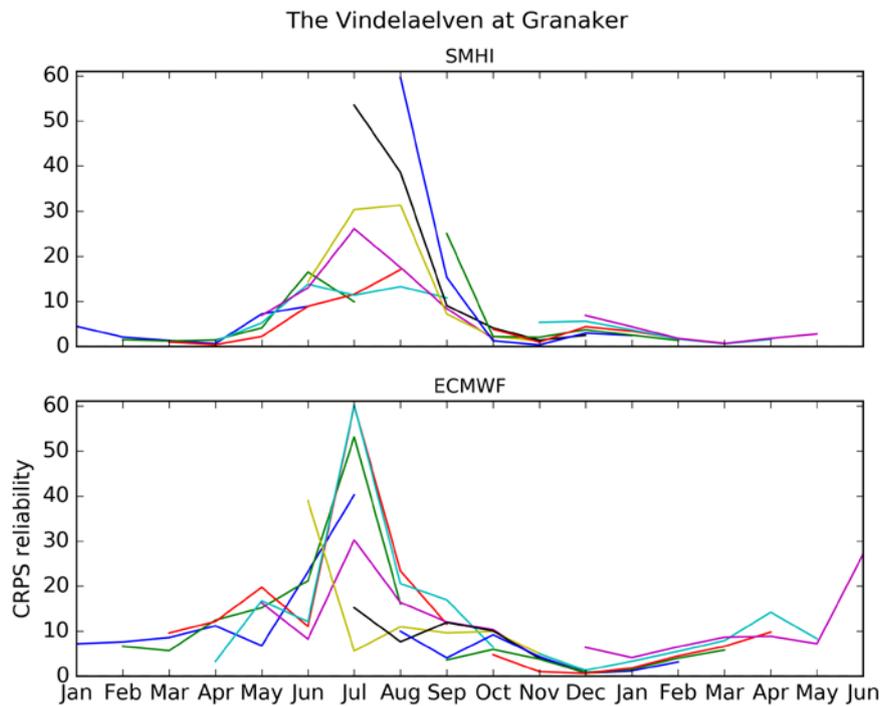


Figure 34 Same as Figure 20 but for the Vindelaelven at Granaker.

In terms of the seasonal discharge forecasts skill, when compared to the observed discharge climatology (i.e., the CRPSS_CLI), the ECMWF forecasts appear less skilful from February-April and in July (see Figure 35 as an example of this general behaviour for the Vindelaelven at Granaker). The skill of the SMHI forecasts depends given the station looked at, but they are generally more skilful than the ECMWF forecasts.



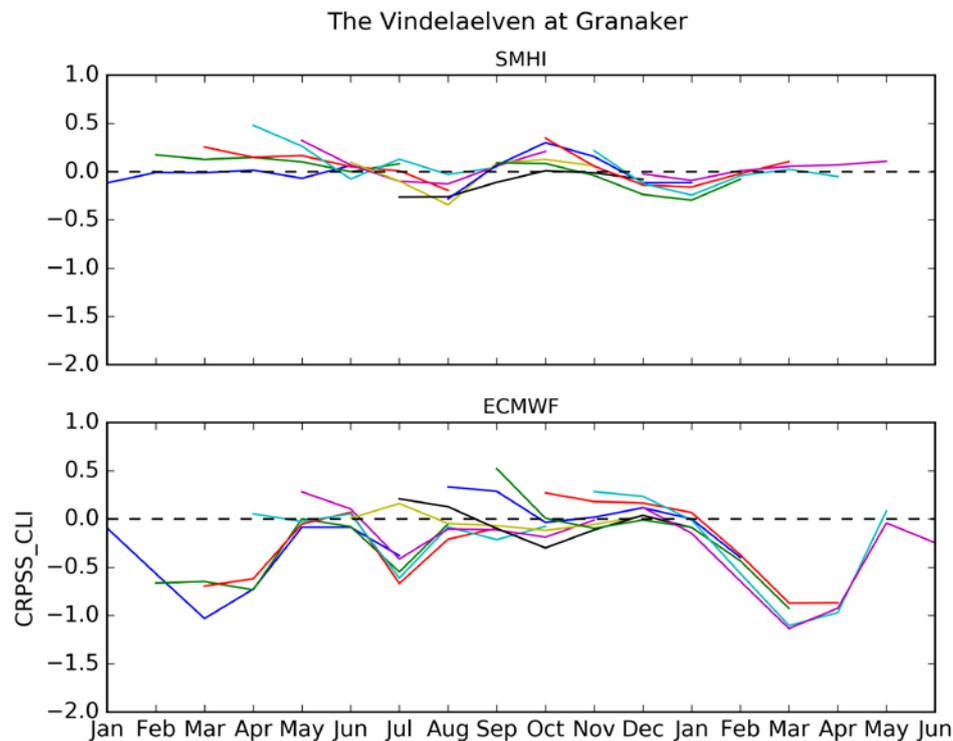


Figure 35 Same as Figure 21 but for the Vindelaelven at Granaker.

The forecasts' accuracy for the lower and the upper terciles of the observed discharge is variable depending on the station and the target season. In general however, both the SMHI and the ECMWF forecasts show a similar accuracy (Brier score of 0.2-0.3 on average over all lead times and all target seasons). For several stations, the ECMWF forecasts have a worst BS33 for DJF for all lead times (not shown). Both systems exhibit a very good accuracy (BS66 close to 0) for the upper tercile for forecasts made for MAM at all lead times (not shown).

4.2 The EPB sensitivity analysis

Figure 36 shows maps of the dominant predictability source (IHC or SCF), the predictability source for which the skill elasticity is highest and which could therefore lead to higher seasonal discharge forecast skill after being improved. The skill elasticities were derived from forecasts produced by the ECMWF seasonal hydrological forecasting system, using the CRPSS calculated against the observed climatology. The results are shown for 74 regions





across Europe, as these regions are the same as the ones used for the seasonal outlook in EFAS. The maps were made for each forecast initialisation date (on the first of each month; each row) and for seven months of lead time. However, only the first three months of lead time (each column) are shown here as the impact of initialisation tends to disappear for lead times exceeding 3 months for most regions in Europe.

From the maps one can see that on average, for the first month of lead time, improving the IHC would lead to a higher discharge forecasting skill. As lead time increases, the relative importance of IHC to SCF decreases and improving the SCF becomes more important to improve the discharge forecasting skill. There are however temporal and spatial variabilities.

For the first month of lead time, the density of regions for which the IHC are relatively more important than the SCF is higher for forecasts starting from May to July, with the largest density in June. This is probably because from May to July, rainfall is low in most parts of Europe, leading to groundwater dominated discharges for most European basins.

For most regions in Scandinavia, the IHC appear to dominate the uncertainty for forecasts started in the winter, with a signal that persists until three months of lead time (and further, not shown). This is maybe due to precipitation falling as snow during those months in these regions, leading to a more groundwater based discharge. Furthermore, a good knowledge of the antecedent snow content will lead to a high skill in spring, when discharge is snowmelt driven in those regions. This is however not the case for windward Scandinavia, where the discharge is mostly sensitive to the SCF. This could be due to weather systems raining out on Scandinavia's western part, leading to a rainfall dominated discharge. Moreover, the ground memory is very low in this part of Scandinavia.

Over the Iberian Peninsula, the IHC dominate the uncertainty for forecast generated in summer (June to September), a signal which persists until three months of lead time. The reason for this pattern is the very dry climate over the Iberian Peninsula during the summer months, leading to a mainly groundwater dominated discharge with long memory over several months.

In central Europe, the eastern side appears to be more IHC dependent for the first month of lead time than the western part. This is probably mostly due to weather systems raining out on central Europe's west coast. The IHC importance in Eastern central Europe could also be due to snowmelt drive discharge in spring.



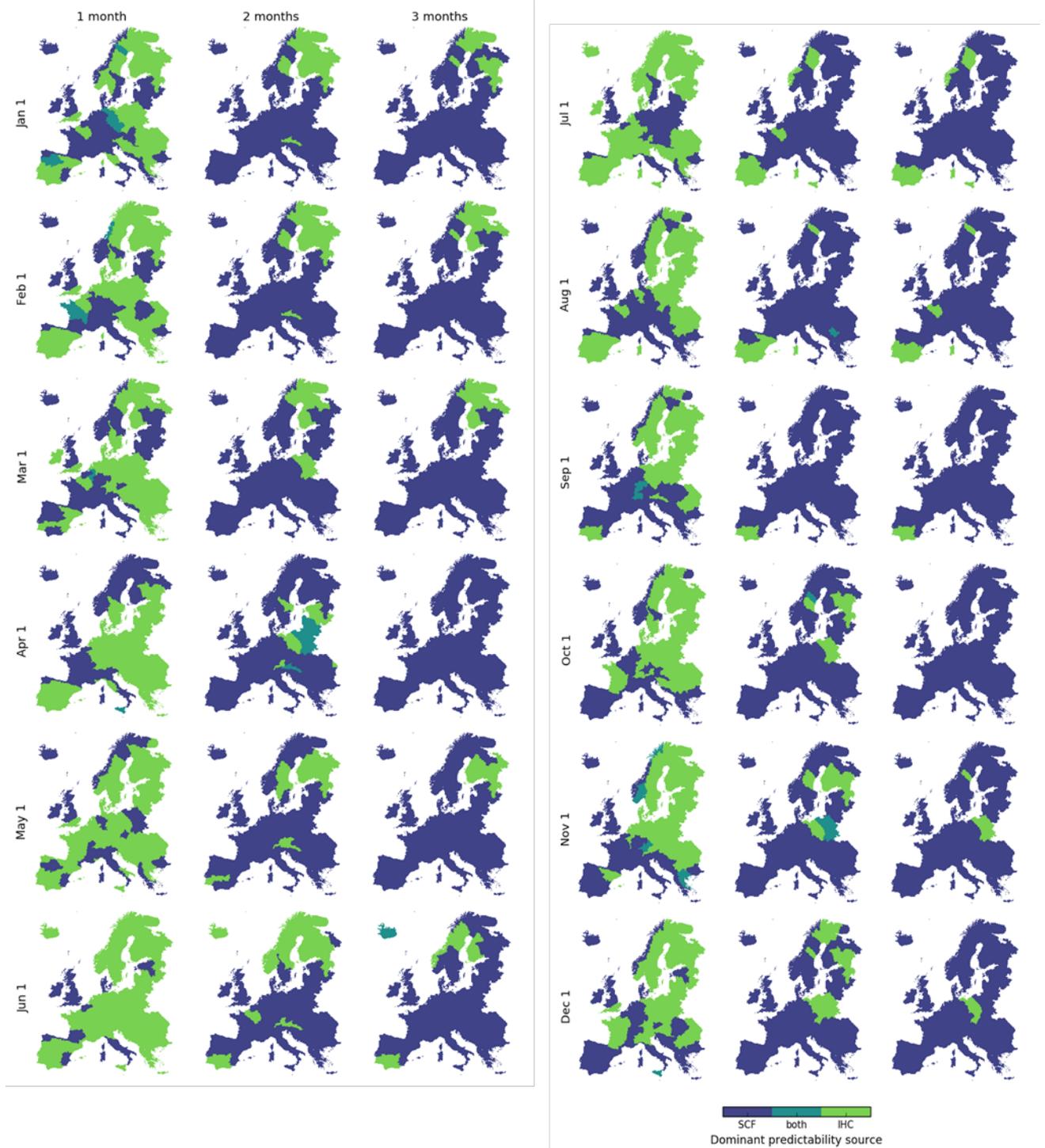


Figure 36 Maps of the dominant predictability source for each forecast initialisation date and the first three months of lead time for the EFAS regions across Europe. Blue [green] colours signify that the SCF [IHC] form the dominant source of predictability.





Figures 37 to 42 display the skill elasticities obtained for several EFAS regions (from the ECMWF forecasts) and for the BfG stations falling in each of those regions (from the BfG forecasts). Results are shown for all forecast initialisation dates and for the first and second months of lead time. The seasonal discharge forecasting skill elasticity to SCF (E_{SCF_i} in blue) and the seasonal discharge forecasting skill elasticity to IHC (E_{IHC_i} in green) indicate the potential to improve the seasonal discharge forecasting skill as a result of improving the quality of those respective predictability sources. When one skill elasticity is larger than the other, it implies that this predictability source has the largest potential to improve the seasonal discharge forecasting skill (for that specific station or region, lead time and forecasting initialisation date) once improved. These figures allow a comparison of the sensitivities of the ECMWF and the BfG seasonal discharge forecasts to the IHC and the SCF. Overall, one can see that the skill elasticities obtained for both the EFAS regions and the BfG stations from the two different forecasts are very similar. There are however slight differences, such as the larger relative importance of the IHC for the forecasts made in the spring for the EFAS region of Figure 37, compared to the corresponding station (The Inn at Passau Ingling). These differences between the skill elasticities for the EFAS regions and the BfG stations could be due to differences between the two systems for which the sensitivity analysis was performed as well as the scale (regional or at a station) at which the analysis was made.



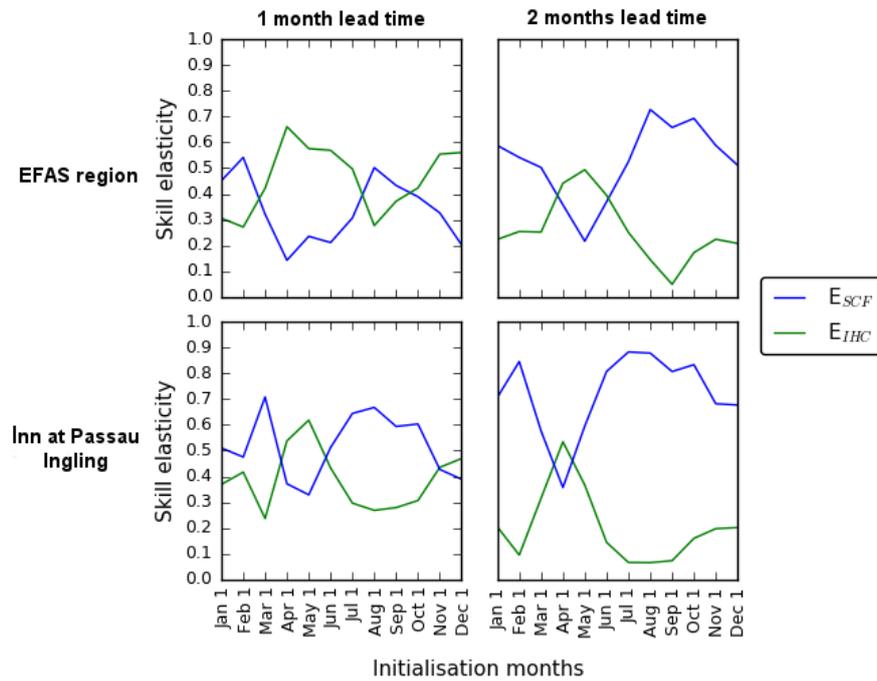


Figure 37 Skill elasticities for (left) the first and (right) the second month of lead time, for (top) the EFAS region and (bottom) the corresponding BfG station for the Inn at Passau Ingling. Skill elasticities are shown for each forecast initialisation month.

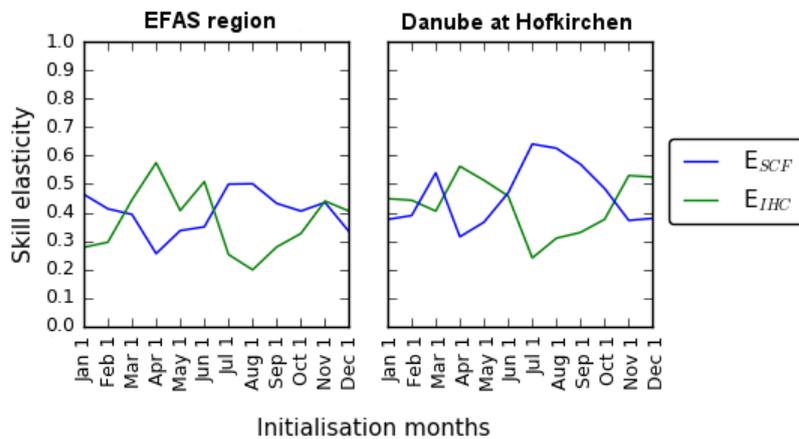


Figure 38 Skill elasticities for (left) the first and (right) the second month of lead time, for (top) the EFAS region and (bottom) the corresponding BfG station for the Danube at Hofkirchen. Skill elasticities are shown for each forecast initialisation month.



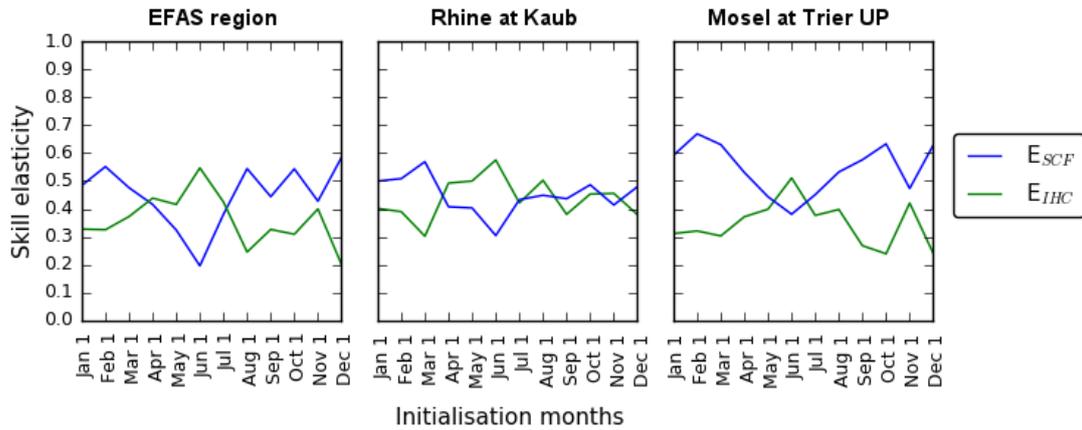


Figure 39 Skill elasticities for (left) the first and (right) the second month of lead time, for (top) the EFAS region and (bottom) the corresponding BfG stations for the Rhine at Kaub and the Mosel at Trier UP. Skill elasticities are shown for each forecast initialisation month.

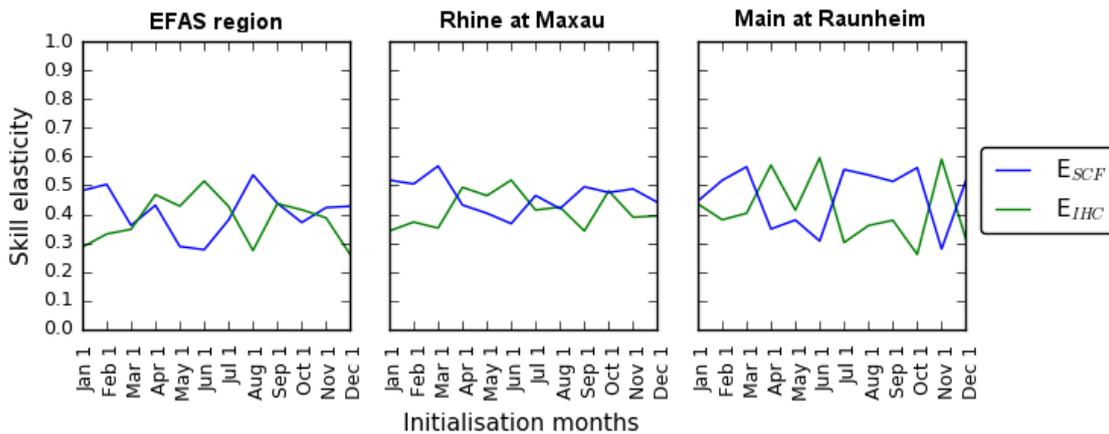


Figure 40 Skill elasticities for (left) the first and (right) the second month of lead time, for (top) the EFAS region and (bottom) the corresponding BfG stations for the Rhine at Maxau and the Main at Raunheim. Skill elasticities are shown for each forecast initialisation month.



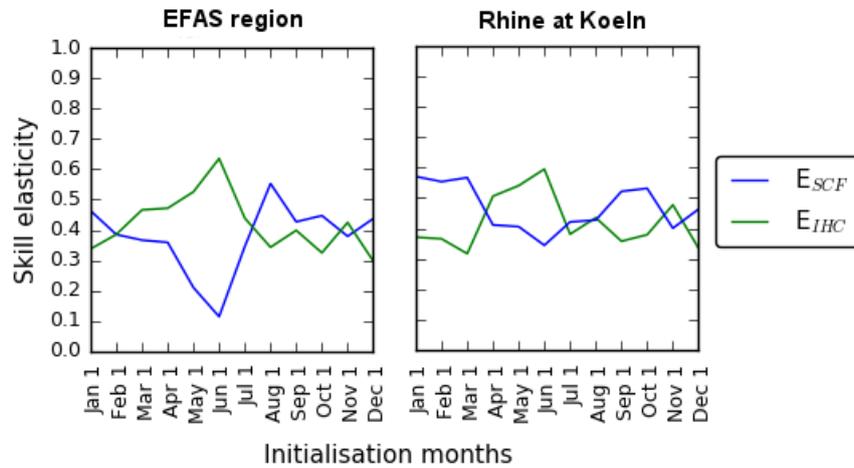


Figure 41 Skill elasticities for (left) the first and (right) the second month of lead time, for (top) the EFAS region and (bottom) the corresponding BfG station for the Rhine at Koeln. Skill elasticities are shown for each forecast initialisation month.

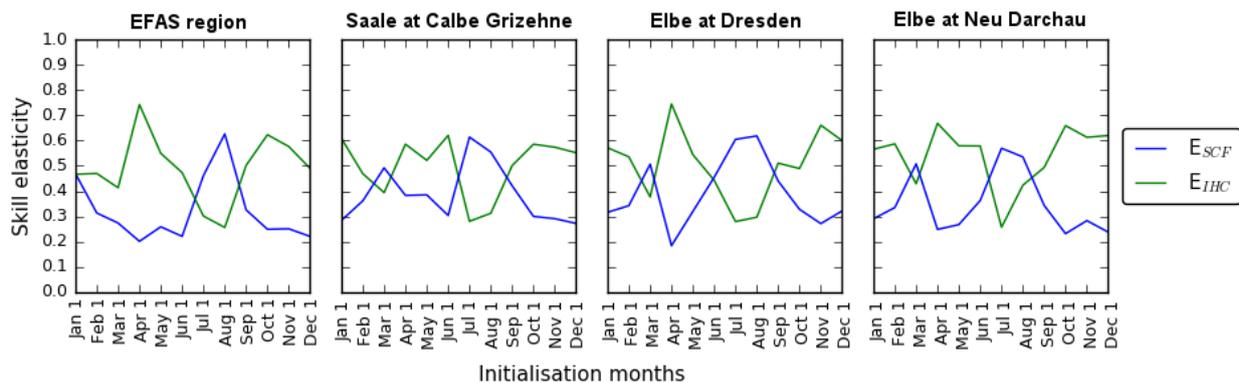


Figure 42 Skill elasticities for (left) the first and (right) the second month of lead time, for (top) the EFAS region and (bottom) the corresponding BfG stations for the Saale at Calbe Grizehne, the Elbe at Dresden and the Elbe at Neu Darchau. Skill elasticities are shown for each forecast initialisation month.

4.3 Comparative analysis

For spatial interpretation of hydrological skill, we investigated potential relationships between predictive skill and physiographic-hydrological-climatic characteristics; hence allowing to identify the key controls of poor/good model skill. First the 15 descriptors (see





Table 3) were analysed for inter-dependence, and one of the highly inter-dependent descriptors was omitted to avoid potential artefacts in the CART regression analysis. Consequently a set of nine significant descriptors was statistically identified for application in the CART analysis, which further allowed us to estimate the descriptors' importance.

Figure 43 shows the ranking of nine descriptors (ranked by importance, with number 1 being the most important descriptor) for all months and lead months. Results show that the dominant descriptors resulting in poor/good model performance are the FlowID (describing the hydrological behaviour of the basin), elevation and remaining bias in temperature (BiasTemp). It is generally expected that remaining biases in temperature will have an impact on the form of precipitation (rainfall or snowfall) during the cold months, and the processes (i.e. changing from (to) snow accumulation to (from) melting). For example, this occurs in northern Europe for April where the mean average temperatures for April is close to 0°C and hence small deviations in the meteorological forecasts will affect the basin response. Elevation (Elev.) is also an important factor. It is expected that the meteorological forecasts are reliable in predicting the climatological variability in highly elevated basins, which are usually snow dominated. Consequently the hydrological regime can be better described in comparison to a rain-fed basin. The basin's hydrological behaviour (FlowID) seems to be the most important descriptor with basins of similar river flow properties achieving similar skill. It is known that river systems experience processes with high memory in comparison to the natural phenomena occurring in the atmosphere. Hence it is expected that hydrological variables (i.e. discharge, runoff, soil moisture) can have higher predictability than meteorological variables (i.e. precipitation, temperature). However, this cannot be linearly translated since the precipitation-discharge process is also not linear, and therefore different systems are expected to respond differently to the meteorological signal.



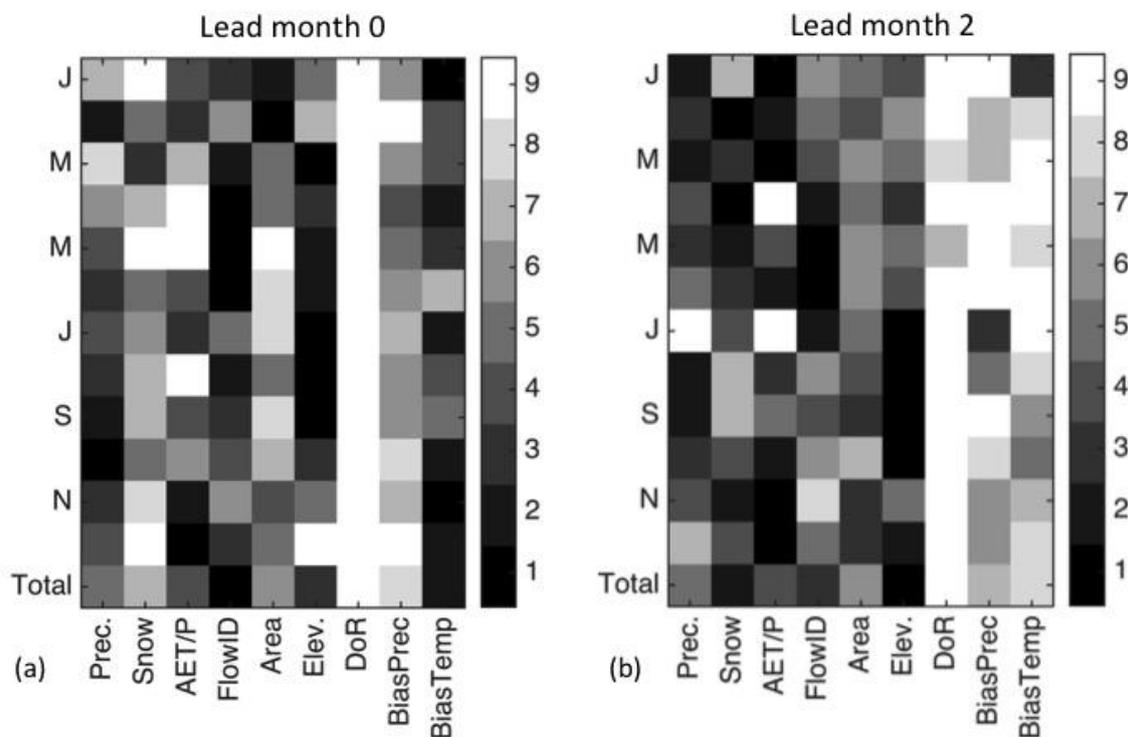


Figure 43 Importance ranking of key descriptors that influence the hydrological forecasting skill over Europe for all months and in lead month: (a) 0, and (b) 2.

To get a better understanding of the basin characteristics that are characterised by a good/poor skill, Figure 44 shows the 11 spatially variable clusters, their distribution of flow signatures, and the distribution of skill in each cluster group. Similarity in catchment behaviour for each class was interpreted and dominant flow generating processes could be distinguished.

Results give a clear separation between basins with poor and good skill. Basins in cluster 5 achieve the highest skill. These basins are characterised by high ranges of baseflow (BFI), low monthly variability (intra-annual variability) (DPar), and high values of low and medium flows (q_{95} and q_{70}). These are properties of basins where short-memory precipitation is aggregated and converted into long-memory discharge. Similar behaviour have the basins in clusters 6, 7 and 9, however not to the distinct level of basins in cluster 5. Basins in cluster 8 and 10 are short-memory rivers characterised by flashy response and high seasonal





variability (DPar and CV). These basins are responding quite fast to the precipitation signal and with strong dynamics (RLD) whilst contribution from base flow is small (BFI). Basins that belong to clusters 1, 2 and 3 perform adequately and are generally characterised by the same flow signatures. These basins are mainly located in the Scandinavian region and also in the central Europe at highly elevated regions. They are distinct for their medium to high slope in their flow distribution (mFDC), which is an indicator of a regime driven by snowmelt.

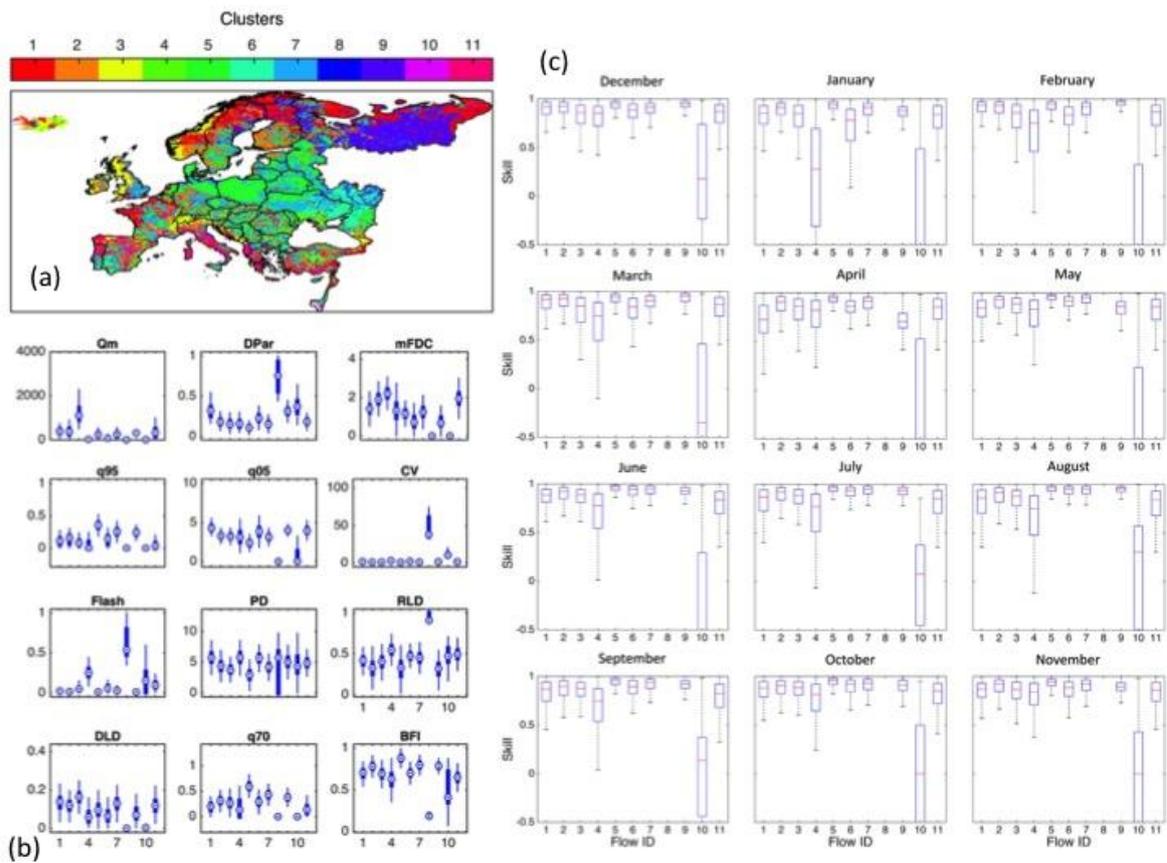


Figure 44 (a) Spatial distribution of hydrologically similar (clusters) basins over Europe, (b) distribution of flow signatures in each cluster group (see also Table 3), and (c) distribution of beta skill in each cluster group



5 Lessons learnt

The intercomparison of seasonal discharge quality from forecasting systems from the ECMWF, the SMHI, the BfG, the FW and the UPV done in a first part of this deliverable is a starting point for this larger task within the IMPREX project. Through this intercomparison, multiple scores of forecast quality were added to the scoreboard developed in WP4 (deliverable 4.1) by the partners of this deliverable for stations within each system's spatial boundaries.

Although the sample of stations for which the intercomparison was made was limited, this task has already revealed several major differences between the seasonal hydrological forecasting systems and their impacts on the relevant water sectors.

The BfG forecasting system overall mostly underestimates the observed discharge for stations shared for the Central European Rivers. This could be problematic for the navigation sector, who is most vulnerable to low flows in summer. These forecasts could potentially lead to an underestimation of the expected river flow in summer and consequently to an underestimation of the capacity of the river and a monetary loss.

The ECMWF forecasting system appears to almost systematically overestimate the spring flow and underestimate the winter flow for the stations shared for Central European Rivers and for the Thames River Basin. The underestimation of the winter discharge could be a problem for the flood protection sector, as it would not flag regions to watch for potential floods in the coming months. For the Segura and Tagus River Basins, the ECMWF forecasts are very accurate in summer. This could be highly beneficial for the agriculture sector in this region, which relies on accurate drought forecasts for the summer. However, the summer flow is highly biased (as well as the winter flow) in the Jucar River Basin, also in Spain. In Sweden, the ECMWF forecasts underestimate the May flow and overestimate the summer river flow. This could be problematic for the hydropower industry, for which there is a particular interest in forecasting the spring flow accurately.





The SMHI forecasting system overestimates the flow in winter and spring for the Central European Rivers and the Thames River Basin. This could be a problem for flood forecasting as it could indicate potential floods in the coming months when none actually occurs. The SMHI is also overall less accurate for the lower tercile of observed discharge for summer for a few Central European River stations. This could be a limitation for the navigation sector as the forecasts would not be able to capture accurately a low extreme event in the summer. The May and summer flow in Sweden appears underestimated by the SMHI forecasts, while the June discharge is overestimated. This could be an issue for the hydropower sector, which needs accurate forecasts especially in spring.

The FW forecasting system overestimates largely the early spring-spring flow for the Tagus River Basin and underestimates slightly the flow during all year for the Segura River Basin. Both biases could be challenging for the agricultural sector.

The UPV forecasts are overall greatly improved by using the EVALHID hydrological model compared to the E-HYPE model, especially in summer. There are however still some biases which need to be overcome before the forecasts can be used operationally for reservoir management purposes in the Jucar River Basin.

For most stations, after one to two months of lead time, using the observed flow climatology leads to more accurate and sharp forecasts than using seasonal hydrological forecasts. This shows that there are still model and meteorological forecast biases which need to be overcome in order to gain a real valuable additional from using seasonal hydrological forecasts operationally for many applications of the water sector.

These results are a starting point, to which it will be possible to add more results along the course of the project. Indeed, as the project proceeds, anyone will be able to upload additional scores, scores for different stations or from a new or modified system to the scoreboard. The latter will enable us to monitor and visualise progresses made throughout the IMPREX project, such as improvements in the seasonal discharge forecast quality as a result of improving the seasonal meteorological forecast quality. Towards this goal, it is in our plans to expand on the work done in this deliverable by adding seasonal discharge forecasting scores from systems using different seasonal meteorological forecasts, such as GloSea5.



The intercomparison results are useful for the multi-modelling of task 4.3 of WP4. The multi-modelling approach could for example use weights for each forecasting system based on the forecasting systems' performances for a certain location, type of event, time of the year. Beyond this deliverable, the results of the intercomparison are valuable for the risk outlook, a deliverable of WP14 within IMPREX. The risk outlook will provide an overview of the hydrological 'risk' for Europe and it will also showcase examples of making hydrological information relevant at a local scale, focusing on selected IMPREX case studies. It is currently under development, so it is not yet fully known what will be included within the tool, but it is likely to show current hydrological status, climatology and seasonal forecasted anomalies as well as sector-specific indicators. An improved understanding of forecasting systems' strengths and weaknesses will enable IMPREX communicate this information in a comprehensive way, by adding information which will help the users know with what level of confidence each forecast can be used.

The EPB sensitivity analysis enabled to highlight which component of the forecast system should be improved in order to improve the seasonal discharge forecasting skill for all forecast initialisation dates and lead time for regions in Europe and individual Central European River stations. These results should be used as an indication of where to concentrate resources in order to obtain the largest improvements in the seasonal discharge forecasting skill. Where the analysis indicated the IHC to be the largest contributors to the errors in seasonal discharge forecasts, data assimilation methods could be used. Where the SCF were highlighted to be the largest contributors to seasonal discharge errors, the SCF used to force the hydrological models should be improved.

Finally, from the comparative analysis of the hydrological skill we spotted the strengths and weaknesses of ensemble seasonal forecasts from ECMWF System 4. We identified links between forecasting skill and different physiographic and hydro-climatic characteristics. CART showed that skill is dependent on the basin's hydrological regime. Elevation and remaining bias in temperature were also identified to be important aspects (dependence of





response at mountainous basins to temperature). The skill seems to be limited for relatively flashy basins experiencing strong flow dynamics over the year (less memory in the system).



6 References

- Arnal, L., A. Wood, E. Stephens, H. Cloke, and F. Pappenberger, 2017: An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity. *J. Hydrometeor.* doi:10.1175/JHM-D-16-0259.1, in press.
- Arribas, A., et al., 2010: The GloSea4 Ensemble Prediction System for Seasonal Forecasting. *Monthly Weather Review*, 139, 6, 1891–1910, doi:10.1175/2010MWR3615.1.
- Berendrecht, W.L., A.H. Weerts, A.A. Veldhuizen, T. Kroon, 2011: An operational drought forecasting system using coupled models for groundwater, surface water and unsaturated zone. IAHS-AISH Publication, Volume 341, 2011, Pages 3-8. 7th International Conference on Calibration and Reliability in Groundwater Modeling: managing groundwater and the environment, ModelCARE 2009, Wuhan, China, September 20-23, 2009.
- Bierkens, M. F. P., and L. P. H. van Beek, 2009: Seasonal Predictability of European Discharge: NAO and Hydrological Response Time. *Journal of Hydrometeorology*, 10, 4, 953–68, doi:10.1175/2009JHM1034.1.
- Blöschl, G., M. Sivapalan, T. Wagener, A. Viglione, and H. Savenije: Runoff prediction in ungauged basins. Synthesis across processes, places and scales, Cambridge University Press, Cambridge, UK, 2013.
- cawcr: Forecast verification, available at <http://cawcr.gov.au/projects/verification/>, last access: 6th January 2017, 2015.
- Church, J. E., 1935: Principles of Snow Surveying as Applied to Forecasting Stream Flow. Vol. 51. 2. *Journal of Agricultural Research*.
- Day, Gerald N., 1985: Extended Streamflow Forecasting Using NWSRFS. *Journal of Water Resources Planning and Management*, 111, 2, 157–70, doi:10.1061/(ASCE)0733-9496(1985)111:2(157).
- Easey, J., C. Prudhomme, and D. M. Hannah, 2006: Seasonal Forecasting of River Flows: A Review of the State-of-the-Art. *IAHS Publication*, 308, 158-162.





Goddard, L., W. E. Baethgen, H. Bhojwani, and A. W. Robertson, 2014: The International Research Institute for Climate & Society: Why, What and How. *Earth Perspectives*, 1, 1, doi:10.1186/2194-6434-1-10.

Gupta, H. V., C. Perrin, G. Blöschl, A. Montanari, R. Kumar, M. Clark, and V. Andréassian, 2014: Large-sample hydrology: a need to balance depth with breadth, *Hydrol. Earth Syst. Sci.*, 18, 2, 463–477, doi:10.5194/hess-18-463-2014.

Helms, D., S. E. Phillips, and P. F. Reich, 2008: The History of Snow Survey and Water Supply Forecasting: Interviews with US Department of Agriculture Pioneers. 8. US Department of Agriculture, Natural Resources Conservation Service, Resource Economics and Social Sciences Division.

Herrera, S., J. Fernández, and J.M. Gutiérrez, 2016: Update of the Spain02 gridded observational dataset for EURO-CORDEX evaluation: assessing the effect of the interpolation methodology. *Int. J. Climatol.*, 36: 900–908. doi:10.1002/joc.4391.

Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, 15, 5, 559–70, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Klein, B., and D. Meissner, 2016: Vulnerability of Inland Waterway Transport and Waterway Management on Hydro-meteorological Extremes. EU Horizon2020 IMPREX Deliverable 9.1. <http://www.imprex.eu/system/files/generated/files/resource/d9-1-imprex-v1.pdf>.

Pagano, T. C., and D. C. Garen, 2005: Integration of Climate Information and Forecasts into Western US Water Supply Forecasts. *Climate Variations, Climate Change, and Water Resources Engineering*, 86–103.

Pechlivanidis, I. G., and B. Arheimer, 2015: Large-scale hydrological modelling by using modified PUB recommendations: the India-HYPE case. *Hydrol. Earth Syst. Sci.*, 19, 4559–4579, doi:10.5194/hess-19-4559-2015.

Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto, 2004: Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. John Wiley & Sons.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, 2008: Global Sensitivity Analysis: The Primer. John Wiley & Sons.



Schaake, J. C., 1978: The National Weather Service Extended Streamflow Prediction Techniques: Description and Applications during 1977. In 3rd Annual Climate Diagnostics Workshop.

Twedt, T. M., J. C. Schaake Jr, and E. L. Peck., 1977: National Weather Service Extended Streamflow Prediction [USA]. In Proceedings Western Snow Conference.

Wang, E., Y. Zhang, J. Luo, F. H. S. Chiew, and Q. J. Wang, 2011: Monthly and Seasonal Streamflow Forecasts Using Rainfall-Runoff Modeling and Historical Weather Data. *Water Resources Research*, 47, 5, W05516, doi:10.1029/2010WR009922.

Wood, A. W., and D. P. Lettenmaier, 2003: Comparing Hydrologic Forecast Uncertainty due to Initial Condition Error versus Climate Forecast Error. In EGS-AGU-EUG Joint Assembly, 1:8162. <http://adsabs.harvard.edu/abs/2003EAEJA.....8162W>.

Wood, A. W., and D. P. Lettenmaier, 2006: A Test Bed for New Seasonal Hydrologic Forecasting Approaches in the Western United States. *Bulletin of the American Meteorological Society*, 87, 12, 1699, doi:10.1175/BAMS-87-12-1699.

Wood, A. W., and Dennis P. Lettenmaier, 2008: An Ensemble Approach for Attribution of Hydrologic Prediction Uncertainty. *Geophysical Research Letters*, 35, 14, doi:10.1029/2008GL034648.

Wood, A. W., T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark, 2016: Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill. *Journal of Hydrometeorology*, 17, 2, 651–68, doi:10.1175/JHM-D-14-0213.1.





7 Annex A – Tabulated overview on hydrological model features

Table 4 Tabulated overview on hydrological model features of SPHY

1. General Information	
Model name	SPHY (S patial P rocesses in H ydrology)
Version	V2.1
Author(s) / First publication	Terink et al. (2015a)
Contact person	Wilco Terink (w.terink@futurewater.nl)
Institute	FutureWater
Website	http://www.sphy.nl/
General modelling objectives	Calculation of river basins water balance
Domain of applicability (catchment types and climate conditions)	The SPHY model has been applied and tested in various studies ranging from real-time soil moisture predictions in flat lands, to operational reservoir inflow forecasting applications in mountainous catchments, irrigation scenarios in the Nile Basin, and detailed climate change impact studies in the snow- and glacier-melt dominated the Himalayan region.
2. Model	



description	
Model type (empirical, conceptual, physically based)	Conceptual/Physically-based model
Continuous or event-based	Continuous
Possible running time steps	24h
Spatial discretization (lumped, semidistributed, distributed)	Spatially distributed leaky bucket type
Short description of model structure detailing main function (evaporation, soil moisture accounting,	SPHY is grid-based and cell values represent averages over a cell, but sub-grid variability is taken into account. A cell can be glacier-free, partially glaciated, or completely covered by glaciers. The cell fraction not covered by glaciers consists of either land covered with snow or land that is free of snow. Land that is free of snow can consist of vegetation, bare soil, or open water. In order to distinct between land cover types at sub-grid level, SPHY calculates and stores the state variables as grid-cell averages. Sub-





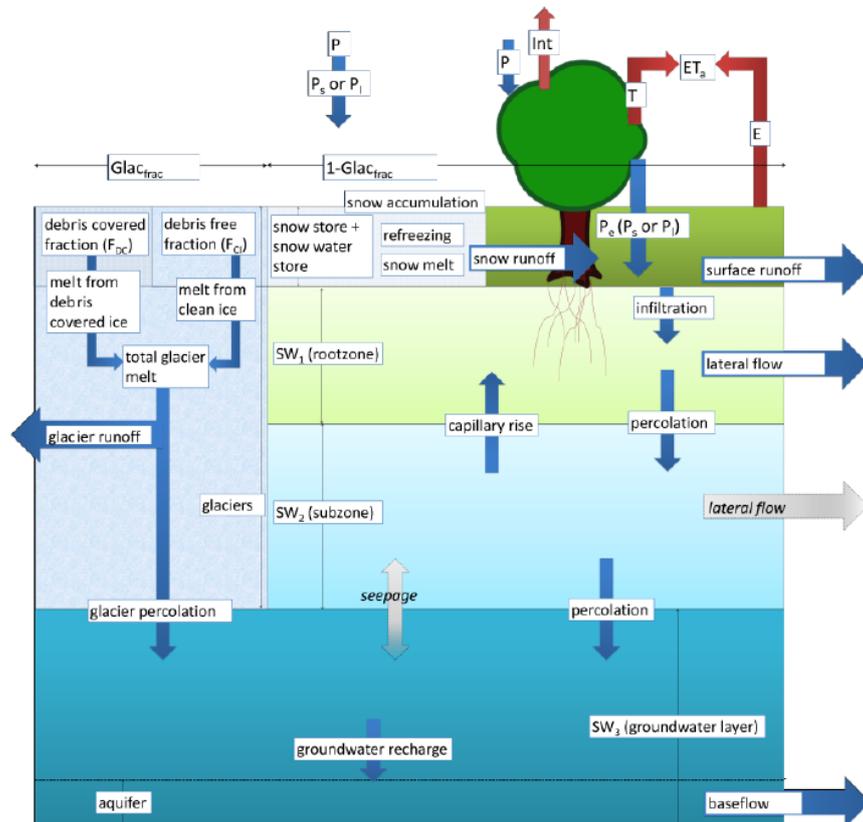
<p>groundwater, routing, snowmelt, etc.)</p>	<p>grid variability is mainly determined by the fractional vegetation coverage, which affects processes such as interception, effective precipitation, and potential evapotranspiration.</p> <p>The land compartment is divided in two upper soil stores and a third groundwater store, with their corresponding drainage components: surface runoff, lateral flow and base flow. SPHY simulates for each cell precipitation in the form of rain or snow, depending on the temperature. Any precipitation that falls on land surface can be intercepted by vegetation and in part or in whole evaporated. The snow storage is updated with snow accumulation and/or snow melt. A part of the liquid precipitation is transformed in surface runoff, whereas the remainder infiltrates into the soil. The resulting soil moisture is subject to evapotranspiration, depending on the soil properties and fractional vegetation cover, while the remainder contributes in the long-term to river discharge by means of lateral flow from the first soil layer, and base flow from the groundwater reservoir.</p> <p>Melting of glacier ice contributes to the river discharge by means of a slow and fast component, being (i) percolation to the groundwater reservoir that eventually becomes base flow, and (ii) direct runoff. The cell-specific runoff, which becomes available for routing, is the sum of surface runoff, lateral flow, base flow, snow melt and glacier melt.</p> <p>If no lakes are present, then the user can choose for a simple flow accumulation routing scheme: for each cell the accumulated amount of material that flows out of the cell into its neighbouring downstream cell is calculated. This accumulated amount is the amount of material in the cell itself plus the amount of material in upstream cells of the cell. For each cell, the following procedure is</p>
--	---



performed: using the local drain direction network, the catchment of a cell is determined which is made up the cell itself and all cells that drain to the cell. If lakes are present, then the fractional accumulation flux routing scheme is used: depending on the actual lake storage, a fraction of that storage becomes available for routing and is extracted from the lake, while the remaining part becomes the updated actual lake storage. The flux available for routing is routed in the same way as in the simple flow accumulation routing scheme.

SPHY enables the user to turn on/off any of the six available modules that are not required: glaciers, snow, groundwater, dynamic vegetation, simple routing, and lake/reservoir routing.

Scheme of model structure





	Source: Terink et al. (2015a)
3. Model parameters	
Distribution of model parameters (yes/no)	Yes
Number of free parameters	Numerous free parameters (Terink et al. (2015a))
Procedure of model parameter estimation (measurement, manual or automatic algorithm, etc.)	<ul style="list-style-type: none"> - Calibration for each sub basin possible - An automatic calibration routine does not belong to the model itself. - For setting-up the model data on streamflows are not necessary. However, to undertake a proper calibration and validation procedure flow data are required. The model could also be calibrated using actual evapotranspiration, soil moisture contents, or snow coverage.
4. Model inputs / Model outputs	
List and characteristics of input variables (type, time step,	As input SPHY requires data on state variables as well as dynamic variables. For the state variables the most relevant are: Digital Elevation Model (DEM), land use type, glacier cover, reservoirs and soil characteristics. The main dynamic variables are climate data such as precipitation, temperature, reference evapotranspiration.



spatial resolution, etc.)	Since SPHY is grid-based optimal use of remote sensing data and global data sources can be made. For example, the Normalized Difference Vegetation Index (NDVI) can be used to determine the Leaf Area Index (LAI) in order to estimate the growth-stage of land cover.
List and characteristics of output variables (type, time step, spatial resolution, etc.)	The SPHY model provides a wealth of output data that can be selected based on the preference of the user. Spatial output can be presented as maps of all the hydrological processes. Maps often displayed as output include actual evapotranspiration, runoff generation (separated by its contributors), and groundwater recharge. These maps can be generated on daily base, but most users prefer to get those at monthly or annual aggregated time periods. Time-series can be generated for each location in the study area. Time-series often used are stream flow under current and future conditions, actual evapotranspiration and recharge to the groundwater.
5. Examples of previous model applications	
Catchments, objectives, etc. Results of existing comparisons with other models	A number of evaluations and applications are documented at the website.





6. List of selected references	
	<ul style="list-style-type: none"> ● Terink, W., S. Khanal. 2016. SPHY: Spatial Processes in Hydrology. Advanced training: input data, sensitivity analysis, model calibration, and scenario analyses. FutureWater Report 159. ● Terink, W., A.F. Lutz, G.W.H. Simons, W.W. Immerzeel, P. Droogers. 2015a. SPHY v2.0: Spatial Processes in Hydrology. Geoscientific Model Development, 8, 2009-2034, doi:10.5194/gmd-8-2009-201 ● Terink, W., A.F. Lutz, W.W. Immerzeel. 2015b. SPHY v2.0: Spatial Processes in Hydrology. Model theory, installation, and data preparation. FutureWater report 142. ● Terink, W., A.F. Lutz, W.W. Immerzeel. 2015c. SPHY: Spatial Processes in Hydrology. Graphical User-Interfaces (GUIs). FutureWater report 144. ● Terink, W., A.F. Lutz, G.W.H. Simons, W.W. Immerzeel. 2015d. SPHY: Spatial Processes in Hydrology. Case-studies for training. FutureWater report 143.

Table 5 Tabulated overview on hydrological model features of HYPE

1. General Information	
Model name	HYPE (H ydrological P redictions for the E nvironment)
Version	v4.13



Author(s) / First publication	Lindström et al. (2010)
Contact person	Charlotta Pers (charlotta.pers@smhi.se) Ilias Pechlivanidis (ilias.pechlivanidis@smhi.se)
Institute	Swedish Meteorological and Hydrological Institute
Website	http://hypecode.smhi.se
General modelling objectives	Calculation/prediction of river basin responses (water quantity and quality)
Domain of applicability (catchment types and climate conditions)	The HYPE model has been applied and tested in different scales, various domains and hydro-climatic conditions. See http://hypeweb.smhi.se
2. Model description	
Model type (empirical, conceptual, physically based)	Conceptual/Process-based model
Continuous or	Continuous





event-based	
Possible running time steps	Daily (also hourly for national operational services)
Spatial discretization (lumped, semidistributed, distributed)	Spatially distributed at the sub-basin scale. Sub-basin resolution depends on the application. In Europe, this is 215 km ² .
Short description of model structure detailing main function (evaporation, soil moisture accounting, groundwater, routing, snowmelt, etc.)	HYPE is most often run at a daily time step and simulates the water flow paths in soil for hydrological response units (HRUs), which are defined by gridded soil and land- use classes and can be divided in up to three layers with a fluctuating groundwater table. The HRUs are further aggregated into sub-basins based on topography. Elevation is also used to get temperature variations within a sub-basin to influence the snowmelt and storage as well as evapotranspiration. Glaciers have a variable surface and volume, while lakes are defined as classes with specified areas and variable volume. Lakes receive runoff from the local catchment and, if located in the sub-basin outlet, also the river flow from upstream sub- basins. On glaciers and lakes, precipitation falls directly on the surface and water evaporates at the potential rate. Each lake has a defined depth below an outflow threshold. The outflow from lakes is determined by a general rating curve unless a specific one is given or if the lake is regulated. Regulated lakes and man-made reservoirs are treated equally but a simple regulation rule can be used, in which the outflow is constant or follows a



seasonal function (as it is often the case with hydropower) for water levels above the threshold. A rating curve for the spillways can be used when the reservoir is full.

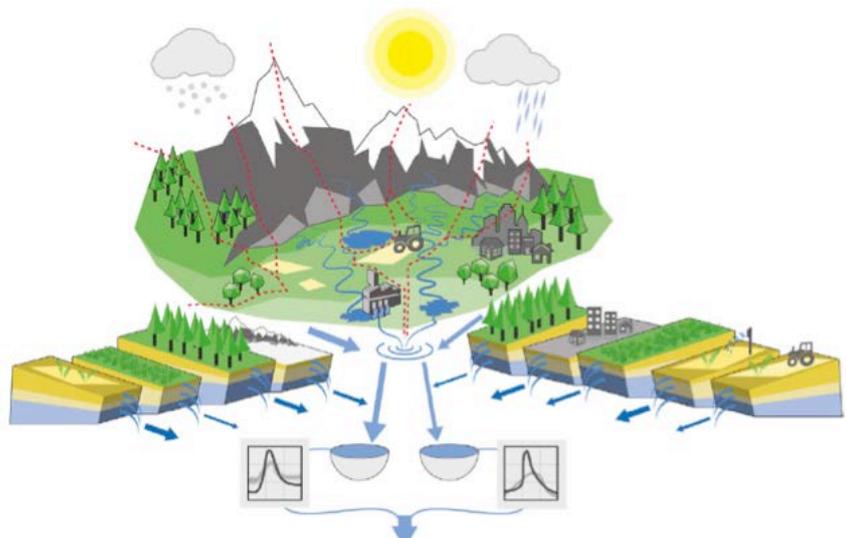
Irrigation

Irrigation is simulated based on crop water demands calculated either with the FAO-56 crop coefficient method (Allen et al., 1998) or relative to a reference flooding level for sub-merged crops (e.g. rice). The demands are withdrawn from rivers, lakes, reservoirs, and/or groundwater within and/or external to the sub-basin where the demands originated and are constrained by the water available at these sources. After subtraction of conveyance losses, the withdrawn water is applied as additional infiltration to the irrigated soils. The agriculture and irrigation data sets (see Table 1) are used to define irrigated area, crop types, growing seasons, crop coefficients, irrigation methods and efficiencies, and irrigation sources. The irrigation parameters regulating water demand and abstraction are usually manually calibrated using discharge stations in irrigation-dominated areas.

River discharge is routed between the sub-basins along the river network and may also pass sub-basins, flow laterally in the soil between sub-basins or interact with a deeper groundwater aquifer in the model.





<p>Scheme of model structure</p>	 <p>Source: Hundecha et al. (2016)</p>
<p>3. Model parameters</p>	
<p>Distribution of model parameters (yes/no)</p>	<p>Yes</p>
<p>Number of free parameters</p>	<p>Numerous free parameters (Lindström et al. (2010))</p>
<p>Procedure of model parameter estimation (measurement, manual or automatic algorithm, etc.)</p>	<p>Many of the parameters are linked to physiographic characteristics in the landscape, such as soil type and depth (soil dependent parameters) or vegetation (land-use-dependent parameters), while others are assumed to be general to the entire domain (general parameters) or specific to a defined region or river (regional parameters). Parameters for each HRU are calibrated for representative gauged basins and then transferred to similar HRUs, which are gridded at a higher resolution than the sub-basins across the whole domain to</p>



	<p>account for spatial variability in soil and land use. Using the distributed HRU approach in the multi-basin concept is thus one part of the regionalisation method for parameter values.</p> <p>Some other parameters, however, are either estimated from literature values and from previous modelling experiences (a priori values) or identified in the (automatic or manual) calibration procedure.</p> <p>Slightly different methods for regionalisation of parameter values have been used when setting up the different HYPE model applications, depending on access to gauging stations, additional data sources, and expert knowledge.</p> <p>An automatic calibration routine based on the differential-evolution Markov-chain (DE-MC) algorithm has also been used.</p> <p>The model parameters can also be constrained using evapotranspiration or snow information.</p>
<p>4. Model inputs / Model outputs</p>	
<p>List and characteristics of input variables</p>	<p>As input HYPE requires data on: Digital Elevation Model (DEM), land use type, glacier cover, reservoirs, irrigation, and soil characteristics. Depending on the application information about crop and vegetation, bifurcations of the river network,</p>





<p>(type, time step, spatial resolution, etc.)</p>	<p>point sources and water extraction, floodplains and regional aquifers can be used.</p> <p>The main dynamic variables are climate data such as precipitation and temperature. Other observations can also be used for different purposes:</p> <p>http://www.smhi.net/hype/wiki/doku.php?id=start:hype_file_reference#observation_data_files</p>
<p>List and characteristics of output variables (type, time step, spatial resolution, etc.)</p>	<p>A list of output variables can be found in:</p> <p>http://www.smhi.net/hype/wiki/doku.php?id=start:hype_file_reference:info.txt:variables</p> <p>The exported time-step depends on the user, i.e. daily, weekly, monthly, and annual.</p> <p>The variables are exported for the basin outlet or represent basin averages.</p>
<p>5. Examples of previous model applications</p>	<p style="background-color: #cccccc; height: 100px;"></p>
<p>Catchments, objectives, etc.</p> <p>Results of existing comparisons with other models</p>	<p>A number of evaluations and applications are documented at the website.</p>
<p>6. List of selected references</p>	<p style="background-color: #cccccc; height: 70px;"></p>
	<ul style="list-style-type: none"> ● Andersson, J.C.M., Pechlivanidis, I.G., Gustafsson, D., Donnelly, C., and Arheimer, B., 2015. Key factors for im-



proving large-scale hydrological model performance. *European Water* 49:77-88.

- Donnelly, C, Andersson, J.C.M. and Arheimer, B., 2016. Using flow signatures and catchment similarities to evaluate a multi-basin model (E-HYPE) across Europe. *Hydr. Sciences Journal* 61(2):255-273, doi: 10.1080/02626667.2015.1027710
- Hundecha, Y., Arheimer, B., Donnelly, C., Pechlivanidis, I., 2016. A regional parameter estimation scheme for a pan-European multi-basin model. *Journal of Hydrology: Regional Studies*, Volume 6, June 2016, Pages 90-111. doi:10.1016/j.ejrh.2016.04.002
- Lindström, G., Pers, C.P., Rosberg, R., Strömqvist, J., and Arheimer, B., 2010. Development and test of the HYPE (Hydrological Predictions for the Environment) model – A water quality model for different spatial scales. *Hydrology Research* 41.3-4:295-319.
- Pechlivanidis, I. G. and Arheimer, B., 2015. Large-scale hydrological modelling by using modified PUB recommendations: the India-HYPE case, *Hydrol. Earth Syst. Sci.*, 19, 4559-4579, doi:10.5194/hess-19-4559-2015.
- Strömqvist, J., Arheimer, B., Dahné, J., Donnelly, C. and Lindström, G., 2012. Water and nutrient predictions in ungauged basins – Set-up and evaluation of a model at the national scale. *Hydrological Sciences Journal* 57(2):229-247.





Table 6 Tabulated overview on hydrological model features of LARSIM

1. General Information	
Model name	LARSIM (Large Area Runoff Simulation Model)
Version	LARSIM Revision 968 (neue Formate)
Author(s) / First publication	Ludwig & Bremicker (2006)
Contact person	LARSIM development community http://www.larsim.info
Institute	LARSIM development community http://www.larsim.info
Website	http://www.larsim.info
General modelling objectives	Continuous simulation of runoff processes in catchments and river networks
Domain of applicability (catchment types and climate conditions)	Largely applied by forecasting centers in Germany, Austria, Luxembourg, Switzerland and the French regions of Alsace and Lorraine, Central Europe
2. Model description	
Model type (empirical, conceptual,	Deterministic conceptual model



physically based)	
Continuous or event-based	Continuous
Possible running time steps	Hourly, daily
Spatial discretization (lumped, semidistributed, distributed)	Distributed
Short description of model structure detailing main function (evaporation, soil moisture accounting, groundwater, routing, snowmelt, etc.)	<p>The main components of the model are routines for interception, evapotranspiration, snow accumulation, compaction and melt, soil water retention, storage and lateral water transport, as well as flood-routing in channels and retention in lakes.</p> <p>Spatial units are grid-based subareas or subareas according to hydrologic subcatchments. Hydrological processes are modelled for each single land use category or alternatively for each land use soil type category in a subarea (Hydrological response unit HRU). HRUs can be further subdivided in elevation zones for snow simulation.</p> <p>Different process descriptions could be selected to model snow</p>



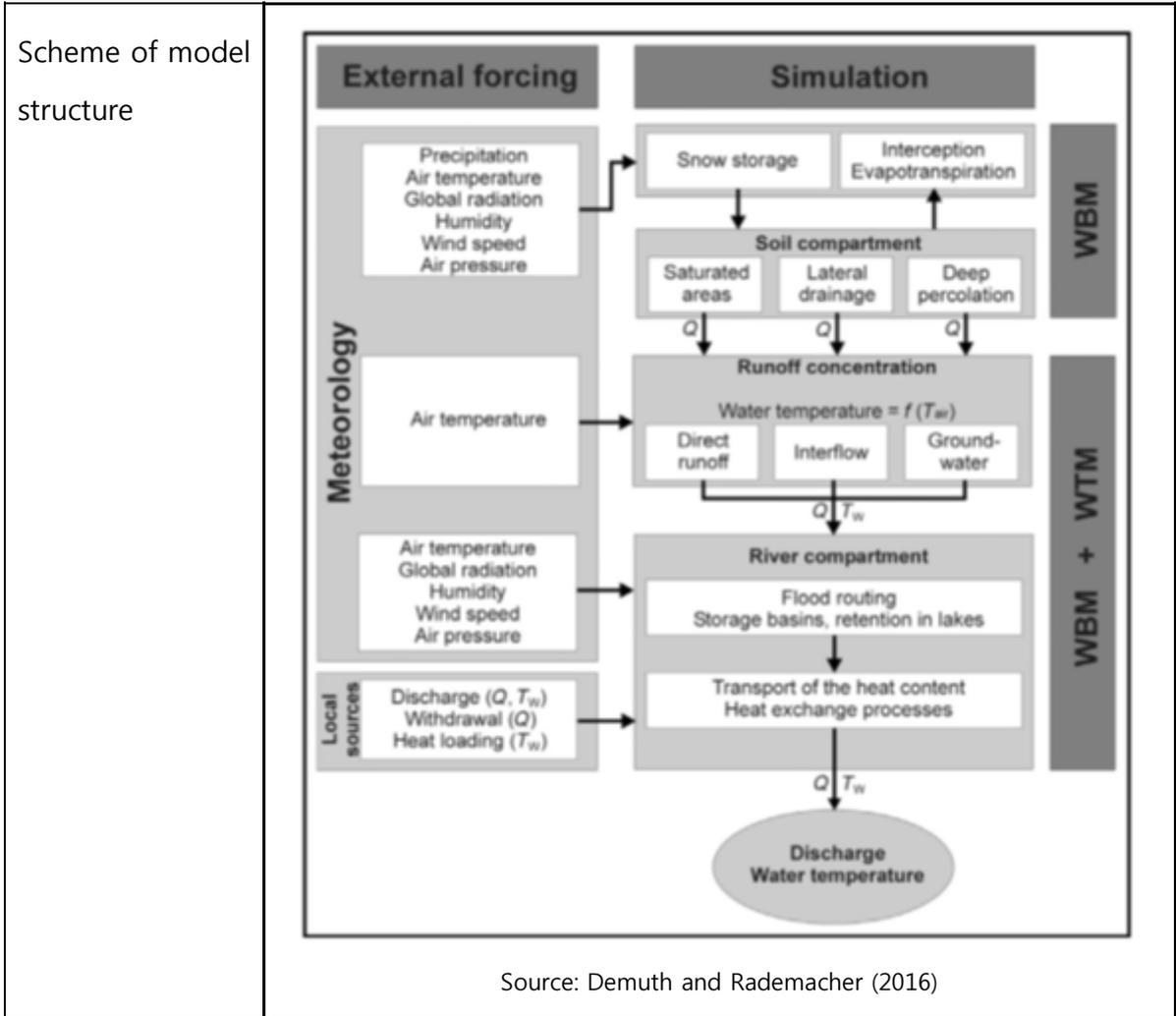


	<p>processes and evaporation. Here the configuration used in this study is described.</p> <p>Snow Routine:</p> <p>Precipitation is divided into rainfall and snowfall using a threshold temperature. On days with temperatures below the threshold, precipitation is supposed to be snow. The consideration of a transition from rain to snow over a temperature interval is possible. Based on a degree-day approach snow melt is computed. Water retention, snow compaction, meltwater outflow is calculated after the snow compaction approach of Bertle. Snow processes could be simulated separately for different elevation zones in the subarea.</p> <p>Soil Routine:</p> <p>The routine mainly controls runoff formation. To simulate the soil storage the Xinanjiang-model is used. Soil water content is calculated by the water balance equation, taking into account the precipitation supply, withdrawal of water through evapotranspiration as well as runoff formation. In the configuration applied here three runoff components are considered: runoff formation on saturated areas towards direct runoff storage, water release from soil storage through lateral drainage towards interflow storage and water release through vertical percolation towards groundwater storage. Saturated areas which control the direct runoff are derived from the soil water storage via the soil-moisture-saturated areas function. Actual evapotranspiration is computed from potential evapotranspiration as a function of soil moisture.</p> <p>Runoff Generation Routine:</p> <p>Runoff concentration from direct runoff storage, interflow storage and groundwater storage of the subareas are calculated by a</p>
--	---



	<p>single linear storage model. The combination of the outflows of these storages results in the total outflow of the subarea.</p> <p>Routing Procedure:</p> <p>The translation and the retention in the channel are calculated in dependency of the channel geometry and the friction of the channel.</p> <p>Lake and Reservoir:</p> <p>Storage effects including operation of dams, lakes and reservoirs can be simulated using different approaches depending on the available data.</p>
--	---





<p>3. Model parameters</p>	
<p>Distribution of model parameters (yes/no)</p>	<p>Yes</p>
<p>Number of free parameters</p>	<p>Numerous free parameters (Ludwig & Bremicker 2006)</p>
<p>Procedure of</p>	<p>- Calibration for each subarea is possible, generally several</p>



model parameter estimation (measurement, manual or automatic algorithm, etc.)	subareas are combined and calibrated together - An automatic calibration routine for some parameters is available
4. Model inputs / Model outputs	
List and characteristics of input variables (type, time step, spatial resolution, etc.)	Depending on the considered process descriptions for potential evapotranspiration and modelling of snow processes different input data sets are required. In the configuration used here daily precipitation, temperature and global radiation are required as input data.
List and characteristics of output variables (type, time step, spatial resolution, etc.)	Numerous possible output variables (Ludwig & Bremicker 2006) e.g. total computed outflow, actual evaporation, soil moisture,... Depending on the variable output is available for subareas, HRUs, combination of several connected subareas, and defined output nodes (e.g. gauges)
5. Examples of previous model applications	





<p>Catchments, objectives, etc.</p> <p>Results of existing comparisons with other models</p>	<p>Operational forecast model, climate change analysis, water balance, water temperature. Applications and publications are documented at the website www.larsim.info. Mesoscale application for the River Rhine (Ebel et al. 2000).</p>
<p>6. List of selected references</p>	
	<ul style="list-style-type: none"> ● Bremicker, M., M. C. Casper & I. Haag (2011): Extrapolationsfähigkeit des Wasserhaushaltsmodells LARSIM auf extreme Abflüsse am Beispiel der Schwarzen Pockau. KW Korrespondenz Wasserwirtschaft 4(8), 445-451 ● Demuth, N. & S. Rademacher (2016): Chapter 5 - Flood Forecasting in Germany — Challenges of a Federal Structure and Transboundary Cooperation A2 - Adams, Thomas E. In: T. C. Pagano (Ed.): Flood Forecasting. Academic Press, Boston, 125-151 ● Ebel, M., K. Ludwig & K. G. Richter (2000): Mesoskalige Modellierung des Wasserhaushaltes im Rheineinzugsgebiet mit LARSIM. Hydrologie und Wasserbewirtschaftung 6, 308-312 ● Haag, I. & A. Luce (2008): The integrated water balance and water temperature model LARSIM-WT. Hydrological Processes 22(7), 1046-1056 ● Ludwig, K. & M. Bremicker (2006): The Water Balance Model LARSIM –Design, Content and Applications. 22. C. Leibundgut, S. Demuth and J. Lange (Eds), Freiburger Schriften zur Hydrologie, Institut für Hydrologie, Universität



	Freiburg im Breisgau, Freiburg, 141 pp.
--	---

Table 7 Tabulated overview on hydrological model features of LISFLOOD

1. General Information	
Model name	LISFLOOD
Version	NA
Author(s) / First publication	De Roo et al. (2000)
Contact person	A.P.J. De Roo
Institute	Joint Research Centre, Space Applications Institute, AIS Unit Environment and Natural Hazards, TP 950, 21020 Ispra (Va), Italy
Website	https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/lisflood-distributed-water-balance-and-flood-simulation-model-revised-user-manual-2013
General modelling objectives	To produce a tool that can be used in large and trans-national catchments for a variety of applications, including: <ul style="list-style-type: none"> • Flood forecasting • Assessing the effects of river regulation measures





	<ul style="list-style-type: none"> Assessing the effects of land-use change Assessing the effects of climate change
Domain of applicability (catchment types and climate conditions)	The LISFLOOD model has been developed for European catchments. It was designed to make the best possible use of several existing databases that contain pan-European information on soils (King et al., 1997; Wösten et al., 1999), land cover (CEC, 1993), topography (Hiederer & de Roo, 2003) and meteorology (Rijks et al., 1998).
2. Model description	
Model type (empirical, conceptual, physically based)	Conceptual/Physically-based model
Continuous or event-based	Continuous
Possible running time steps	24h
Spatial discretization (lumped, semidistributed, distributed)	Spatially distributed
Short description	The figure below gives an overview of the structure of the

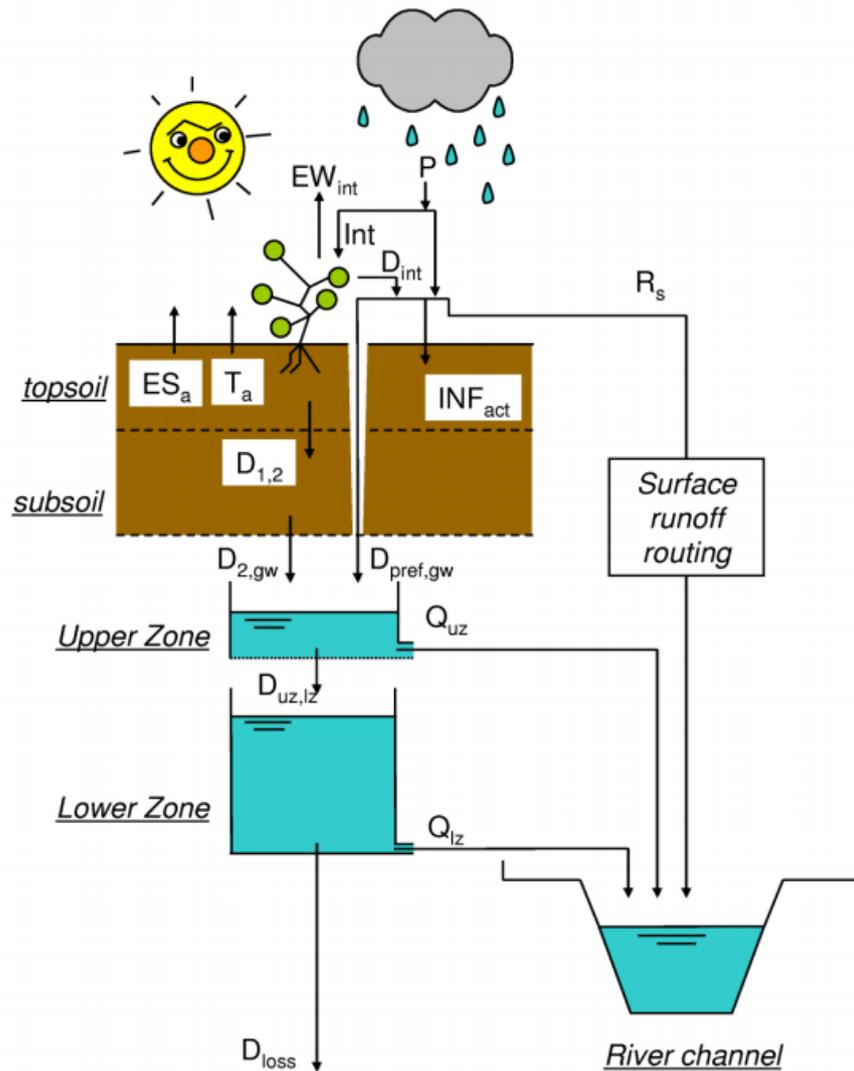


<p>of model structure detailing main function (evaporation, soil moisture accounting, groundwater, routing, snowmelt, etc.)</p>	<p>LISFLOOD model. Basically, the model is made up of the following components:</p> <ul style="list-style-type: none"> • a 2-layer soil water balance sub-model • sub-models for the simulation of groundwater and subsurface flow (using 2 parallel interconnected linear reservoirs) • a sub-model for the routing of surface runoff to the nearest river channel • a sub-model for the routing of channel flow (not shown in the Figure) <p>The processes that are simulated by the model include snow melt (not shown in the figure), infiltration, interception of rainfall, leaf drainage, evaporation and water uptake by vegetation, surface runoff, preferential flow (bypass of soil layer), exchange of soil moisture between the two soil layers and drainage to the groundwater, sub-surface and groundwater flow, and flow through river channels.</p>
---	--





Scheme of model structure



Overview of the LISFLOOD model. P = precipitation; Int = interception; EW_{int} = evaporation of intercepted water; D_{int} = leaf drainage; ES_a = evaporation from soil surface; T_a = transpiration (water uptake by plant roots); INF_{act} = infiltration; R_s = surface runoff; $D_{1,2}$ = drainage from top- to subsoil; $D_{2,gw}$ = drainage from subsoil to upper groundwater zone; $D_{pref,gw}$ = preferential flow to upper groundwater zone; $D_{uz,lz}$ = drainage from upper- to lower groundwater zone; Q_{uz} = outflow from upper groundwater zone; Q_{lz} = outflow from lower groundwater zone; D_{loss} = loss from lower



	groundwater zone. Note that snowmelt is not included in the Figure (even though it is simulated by the model).
3. Model parameters	
Distribution of model parameters (yes/no)	Yes
Number of free parameters	Numerous free parameters (http://publications.jrc.ec.europa.eu/repository/bitstream/JRC78917/lisflood_2013_online.pdf).
Procedure of model parameter estimation (measurement, manual or automatic algorithm, etc.)	"A calibration exercise completed in 2013 (Zajac et al., 2013) produced Europe wide parameter maps based on the estimation of parameter values for 693 catchments. Estimation was carried out using the Standard Particle Swarm 2011 (SPSO-2011) algorithm (Zambrano-Bigiarini and Rojas, 2013) and a root mean squared error criteria. For 659 of these a set of 9 parameters that control snowmelt, infiltration, preferential bypass flow through the soil matrix, percolation to the lower ground water zone, percolation to deeper groundwater zones, residence times in the soil and subsurface reservoirs and river routing, were estimated by calibrating the model against historical records of river discharge. For the remaining 34 catchments the option to represent





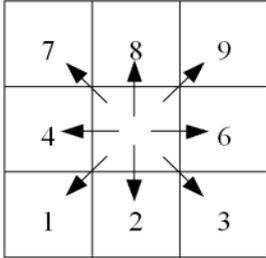
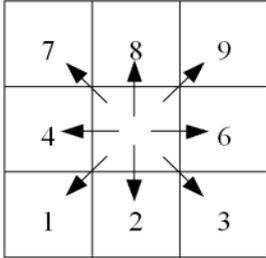
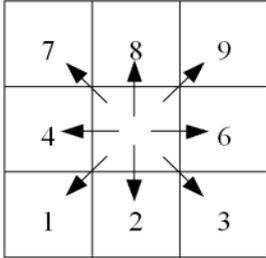
	reservoirs was used requiring the calibration of four additional parameters related to reservoir operation; though neglecting the calibration of the deepest groundwater store resulted in 12 calibration parameters for these catchments." (Smith et al., 2016)																																				
4. Model inputs / Model outputs																																					
List and characteristics of input variables (type, time step, spatial resolution, etc.)	<table border="1"> <thead> <tr> <th colspan="4" data-bbox="451 719 1350 775"><i>Table A12.1 LISFLOOD input maps (continued on next pages)</i></th> </tr> <tr> <th colspan="4" data-bbox="451 775 1350 797">GENERAL</th> </tr> <tr> <th data-bbox="451 797 619 819">Map</th> <th data-bbox="619 797 794 819">Default name¹⁵</th> <th data-bbox="794 797 991 819">Units, range</th> <th data-bbox="991 797 1350 819">Description</th> </tr> </thead> <tbody> <tr> <td data-bbox="451 819 619 875">MaskMap</td> <td data-bbox="619 819 794 875">area.map</td> <td data-bbox="794 819 991 875">Unit: - Range: 0 or 1</td> <td data-bbox="991 819 1350 875">Boolean map that defines model boundaries</td> </tr> <tr> <th colspan="4" data-bbox="451 875 1350 898">TOPOGRAPHY</th> </tr> <tr> <th data-bbox="451 898 619 920">Map</th> <th data-bbox="619 898 794 920">Default name</th> <th data-bbox="794 898 991 920">Units, range</th> <th data-bbox="991 898 1350 920">Description</th> </tr> <tr> <td data-bbox="451 920 619 1491">Ldd</td> <td data-bbox="619 920 794 1491">ldd.map</td> <td data-bbox="794 920 991 1491">U.: flow directions R.: $1 \leq \text{map} \leq 9$</td> <td data-bbox="991 920 1350 1491">local drain direction map (with value 1-9); this file contains flow directions from each cell to its steepest downslope neighbour. Ldd directions are coded according to the following diagram:  This resembles the numeric key pad of your PC's keyboard, except for the value 5, which defines a cell without local drain direction (pit). The pit cell at the end of the path is the outlet point of a catchment.</td> </tr> <tr> <td data-bbox="451 1491 619 1536">Grad</td> <td data-bbox="619 1491 794 1536">gradient.map</td> <td data-bbox="794 1491 991 1536">U.: [m m⁻¹] R.: map > 0 !!!</td> <td data-bbox="991 1491 1350 1536">Slope gradient</td> </tr> <tr> <td data-bbox="451 1536 619 1592">Elevation Stdev</td> <td data-bbox="619 1536 794 1592">elvstd.map</td> <td data-bbox="794 1536 991 1592">U.: [m] R.: map ≥ 0</td> <td data-bbox="991 1536 1350 1592">Standard deviation of elevation</td> </tr> </tbody> </table>	<i>Table A12.1 LISFLOOD input maps (continued on next pages)</i>				GENERAL				Map	Default name ¹⁵	Units, range	Description	MaskMap	area.map	Unit: - Range: 0 or 1	Boolean map that defines model boundaries	TOPOGRAPHY				Map	Default name	Units, range	Description	Ldd	ldd.map	U.: flow directions R.: $1 \leq \text{map} \leq 9$	local drain direction map (with value 1-9); this file contains flow directions from each cell to its steepest downslope neighbour. Ldd directions are coded according to the following diagram:  This resembles the numeric key pad of your PC's keyboard, except for the value 5, which defines a cell without local drain direction (pit). The pit cell at the end of the path is the outlet point of a catchment.	Grad	gradient.map	U.: [m m ⁻¹] R.: map > 0 !!!	Slope gradient	Elevation Stdev	elvstd.map	U.: [m] R.: map ≥ 0	Standard deviation of elevation
<i>Table A12.1 LISFLOOD input maps (continued on next pages)</i>																																					
GENERAL																																					
Map	Default name ¹⁵	Units, range	Description																																		
MaskMap	area.map	Unit: - Range: 0 or 1	Boolean map that defines model boundaries																																		
TOPOGRAPHY																																					
Map	Default name	Units, range	Description																																		
Ldd	ldd.map	U.: flow directions R.: $1 \leq \text{map} \leq 9$	local drain direction map (with value 1-9); this file contains flow directions from each cell to its steepest downslope neighbour. Ldd directions are coded according to the following diagram:  This resembles the numeric key pad of your PC's keyboard, except for the value 5, which defines a cell without local drain direction (pit). The pit cell at the end of the path is the outlet point of a catchment.																																		
Grad	gradient.map	U.: [m m ⁻¹] R.: map > 0 !!!	Slope gradient																																		
Elevation Stdev	elvstd.map	U.: [m] R.: map ≥ 0	Standard deviation of elevation																																		



Table A12.1 LISFLOOD input maps (continued from previous page)			
LAND USE - fraction maps			
Map	Default name	Units, range	Description
Fraction of water	fracwater.map	U.: [-] R.: $0 \leq \text{map} \leq 1$	Fraction of inland water for each cell. Values range from 0 (no water at all) to 1 (pixel is 100% water)
Fraction of sealed surface	fracsealed.map	U.: [-] R.: $0 \leq \text{map} \leq 1$	Fraction of impermeable surface for each cell. Values range from 0 (100% permeable surface - no urban at all) to 1 (100% impermeable surface).
Fraction of forest	fracforest.map	U.: [-] R.: $0 \leq \text{map} \leq 1$	Forest fraction for each cell. Values range from 0 (no forest at all) to 1 (pixel is 100% forest)
Fraction of other land cover	fracother.map	U.: [] R.: $0 \leq \text{map} \leq 1$	Other (agricultural areas, non-forested natural area, pervious surface of urban areas) fraction for each cell.
LAND COVER depending maps			
Map	Default name	Units, range	Description
Crop coef. for forest	cropcoef_forest.map	U.: [-] R.: $0.8 \leq \text{map} \leq 1.2$	Crop coefficient for forest
Crop coef. for other	cropcoef_other.map	U.: [-] R.: $0.8 \leq \text{map} \leq 1.2$	Crop coefficient for other
Crop group number for forest	crgrnum_forest.map	U.: [-] R.: $1 \leq \text{map} \leq 5$	Crop group number for forest
Crop group number for other	crgrnum_other.map	U.: [-] R.: $1 \leq \text{map} \leq 5$	Crop group number for other
Manning for forest	mannings_forest.map	U.: [-] R.: $0.2 \leq \text{map} \leq 0.4$	Manning's roughness for forest
Manning for other	mannings_other.map	U.: [-] R.: $0.01 \leq \text{map} \leq 0.3$	Manning's roughness for other
Soil depth for forest for layer1	soildep1_forest.map	U.: [mm] R.: $\text{map} \geq 50$	Forest soil depth for soil layer 1 (rooting depth)
Soil depth for other for layer2	soildep1_other.map	U.: [mm] R.: $\text{map} \geq 50$	Other soil depth for soil layer 1 (rooting depth)
Soil depth for forest for layer2	soildep2_forest.map	U.: [mm] R.: $\text{map} \geq 50$	Forest soil depth for soil layer 2
Soil depth for other for layer2	soildep2_other.map	U.: [mm] R.: $\text{map} \geq 50$	Other soil depth for soil layer 2





Table A12.1 LISFLOOD input maps (continued from previous page)			
SOIL HYDRAULIC PROPERTIES (depending on soil texture)			
Map	Default name	Units, range	Description
ThetaSat1 for forest	thetas1_forest.map	U: [-] R: 0 < map < 1	Saturated volumetric soil moisture content layer 1
ThetaSat1 for other	thetas1_other.map	U: [-] R: 0 < map < 1	Saturated volumetric soil moisture content layer 1
ThetaSat2 for forest and other	thetas2.map	U: [-] R: 0 < map < 1	Saturated volumetric soil moisture content layer 2
ThetaRes1 for forest	thetar1_forest.map	U: [-] R: 0 < map < 1	Residual volumetric soil moisture content layer 1
ThetaRes1 for other	thetar1_other.map	U: [-] R: 0 < map < 1	Residual volumetric soil moisture content layer 1
ThetaRes2 for forest and other	thetar2.map	U: [-] R: 0 < map < 1	Residual volumetric soil moisture content layer 2
Lambda1 for forest	lambda1_forest.map	U: [-] R: 0 < map < 1	Pore size index (λ) layer 1
Lambda1 for other	lambda1_other.map	U: [-] R: 0 < map < 1	Pore size index (λ) layer 1
Lambda2 for forest and other	lambda2.map	U: [-] R: 0 < map < 1	Pore size index (λ) layer 2
GenuAlpha1 for forest	alpha1_forest.map	U: [-] R: 0 < map < 1	Van Genuchten parameter α layer 1
GenuAlpha1 for other	alpha1_other.map	U: [-] R: 0 < map < 1	Van Genuchten parameter α layer 1
GenuAlpha2 for forest and other	alpha2.map	U: [-] R: 0 < map < 1	Van Genuchten parameter α layer 2
Sat1 for forest	ksat1_forest.map	U: [cm day ⁻¹] R: 1 ≤ map ≤ 100	Saturated conductivity layer 1
Sat1 for other	ksat1_other.map	U: [cm day ⁻¹] R: 1 ≤ map ≤ 100	Saturated conductivity layer 1
Sat2 for forest and other	ksat2.map	U: [cm day ⁻¹] R: 1 ≤ map ≤ 100	Saturated conductivity layer 2



Table A12.1 LISFLOOD input maps (continued from previous page)			
CHANNEL GEOMETRY			
Map	Default name	Units, range	Description
Channels	chan.map	U: [-] R: 0 or 1	Map with Boolean 1 for all channel pixels, and Boolean 0 for all other pixels on MaskMap
ChanGrad	changrad.map	U: [m m ⁻¹] R: map > 0 !!!	Channel gradient
ChanMan	chanman.map	U: [-] R: map > 0	Manning's roughness coefficient for channels
ChanLength	chanleng.map	U: [m] R: map > 0	Channel length (can exceed grid size, to account for meandering rivers)
ChanBottomWidth	chanbw.map	U: [m] R: map > 0	Channel bottom width
ChanSdXdY	chans.map	U: [m m ⁻¹] R: map ≥ 0	Channel side slope Important: defined as horizontal divided by vertical distance (dx/dy); this may be confusing because slope is usually defined the other way round (i.e. dy/dx)!
ChanDepth Threshold	chanbnkf.map	U: [m] R: map > 0	Bankfull channel depth
METEOROLOGICAL VARIABLES			
Map	Default prefix	Units, range	Description
Precipitation Maps	pr	U: [mm day ⁻¹] R: map ≥ 0	Precipitation rate
TavgMaps	ta	U: [°C] R: -50 ≤ map ≤ +50	Average <i>daily</i> temperature\
E0Maps	e	U: [mm day ⁻¹] R: map ≥ 0	Daily potential evaporation rate, free water surface
ES0Maps	es	U: [mm day ⁻¹] R: map ≥ 0	Daily potential evaporation rate, bare soil
ET0Maps	et	U: [mm day ⁻¹] R: map ≥ 0	Daily potential evapotranspiration rate, reference crop
DEVELOPMENT OF VEGETATION OVER TIME			
Map	Default prefix	Units, range	Description
LAIMaps for forest	lai_forest	U: [m ² m ⁻²] R: map ≥ 0	Pixel-average Leaf Area Index for forest
LAIMaps for other	lai_other	U: [m ² m ⁻²] R: map ≥ 0	Pixel-average Leaf Area Index for other
DEFINITION OF INPUT/OUTPUT TIMESERIES			
Map	Default name	Units, range	Description
Gauges	outlets.map	U: [-] R: For each station an individual number	Nominal map with locations at which discharge timeseries are reported (usually correspond to gauging stations)
Sites	sites.map	U: [-] R: For each station an individual number	Nominal map with locations (individual pixels or areas) at which timeseries of intermediate state and rate variables are reported (soil moisture, infiltration, snow, etcetera)





	<i>Table A12.2 Optional maps that define grid size</i>			
	Map	Default name	Units, range	Description
	PixelLengthUser	pixleng.map	U.: [m] R.: map > 0	Map with pixel length
	PixelAreaUser	pixarea.map	U.: [m] R.: map > 0	Map with pixel area
	<i>Table A12.3 LISFLOOD input tables</i>			
	LAND USE			
	Table	Default name	Description	
	Day of the year -> LAI	LaiOfDay.txt	Lookup table: Day of the year -> LAI map	
List and characteristics of output variables (type, time step, spatial resolution, etc.)	<i>Table A13.1 LISFLOOD default output time series</i>			
	RATE VARIABLES AT GAUGES			
	Description	Units	Settings variable	File name
	^{1,2} channel discharge	m ³ s ⁻¹	disTS	dis.tss
	NUMERICAL CHECKS			
	Description	Units		File name
	² cumulative mass balance error	m ³	WaterMassBalanceTSS	mbError.tss
	² cumulative mass balance error, expressed as mm water slice (average over catchment)	mm	MassBalanceMM TSS	mbErrorMm.tss
	² number of sub-steps needed for channel routing	-	NoSubStepsChan	NoSubStepsChannel.tss
	² number of sub-steps needed for gravity-based soil moisture routine	-	StepsSoilTS	steps.tss
	¹ Output only if option 'InitLisflood' = 1 (pre-run)			
² Output only if option 'InitLisflood' = 0				



Table A13.2 LISFLOOD optional output time series (only 'InitLisflood' = 0) (continued on next pages)			
STATE VARIABLES AT SITES (option repStateSites)¹⁶			
Description	Units	Settings variable	Default name
depth of water on soil surface	mm	WaterDepthTS	wDepth.tss
depth of snow cover on soil surface (pixel-average) ¹⁷	mm	SnowCoverTS	snowCover.tss
depth of interception storage	mm	CumInterceptionTS	cumInt.tss
soil moisture content upper layer	mm ³ / mm ³	Theta1TS	thTop.tss
soil moisture content lower layer	mm ³ / mm ³	Theta2TS	thSub.tss
storage in upper groundwater zone	mm	UZTS	uz.tss
storage in lower groundwater zone	mm	LZTS	lz.tss
number of days since last rain	days	DSLRTS	dslr.tss
frost index	°C days ⁻¹	FrostIndexTS	frost.tss
Table A13.2 LISFLOOD optional output time series (continued from previous page)			
RATE VARIABLES AT SITES (option repRateSites)¹⁸			
Description	Units	Settings variable	Default name
rain (excluding snow)	mm/timestep	RainTS	rain.tss
snow ¹⁹	mm/timestep	SnowTS	snow.tss
snow melt	mm/timestep	SnowmeltTS	snowMelt.tss
actual evaporation	mm/timestep	ESActTS	esAct.tss
actual transpiration	mm/timestep	TaTS	tAct.tss
rainfall interception	mm/timestep	InterceptionTS	interception.tss
evaporation of intercepted water	mm/timestep	EWIntTS	ewIntAct.tss
leaf drainage	mm/timestep	LeafDrainageTS	leafDrainage.tss
infiltration	mm/timestep	InfiltrationTS	infiltration.tss
preferential (bypass) flow	mm/timestep	PrefFlowTS	prefFlow.tss
percolation upper to lower soil layer	mm/timestep	PercolationTS	dTopToSub.tss
percolation lower soil layer to subsoil	mm/timestep	SeepSubToGWTS	dSubToUz.tss
surface runoff	mm/timestep	SurfaceRunoffTS	surfaceRunoff.tss
outflow from upper zone	mm/timestep	UZOutflowTS	qUz.tss
outflow from lower zone	mm/timestep	LZOutflowTS	qLz.tss
total runoff	mm/timestep	TotalRunoffTS	totalRunoff.tss
percolation from upper to lower zone	mm/timestep	GwPercUZLZTS	percUZLZ.tss
loss from lower zone	mm/timestep	GwLossTS	loss.tss





TIME SERIES, AVERAGE UPSTREAM OF GAUGES				
METEOROLOGICAL INPUT VARIABLES (option repMeteoUpsGauges)				
Description	Units	Settings variable	Default name	
precipitation	mm/timestep	PrecipitationAvUpsTS	precipUps.tss	
potential reference evapotranspiration	mm/timestep	ETRefAvUpsTS	etUps.tss	
potential evaporation from soil	mm/timestep	ESRefAvUpsTS	esUps.tss	
potential open water evaporation	mm/timestep	EWRefAvUpsTS	ewUps.tss	
average daily temperature	°C	TavgAvUpsTS	tAvgUps.tss	
<i>Table A13.2 LISFLOOD optional output time series (continued from previous page)</i>				
STATE VARIABLES (option repStateUpsGauges)				
Description	Units	Settings variable	Default name	
depth of water on soil surface	mm	WaterDepthAvUpsTS	wdepthUps.tss	
depth of snow cover on	mm	SnowCoverAvUpsTS	snowCoverUps.tss	
depth of interception storage	mm	CumInterceptionAvUpsTS	cumInterceptionUps.tss	
soil moisture upper layer	mm ³ / mm ³	Theta1AvUpsTS	thTopUps.tss	
soil moisture lower layer	mm ³ / mm ³	Theta2AvUpsTS	thSubUps.tss	
groundwater upper zone	mm	UZAvUpsTS	uzUps.tss	
groundwater lower zone	mm	LZAvUpsTS	lzUps.tss	
number of days since last rain	Days	DSLRAvUpsTS	dslrUps.tss	
frost index	°C days ⁻¹	FrostIndexAvUpsTS	frostUps.tss	
<i>Table A13.2 LISFLOOD optional output time series (continued from previous page)</i>				
RATE VARIABLES (option repRateUpsGauges)				
Description	Units	Settings variable	Default name	
rain (excluding snow)	mm/timestep	RainAvUpsTS	rainUps.tss	
snow ²⁰	mm/timestep	SnowAvUpsTS	snowUps.tss	
snow melt	mm/timestep	SnowmeltAvUpsTS	snowMeltUps.tss	
actual evaporation	mm/timestep	ESActAvUpsTS	esActUps.tss	
actual transpiration	mm/timestep	TaAvUpsTS	tActUps.tss	
rainfall interception	mm/timestep	InterceptionAvUpsTS	interceptionUps.tss	
evaporation of intercepted water	mm/timestep	EWIntAvUpsTS	ewIntActUps.tss	
leaf drainage	mm/timestep	LeafDrainageAvUpsTS	leafDrainageUps.tss	
infiltration	mm/timestep	InfiltrationAvUpsTS	infiltrationUps.tss	
preferential (bypass) flow	mm/timestep	PrefFlowAvUpsTS	prefFlowUps.tss	
percolation upper to lower soil layer	mm/timestep	PercolationAvUpsTS	dTopToSubUps.tss	
percolation lower soil layer to subsoil	mm/timestep	SeepSubToGWAUpsTS	dSubToUzUps.tss	
surface runoff	mm/timestep	SurfaceRunoffAvUpsTS	surfaceRunoffUps.tss	
outflow from upper zone	mm/timestep	UZOutflowAvUpsTS	qUzUps.tss	
outflow from lower zone	mm/timestep	LZOutflowAvUpsTS	qLzUps.tss	
total runoff	mm/timestep	TotalRunoffAvUpsTS	totalRunoffUps.tss	
percolation upper to lower zone	mm/timestep	GwPercUZLZAvUpsTS	percUZLZUps.tss	
loss from lower zone	mm/timestep	GwLossTS	lossUps.tss	
WATER LEVEL IN CHANNEL (option repWaterLevelTs)				
Description	Units	Settings variable	Default name	
water level in channel	m (above channel bottom)	WaterLevelTS	waterLevel.tss	
OUTPUT RELATED TO LOWER ZONE INITIALISATION				
Description	Option		Settings variable	Default name
average inflow into lower zone	repLZAvInflowSites	mm day ⁻¹	LZAvInflowTS	lzAvIn.tss
average inflow into lower zone	repLZAvInflowUpsGauges	mm day ⁻¹	LZAvInflowAvUpsTS	lzAvInUps.tss



Table A13.3 LISFLOOD default output maps			
AVERAGE RECHARGE MAP (for lower groundwater zone) (option InitLisflood)			
Description	Units	File name	Domain
¹ average inflow to lower zone	mm day ⁻¹	lzavin.map	other fraction
¹ average inflow to lower zone (forest)	mm day ⁻¹	lzavin_forest.map	forest fraction
INITIAL CONDITION MAPS at defined time steps²¹ (option repStateMaps)			
Description	Units	File name ²²	Domain
² waterdepth	mm	wdepth00.xxx	whole pixel
² channel cross-sectional area	m ²	chcro000.xxx	channel
² days since last rain variable	days	dslr0000.xxx	other pixel
² snow cover zone A ²³	mm	scova000.xxx	snow zone A (1/3 rd pixel)
² snow cover zone B	mm	scovb000.xxx	snow zone B (1/3 rd pixel)
² snow cover zone C	mm	scovc000.xxx	snow zone C (1/3 rd pixel)
² frost index	°C days ⁻¹	frost000.xxx	other pixel
² cumulative interception	mm	cumi0000.xxx	other pixel
² soil moisture upper layer	mm ³ /mm ³	thtop000.xxx	other fraction
² soil moisture lower layer	mm ³ /mm ³	thsub000.xxx	other fraction
² water in lower zone	mm	lz000000.xxx	other fraction
² water in upper zone	mm	uz000000.xxx	other fraction
² days since last rain variable (forest)	days	dslF0000.xxx	forest pixel
² cumulative interception (forest)	mm	cumF0000.xxx	forest pixel
² soil moisture upper layer (forest)	mm ³ /mm ³	thFt0000.xxx	forest fraction
² soil moisture lower layer (forest)	mm ³ /mm ³	thFs0000.xxx	forest fraction
² water in lower zone (forest)	mm	lzF00000.xxx	forest fraction
² water in upper zone (forest)	mm	uzF00000.xxx	forest fraction
² water in depression storage (sealed)	mm	cseal000.xxx	sealed fraction
¹ Output only if option 'InitLisflood' = 1 (pre-run)			
² Output only if option 'InitLisflood' = 0			





Table A13.4 LISFLOOD optional output maps (only 'InitLisflood' = 0) (continued on next page)				
DISCHARGE AND WATER LEVEL				
Description	Option	Units	Settings variable	Prefix
discharge	repDischargeMaps	m ³ s ⁻¹	DischargeMaps	dis
water level	repWaterLevelMaps	m (above channel bottom)	WaterLevelMaps	wl
METEOROLOGICAL INPUT VARIABLES				
Description	Option		Settings variable	Prefix
precipitation	repPrecipitationMaps	mm	PrecipitationMaps	pr
potential reference evapotranspiration	repETRefMaps	mm	ETRefMaps	et
potential evaporation from soil	repESRefMaps	mm	ESRefMaps	es
potential open water evaporation	repEWRefMaps	mm	EWRefMaps	ew
average daily temperature	repTavgMaps	mm	TavgMaps	tav
STATE VARIABLES²⁴				
Description	Option		Settings variable	Prefix
depth of water on soil surface	repWaterDepthMaps	mm	WaterDepthMaps	wdep
channel cross-sectional area	repChanCrossSectionMaps	m ²	ChanCrossSectionMaps	chcro
depth of snow cover on soil surface	repSnowCoverMaps	mm	SnowCoverMaps	scov
depth of interception storage	repCumInterceptionMaps	mm	CumInterceptionMaps CumInterceptionForestMaps	cumi cumF
soil moisture content upper layer	repTheta1Maps	mm ³ / mm ³	Theta1Maps Theta1ForestMaps	thtop thFt
soil moisture content lower layer	repTheta2Maps	mm ³ / mm ³	Theta2Maps Theta2ForestMaps	thsub thFs
storage in upper groundwater zone	repUZMaps	mm	UZMaps UZForestMaps	uz uzF
storage in lower groundwater zone	repLZMaps	mm	LZMaps LZForestMaps	lz lzF
number of days since last rain	repDSLRLMaps	days	DSLRLMaps DSLRLForestMaps	dslr dslF
frost index	repFrostIndexMaps	°C days ⁻¹	FrostIndexMaps	frost
RATE VARIABLES²⁵				
Description	Option		Settings variable	Prefix
rain (excluding snow)	repRainMaps	mm/timestep	RainMaps	rain
snow ²⁶	repSnowMaps	mm/timestep	SnowMaps	snow
snow melt	repSnowMeltMaps	mm/timestep	SnowMeltMaps	smelt
actual evaporation	repESActMaps	mm/timestep	ESActMaps	esact
actual transpiration	repTaMaps	mm/timestep	TaMaps	tact
rainfall interception	repInterceptionMaps	mm/timestep	InterceptionMaps	int
evaporation of intercepted water	repEWIntMaps	mm/timestep	EWIntMaps	ewint
leaf drainage	repLeafDrainageMaps	mm/timestep	LeafDrainageMaps	ldra
infiltration	repInfiltrationMaps	mm/timestep	InfiltrationMaps	inf
Table 12.4 LISFLOOD optional output maps (continued from previous page)				
preferential (bypass) flow	repPrefFlowMaps	mm/timestep	PrefFlowMaps	pflow
percolation upper to lower soil layer	repPercolationMaps	mm/timestep	PercolationMaps	to2su
percolation lower soil layer to subsoil	repSeepSubToGWMaps	mm/timestep	SeepSubToGWMaps	su2gw
surface runoff	repSurfaceRunoffMaps	mm/timestep	SurfaceRunoffMaps	srun
outflow from upper zone	repUZOutflowMaps	mm/timestep	UZOutflowMaps	quz
outflow from lower zone	repLZOutflowMaps	mm/timestep	LZOutflowMaps	qlz
total runoff	repTotalRunoffMaps	mm/timestep	TotalRunoffMaps	trun
percolation upper to lower zone	repGwPercUZLZMaps	mm/timestep	GwPercUZLZMaps	uz2lz
loss from lower zone	repGwLossMaps	mm/timestep	GwLossMaps	loss
5. Examples of previous model applications				



<p>Catchments, objectives, etc.</p> <p>Results of existing comparisons with other models</p>	
<p>6. List of selected references</p>	
	<ul style="list-style-type: none"> ● De Roo, A. P. J., C. G. Wesseling, and W. P. A. Van Deursen, 2000: Physically Based River Basin Modelling within a GIS: The LISFLOOD Model. <i>Hydrological Processes</i>, 14, 11-12, 1981–92. doi:10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F. ● “LISFLOOD - Distributed Water Balance and Flood Simulation Model - Revised User Manual 2013 - EU Science Hub - European Commission.” 2013. EU Science Hub. October 14. https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/lisflood-distributed-water-balance-and-flood-simulation-model-revised-user-manual-2013. ● Smith, Paul, Florian Pappenberger, Fredrik Wetterhall, Jutta Thielen, Blazej Krzeminski, Peter Salamon, Davide Muraro, Milan Kalas, and Calum Baugh. 2016. On the Operational Implementation of the European Flood Awareness System (EFAS). European Centre for Medium-Range Weather Fore-





	<p>casts.</p> <p>http://www.ecmwf.int/sites/default/files/elibrary/2016/16337-operational-implementation-european-flood-awareness-system-efas.pdf.</p> <ul style="list-style-type: none">● Bódis, K., 2009. Development of a data set for continental hydrologic modelling.● De Roo, A., Thielen, J., Gouweleeuw, B., 2003. LISFLOOD, a Distributed WaterBalance, Flood Simulation, and Flood Inundation Model, User Manual version 1.2. Internal report, Joint Research Center of the European Communities, Ispra, Italy, 74 pp.● Hock, R., 2003. Temperature index melt modelling in mountain areas. <i>Journal of Hydrology</i>, 282(1-4), 104–115.● Van Der Knijff, J., De Roo, A., 2006. LISFLOOD – Distributed Water Balance and Flood Simulation Model, User Manual. EUR 22166 EN, Office for Official Publications of the European Communities, Luxembourg, 88 pp.● Van der Knijff, J., 2008. LISVAP– Evaporation Pre-Processor for the LISFLOOD Water Balance and Flood Simulation Model, Revised User Manual. EUR 22639 EN/2, Office for Official Publications of the European Communities, Luxembourg, 31 pp.● Van Der Knijff, J., De Roo, A., 2008. LISFLOOD – Distributed Water Balance and Flood Simulation Model, Revised User Manual. EUR 22166 EN/2, Office for Official Publications of the European Communities, Luxembourg, 109 pp.● Van der Knijff, J. M., Younis, J. and de Roo, A. P. J.: LISFLOOD: A GIS-based distributed model for river basin scale water balance and flood simulation, <i>Int. J. Geogr. Inf. Sci.</i>, 24(2), 189–212, 2010.
--	--



Table 8 Tabulated overview on hydrological model features of wflow_w3ra

1. General Information	
Model name	wflow_w3ra (World Wide Water Resources Analysis) + wflow_routing
Version	v1
Author(s) / First publication	Van Dijk et al. (2013)
Contact person	Albrecht Weerts (albrecht.weerts@deltares.nl) Jaap Schellekens (jaap.schellekens@deltares.nl)
Institute	Deltares
Website	https://github.com/openstreams/wflow
General modelling objectives	Calculation/prediction of hydrological water resources
Domain of applicability (catchment types and climate conditions)	The W3RA model has been applied on global scale
2. Model description	





Model type (empirical, conceptual, physically based)	Conceptual/Process-based model
Continuous or event-based	Continuous
Possible running time steps	Daily (also hourly)
Spatial discretization (lumped, semidistributed, distributed)	Spatially distributed at the sub-basin scale. Sub-basin resolution depends on the application. In Europe, this is 215 km ² .
Short description of model structure detailing main function (evaporation, soil moisture accounting, groundwater, routing, snowmelt, etc.)	<p style="text-align: center;">World Wide Water Resources Assessment (W3RA) Rainfall-Runoff Model of the ANU Based on the BoM Australian Water Resources Assessment (AWRA) model Estimates water storage in three soil layers, shallow groundwater & streams Incorporates HBV-96 snow module 0.5° Resolution (~50 km) , Soon to be 5 km River Network derived from NASA SRTM 90 m resolution digital elevation data and HydroSHEDS</p> <p style="text-align: center;">wflow Deltares River Routing Model Run at the resolutions of the Rainfall-Runoff Models</p> <p>from Emmerton et al. (2016)</p>





parameter estimation (measurement, manual or automatic algorithm, etc.)	characteristics in the landscape (see van Dijk et al. 2013 and references therein)
4. Model inputs / Model outputs	
List and characteristics of input variables (type, time step, spatial resolution, etc.)	The main dynamic variables are climate forcing data such as precipitation, temperature and potential evaporation. Other observations can also be used for different purposes.
List and characteristics of output variables (type, time step, spatial resolution, etc.)	The exported time-step depends on the user, i.e. hourly, daily. The variables (states and fluxes) can be exported for the whole grid or selected gauge locations .
5. Examples of previous model applications	



Catchments, objectives, etc. Results of existing comparisons with other models	Schellekens, J., Dutra, E., Martínez-de la Torre, A., Balsamo, G., van Dijk, A., Sperna Weiland, F., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., Fink, G., Flörke, M., Peßenteiner, S., van Beek, R., Polcher, J., Beck, H., Orth, R., Calton, B., Burke, S., Dorigo, W., and Weedon, G. P.: A global water resources ensemble of hydrological models: the earth2Observe Tier-1 dataset, <i>Earth Syst. Sci. Data Discuss.</i> , doi:10.5194/essd-2016-55, in review, 2016.
6. List of selected references	
	Van Dijk, Pea-Arancibia, J. L., Wood, E. F., Sheffield, J., & Beck, H. E. (2013). Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide. <i>Water Resources Research</i> , 49(5), 2729–2746. http://doi.org/10.1002/wrcr.20251

Table 9 Tabulated overview on hydrological model features of wflow_hbv

1. General Information	
Model name	wflow_hbv
Version	V1
Author(s) / First publication	Lindstrom et al. (1997), see also Rakovec et al. (2012, 2015)





Contact person	Albrecht Weerts (albrecht.weerts@deltares.nl)
Institute	Deltares
Website	https://github.com/openstreams/wflow
General modelling objectives	Calculation/prediction of hydrological water resources
Domain of applicability (catchment types and climate conditions)	wflow_hbv has been applied on catchment to global scale
2. Model description	
Model type (empirical, conceptual, physically based)	Conceptual/Process-based model
Continuous or event-based	Continuous
Possible running time steps	Hourly, daily
Spatial discretization (distributed)	spatially distributed, grid size determined by end user (grid size for Rhine 1.44 km ²)



<p>Short description of model structure detailing main function (evaporation, soil moisture accounting, groundwater, routing, snowmelt, etc.)</p>	<p>The Hydrologiska Byrans Vattenbalansavdelning (HBV) model was introduced back in 1972 by the Swedish Meteorological and Hydrological Institute (SMHI). The HBV model is mainly used for runoff simulation and hydrological forecasting. The model is particularly useful for catchments where snow fall and snow melt are dominant factors, but application of the model is by no means restricted to these type of catchments.</p> <p>The wflow_hbv model is based on the HBV-96 model. However, the hydrological routing represent in HBV by a triangular function controlled by the MAXBAS parameter has been removed. Instead, the kinematic wave function is used to route the water downstream. All runoff that is generated in a cell in one of the HBV reservoirs is added to the kinematic wave reservoir at the end of a timestep. There is no connection between the different HBV cells within the model. Wherever possible all functions that describe the distribution of parameters within a subbasin have been removed as this is not needed in a distributed application/</p> <p>A catchment is divided into a number of grid cells. For each of the cells individually, daily runoff is computed through application of the HBV-96 of the HBV model. The use of the grid cells offers the possibility to turn the HBV modelling concept, which is originally lumped, into a distributed model.</p>
---	---





	<p>Adding lakes and reservoirs is also possible</p>
<p>Scheme of model structure</p>	
<p>3. Model parameters</p>	
<p>Distribution of model parameters (yes/no)</p>	<p>Yes</p>
<p>Number of free parameters</p>	<p>Numerous free parameters</p>
<p>Procedure of model parameter estimation (measurement, manual or automatic algorithm, etc.)</p>	<p>The parameter are obtained from lumped daily model calibrated using GLUE from upstream to downstream (ref)</p>
<p>4. Model inputs / Model outputs</p>	

<p>List and characteristics of input variables (type, time step, spatial resolution, etc.)</p>	<p>The main dynamic variables are climate forcing data such as precipitation, temperature and potential evaporation. For the daily model the other observations can also be used for different purposes (for instance DA using OpenDA).</p>
<p>List and characteristics of output variables (type, time step, spatial resolution, etc.)</p>	<p>The exported time-step depends on the user, i.e. hourly, daily. The variables (states and fluxes) can be exported for the whole grid or selected gauge locations .</p>
<p>5. Examples of previous model applications</p>	
<p>Catchments, objectives, etc. Results of existing comparisons with other models</p>	<p>Many (e.g. Rhine, Meuse etc)</p>
<p>6. List of selected references</p>	





- Lindstrom, G., Johansson, B., Persson, M., Gardelin, M., and Bergstrom, S.: Development and test of the distributed HBV-96 hydrological model, *J. Hydrol.*, 201, 272–288, 1997
- Rakovec, O., Weerts, A. H., Hazenberg, P., Torfs, P. J. J. F., and Uijlenhoet, R.: State updating of a distributed hydrological model with Ensemble Kalman Filtering: effects of updating frequency and observation network density on forecast accuracy, *Hydrol. Earth Syst. Sci.*, 16, 3435–3449, doi:10.5194/hess-16-3435-2012, 2012
- Operational aspects of asynchronous streamflow assimilation for improved flood forecasting, O. Rakovec, A. H. Weerts, J. Sumihar, and R. Uijlenhoet, *HESS*, 19(6), 2911-2924 doi:10.5194/hess-19-2911-2015.





A8: Flexible operational seasonal river flow forecasting

This poster presents a co-author contribution arising through collaboration during this PhD, summarised in Chapter 4, Sect. 4.4.

L.A. conceived the experiment, designed and carried out the analysis.

Flexible operational seasonal river flow forecasting



Louise Arnal^{1,2}, Florian Pappenberger^{2,3}, Paul Smith², Hannah Cloke^{1,4}, Liz Stephens¹

I.l.s.arnal@pgr.reading.ac.uk; louise.arnal@ecmwf.int



¹Department of Geography and Environmental Science, University of Reading, UK - ²ECMWF, UK - ³School of Geographical Sciences, University of Bristol, UK - ⁴Department of Meteorology, University of Reading, UK

Introduction

Many operational dynamic seasonal hydrological forecasting systems are static. Meaning that a single forecasting method is used in time and space. However, predictability sources vary with time and space, and this change should be reflected in the way we forecast the river flow from one season to the next. Could this be done by combining several different forecasting methods operationally?

Aim: Present the concept of a new flexible operational seasonal river flow forecasting system.

1. What are the predictability sources in my river basin?

The EPB (End Point Blending) is a cheap and easily implementable sensitivity analysis method*. It shows the dominant predictability source (the Initial Hydrological Conditions [IHCs] or the Seasonal Climate Forecasts [SCFs]) of a seasonal river flow forecasting system's forecast skill.

Method

New ensemble hindcasts are produced by combining the ESP and reverse-ESP seasonal river flow hindcasts with the climatology and the 'perfect' hindcasts (see Fig. 1). The number of members selected from each of these four data sources varies given the level of IHC vs. SCF skill that we want to reflect (points in Fig. 2).

E.g., A new ensemble hindcast with 50% SCF skill and 100% (or perfect) IHC skill → 50% ESP and 50% 'perfect'.

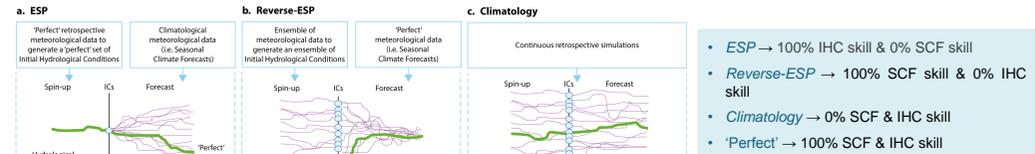


Figure 1 Schematic of the hindcasts needed for the EPB

Visualising the results

Skill surface plots (see Fig. 2) show the variations in the (new) river flow hindcast skill (here in terms of R²) with varying levels of IHC and SCF skill. This indicates the main predictability source of the river flow hindcasts for each river basin – forecast starting date – forecast lead time and aggregation period :

- horizontal skill gradients → SCF
- vertical skill gradients → IHC

The information contained in the skill surface plots can be summarised in two single values: the IHC and SCF skill elasticities (E_{IHC} and E_{SCF} ; see Fig. 3).

$$E_{IHC} [E_{SCF}] = \frac{\Delta \text{hindcast skill}}{\Delta \text{IHC} [\text{SCF} \text{ skill}]}, \text{ measured in the skill surface plots.}$$

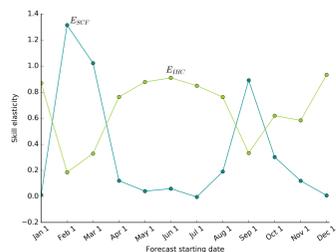


Figure 3 Skill elasticities

- $E_{IHC} [E_{SCF}] > 0$ → positive improvements in the river flow hindcast skill from improving the IHC [SCF] skill
- $E = 0$ → no improvements in the hindcast skill
- $E < 0$ → negative improvements in the hindcast skill

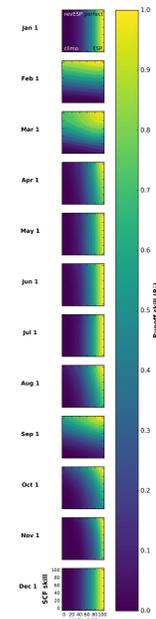


Figure 2 Skill surface plots

2. Flexible seasonal forecasting: concept

How can we use the results of the EPB to improve future seasonal river flow forecasts?

- Invest in improving the SCF by using e.g., SCF instead of climatological meteorological data (Seas. vs. ESP)
- Invest in improving the IHC through e.g., data assimilation (DA)
- Invest in improving both, interchangeably → flexible seasonal forecasting

Concept

Allow for the use of a different seasonal river flow forecasting method within the existing system, depending on the dominant predictability source for a given river basin – forecast starting date – forecast lead time and aggregation period. The idea is to use an operationally cheap yet skilful forecasting system.

→ Each new forecast would be produced following the decision tree displayed on Fig. 4.

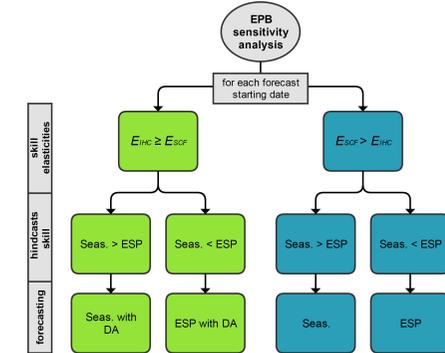


Figure 4 Decision tree to produce the flexible seasonal river flow forecasts for a given river basin – forecast lead time and aggregation period. Seas. refers to a seasonal river flow forecast produced by forcing the hydrological model with SCF

3. Example: proof of concept

The concept was tested on the EFAS (European Flood Awareness System) seasonal river flow forecasts (ECMWF's System 4 SCF run through the LISFLOOD hydrological model):

- 3-monthly discharge averages with 0 lead time,
- discharge aggregated over 74 European regions,
- verification score: CRPSS of the forecast against the climatology.

The EPB was run from 02-1990 to 12-2013 and the flexible seasonal forecasting method was used to forecast river flows over Europe from 01-2014 to 12-2014.

N.B. The results do not include any data assimilation yet.

Results

The map (see Fig. 5) shows the most skilful seasonal river flow forecasting system for 2014. The flexible seasonal forecasting system is more skilful in 21 regions, vs. 16 for the Seas. and 37 for the ESP.

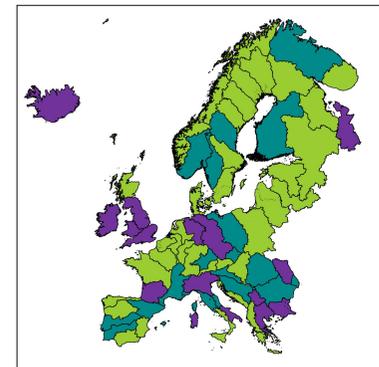


Figure 5 Map of the most skilful seasonal river flow forecasting system for 74 European regions for the year 2014

The range of CRPSS values obtained for the flexible seasonal forecasting system are very similar to the ESP and better than the seasonal forecasting system (see Fig. 6).

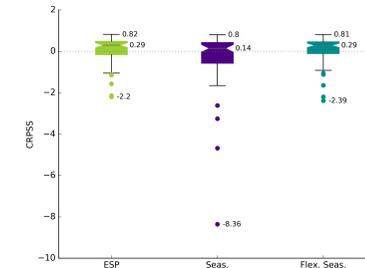


Figure 6 Boxplots of the CRPSS obtained for the 3 forecasting systems for 74 European regions for the year 2014

Take-home messages & references

This new flexible seasonal river flow forecasting system has the potential to be **more skilful** than the ESP and seasonal forecasts and **cheaper** to run operationally compared to running a seasonal forecasting system with DA.



*Arnal, L., A. Wood, E. Stephens, H. Cloke, and F. Pappenberger, 2017: An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity. J. Hydrometeor. doi:10.1175/JHM-D-16-0259.1, in press.

A9: “Are we talking just a bit of water out of bank? Or is it Armageddon?” Front line perspectives on transitioning to probabilistic fluvial flood forecasts in England

This paper presents the submitted (currently in review) version of Chapter 5, Sect. 5.2 of this thesis, with the following reference:

Arnal, L., L. Anspoks, S. Manson, J. Neumann, T. Norton, E. Stephens, L. Wolfenden, and H. L. Cloke, 2019: Are we talking just a bit of water out of bank? Or is it Armageddon? Front line perspectives on transitioning to probabilistic fluvial flood forecasts in England, *Geosci. Commun. Discuss.*, doi:10.5194/gc-2019-18, in review*

* ©2019. The Authors. Geoscience Communication, a journal of the European Geosciences Union published by Copernicus. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided that the original work is properly cited.



“Are we talking just a bit of water out of bank? Or is it Armageddon?” Front line perspectives on transitioning to probabilistic fluvial flood forecasts in England

Louise Arnal^{1,2}, Liz Anspoks³, Susan Manson³, Jessica Neumann¹, Tim Norton³, Elisabeth Stephens¹,
5 Louise Wolfenden³, Hannah Louise Cloke^{1,4,5}

¹University of Reading, UK

²European Centre for Medium-Range Weather Forecasts, UK

³Environment Agency, UK

⁴Uppsala University, Sweden

10 ⁵Centre of Natural Hazards and Disaster Science, Sweden

Correspondence to: Louise Arnal (l.l.s.arnal@pgr.reading.ac.uk)

Abstract. The inclusion of uncertainty in flood forecasts is a recent, important yet challenging endeavour. In the chaotic and far from certain world we live in, probabilistic estimates of potential future floods are vital. By showing the uncertainty surrounding a prediction, probabilistic forecasts can give an earlier indication of potential future floods, increasing the amount
15 of time we have to prepare. In practice, making a binary decision based on probabilistic information is challenging. The Environment Agency (EA), responsible for managing risks of flooding in England, is in the process of a transition to probabilistic fluvial flood forecasts. A series of interviews were carried out with EA decision-makers (i.e. duty officers) to understand how this transition might affect their decision-making activities. The interviews highlight the complex and evolving landscape (made of alternative ‘hard scientific facts’ and ‘soft values’) in which EA duty officers operate, where forecasts play
20 an integral role in decision-making. While EA duty officers already account for uncertainty and communicate their confidence in the system they use, they view the transition to probabilistic flood forecasts as both an opportunity and a challenge in practice. Based on the interview results, recommendations are made to the EA to ensure a successful transition to probabilistic forecasts for flood early warning in England.

We believe that this paper is of wide interest for a range of sectors at the intersection between geoscience and society. A
25 glossary of technical terms is highlighted by asterisks in the text and included in Appendix A.

1 Introduction

One of the most recent and significant challenges in hydrology has been the inclusion of uncertainty information in flood forecasts. We live in a world where it is currently impossible to say with 100% certainty how the weather will evolve in the following days to months, or by how much exactly a river level is expected to change. This is due to the inaccurate measurement
30 of hydro-meteorological observations*, errors in the mathematical models used to produce these forecasts (due to scientific and technical limitations) and, most importantly, nature’s intrinsic chaos* (Lorenz, 1969; Buizza, 2008). In this world, probabilistic estimates of potential future floods are vital. Probabilistic forecasts* give a range of likely possible future outcomes, contrary to deterministic forecasts*, which indicate a single future possibility (Buizza, 2008). Probabilistic flood forecasts are generally produced by forcing* a hydrological model* with an ensemble* of future meteorological scenarios
35 (Cloke and Pappenberger, 2009). By giving an idea of the uncertainty surrounding a prediction, probabilistic forecasts can give an earlier indication of potential future extreme events, such as floods, increasing the amount of time decision-makers have to prepare (Buizza, 2008; Stephens and Cloke, 2014).

In practice however, probabilistic forecasts can be challenging to use for operational decision-making*, given their uncertain nature (Nicholls, 1999; Cloke and Pappenberger, 2009; Demeritt et al., 2010; Nobert et al., 2010; Ramos et al., 2010; Stephens



40 et al., 2019). Having to translate a range of possible outcomes into a binary decision (such as sending out a flood warning) is intricate and requires careful interpretation, an understanding of probabilities, risk*, uncertainty* (Dessai and Hulme, 2004) and of the systems modelled. Furthermore, probabilistic forecasts are designed to capture scenarios that may not always realise, which in turn could lead to false alarms*. Decision-making can be based on a set of rules, such as threshold exceedance (Dale et al., 2013). It is, for example, possible to take decisions (e.g. send a flood warning) when a pre-defined threshold is reached

45 with a minimum forecast probability (Thielen et al., 2009). However, the decision-making process is generally based on, and influenced, by several additional factors. These include the type of event considered (e.g. a localised small flood event vs a large scale extreme flood event), the costs of taking action vs not taking action, experience of past events, the decision-maker's trust in the forecast (which can be built up over time), their risk aversion, and the cultural context in which decisions are made (Cloke et al., 2009; Arnal et al., 2016; Neumann et al., 2018).

50 The Environment Agency (EA)* is responsible for managing risks of flooding in England and their flood incident management strategy* is often shaped by major flood events (Werner et al., 2009; Stephens and Cloke, 2014; Pilling et al., 2016). In the 1990s and early 2000s, the UK policy shifted from a 'flood defence' to a 'flood risk management' strategy, on the back of the 1998 and 2000 floods (McEwen et al., 2012), which has led to more forecast-based decision-making. The summer 2007 UK floods boosted the development of the National Flood Forecasting System and the Flood Forecasting Centre (FFC*; a UK Met

55 Office and EA partnership), with the aim to improve national flood warning services (Pitt, 2008; Stephens and Cloke, 2014). The winter 2013/14 UK floods further demonstrated the value of the FFC and the use of ensemble surge forecasts* for flood preparedness* (Stephens and Cloke, 2014). It was also during the 2013/14 floods that the EA started using two fluvial (or river) flood scenarios* (a reasonable worst case* and a best estimate*, instead of a single prediction) for flood incident management. Following this, Defra (the UK government Department for Environment, Food & Rural Affairs)* published a

60 National Flood Resilience Review (NFRR) in 2016 (HM Government, 2016; House of Commons - Environment, Food and Rural Affairs Committee, 2016). This review aimed at understanding and increasing the UK's resilience to river and coastal flooding from extreme weather over the next ten years. The NFRR recommends a better integration of probabilistic weather forecasts into flood forecast products, for an improved characterisation of uncertainty and an enhanced communication of flood risk and likelihood to inform a range of flood management measures*.

65 While catastrophic events can foster the uptake of state-of-the-art science (e.g. probabilistic forecasts) for decision-making, achieving a complete and successful transition relies on many elements. For example, the use of ensemble surge forecasts in 2013/14 might not have been possible without the prior shift to a flood risk management mindset and the creation of the FFC. Moreover, we do not want to be in a situation where we require a catastrophic event in order to begin implementing the best science into risk management practice; it is vital to understand a country's and institution's cultural landscape to ensure that

70 science is not being under- or misused (Golding et al., 2017). In the case of probabilistic forecasts, making sure that they add value rather than uncertainty to operational decision-making is key (Nobert et al., 2010). Interviews can be an effective method to capture an institution's complex cultural landscape (Schoenberger, 1991; Pagano et al., 2004). They can provide interviewees with an understanding of the world (in this case the institution world) from the perspective of the informants, shedding light on their unique perceptions and information only known to them (Sivle et al., 2014).

75 As outlined by the NFRR, the EA is in the process of a transition to probabilistic fluvial flood forecasts, from the two flood scenarios they currently use operationally (Orr and Twigger-Ross, 2009; Sene et al., 2009). To capture the EA's forecasting practice landscape and understand how this transition might affect their flood decision-making activities, a series of interviews were carried out in the summer 2018 with EA 'Monitoring and Forecasting Duty Officers' (MFDOs) and 'Flood Warning Duty Officers' (FWDOs). These two roles are at the heart of the EA's flood risk management decision-making chain. The

80 outcomes of these interviews were used as a basis for this paper, with the aim to highlight the potential opportunities and challenges that this transition might translate to for the duty officers, ahead of it happening.



2 Context: the Environment Agency's flood incident management strategy

The Environment Agency (EA) is an executive non-departmental public body, sponsored by Defra. The EA has an operational responsibility to manage risks of flooding from rivers and the sea in England, by warning and informing the public and businesses about impending floods. Flood warnings are sent with a 2-hour minimum lead time*, however, different lead times have recently been introduced to take into account the type of flooding and catchment characteristics*; i.e. flash flooding vs slow responding catchment. Under the Flood and Water Management Act 2010 (DEFRA, 2010), the EA takes a lead role on river and coastal flooding, whilst lead local flood authorities take a lead role on local flood risk (which covers flooding from other sources, including surface water, groundwater and minor watercourses). The EA also has a strategic overview role for all sources of flooding and works with lead local flood authorities by providing guidance, knowledge and support in responding to surface water flooding. The following schematic (Fig. 1) displays the EA's institutional landscape, with a particular focus on the flood incident management (FIM) information flow to and from MFDOs and FWDOs.

Historically, the EA was structured as a national body, delivering its work across England in six operational regional boundaries (i.e. regional boundaries were political delineations and were roughly aligned with the regional development fund boundaries). On 1st April 2014, the EA changed its operating structure to adopt area boundaries (i.e. broadly based on catchment delineations, but some catchments span different areas, especially at the borders with Wales and Scotland). These were aligned in 2016 with the Natural England (non-departmental public body, sponsored by Defra, and responsible for ensuring the protection and improvement of England's natural environment) boundaries. The EA is now operating over 14 areas with 7 forecasting centres (hereafter referred to as 'centres'; see Fig. 2).

To help manage flood risk, the EA receive hydro-meteorological forecasts* produced by the Flood Forecasting Centre (FFC; see Fig. 1) on a daily basis (more or less frequently depending on the forecasting product* – see Sect. 4.1.1). The FFC is a partnership between the EA and the UK Met Office. It combines the hydrological and meteorological expertise from both institutes to provide hydro-meteorological forecasting products (for all natural forms of flooding, including river, surface water, coastal and groundwater flooding) to emergency responders: category 1 (e.g. police services, fire and rescue authorities, including the EA for England), category 2 (e.g. utilities, telecommunications, transport providers, Highways Agency), Natural Resources Wales (for Wales) and the Met Office (for England and Wales).

The EA's FIM is based on the principle: 'think big, act early, be visible' (EA, 2018). This is part of a wider move from incident response to risk anticipation, with the aim to ensure that resources are put in place early and that the EA is prepared to scale-up or -down (i.e. preparations for measures implemented or not closer to the potential incident; e.g. expanded incident rotas with duty officers on standby, instigating requests for mutual aid to a different area, requests for equipment to support preventative and/or repair work, such as temporary barriers and pumps). As part of this strategy, the FFC forecasts are currently (and since the UK winter floods of 2013/14) used to produce two deterministic fluvial flood scenarios with a five-day lead time at the EA, a 'Best Estimate' and a 'Reasonable Worst Case'.

Several internal documents have been written to give guidance on how to use these scenarios to support decision-making for FIM activities, in line with the EA's principle. In summary, the Reasonable Worst Case gives an indication of what 'could' happen and should be used for preparation, information and response to flooding. The Best Estimate gives an indication of what 'should' happen and should be used as the basis for planning for warning. Together, the two scenarios provide the scale and size of the incident for planning and response preparations (FFC, 2017).

According to research done in the Thames river basin (UK), New et al. (2007) showed that probabilistic forecasts provide more informative results (enabling the potential risks of impacts to be quantified) than a scenario-based approach. The transition to the two scenarios can be seen as a stepping stone towards probabilistic fluvial flood forecasts. Ultimately, the EA would like to: 1) quantify uncertainty and communicate flood risk in a clear manner internally and externally, and 2) make decisions around incident preparation and escalation, operational activities and flood warnings effectively, intelligently and accurately. While the EA acknowledges that a potential benefit of probabilistic flood forecasts is the possibility to give earlier



125 warnings, they question the extent to which probabilistic forecasts would reduce scientific and decision uncertainties in a FIM
context (Orr and Twigger-Ross, 2009).

While work has already been done by the EA to investigate the technical feasibility of a transition to probabilistic fluvial flood
forecasts (Orr and Twigger-Ross, 2009; Sene et al., 2009, Dale et al., 2013), this paper focuses on exploring the perceptions
of the EA duty officers on the subject matter. This work is important as it will ensure the appropriate use of fluvial flood
130 probabilistic forecasts for FIM decision-making activities, once operational. It should be noted that the EA already uses coastal
flood probabilistic forecasts (Flowerdew et al., 2009); this work focuses on fluvial flooding. To this end, a series of interviews
were carried out with EA ‘Monitoring and Forecasting Duty Officers’ (MFDOs) and ‘Flood Warning Duty Officers’ (FWDOs),
as they are the two roles at the heart of the EA’s internal forecast-led decision-making, building on the exchange between the
MFDOs and the FFC (see Fig. 1; more information about their respective roles in Sect. 3.1 and 4.1).

135 3 Methods

3.1 Participants

The EA has several MFDO and FWDO roles, fulfilled by a number of different people. These are voluntary roles, added to
the staff’s day-to-day job, for which they follow relevant training. MFDOs receive, process and communicate forecast
information to FWDO’s, who are responsible for interpreting the information and working out the potential impacts on the
140 ground. The duty officers’ schedules are predetermined by a rota, and duty officers are on call for a period of one week at a
time. During times of increased flood risk, when more forecasting or warning activities are required, additional rostering takes
place. Duty officers receive a range of forecasts (nowcasting* products to monthly outlooks*) and are aware of potential
situations from a month out. Five days ahead is when the activity really starts to build and is the focus of these interviews.

A total of six EA MFDOs and FWDOs from three different EA centres (one pair per centre) were interviewed to capture a
range of perspectives in relation to this topic, following best practice (Sivle et al., 2014; participant information sheet provided
as supplementary material). Forecasting and decision-making varies between EA centres due to different management
approaches and different types of geography and catchment response. To protect anonymity, the three centres where interviews
were carried out are shown in terms of the wider area they are responsible for: 1) the Yorkshire area (YOR) in the North (area
3), 2) the Thames area (THM) in the South East (area 11), and 3) the Solent and South Downs area (SSD) in the South East
150 (area 14) (Fig. 2).

MFDOs and FWDOs were interviewed in pairs as they are used to working together and the information they use sits between
these two roles. The thought was that by talking to the MFDOs alone we would lose the element of “and so what?”, while
talking to the FWDOs alone we would lose all the expertise about forecasting. All MFDOs and FWDOs interviewed had
several years of experience and so were able to describe how current practice would change with a different type of forecast.

155 Participants were selected by EA study co-developer I1 to meet the above criteria. For the purpose of anonymity, the
interviewees will thereafter be reported using codes. The three MFDOs interviewed will be referred to as MFDO1, MFDO2
and MFDO3, and the three FWDOs interviewed as FWDO1, FWDO2 and FWDO3 (interviewed pairs are however represented
by the same number). As well as those from the MFDOs and FWDOs, quotes from two EA study co-developers are reported
in this paper, I1 and I2, who helped the interviewer (Louise Arnal) by providing some context about the EA’s organisational
160 landscape, forecasting systems and MFDO and FWDO roles prior to the three interviews.

3.2 Interviews

By design, qualitative, semi-structured interviews are used to understand interviewees’ perspectives, allowing the exploration
of a research question that does not necessitate quantifying information and creating generalisations from the interview
transcripts. The strength of such studies (compared to other survey methods) is that they are more sensitive to historical and



165 institutional complexity and can capture the influence of local context (Schoenberger, 1991; Pagano et al., 2004). Moreover,
they are flexible, allowing the interviewer to remodel questions throughout an interview and from one interview to the next,
to follow up on new information discovered (Sivle et al., 2014).

A fixed set of open-ended questions were prepared in advance to guide the discussion and allow for comparability across all
three interviews. To prompt discussion, all three MFDO and FWDO pairs were asked the same opening question: “Could you
170 please walk me through what you would do ahead of a potential flood event?” The following questions were also prepared in
advanced, but their order was changed, or they were skipped depending on whether the interviewees had already answered
them:

- “Could you tell me about the uncertainties in the information you said you used in this context?”
- “How do you deal with these uncertainties?”
- 175 • “Could you tell me about how you communicate these uncertainties to each other?”
- “How would your job be influenced by a transition to probabilistic forecasts?”

Each interview lasted between 30 minutes and 1 hour 30 minutes. All interviews were conducted and digitally recorded by the
first author (Louise Arnal) in meeting rooms at the corresponding EA centres.

3.3 Data analysis

180 All interviews were transcribed verbatim and transcripts were analysed qualitatively with respect to three main research
questions. These research questions provide the structure for the results’ section of this paper (Sect. 4).

- 1) What are the MFDOs’ and FWDOs’ roles and how do they interact with one another?
- 2) Where are the forecasts currently situated amidst their decision-making process?
- 3) Considering how the duty officers communicate confidence with one another at present, what might be the potential
185 impacts of a transition to probabilistic forecasts on their roles?

Although interpretations might have been communicated by many interviewees, no frequencies are provided as quantitative
generalisations cannot be inferred from this small and purposive sample. Following best practice, the results contain a mix of
interviewees’ perspectives, supported by quotes, and further interpretation of the interview transcripts by the authors,
identifiable throughout the text (Davies et al., 2014).

190 4 Results

4.1 Roles and interactions between EA duty officers

Below, we summarise the MFDOs’ and FWDOs’ roles in an incident response context, using the interviewees’ responses to
the question: “Could you please walk me through what you would do ahead of a potential flood event?” It is worth noting that
all interviewed pairs suggested the MFDO answer that question before the FWDO, indicating that the decision-making process
195 starts with the MFDO.

*“My role’s an MFDO so generally if there’s a flood event coming I should know before the FWDO, in theory”
[MFDO2]*

4.1.1 The role of Monitoring and Forecasting Duty Officers

200 **“Ramping up to a flood event, the MFDO gathers that information, processes it and filters it, and
passes that along to the area staff [FWDO].” [MFDO2]**

What information do they use?



The MFDOs regularly receive FFC (Flood Forecasting Centre) national and county scale (i.e. area sub-divisions) flood risk forecasts and produce catchment/local scale flood forecasts, which they communicate with the FWDOs (see Fig. 3). The FFC generates three types of products:

- 205
- **Outlook products** – annual, seasonal and monthly assessments of flood risk;
 - **Flood Guidance Statement (FGS)*** – a five-day forecast of flood risk for all sources of flooding, for England and Wales, at a county scale (see Appendix B, Fig. (a) for an example);
 - **Hydro-Meteorological Services*** – detailed products communicating flood forecast data, currently comprising Hydro-Meteorological Guidance, Forecast Meteorological Data and Heavy Rainfall Alerts (see Appendix B, Fig. (b),

210 (c) and (d), respectively, for examples).

How do they use this information?

Based on this suite of information, the MFDOs decide whether they want to run the hydrological forecasting model, which sits in a separate system called the National Flood Forecasting System (NFFS; see Appendix B, Fig. (e) for an example). The decision can be triggered by the colours shown on the FGS (which communicates flood risk as a combination of likelihood and impact; e.g. high flood risk values on the FGS are more likely to lead to the MFDOs running the hydrological model). The

215 NFFS allows users to explore the observed data (i.e. river levels and rainfall) and run hydrological and hydraulic models*. These models, forced with the FFC's deterministic weather forecast, provide a single trace of past and future (i.e. for the next five days) river level for specific areas. This initial forecast scenario is usually referred to as the 'Best Estimate' scenario, showing what 'should' happen. What 'could' happen (i.e. the 'Reasonable Worst Case' scenario) may not always be run.

220 *"If there's uncertainty in the forecast like if there's showers [...] especially when they're thundery and they can give you really high totals in a very short space of time that's when you start to run 'What If' scenarios" [MFDO1]*

'What If' scenarios (i.e. 'Reasonable Worst Case' scenarios) are additional forecasts run by the MFDOs by manually modifying the FFC's deterministic weather forecast (usually through the use of predefined factors applied over an entire catchment; e.g. 200% of catchment rainfall totals in the next 6 hours). They then run this 'modified weather forecast' through

225 the hydrological/hydraulic models to obtain a new river level forecast scenario, often referred to as the 'Reasonable Worst Case' scenario. The MFDOs choose which What If scenario to run based on the FFC Hydrometeorological Guidance and their own expert knowledge, to estimate the likelihood of both scenarios (the 'Best Estimate' and the 'Reasonable Worst Case').

"[The FFC] might give us a number of different scenarios and we tend to pick the worst one and then see what that does" [MFDO1]

230 A critical part of the MFDOs' role is to interpret the different forecasting products, which might sometimes be inconsistent (e.g. differences between the national and local scale pictures). The MFDOs usually do this by applying expert judgement based on knowledge of model performance and catchment response* to make a coherent story and put the information into context for the FWDOs.

The MFDOs decide when to pass the information on to the FWDOs, generally waiting for the forecast to be confident* before

235 flagging a situation. The exact content of the communication depends on each pair, but usually contains information about the scale of the event and their confidence in the forecast.

"Which scenario is going through which threshold [and] how likely that is to happen" [MFDO1]. "Approximate [...] scale of the event [...] are we talking just a bit of water out of bank? Or is it Armageddon?" [MFDO2]

The conversation can sometimes be bilateral, and the MFDOs might ask questions to the FWDOs.



240 “Can they provide information [...] in terms of local sensitivity [...] and are works going on in that catchment? Is there
a gauge out of play?” [MFDO2]

4.1.2 The role of Flood Warning Duty Officers

“The role of the FWDO is to make sense of all that forecasting information and try and work out
potentially what the impacts could be of that on the ground and then make decisions as to whether
245 or not [they] issue flood alerts, flood warnings or severe flood warnings.” [FWDO1]

What information do they use?

The FWDOs’ role is to combine several different types of information to decide whether to issue a flood alert or warning (see
Fig. 3). The information available to them includes:

- **The processed hydro-meteorological forecast and interpretation from the MFDOs**
- 250 • **Factors within the catchment** that could influence river levels (e.g. blockage from a tree fallen down). This is ad-
hoc information and comes from a variety of sources, including: information gathered from community contacts
(flood wardens*, flood action groups*, etc.), from EA staff and duty officers, hydrometric data/CCTV images, details
of consented works (i.e. work going on in a channel);
- **The situation on nowcasting meteorological products** (e.g. rainfall radar);
- 255 • **Information about the communities that might be affected** (e.g. have they been affected by many floods in the
past);
- **Expert knowledge about catchment response.**

How do they use this information?

The FWDOs assess these various sources of information (e.g. in terms of their accuracy) to make a decision, knowing that
260 they do not necessarily have all the information to make a judgement call.

*“I look at the river level forecasts and then what I want to know from the MFDO is, does this account for the rain we’ve
had? So, do you think this is likely to change? Is the forecast I’m seeing on my screen a good river level forecast? Or
do we think it’s not picked something up properly?” [FWDO2]*

According to an internal document on using the two flood scenarios in practice, the *Best Estimate* should be used as a basis to
265 issue flood alerts or warnings. However, both scenarios are currently used for incident planning activities (e.g. resources
needed for response) and communication with responders and communities, while flood alerts and warnings are mostly issued
based on nowcasting products. This discrepancy could be due to the challenges associated with forecast accuracy* and lead
time, specifically for surface water flooding* and rapid-response catchments*. This document does however encourage the
use of the two scenarios for planning and flood warning activities whenever possible, in combination with expert judgement.

270 “The scenarios are planning scenarios and at some point [...] we move into operational now type forecasting. So
normally we’d issue a flood warning with anywhere between 30 minutes to [...] six hours lead time, whereas these
scenarios are generally two to five days ahead. So you wouldn’t normally [...] come up with a simple statement that
will issue flood warnings based on the best estimate [...] and at some point we transition into something that’s more
now that we use for operational decision making” [11]

275 4.1.3 Communication between MFDOs and FWDOs



280 **“The FWDO shouldn’t even really be thinking about anything until they’ve had a phone call from the MFDO [...]. Some FWDOs do go a bit more proactive than that, I think particularly the ones with the forecasting backgrounds almost can’t help themselves looking into it. And it depends on personality as well, some people hate the idea of being surprised by anything. But it does also depend on the MFDO.” [FWDO2]**

There is usually a constant exchange of information between MFDOs and FWDOs, even when no major event is on the horizon. However, more recently, the level of activity in preparation for a potential event has increased. Since 2007 (this corresponds vaguely with the summer 2007 floods), the lead time for which forecasts are shown and on which MFDOs and FWDOs can take action has increased from a few days to a few months ahead (based on the FFC’s outlook products mentioned in Sect. 4.1.1). This is consistent with findings from Neumann et al. (2018), who report that the EA currently uses long-range* (i.e. seasonal) hydrological forecasts mainly as supporting information, while relying on the shorter-range forecasts* for action.

285 *“So even from a month out now we’re starting to become aware of potential situations [...], but [...] because [...] most of our products [...] are [...] based on that five-day forecast [...] that’s when the activity really starts to build” [MFDO1]*

290 The communication between MFDOs and FWDOs varies across people and EA centres. Factors that might influence communication – in terms of its trigger, frequency and content – include the duty officers’ personality, day-to-day job and level of experience. Some FWDOs are more proactive than others in obtaining the information needed to make a decision; some might wait to be contacted by the MFDOs with a processed forecast, and others monitor the situation on a daily basis (see quote from FWDO2 above). In some cases, the FWDO might contact the MFDO first to get more details about an area of concern to them.

295 *“[...] and [...] then it’s [...] liaising with regional forecasting [the MFDOs] so they can give us any more detail or certainty or if we’re concerned about an area they can watch it a bit more for us [the FWDOs]” [FWDO3]*

Duty officers’ level of experience can also influence the content and interpretation of the conversation. Knowing each other helps interpret and gauge the confidence from each other’s language, which MFDO2 refers to as ‘nuanced communication’.

300 Working with new duty officers can lead to misinterpretations and you might have to justify your position further and prompt them to obtain the information you need.

305 *“I’ve known [FWDO1] for quite a while so when I’m on duty with [them] [...] I can sense [...] what sort of questions [they] want to ask, where [they’re] coming from. I think with less experienced duty officers it’s often more tricky to do that. So [...] the verbal communication that you go into with [FWDO1] for example might be a bit brief probably because I know that [they’ve] understood the message and interpreted the message well, whereas a new duty officer you might be spelling out [...] your position more, spending more time explaining why the uncertainty is such and how that may impact on the ground” [MFDO1]*

310 *“Knowing each other is really important because if I know it’s [MFDO2] on duty [they’ve] probably put that interpretation on already. If I get someone who’s reading off the screen, I put the interpretation on and if we misjudge that and we both put it on we could end up getting it too low” [FWDO2]*

Other factors that influence communication include the context of the event, duty officers’ geographical proximities and a centre’s practice. In some areas, the FWDOs will make the final call of warning the public or not, while in other areas, the MFDOs will tell the FWDOs when they need to issue a warning. In addition, MFDOs and FWDOs do not always sit in the



315 same building or town. MFDOs work from forecasting centres, while FWDOs are based in Area offices or Area incident rooms, which influences their (mode of) communication (in person vs via phone or emails).

“If these people [the FWDOs] were sitting geographically with these other people [the MFDOs], I think you’d get a better service” [11]

4.2 The forecast, a small cog in a much bigger wheel

320 **“Forecasting’s really important. It is, it should be really central to what we do [...] but actually it’s a small cog in the middle of a much bigger wheel.” [11]**

Forecasting supports incident response by providing a critical piece of information. However, duty officers have to consider a range of other sources of information and factors when making risk-based decisions.

“We always implore people to try and look at different sources of information” [12]

325 These additional sources of information include river level correlations*, model performance*, local knowledge (i.e. knowledge of how a certain catchment behaves), personal experience, and internal and external considerations (see Fig. 4). This section gives a more detailed overview of these factors and their relevance for decision-making.

4.2.1 River level correlations and model performance

330 **“The MFDO will be looking at how much rain is falling compared to what was forecast. You can check the river levels on the telemetry sites*, so you can see how fast they’re responding compared to the model and you can start to gauge how that catchment’s responding compared to what you thought it would do” [MFDO1]**

335 MFDOs might use several products to gain an understanding of model and forecast performance while the event unravels. More basic forecasting methods, like river level correlation tables, complement forecast information and aid the decision-making process. These correlations are based on a linear regression between peak levels upstream and downstream of a station. However, discrepancies between the forecasts and correlations can call into question the forecast accuracy.

“If the model says you’re going to get flooding, the correlation says we’re going to get flooding, we’ve had more rainfall than any previous event, you know that that decision’s [...] a clear one. If the model says flooding, the correlation says no you’re fine, and we’ve had somewhere in the middle in terms of rainfall, that’s when it gets difficult, because those borderline calls are really tricky to make” [12]

340 The MFDOs’ knowledge of the hydraulic/hydrological model performances, for certain types of events and catchments, is also key in interpreting the forecast. This is based on performance measures*, local feedback from real time river gauges*, experience and target lead times (i.e. the theoretical maximum lead time you have to send out a flood warning for a catchment before it floods, based on catchment size, gauge location and flood risk in that catchment). For certain types of events, such as convective rainfall events*, for which the duty officers know models are still limited, they might decide to issue a warning based on the ‘Reasonable Worst Case’, although it is *“technically against procedure” [MFDO2]*.

The FFC meteorological products also communicate some sort of confidence, which the MFDOs can use to complement the hydrological models’ performance information.

4.2.2 Local knowledge and personal experience



350 **“Whilst we are very data reliant on the information coming through, there’s also that experience
that you know that certain watercourses are very slow responding and [...] no matter how much
money we spend on your forecast, it’s always not very good, you always delay it by a day and drop
the peak by a bit. [...] Data is very important but that local experience is as important if not more so
in certain circumstances” [MFDO2]**

Local knowledge and personal experience are key ingredients for judgement, an important component of the decision-making
355 process. This means duty officers can react appropriately to an event and add confidence to the forecast. As MFDO2 put it,
“*experience is the unwritten part of the value that each role has*”.

Local knowledge is so important to decision-making that the interviewees believe it cannot be replaced by training, written
material or fully automated systems.

360 *“Some areas have very set triggers for a severe flood warning whereas other areas may just take it on a feel. [...] And
each area has done it for a good reason, it’s the local reasons for doing that but it isn’t nationally consistent” [MFDO2]*

*“We have in the past looked at automated warnings [...], we can’t automate them [...], there’s a lot of personal
interpretation and judgement [that] goes into it, and if a computer just hits a level and issues a warning, it’s going to
go wrong” [FWDO2]*

365 This also manifests itself in perceptions about how successfully duty officers can transfer to other centres or areas to help
during an important flood event.

370 *“One of the things we’re trying to do at the moment is to get mutual aid sorted out so that if a flood event happens in
[some of the Northern areas] and their MFDOs [...] or the FWDOs are very [...] stretched [...] we can go [...] there,
use their tools, their systems and do the same job. But whenever we’ve tried it the local knowledge is the key thing. Like
knowing that this river responds particularly quickly and that we need to deal with it first before we move on to other
ones that’s the sort of thing that even if you’re picking it up whilst you’re working in a different centre it’s affecting
your ability to deliver the role at the time” [MFDO1]*

Duty officers have access to tangible information about past flood events that can be useful for placing model information into
context. The ‘Flood Intelligence Files’ compile information (e.g. highest events on record, what rainfall led to them, what the
catchment state was at the time and any known impacts) for every gauge the EA is providing forecasts for.

375 How information is interpreted, risk appetite and past experience, can all affect decisions taken. There is the danger of
following instincts too much and becoming biased towards issuing too many (i.e. risk-averse) or not enough warnings (i.e.
risk-hungry), while in some cases decisions might never be forecast-led.

*“Since the Boxing Day floods I think the next level of flooding after that there was some discrepancies amongst the
area responses [...] they were a bit [...] jumpy [...] to not be caught out again which is understandable” [MFDO2]*

380 *“these kind of decisions about do we need to draw up a roster, do we need to be in the office overnight, a lot of that has
probably been done on gut feel, probably this FWDO being the advisor. [...] Do we need to do whatever based on
judgement, experience, feel for it. [...] I wouldn’t expect these people to actually be looking at any forecast and saying,
based on this I will do” [I1]*

4.2.3 Internal and external considerations



385 **“There are lots of external pressures as well, particularly as FWDO you can come under pressure
from all different types of sources to make decisions and perhaps not based on the evidence that
you’ve got for political reasons, [...] reputational reasons, organisation, in terms of being seen to be
active, seen to [...] act early” [FWDO1]**

Decisions are not only dictated by the science, local knowledge or personal experience and differences, but might have to
390 respond to internal and external considerations, especially during major events.

At an internal level, some areas and duty officers might be more forecast-led while others are more reliant on a nowcasting
type approach. Discrepancies amongst the area responses are partially due to historical differences across the different areas
and EA centres.

395 *“There are definite differences between areas and [...] between individual staff, so [town X] are far more likely to issue
flood alerts [...] purely on rainfall than [town Y] is, [town Y] will generally wait for a river level to rise and that
develops I suppose out of slight historical differences and personalities involved” [FWDO2]*

400 *“Some other areas will issue messages based on forecast whereas, we were always told to base it on what’s happening,
so we kind of wait to see if the rain comes in and then if anything happens issue. And we get marked on messages that
we send out, so one of the things is the timeliness and if you’ve issued one, did it actually flood afterwards? So if you’re
obviously issuing on a forecast, then you’re probably going to get scored low because it doesn’t always happen, so it’s
difficult” [FWDO3]*

There are exceptions to these procedures and FWDO3 mentions the possibility of issuing flood alerts based on the forecast
when the impact is expected to occur overnight or if the forecast displays “rarely high confidence” of rainfall and “if it’s a
more prolonged event” and “you know the catchment’s already wet”.

405 The EA’s principle, ‘think big, act early, be visible’, is an example of an internal consideration, which might influence the
duty officers’ decision-making (EA, 2018). In what ways does the EA’s statutory warning responsibilities and principle
influence decision-making? Does ‘act early’ put the forecast in first place while ‘think big’ and ‘be visible’ move it to a
secondary position?

410 *“Our mantra to incident response is think big, act early so sometimes [...] there is a danger that you’re over responding.
Somewhere you’re issuing alerts and warnings when actually the risk is low. So I think the role of the FWDO is to
assimilate all that information, forecasting information and using it to help inform the instant response but also manage
expectations” [FWDO1]*

There is usually a political element (external consideration) to the response immediately following a very major flood, as the
EA puts a greater focus on demonstrating to communities and the government that they are being proactive in warning,
415 informing, etc. There is also the need for the EA to align its message with actions of lead local flood authorities and responders
and to think about public response.

*“It’s managing expectations internally in terms of operational response and how this is going to potentially play out
which [...] can still be quite hard to do but it’s even harder to do it externally with [the] mood of the public or even
some of our professional partners, so local authorities are also obviously geared up to respond to flooding” [FWDO1]*

420 To conclude this section, it is evident that the duty officers have to take different sources of information, besides the forecast,
into consideration to make a decision. However, the forecast helps determine the timing of warning and response activities.
Because the forecast plays a seemingly small part in a much bigger system, could that mean that the transition to a different



type of forecast will have very minor impacts on the duty officers? Or on the contrary, could it unsettle this very complex machine?

425 4.3 What could a transition to probabilistic forecasting mean in practice?

4.3.1 Current practice: communicating confidence for decision-making at the EA

“Uncertainty is present in everything that we do and every bit of communication, [...] I don’t think I’ve ever been able to say something with 100% confidence, ever.” [MFDO2]

430 We have previously touched on the factors and uncertainties duty officers have to work with, including uncertainties in: the weather (and how it cascades down to hydrological response), model performance, the different spatial scales of response (local vs national), the situation on the ground (e.g. soil conditions prior to an event and river blockages), EA staff decisions and actions, and the public’s reaction to warnings.

435 Duty officers currently adapt the language they use to communicate these uncertainties internally and externally, based on their confidence level. According to internal EA guidelines, the language used should change according to the scenario used so that duty officers *“get used to the [...] way they’re working around scenarios and probabilistic forecasting”* [I1].

“If messages around a ‘Reasonable Worst Case’ use, could or [...] is possible; if it’s a ‘Best Estimate’ use, we expect, it’s probable” [I1]

Between the MFDOs and the FWDOs, confidence and uncertainty appears to always be (based on these interviews) communicated, usually using the two flood forecast scenarios.

440 *“I don’t think we can withhold uncertainty. One, the key role for MFDO is providing the forecast. So it’s getting the forecast as accurate as you can and then communicating it in the clearest way possible. So that’s often about interpreting the uncertainty and communicating it. So we often use the ‘Reasonable Worst Case’ and the ‘Best Estimate’ to do that”* [MFDO1]

445 Messages to the public are also worded with care to communicate the appropriate level of risk and prompt appropriate response and also contain some information about confidence and uncertainty. These messages are usually free-text messages and will therefore vary from across FWDOs.

“The message starts off with this flood warning has been issued for this place then it runs on after a while into detail which is where you can communicate those shades of grey” [FWDO2]

450 However, not all uncertainties are critical, and local knowledge and experience are key for the *“interpretation of the uncertainties”* [FWDO2] and their impact on the ground.

“Uncertainty from the forecasting point of view is always prevalent but understanding how it will impact the [...] area’s reaction is kind of the key thing” [MFDO2]

455 There is currently space for the communication of confidence at the EA and externally. This is a step towards probabilistic forecasting. But how big of a step is it? And how big of a step is still needed to reach that full transition to probabilistic flood forecasts?

4.3.2 The duty officers’ perceived opportunities and challenges

“Whether it creates as many problems as it solves, maybe” [I2]



The transition to probabilistic forecasts is a significant evolution, which generates mixed feelings amongst the duty officers. It is undeniable that this transition will bring changes at the EA; as *FWDO2* put it, “*probabilistic forecasting is kind of a fresh start for everyone*”. This section presents the interviewees’ perspectives on the changes that will ensue from this transition, in terms of perceived opportunities (left wordcloud on Fig. 5), challenges (right wordcloud on Fig. 5) and neutral changes. Table 1 outlines these perspectives, split into six main topics and supported by quotes reported in Appendix C. Some of quotes reported in Appendix C might sound very extreme, which could be partly due to the way the questions that prompted them were phrased. However, it could also reflect personal resistance and should be explored further.

465 **5 Discussion and recommendations**

465 **5.1 Considerations for a successful transition to probabilistic forecasts**

Probabilistic forecasts have a great potential to capture extreme events (Stephens and Cloke, 2014), and their benefits (compared to deterministic forecasts) for flood warning are evident (Verkade and Werner, 2011; Pappenberger et al., 2015). However, despite the increasing lead times at which we can confidently predict floods, the uncertainty inherent in the chaotic natural system being modelled grows with increasing lead times, posing new problems. As science and decision-making are both individually progressing, adapting to their respective internal and external changes, there still lacks an ideal framework for the incorporation of new and ‘uncertain’ science in decision-making practices, and, respectively, the uptake of decision-makers’ perspectives in the design of scientific practice. Here, results from this study and relevant literature are joined to put forward elements that should be considered for a successful transition to probabilistic forecasts for flood warning in England. From these interviews and previous EA studies, it is apparent that forecasts are one element in the complex decision-making landscape within which EA duty officers operate (Orr and Twigger-Ross, 2009; Dale et al., 2014). This landscape includes alternative ‘hard scientific facts’ (e.g. correlations, model performance and local knowledge to an extent), and ‘soft values’ (dependent on culture and context, personal experience and internal and external considerations) (Morss et al., 2005; Cloke et al., 2009; Arnal et al., 2016; Neumann et al., 2018). Morss et al. (2005) found that “although flood management practitioners might appreciate more certain hydro-meteorological information, scientific uncertainty is often swamped by other factors [e.g. community perception, time, money and resource constraints] and thus is not a high priority.” When uncertainties are evident and decision stakes are high, as is the case for the uncertainty communicated by probabilistic forecasts for flood incident management, traditional decision-making pathways could become ineffective and soft values might become more important than hard scientific facts (Funtowicz and Ravetz, 1993). In this specific study for example, an uncertain probabilistic forecast could lead to some duty officers reverting to the ‘Best Estimate’ and the river level correlations to make a decision, ignoring low probabilities of extreme events which could have ultimately led to an earlier flood warning.

Facing constantly evolving soft values, some decision-makers may find familiarity with the scientific methods they use reassuring, reducing their personal willingness to adopt new scientific methods (Morss et al., 2005; Ishikawa et al., 2011). This personal willingness was captured in the range of responses (perceived challenges and opportunities) obtained during the interviews. An institute’s operating practice should reflect the complex landscape in which decision-makers operate, where the forecast plays an integral role in decision-making. To this end, the co-design of forecasting systems by both forecasters and users is necessary.

To do that, clear communication between forecasters and users is needed. However, language is perhaps one of the biggest barriers between scientists and decision-makers. It has been observed that “the way scientists referred to and discussed uncertainty sometimes confused practitioners” (Morss et al., 2005). Similarly, there is a lot of research done on the impacts of graphical representation of uncertainty in hazard forecasts on decision-making. These have shown that great care has to be taken when designing and communicating uncertain information, as it can impact the nature of the actions taken (Bruen et al., 2010; Joslyn and Savelli, 2010; Stephens et al., 2012; Pappenberger et al., 2013; Sivle et al., 2014).



500 There is the common misconception amongst the scientific community that decision-makers want 100% certain information (Demeritt et al., 2013; Michaels, 2014). In reality, as shown in this paper, decision-makers appreciate that scientific information is uncertain, not unlike other types of information they use. Decision-makers want to see that uncertainty, which they do not necessarily perceive as a barrier to use (Morss et al., 2005; Bruen et al., 2010; Neumann et al., 2018). One reason for this misconception might be the different ways scientists and decision-makers approach forecast uncertainty. Scientists see (the reduction of) forecast uncertainty as an end goal and “often deal with uncertainty by attempting to reduce, quantify, analyze, and/or assess it”. Decision-makers “view uncertainty as an unavoidable factor [...] all information about the future is uncertain [and] they must make decisions under uncertainty every day, in a complex, evolving social, institutional, and political environment” (Morss et al. 2005).

In this complex evolving landscape, decision-makers deal with forecast uncertainty similarly to other uncertainties they might face: under time and resources constraints. They assess the total uncertainty there is (the forecast uncertainty might sometimes be negligible compared to all the other factors at stake) in terms of its potential effect on the decision-making process and outcome (Morss et al., 2005). As mentioned by a few EA duty officers, uncertainty is prevalent in everything that they do, and the key is understanding what the impact of these uncertainties will be on the ground. It is crucial to develop a methodology for decision-makers to be able to use (forecast) uncertainty information optimally. A solution that does not require any additional time- and resource-consuming complex analyses, given the high stakes and strict deadlines decision-makers have to work with. Smith et al. (2018) argue that if there was a “greater involvement of decision-makers in the design and execution of uncertainty analyses”, “more purposeful evaluation and communication of uncertainty would certainly result”. This remains an open challenge to be tackled.

By design, probabilistic forecasts might contain some realisations that capture scenarios which do not always realise. This may lead to false alarms. Institutions can have specific risk perceptions and flood management priorities: seeking to avoid false alarms, or on the contrary, seeking to avoid missed flood events*, and the minimum/maximum lead time at which they (have to) issue flood warnings. This cultural landscape within which decision-makers operate may have an impact on the decision-making outcome (as discussed in Sect. 4.2.3) and an institution’s uptake of probabilistic flood forecasts in practice (Nobert et al., 2010; Ishikawa et al., 2011; McEwen et al., 2012; Demeritt et al., 2013; Michaels, 2014). A transition to probabilistic flood forecasts should be reflected in an institution’s wider flood management priorities. This could be done, for example, by changing their internal communication pathways or their warning procedures (e.g. lead times at which they operate).

Very often however, the ability of an institution to pick up new information and methods is not only down to them, but could be influenced by the wider socio-political context and other key actors in the decision-making web (e.g. the government, local authorities, regulations and guidelines), additionally to the populations at risk and the way they respond to flood warnings (Dessai and Hulme, 2004; Morss et al., 2005; Parker et al., 2009). This is reflected in the interviewed EA duty officers’ perceived challenges regarding ‘Language & communication’ and ‘Binary decision-making’ (Sect. 4.3.2). In the face of a socio-political context that is demanding ever more precise information and with the rise of a post-factual society, the general trust in science might be a limiting factor to the uptake of new science and institutions should trust their capacity to use uncertain probabilistic information (Soares and Dessai, 2015; Golding et al., 2017; Knudsen and de Bolsée, 2019).

It is also important to note that “moving to probabilistic forecasting from deterministic forecasting may trigger an institutional shift in who is responsible for decision making under uncertainty” (Michaels, 2014). Because making a decision based on probabilistic information is more nuanced than using deterministic information, the outcome will determine who will be ‘blamed’ and this ownership of the uncertainty judgment might have implications on the forecasters-users relationships (Michaels, 2014). This relates to some of the interviewed duty officers’ fears of a transition to probabilistic forecasts at the EA, as it might move “the burden of making a decision further down the tree” (Sect. 4.3.2). In this context, a framework to engage with all key actors of the decision-making web ahead of and during a transition to probabilistic forecasts appears crucial. Ramos et al. (2010) advocated the use of integrated platforms to allow a continuous exchange between scientists and



545 decision-makers in real-time. Similar studies on the provision of climate services have identified the lack of user engagement as a great limiting factor of the uptake of climate information in practice (Golding et al. 2017). It is evident that a transition to probabilistic forecasts is not only a scientific endeavour and feasibility studies should include other disciplines, such as social-science.

5.2 Recommendations to the EA

550 In light of the findings of this study, and other relevant studies, we make a list of recommendations to support the uptake of probabilistic forecasts at the EA. These ten recommendations are high priority actions for the EA as an institution. The service, role owners and those responsible for ensuring a quality service delivery should ensure that these recommendations are pursued, alongside technical work around the transition. Please note that these recommendations are not ranked in priority order for the EA, as some of these will be quicker and easier to implement and to demonstrate progress on.

- 555 1) Communicate (via engagement campaigns, videos, email newsletters, social media updates and webinars, etc.) with all key players in the decision-making chain (as well as external players such as the emergency responders and the public) to ensure that they are all aware that the transition to probabilistic forecasts will become operational practice.
- 560 2) Give appropriate and custom designed internal training to all key players (Nobert et al., 2010). Duty officers must receive training on how to make decisions based on probabilistic forecasts (for example in the form of decision-making activities and serious games - see the HEPEx¹ and the Red Cross Climate Centre² resources for inspiration).
- 565 3) Expand existing EA communication structures to allow the co-design of the new products between forecast producers and users (Morss et al., 2005; Smith et al., 2018). Everyone using the forecasting products and systems at the EA should have the chance to have a say in how the system will look and function through a mutual design strategy. If the new system does not reflect the complex landscape in which duty officers operate (a mix of 'hard scientific facts' and 'soft values'), probabilistic forecasts might end up being under- or misused.
- 570 4) Reach out to the community of practice in hydrological probabilistic forecasting, such as HEPEx³ (community of international experts in the field of probabilistic hydrological forecasting and decision-making) and connect with institutes which have already gone through such a transition to gain insights and share best practice, as some elements might be transferrable (Nobert et al., 2010; Dale et al., 2014). This could be done through organised workshops, webinars and the establishment of an advisory group.
- 575 5) The way probabilistic information will be translated into meaningful content and communicated to the emergency responders and the public requires careful thought and design (Bruen et al., 2010; Joslyn and Savelli, 2010; Stephens et al., 2012; Pappenberger et al., 2013; Sivle et al., 2014). To this end, an interdisciplinary approach between forecasters and social-scientists would be greatly valuable as social-science can offer insights into the human response to warning messages. A tailored and inter-disciplinary study of the forecasting products using probabilistic information and used in the decision-making process is urgently required.

¹ hepex.irstea.fr/resources/hepex-games

² www.climatecentre.org/resources-games/games

³ hepex.irstea.fr



- 580 6) a) The EA's heterogeneity at the national level should be accounted for and addressed. Given the heterogeneity of
the EA at a national level and the areas' diversity in terms of history and catchment response, we do not expect
probabilistic forecasts to be welcomed similarly in all the EA centres. Efforts will therefore have to be made by the
EA to achieve a simultaneous and homogeneous transition in all its centres.
- 585 b) Furthermore, the design of the new forecasting system should be homogenised at the national level (to allow for
staff movement during major flood events), while accounting for the heterogeneity of local conditions, existing
dynamics and institutional practices. This could be achieved through the co-design of the forecasting system with
local duty officers (see recommendation 3).
- 590 7) Be prepared to move towards lead times that reflect the probabilistic forecast predictability. The optimal lead time to
trigger action depends on both the probabilistic flood forecast quality and the actions' operational implementation
time (Bischiniotis et al., 2019). While the EA operates with pre-defined lead times for each specific activity (e.g. it
takes x hours/days to move equipment from A to B, or to deploy temporary defences), probabilistic forecasts could
in theory provide earlier indications of potential future floods, giving the EA more time to prepare ahead of a flood
595 event. To utilise probabilistic forecasts to their full potential, tailored studies should be performed during the EA
system's co-design to adjust lead times (for planning and warning) on the probabilistic products and event types, with
ample time for testing by the EA duty officers.
- 600 8) Under no circumstances should the old system be switched off as soon as the probabilistic system is operational.
There should be a reasonable period of overlap between the two systems in order to give everyone some time to
gradually adapt (Funtowicz and Ravetz, 1993). During that time of overlap, end-user feedback should be collected
(Thielen et al., 2006). To avoid situations where the probabilistic forecast and the two scenarios show contrasting
results, the new operating procedures need to specify that the probabilistic forecasts should be looked at first.
- 605 9) Update the duty officers' operating procedures. Clear guidelines should be provided to the duty officers on how to
make a decision based on the new probabilistic products. These guidelines should include information such as: the
various sources of information available to them for making a decision, how to interpret a probabilistic forecast, the
forecast confidence at which certain decisions and actions should be made and the language that should be used.
- 610 10) Document this transition (in writing or through documentary-style interviews, etc.) to help other institutes and future
transitions at the EA (Pielke, 1997). While this paper investigates how things might change, post-transition evaluation
should seek to answer the question: "How did we do?"

Many of these recommendations are however general and could be applicable to other institutes and types of information.

615 **6 Conclusions**

The Environment Agency (EA) is in the process of a transition to probabilistic fluvial flood forecasts, from the two flood
scenarios they currently use operationally for flood warning and incident management activities in England. State-of-the-art
probabilistic forecasts can give an earlier indication of potential future extreme events, such as floods, increasing the amount
of time decision-makers have to prepare. A series of interviews were carried out with EA 'Monitoring and Forecasting Duty
620 Officers' (MFDOs) and 'Flood Warning Duty Officers' (FWDOs), two roles at the heart of the EA's flood risk management



625 decision-making chain. The aim was to understand how an operational transition to probabilistic flood forecasts might affect
their decision-making activities. Overall, none of the interviewed duty officers mentioned concerns about impacts of this
transition on their two roles' interaction. Perceived challenges lie mostly outside of their roles and relate to: communication
with emergency responders and the public, translating uncertain information into a binary decision and the speed of the
transition. Ten high priority recommendations were made to the EA to ensure a successful transition. They include: i)
communicating with all key players in the decision-making chain (as well as emergency responders and the public) to ensure
that they are all aware that this transition will become operational practice, ii) facilitating the co-design of the new products
by forecasters and users and collecting end-user feedback during a reasonable period of overlap between the two systems, iii)
630 employing an inter-disciplinary approach to translate probabilistic information into meaningful content for communication
with emergency responders and the public, and iv) being prepared to adapt the EA's overarching warning and incident planning
strategy to reflect this transition. It is vital for these recommendations to be followed to ensure that state-of-the-art science is
used to its fullest potential for risk management practice and is not being under- or misused.

635 **Author contributions.** H.L.C., L.An. and S.M. posed the original question. L.An., S.M., T.N. and L.W. brought L.Ar. up to
speed about the EA and their decision-making practices. T.N. identified the interviewees. L.Ar., H.L.C., T.N. and E.S. designed
the interviews. L.Ar. carried out the interviews and analysed the interview transcripts. L.Ar., J.N. and H.L.C. wrote the paper.
H.L.C., S.M., J.N. and T.N. commented on the manuscript.

640 **Competing interests.** The authors declare that they have no conflict of interest.

Disclaimer. The information and findings in this paper are based on interviewees with six EA duty officers. They should not
be taken as representing the views or practice of the EA as a whole.

645 **Acknowledgements.** This work was funded by the EU Horizon 2020 IMPREX project (www.imprex.eu) (641811) and the
joint Flood and Coastal Erosion Risk Management Research and Development Programme. We thank all the interviewees who
dedicated some time to this work. We would also like to thank Stuart Hyslop at the EA for the background he provided in
preparation for the interviews, and for his support in the organisation of the interviews.

References

- 650 Arnal, L., Ramos, M.-H., Coughlan de Perez, E., Cloke, H. L., Stephens, E., Wetterhall, F., van Andel, S. J., and
Pappenberger, F.: Willingness-to-Pay for a Probabilistic Flood Forecast: A Risk-Based Decision-Making Game,
Hydrol. Earth Syst. Sci., 20, 3109–28, <https://doi.org/10.5194/hess-20-3109-2016>, 2016.
- Bischiñotis, K., van den Hurk, B., Coughlan de Perez, E., Veldkamp, T., Guimarães Nobre, G., and Aerts, J.: Assessing
Time, Cost and Quality Trade-Offs in Forecast-Based Action for Floods, Int. J. Disast. Risk Re., 40, 101252,
<https://doi.org/10.1016/j.ijdr.2019.101252>, 2019.
- 655 Bruen, M., Krahe, P., Zappa, M., Olsson, J., Vehvilainen, B., Kok, K., and Daamen, K.: Visualizing Flood Forecasting
Uncertainty: Some Current European EPS Platforms-COST731 Working Group 3, Atmos. Sci. Lett., 11, 92–99,
<https://doi.org/10.1002/asl.258>, 2010.
- Buizza, R.: The Value of Probabilistic Prediction, Atmos. Sci. Lett., 9, 36–42, <https://doi.org/10.1002/asl.170>, 2008.
- Cloke, H. L., and Pappenberger, F.: Ensemble Flood Forecasting: A Review, J. Hydrol., 375, 613–26,
660 <https://doi.org/10.1016/j.jhydrol.2009.06.005>, 2009.
- Cloke, H. L., Thielen, J., Pappenberger, F., Nobert, S., Bálint, G., Edlund, C., Koistinen, A., de Saint-Aubin, C.,



- Sprokkereef, E., Viel, C., Salamon, P., and Buizza, R.: Progress in the Implementation of Hydrological Ensemble Prediction Systems (HEPS) in Europe for Operational Flood Forecasting, ECMWF Newsletter No. 121, Autumn, Reading, UK, 20-24, 10.21957/bn6mx5nxfq, 2009.
- 665 Dale, M., Ji, Y., Wicks, J., Mylne, K., Pappenberger, F., and Cloke, H. L.: Applying Probabilistic Flood Forecasting in Flood Incident Management, Environment Agency Technical Report, Project No. SC090032, Bristol, UK, 97 pp., 2013.
- Dale, M., Wicks, J., Mylne, K., Pappenberger, F., Laeger, S., and Taylor, S.: Probabilistic Flood Forecasting and Decision-Making: An Innovative Risk-Based Approach, *Nat. Hazards*, 70, 159–72, <https://doi.org/10.1007/s11069-012-0483-z>, 2014.
- 670 Davies, A., Hoggart, K., and Lees, L.: *Researching Human Geography*, 1st edition, Routledge, London, 384 pp., 2014.
- Demeritt, D., Nobert, S., Cloke, H. L., and Pappenberger, F.: The European Flood Alert System and the Communication, Perception, and Use of Ensemble Predictions for Operational Flood Risk Management, *Hydrol. Process.*, 27, 147–57, <https://doi.org/10.1002/hyp.9419>, 2013.
- Demeritt, D., Nobert, S., Cloke, H. L., and Pappenberger, F.: Challenges in Communicating and Using Ensembles in Operational Flood Forecasting, *Meteorol. Appl.*, 17, 209–22, <https://doi.org/10.1002/met.194>, 2010.
- 675 Department for Environment Food and Rural Affairs: Flood and Water Management Act 2010, UK Public General Acts, 1–84, 2010.
- Dessai, S., and Hulme, M.: Climate Policy Does Climate Adaptation Policy Need Probabilities? Does Climate Adaptation Policy Need Probabilities?, *Clim. Policy*, 4, 107–28, <https://doi.org/10.1080/14693062.2004.9685515>, 2004.
- 680 Environment Agency: *Creating a Better Place - Our Ambition to 2020*, 2018.
- Flood Forecasting Centre: *Flood Guidance Statement User Guide*, 2017.
- Flowerdew, J., Horsburgh, K., and Mylne, K.: Ensemble Forecasting of Storm Surges, *Mar. Geod.*, 32, 91–99, <https://doi.org/10.1080/01490410902869151>, 2009.
- Funtowicz, S. O., and Ravetz, J. R.: Science for the Post-Normal Age, *Futures*, 25, 739–55, [https://doi.org/10.1016/0016-3287\(93\)90022-L](https://doi.org/10.1016/0016-3287(93)90022-L), 1993.
- 685 Golding, N., Hewitt, C., Zhang, P., Bett, P., Fang, X., Hu, H., and Nobert, S.: Improving User Engagement and Uptake of Climate Services in China, *Climate Services*, 5, 39–45, <https://doi.org/10.1016/j.cliser.2017.03.004>, 2017.
- HM Government: *National Flood Resilience Review*, 2016.
- House of Common - Environment Food and Rural Affairs Committee: *Future Flood Prevention - Second Report of Session 2016-17*, 2016.
- 690 Joslyn, S., and Savelli, S.: Communicating Forecast Uncertainty: Public Perception of Weather Forecast Uncertainty, *Meteorol. Appl.*, 17, 180–95, <https://doi.org/10.1002/met.190>, 2010.
- Lorenz, E. N.: The Predictability of a Flow Which Possesses Many Scales of Motion, *Tellus*, 21, 289–307, <https://doi.org/10.3402/tellusa.v21i3.10086>, 1969.
- 695 McEwen, L. J., Krause, F., Jones, O., and Garde Hansen, J.: Sustainable Flood Memories, Informal Knowledge and the Development of Community Resilience to Future Flood Risk, *WIT Trans. Ecol. Envir.*, 159, 253–64, <https://doi.org/10.2495/FRIAR120211>, 2012.
- Michaels, S.: Probabilistic Forecasting and the Reshaping of Flood Risk Management, *Journal of Natural Resources Policy Research*, 7, 41–51, <https://doi.org/10.1080/19390459.2014.970800>, 2014.
- 700 Morss, R. E., Wilhelmi, O. V., Downton, M. W., and Grunfest, E.: Flood Risk, Uncertainty, and Scientific Information for Decision Making: Lessons from an Interdisciplinary Project, *B. Am. Meteorol. Soc.*, 86, 1593–1602, <https://doi.org/10.1175/BAMS-86-11-1593>, 2005.
- Mulder, K. J., Lickiss, M., Harvey, N., Black, A., Charlton-Perez, A., Dacre, H., and McCloy, R.: Visualizing Volcanic Ash Forecasts: Scientist and Stakeholder Decisions Using Different Graphical Representations and Conflicting Forecasts,



- 705 Weather Clim. Soc., 9, 333–48, <https://doi.org/10.1175/WCAS-D-16-0062.1>, 2017.
- Neumann, J. L., Arnal, L., Emerton, R. E., Griffith, H., Hyslop, S., Theofanidi, S., and Cloke, H. L.: Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity, *Geosci. Commun.*, 1, 35–57, <https://doi.org/10.5194/gc-1-35-2018>, 2018.
- New, M., Lopez, A., Dessai, S., and Wilby, R.: Challenges in Using Probabilistic Climate Change Information for Impact Assessments: An Example from the Water Sector, *Philos. T. Roy. Soc. A*, 365, 2117–31, <https://doi.org/10.1098/rsta.2007.2080>, 2007.
- 710 Neville, N.: Cognitive Illusions, Heuristics, and Climate Prediction, *B. Am. Meteorol. Soc.*, 80, 1385–97, [https://doi.org/10.1175/1520-0477\(1999\)080<1385:CIHACP>2.0.CO;2](https://doi.org/10.1175/1520-0477(1999)080<1385:CIHACP>2.0.CO;2), 1999.
- Nobert, S., Demeritt, D., and Cloke, H. L.: Informing Operational Flood Management with Ensemble Predictions: Lessons from Sweden, *J. Flood Risk Manag.*, 3, 72–79, <https://doi.org/10.1111/j.1753-318X.2009.01056.x>, 2010.
- Orr, P., and Twigger-Ross, C.: Communicating Risk and Uncertainty in Flood Warnings: A Review of Defra/Environment Agency FCERM Literature, Environment Agency Science Report, Project No. SC070060/SR2, Bristol, UK, 60 pp., 2009.
- Pagano, T. C., Hartmann, H. C., and Sorooshian, S.: Seasonal Forecasts and Water Management in Arizona: A Case Study of the 1997–98 El Niño Event, 29th Annual Water Resources Planning and Management Conference, 21, 1–11, [https://doi.org/10.1061/40430\(1999\)227](https://doi.org/10.1061/40430(1999)227), 2004.
- 720 Pappenberger, F., Stephens, E., Thielen, J., Salamon, P., Demeritt, D., van Andel, S. J., Wetterhall, F., and Alfieri, L.: Visualizing Probabilistic Flood Forecast Information: Expert Preferences and Perceptions of Best Practice in Uncertainty Communication, *Hydrol. Process.*, 27, 132–46, <https://doi.org/10.1002/hyp.9253>, 2013.
- 725 Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S., and Thielen, J.: The Monetary Benefit of Early Flood Warnings in Europe, *Environmental Science & Policy*, 51, 278–91, <https://doi.org/10.1016/j.envsci.2015.04.016>, 2015.
- Parker, D. J., Priest, S. J., and Tapsell, S. M.: Understanding and Enhancing the Public's Behavioural Response to Flood Warning Information, *Meteorol. Appl.*, 114, 103–14, <https://doi.org/10.1002/met.119>, 2009.
- 730 Pielke, R. A. Jr.: Asking the Right Questions: Atmospheric Sciences Research and Societal Needs, *B. Am. Meteorol. Soc.*, 78, 255–255, [https://doi.org/10.1175/1520-0477\(1997\)078<0255:ATRQAS>2.0.CO;2](https://doi.org/10.1175/1520-0477(1997)078<0255:ATRQAS>2.0.CO;2), 1997.
- Pilling, C., Dodds, V., Cranston, M., Price, D., Harrison, T., and How, A.: Chapter 9 - Flood Forecasting — A National Overview for Great Britain, *Flood Forecasting - A Global Perspective*, edited by: Adams, T. E., III, and Pagano, T. C., Academic Press, 201–47, <https://doi.org/10.1016/B978-0-12-801884-2.00009-8>, 2016.
- 735 Pitt, M.: Learning Lessons from the 2007 Floods, *The Pitt Review*, London: Cabinet Office, 1–205, 2008.
- Ramos, M.-H., Mathevet, T., Thielen, J., and Pappenberger, F.: Communicating Uncertainty in Hydro-Meteorological Forecasts: Mission Impossible?, *Meteorol. Appl.*, 17, 223–35, <https://doi.org/10.1002/met.202>, 2010.
- Schoenberger, E.: The Corporate Interview as a Research Method in Economic Geography, *The Professional Geographer*, 43, 180–89, <https://doi.org/10.1111/j.0033-0124.1991.00180.x>, 1991.
- 740 Sene, K., Weerts, A., Beven, K., Moore, R. J., Whitlow, C., and Young P.: Risk-Based Probabilistic Fluvial Flood Forecasting for Integrated Catchment Models - Phase 1 Report, Environment Agency Science Report, Project No. SC080030/SR1, Bristol, UK, 179 pp., 2009.
- Sivle, A. D., Kolstø, S. D., Hansen, P. J. K., and Kristiansen, J.: How Do Laypeople Evaluate the Degree of Certainty in a Weather Report? A Case Study of the Use of the Web Service Yr.No., *Weather Clim. Soc.*, 6, 399–412, <https://doi.org/10.1175/WCAS-D-12-00054.1>, 2014.
- 745 Smith, K. A., Wilby, R. L., Broderick, C., Prudhomme, C., Matthews, T., Harrigan, S., and Murphy, C.: Navigating Cascades of Uncertainty — As Easy as ABC? Not Quite...., *Journal of Extreme Events*, 5, 1850007,



<https://doi.org/10.1142/S2345737618500070>, 2018.

750 Stephens, E., and Cloke, H. L.: Improving Flood Forecasts for Better Flood Preparedness in the UK (and Beyond), *Geogr. J.*,
180, 310–16, <https://doi.org/10.1111/geoj.12103>, 2014.

Stephens, E. M., Edwards, T. L., and Demeritt, D.: Communicating Probabilistic Information from Climate Model
Ensembles-Lessons from Numerical Weather Prediction, *WIREs Clim. Change*, 3, 409–26,
<https://doi.org/10.1002/wcc.187>, 2012.

755 Thielen, J., Bartholmes, J., Ramos, M.-H., and de Roo, A.: The European Flood Alert System – Part 1: Concept and
development, *Hydrol. Earth Syst. Sci.*, 13, 125-140, <https://doi.org/10.5194/hess-13-125-2009>, 2009.

Thielen, J., Bartholmes, J., and Ramos, M.-H.: The Benefit of Probabilistic Flood Forecasting on European Scale – Results
of the European Flood Alert System for 2005/2006, European Commission, Ispra, Italy, 99 pp., 2006.

Verkade, J. S. and Werner, M. G. F.: Estimating the benefits of single value and probability forecasting for flood warning,
Hydrol. Earth Syst. Sci., 15, 3751-3765, <https://doi.org/10.5194/hess-15-3751-2011>, 2011.

760 Werner, M., Cranston, M., Harrison, T., Whitfield, D., and Schellekens, J.: Recent Developments in Operational Flood
Forecasting in England, Wales and Scotland, *Meteorol. Appl.*, 16, 13–22, <https://doi.org/10.1002/met.124>, 2009.

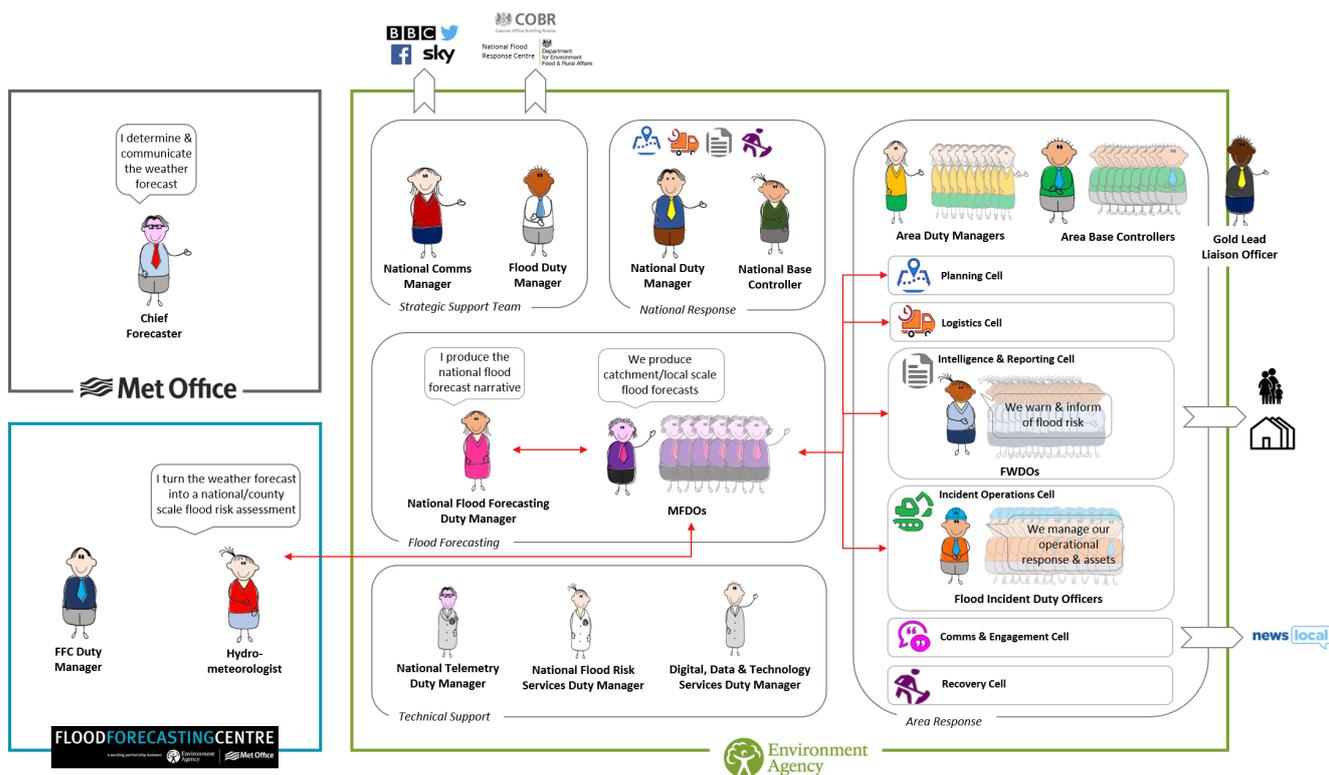


Figure 1: Schematic of the EA's institutional landscape and the FIM information flow between MFDOs, FWDOs and first-degree contact points (red arrows) (source: EA).



- North**
- 1 North East (NEA)
- 2 Cumbria and Lancashire (CLA)
- 3 Yorkshire (YOR)
- 4 Greater Manchester, Merseyside and Cheshire (GMC)
- West and Central**
- 5 Lincolnshire and Northamptonshire (LNA)
- 6 East Midlands (EMD)
- 7 West Midlands (WMD)
- 8 Wessex (WSX)
- 9 Devon, Cornwall and the Isles of Scilly (DCS)
- South East**
- 10 East Anglia (EAN)
- 11 Thames (THM)
- 12 Hertfordshire and North London (HNL)
- 13 Kent, South London and East Sussex (KSL)
- 14 Solent and South Downs (SSD)

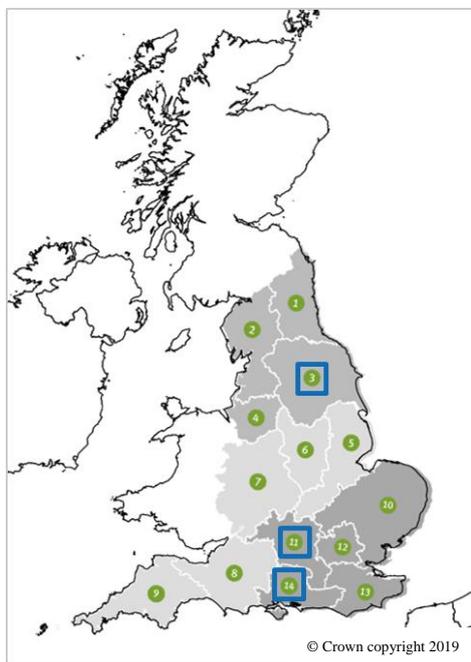


Figure 2: Map showing the geographical areas of the EA's operations (green numbered areas), highlighting the three areas which the centres where interviews were carried out are responsible for (blue boxes) (source: EA). The works published in this journal are distributed under the Creative Commons Attribution 4.0 License. This licence does not affect the Crown copyright work, which is re-usable under the Open Government Licence (OGL). The Creative Commons Attribution 4.0 License and the OGL are interoperable and do not conflict with, reduce or limit each other.

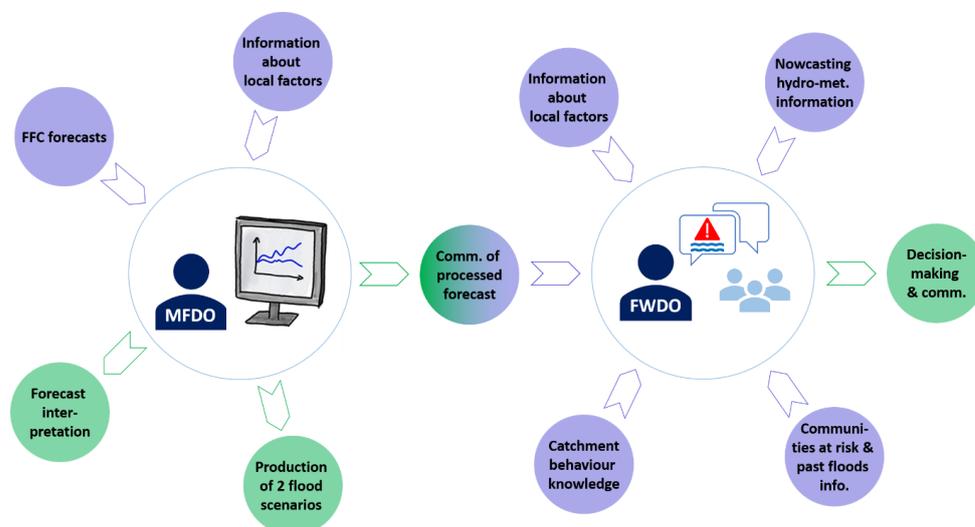


Figure 3: Roles and interactions between EA duty officers. Blue arrows and circles are for incoming information and green arrows and circles relate to outputs from either of the duty officers.

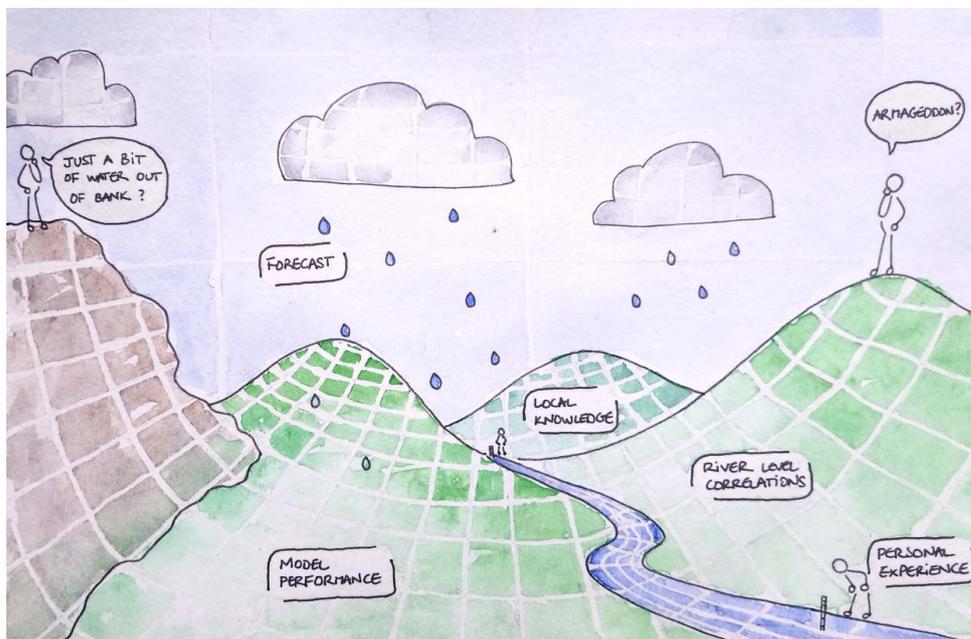


Figure 4: Complex decision-making landscape in which EA duty officers operate.

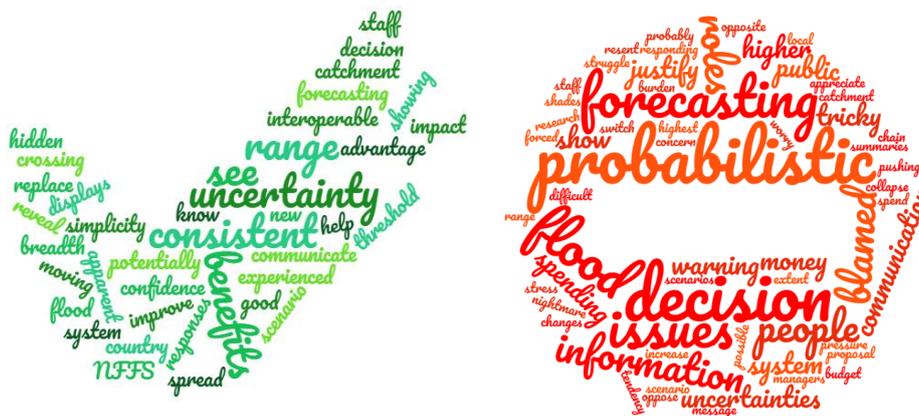


Figure 5: Wordclouds of perceived opportunities (left) and challenges (right), based on the interview transcripts.



Table 1: Interviewees’ perceived opportunities, challenges and neutral changes associated with a transition to probabilistic forecasts. Perspectives are split into six main topics (rows). Supporting quotes can be found in Appendix C.

Language and communication	<p>Most interviewees agreed that this will probably be the biggest change. Some said they thought it might improve long-term communication and increase the MFDOs’ credibility and confidence (quote O1). This was also found by Thielen et al. (2006). Others believe that there is a potential for misunderstanding and that a lot more work is still needed on this topic (quotes C1 and C2).</p>
Uncertainty	<p>Probabilistic forecasts contain uncertainty which they openly display. Some interviewees thought that this would materialise the forecast uncertainty, otherwise sometimes hidden with the two scenarios (quote O2). This is in line with the EA’s 2009 science report (Sene et al., 2009). Many interviewees however questioned whether probabilistic forecasts would really help tackle the uncertainty they deal with while on duty (quotes N1 and C3).</p>
The forecasting system	<p>Some interviewees mentioned that the two scenarios, and the What If scenarios used to produce them, were sometimes challenging to play with and required a lot of expert judgment, thus making them inconsistent nation-wide. There were hints that a few MFDOs thought probabilistic forecasts might lead to more consistency across the EA centres (quote O3). It was however clear from the interviews that things will need to change slowly to give duty officers time to build confidence in the new system (quote C4).</p>
Decision-making	<p>A few interviewees mentioned the fact that probabilistic forecasts will not solve the fundamental need of decision-making to be binary and saw this as a challenge (quotes C5 and C6). Others saw this as an opportunity for early warning and long-term planning (quotes O4 and O5).</p>
Duty officers’ roles	<p>This transition was seen neither as an opportunity nor as a challenge by and for the MFDOs. They simply stated how things might change for them (quotes N2 and N3). A few of the FWDOs however thought that this might push more of the interpretation on to them (quotes C7 and C8). It is worth noting is that none of the interviewees mentioned worries concerning potential impacts of this future transition on the communication and interaction between duty officers. The worries seem to mostly lie outside of their interaction (quote O6).</p>
New staff training	<p>An interviewee mentioned that probabilistic forecasts could help with new staff training, by increasing their understanding of catchment response (quote O7).</p>



Appendix A. Glossary of terms.

Best Estimate	A forecaster’s assessment of the most likely rainfall, river and groundwater levels, and coastal conditions, and their impacts.
Catchment characteristics and response	Catchment characteristics are the features that describe a river basin (i.e. the area of land drained by a river), such as its location, size, vegetation cover, soil type and topography. They partially define the catchment response, the catchment’s reaction when subjected to a rainfall event (e.g. how fast the water level increases after a rainfall event).
Chaos	The property of a complex system, like the weather, whose behaviour is so unpredictable that it appears random. This is due to the system’s sensitivity to small changes in conditions.
Confident	A forecaster’s expert judgement of how certain they are that the forecast is right.
Convective rainfall events	The sun heats the ground, warming up the air above it. This causes the air to rise. As the air rises it cools and condenses, forming water droplets that organise into clouds and lead to rainfall. Convective rainfall events can lead to thunderstorms.
Department for Environment, Food and Rural Affairs (Defra)	UK government department responsible for safeguarding the UK’s natural environment and supported by 33 agencies and public bodies, including the Environment Agency (EA). www.gov.uk/government/organisations/department-for-environment-food-rural-affairs
Deterministic forecasts	Refers to a forecast which gives a single possible outcome of the future rainfall, river and groundwater levels and coastal conditions.
Ensemble	Instead of running a single deterministic forecast, computer models can run a forecast several times, using slightly different inputs to account for uncertainties in the forecasting process. The complete set of forecasts is called an ‘ensemble’, and each individual forecast within it are ‘ensemble members’. Each ensemble member represents a different possible scenario of future rainfall, river and groundwater levels and coastal conditions. Each scenario is equally likely to occur.
Environment Agency (EA)	An executive non-departmental public body sponsored by Defra. The EA has an operational responsibility to manage risks of flooding from rivers and the sea in England, by warning and informing the public and businesses about impending floods. www.gov.uk/government/organisations/environment-agency
False alarms	A warning given ahead of an event (e.g. flood) that does not ultimately occur.
Flood action groups	Cores of local people who act as representative voices for their wider community. They work alongside agencies and authorities and meet on a regular basis with the aim of reducing their community’s flood risk and improving its resilience to flooding.
Flood Forecasting Centre (FFC)	A partnership between the Environment Agency and the UK Met Office. It provides a UK-wide 24/7 hydro-meteorological service to emergency responders to better prepare for flooding (river, surface water, tidal/coastal and groundwater). www.ffc-environment-agency.metoffice.gov.uk
Flood Guidance Statement (FGS)	A daily flood risk forecast for the UK, produced by the FFC (in collaboration with the EA and Natural Resources Wales) to assist with strategic, tactical and operational planning decisions. It gives a flood risk assessment shown by county and unitary authority across England and Wales over the next five days for all types of natural flooding (coastal/tidal, river, groundwater and surface water). The FGS is issued by the FFC every day at 10:30am and at other times, day or night, if the flood risk assessment changes.



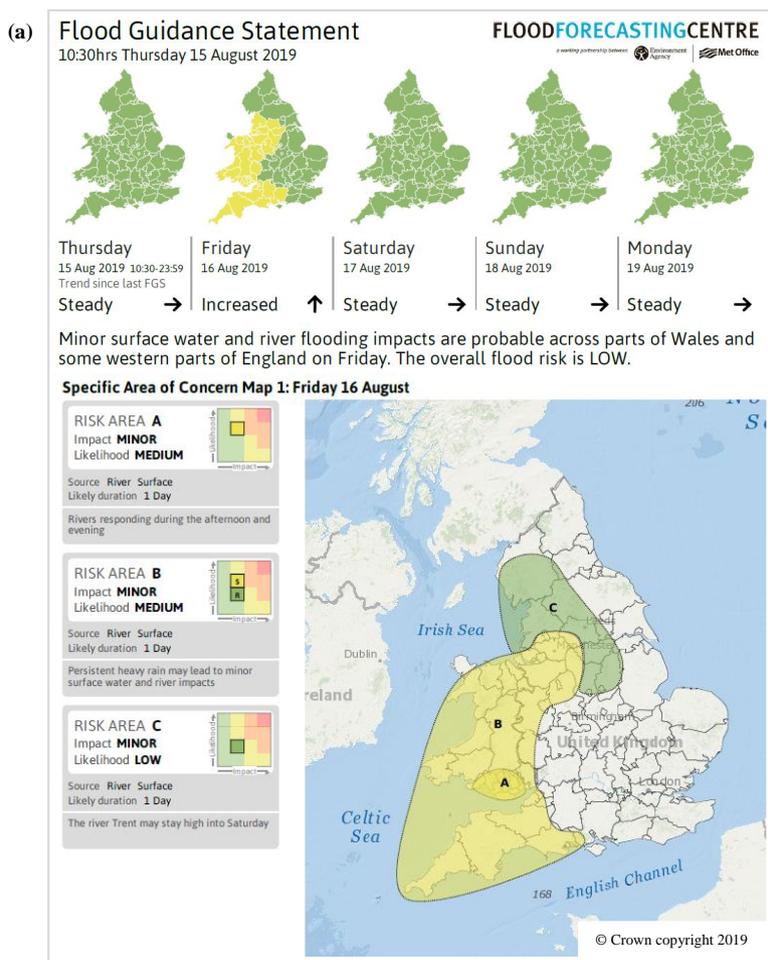
Flood incident management strategy	<p>www.ffc-environment-agency.metoffice.gov.uk/services/FGS_User_Guide.pdf An institute's priorities for preparing for and responding to flood events.</p>
Flood management measures	<p>Solutions to reduce the impacts that floods pose to humans and the environment. They can be natural (e.g. planting vegetation to retain extra water in the ground) or engineered (e.g. flood barriers).</p>
Flood preparedness	<p>Measures taken to prepare for and reduce the effects of a flood event.</p>
Flood scenarios	<p>Possible future development of a flood event and its associated likelihood.</p>
Flood wardens	<p>Volunteers from local communities who have the responsibility to monitor watercourses in the area they cover and contact local authorities with up to date information.</p>
Forcing	<p>The action of inputting information into a computer model to produce a forecast.</p>
Forecast accuracy	<p>The level of agreement between the forecast and the truth (i.e. what is observed in reality).</p>
Forecasting product	<p>A comprehensive and tailored overview (i.e. in the form of text, graphics and/or tables, etc.) of the forecast.</p>
Hydraulic model	<p>Mathematical model of the movement of water in a system (e.g. a river).</p>
Hydrological model	<p>Simplified model of a real-world system that describes the water cycle.</p>
Hydro-meteorological observations and forecasts	<p>Hydro-meteorology is a branch of meteorology and hydrology that studies the transfer of water and energy between the land surface and the lower atmosphere. Hydro-meteorological observations include observations of meteorological (e.g. temperature and rainfall) and hydrological variables (e.g. river and groundwater levels). Hydro-meteorological forecasts are forecasts that predict the evolution of meteorological and hydrological variables in time.</p>
Hydro-Meteorological Services	<p>Hydro-meteorological forecasting* products* produced by the FFC and issued daily (Hydro-Meteorological Guidance), twice daily (Forecast Meteorological Data) or whenever required (Heavy Rainfall Alerts).</p>
Lead time	<p>The length of time between when the forecast is made and the occurrence of the event (e.g. flood) being predicted.</p>
Long-range forecasts	<p>Forecasts which cover a period of time from a month to more than a season.</p>
Missed flood events	<p>A flood for which no warning was given ahead of it happening.</p>
Model performance	<p>The level of agreement between the model's outputs and their observations in reality. The difference between a model output and its respective observation is the error. The lower the error, the greater the model performance.</p>
Nowcasting	<p>Extrapolating from the latest observations (e.g. radar rainfall) to forecast the evolution of, for example the weather, in the next couple of hours.</p>
Operational decision-making	<p>Decision-making based on real-time information to resolve imminent situations.</p>
Outlook	<p>Refers to a forecasting product* based on long-range forecasts* (i.e. monthly to seasonal).</p>
Performance measures	<p>Metrics that characterise the quality of a forecast or a model compared to observations.</p>
Probabilistic forecasts	<p>While a deterministic model gives a single possible outcome for an event, a probabilistic model gives a probability distribution as a solution, indicating the likelihood of each scenario to occur. Probabilistic and ensemble forecasts are sometimes used interchangeably (see 'Ensemble').</p>



Rapid-response catchments	Catchments and rivers that respond quickly to rainfall events.
Real-time river gauges	Instruments that measure a river's characteristics (e.g. flow or water level) and communicate these data in real-time remotely.
Reasonable Worst Case	A forecaster's assessment of the potential upper range of rainfall, river and groundwater levels, and coastal conditions, and their impacts.
Risk	A combination of likelihood and impact of an event.
River level correlations	Mathematical characterisation of the river level at one point of the river with respect to another point on the river. This can be used to estimate the river level at a point on the river if the river level upstream is known.
Short-range forecasts	Forecasts which cover a period of time from a couple of a hours to a couple of weeks.
Surface water flooding	Flooding caused when the volume of rainwater falling does not drain away through the river network and other drainage systems, or infiltrate into the ground, but lies on or flows over the ground.
Surge forecasts	Forecasts of the rise of water along coastlines.
Telemetry sites	Sites where instruments collect measurements automatically and transmit it remotely (see 'Real-time river gauges')
Uncertainty	Having limited knowledge or understanding of our environment, it is impossible to characterise and predict its evolution with 100% certainty. All forecasts are uncertain, and that uncertainty amplifies with lead time*. Ensemble* or probabilistic forecasting* can be used to represent the forecast uncertainty.



Appendix B. Visual examples of operational products used by EA MFDOs and FWDs: (a) Flood Guidance Statement, (b) Hydro-Meteorological Guidance, (c) Forecast Meteorological Data, (d) Heavy Rainfall Alert, and (e) National Flood Forecasting System (source: EA). The works published in this journal are distributed under the Creative Commons Attribution 4.0 License. This licence does not affect the Crown copyright work, which is re-usable under the Open Government Licence (OGL). The Creative Commons Attribution 4.0 License and the OGL are interoperable and do not conflict with, reduce or limit each other.





(b) Daily Hazard Assessment

Issued 14:02 on Wednesday, 14 August 2019

The Daily Hazard Assessment is intended to provide an 'at a glance' top level overview only. The links provided to the relevant Partner Organisations should then be used to obtain further

**Hazards Five Day Summary – FLOOD: YELLOW,
THUNDERSTORM: YELLOW**

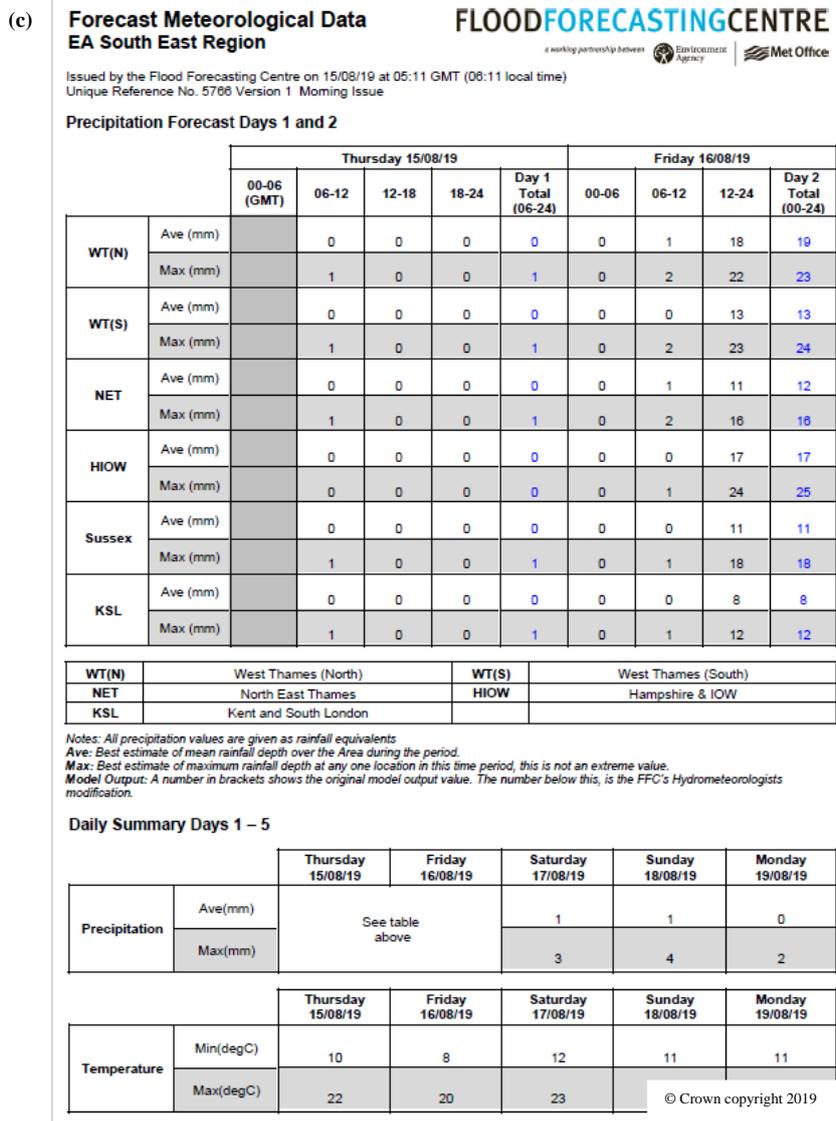
FLOOD: Significant surface water and river flooding impacts are possible but not expected across central England on Wednesday and into Thursday morning. The overall flood risk is LOW.

THUNDERSTORM: Heavy showers and thunderstorms may cause flooding and transport disruption for parts of central and eastern England on Wednesday afternoon.

Hazards Five Day Summary Maps

Wednesday 14-Aug 1400 – 2359	Thursday 15-Aug 0000 – 2359	Friday 16-Aug 0000 – 2359	Saturday 17-Aug 0000 – 2359	Sunday 18-Aug 0000 - 2359
------------------------------------	-----------------------------------	---------------------------------	-----------------------------------	---------------------------------







(d) **Heavy Rainfall Alert
 EA South East Region (Summer)**

FLOODFORECASTINGCENTRE

a working partnership between Environment Agency | Met Office

Issued by the Flood Forecasting Centre on 19/07/19 at 16:51 GMT (17:51 local time)
 Unique Alert Reference No. 2817_SOUTHEAST_795 Version 1

ORIGINAL

Start of meteorological event: 0800 GMT on 20/07/19

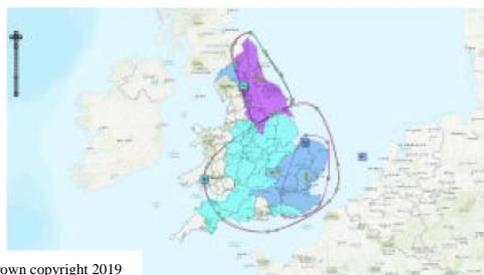
End of meteorological event: 2100 GMT on 20/07/19

Summary of Alert Criteria Met

Alert Criteria	HRA Areas covered	Confidence
10 mm (or more) in 1 hours (or less)	West Thames (North)	L
	West Thames (South), North East Thames, Kent and South London	M
30 mm (or more) in 6 hours (or less)	Sussex, Kent and South London	L
30 mm (or more) in 12 hours (or less)	West Thames (South), North East Thames	L

Notes:

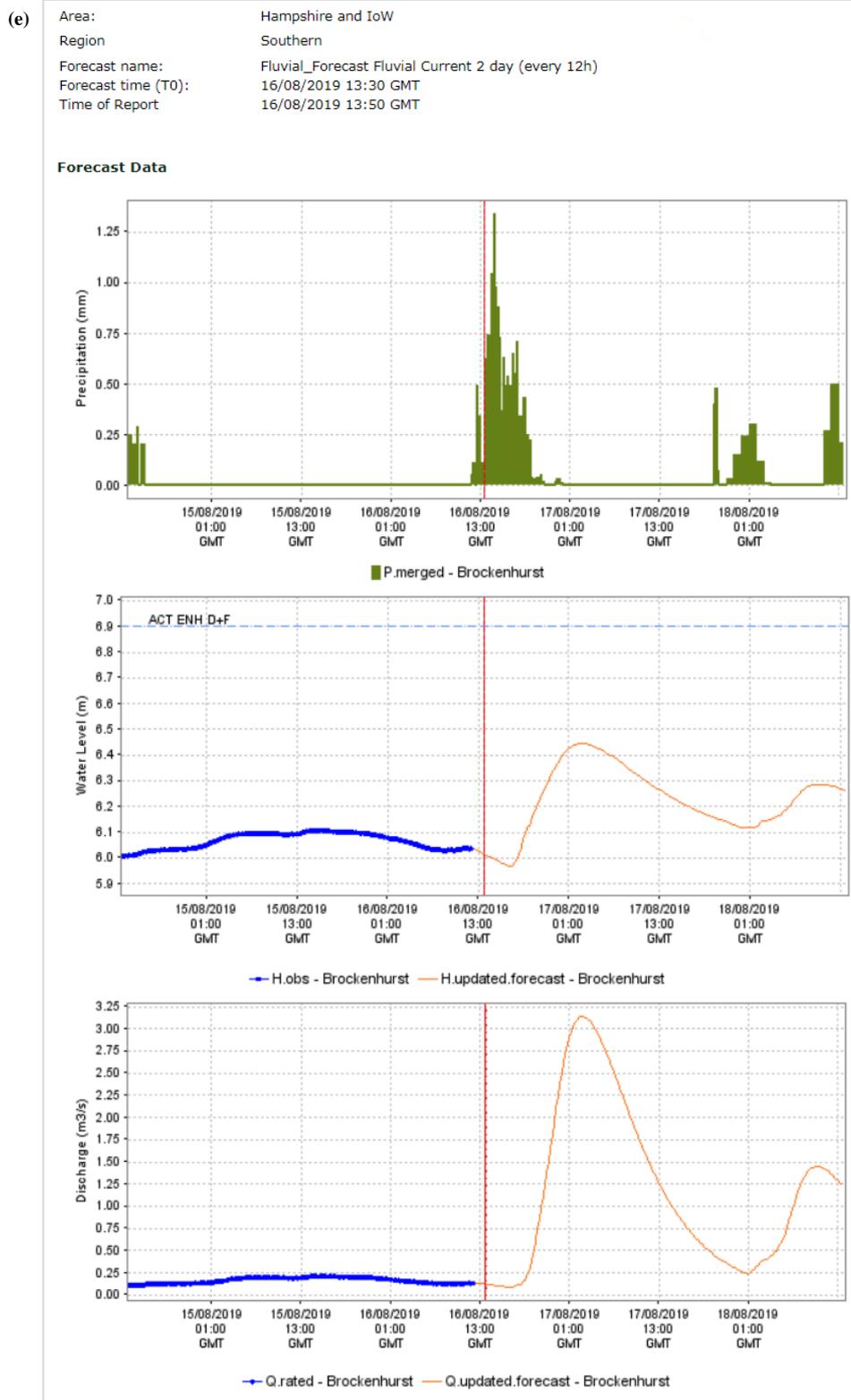
- **Confidence:** The probability of this threshold being achieved anywhere in the specific HRA Area within the time periods outlined by the Heavy Rainfall Alert. H = more than 60%; M = 40 – 60%; and L = 20 – 40%
- Issue of a Heavy Rainfall Alert means the probability of rainfall thresholds being met or exceeded during the meteorological event is within the bands indicated by the confidence levels above.
- All Alert criteria should be defined in this table. If it is predicted that some criteria will not be exceeded, these boxes should be greyed out



Each HRA Area is coloured according to the probability of its threshold being breached:

- Low (20 - 39%)
- Medium (40-59%)
- High (>= 60%)

© Crown copyright 2019





Appendix C. Interviewees' quotes in relation to their perceptions (opportunities, neutral and challenges) associated with a transition to probabilistic forecasts.

Opportunities

- O1:** *"If you've got a huge spread then you know that there's a very wide range of impact potentially, but if [...] everything's within a couple of centimetres of each other, it gives you a lot more confidence in saying, no I think we're going, we're not going to see a threshold crossing. So [...] it will help decision making I think" [MFDO3]*
- O2:** *"I think in a good way [...] it will [...] reveal the uncertainty that's hidden by apparent simplicity" [I1]*
- O3:** *"The new flood forecasting system is being developed at the moment so it's going to replace the NFFS. [The] benefits to that I suppose [...] are that if we can look to be more consistent across the country in even simple things like what displays look like [...] we're more interoperable if we need to" [MFDO1]*
- O4:** *"I think in an incident I'm happy that that's [...] a useful range of things to know, like you said, you probably warn for the lowest one and plan for the highest one and we can interpret between them" [FWDO2]*
- O5:** *"We're talking about some of these decisions that have got a long lead time, we're going to move people around the country, we're going to move equipment. It takes a long time to do that" [I1]*
- O6:** *"Between us [duty officers], it's probably OK because we've got that understanding of the roles" [FWDO3]*
- O7:** *"I can see some benefits to it, especially when you've got less experienced staff [...], you're almost [...] showing them the breadth of what a catchment could do given a range of responses" [MFDO2]*

Neutral

- N1:** *"Uncertainties are very tricky to deal with, whether probabilistic forecasting and a switch to that is going to help?" [MFDO2]*
- N2:** *"I think the MFDO role won't change, it will still be to communicate a forecast but the [...] wording of the forecast may change slightly" [MFDO1]*
- N3:** *"I think from our point of view it will just mean a bit more interpretation of forecasts and then [...] just a slightly different way of passing it on [...]. But I don't think it will change the process" [MFDO3]*



Challenges

C1: *“All the comms research we hear about generally says [...] the public message has to be as simple as possible, so that is working the opposite way to any proposal for probabilistic forecasting” [FWDO2]*

C2: *“A lot of local authorities standing their staff up, putting them on standby for a weekend is quite a big budget thing [...]. So [...] if we say, it is going to flood, they can justify the spend on it [...]. If we pass it on as shades of grey, a lot of them, they'll appreciate the information but some of them would actually resent having the decision forced on them because they will struggle to then justify doing something or they'll be blamed, either way, blamed for spending money if it doesn't happen and blamed for not spending enough if it does happen.” [FWDO2]*

C3: *“That would be my concern that it's even more information and more uncertainty and it's kind of like, well what do you do with this information? And which bit do you communicate to who?” [FWDO3]*

C4: *“It is something to bear in mind with that if probabilistic forecasting put too much pressure and stress on decision making on the people in these roles, the system probably would just collapse, people would walk away” [FWDO2]*

C5: *“You're still going to have this overriding issue with fast responding catchment where one scenario says we might need to issue a flood warning but 99 of them say no. Someone has to make a decision” [MFDO1]*

C6: *“I think still for a lot of people the question they [...] want answered is am I going to flood?” [I2]*

C7: *“I think my role is going to be the one where it has to stop and it can't be probabilistic because it [...] does come to a yes or no, issue it, don't issue it. So to some extent, probabilistic forecasting does feel like everyone else just pushing things down the line saying you make the decision, [...], we have to make the decision because we're the last ones on the line” [FWDO2]*

C8: *“Having probabilistic forecasting just moves the burden of making a decision further down the tree” [MFDO2]*

Participant Information Sheet

Investigators: Louise Arnal, Professor Hannah Cloke & Dr Elisabeth Stephens

Contact details: Louise Arnal

l.l.s.arnal@pgr.reading.ac.uk

Event: Change in operational forecast production and its impact on the Environment Agency (EA) monitoring and forecasting duty officers (MFDOs) & flood warning duty officers (FWDOs)? – Interviews with EA MFDOs & FWDOs

Undertaken as part of the EU Horizon 2020 IMPREX Project (<http://www.imprex.eu/>) (641811). This event is being co-organised with the Environment Agency (EA) and with support provided by the EA and the University of Reading.

Thank you for your interest in taking part in this series of interviews held in July to August 2018.

Your contribution to this interview is entirely voluntary and you have the right to withdraw from the interview at any time, without needing to provide a reason. You can do this by contacting me using the details above now, or at any time in the future.

Please note that if you choose to withdraw after the interviews have taken place, you will need to inform Louise Arnal before 18th August 2018 to ensure that all the information you provided is removed from any output. After this date, it may not be possible to guarantee that your information is removed.

Aims of the interviews

The Environment Agency (EA) has been using a single flood forecast scenario for nation-wide incident planning. However, the EA is currently changing their practice in-house to using two flood forecast scenarios (a best estimate and a reasonable worst case scenario), thereby communicating uncertainty in the flood forecasts. To support the use of these scenarios in operational decision-making, the EA would like to investigate the challenges faced by their MFDOs & FWDOs during this transition period, as well as the opportunities in making use of uncertainty information.

The main objectives of this research are:

- To understand how the transition to two flood forecast scenarios affects the EA MFDOs & FWDOs in their day-to-day job (opportunities vs challenges).
- To find ways to support EA MFDOs & FWDOs through this transition and future ones at the EA.
- To draw general guidelines from this case study, which could be used to support MFDOs & FWDOs in other institutes across the world during similar transitions.

How we have selected people to participate

We have invited up to 10 participants who work as MFDOs or FWDOs at different EA centres and who have all worked with single flood forecast scenarios in the past. The interviews are done with pairs of MFDOs & FWDOs, to have a more complete picture of the flow of information between those 2 roles.

On the day

The semi-structured interviews will be held throughout July and August 2018 and one interview should take approximately 1-2 hours.

All interviews will be recorded by Louise Arnal.

About the information you supply

The outputs from the interviews will be used to generate a short report, and for a published scientific journal article. Outputs may also form part of a PhD thesis, conference presentation and other output delivered as part of the IMPREX project.

All information will be treated in the strictest confidence. Verbal communication data recorded by Louise Arnal shall be subsequently anonymised (using codes to identify participants).

All the information shall be securely stored (see below) by the research lead (Louise Arnal) and may be shared with co-researchers (Professor Hannah Cloke, Dr Elisabeth Stephens and the EA commissioning team) as required.

To protect participant confidentiality, the University of Reading will keep any information you supply in a password protected secure system and in a separate place to any typed transcripts or hand-written notes until December 2019, when it will be destroyed.

All published output will include a disclosure statement that the information provided has been collated based on informal interviews and personal opinion and must not be taken as being representative of the views or practice of a particular organisation or institution.

We please ask that participants do not discuss the content of the discussion outside of the interview (i.e. Chatham House Rules).

This project has been subject to ethical review by the School of Archaeology, Geography and Environment Science and has been given a favourable ethical opinion for conduct. The research will be carried out by the Investigators listed above.

If you would like to know more about the information you provide or have any questions about the interview, please contact Louise Arnal using the details above.

Yours faithfully,

Louise Arnal