University of Reading

Post-processing large-scale river discharge forecasts at ungauged locations

Gwyneth Rhian Matthews

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Meteorology

School of Mathematical, Physical and Computational Sciences

June 2025

Declaration

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

Gwyneth Rhian Matthews

Abstract

Reliable river discharge forecasts are crucial for effective water resource management and flood risk mitigation. However, uncertainty in the forecasts is inevitable due to limitations in hydrological process understanding, errors in the meteorological forcings, observation uncertainty, and computational constraints. For flood forecasting, this forecast uncertainty can lead to ineffective or delayed preparatory actions. The primary aim of this thesis is to improve the actionable information from ensemble river discharge forecasts at gauged and ungauged locations using post-processing – a technique used to statistically correct the forecasts and reduce the uncertainty. Post-processing is already part of the forecasting system of the Copernicus Emergency Management Service's European Flood Awareness System (EFAS). The skill of the EFAS operational at-gauge post-processing method is evaluated, finding that post-processing improves the skill of the forecasts particularly for large rivers for which hydrological errors dominate. Barriers to the use of the operational post-processed forecasts are identified via a co-production workshop, including lack of local relevance, and difficulty accessing the forecasts and associated documentation. A key limitation of the operational post-processing method is that it is only applicable at gauged locations. To overcome this limitation, a data-assimilation-inspired technique is developed to propagate observation information along the river network. Combined with the at-gauge post-processing method, this new information propagation technique allows the correction of river discharge forecasts at ungauged locations. This new post-processing method was evaluated and found to improve forecasts up to a 5-day lead-time. The new method is computationally efficient, adapts to the flow situation, and is applicable to any ensemble river discharge forecast. By improving the skill of the forecast at ungauged locations, this work aims to support more informed decision-making in flood risk management and water resource planning, ultimately helping to protect people, infrastructure, and economies from hydrological extremes.

Acknowledgements

First and foremost, I would like to thank my supervisory dream team — Sarah Dance, Hannah Cloke, and Christel Prudhomme — for believing in me from the start and guiding me over the past half-decade. Through your kind honesty, you have taught me the importance of balancing pragmatism and perfection, work and life, details and the big picture — even if I'm still working on mastering it in practice. Thank you for the incalculable opportunities and for supporting me through the twists and turns of my PhD journey, including each switch in study mode and every misspelling of *ungauged*. Your patience, compassion, and brilliance as researchers will continue to inspire me long after this PhD.

I owe a massive thank you to the members of my research groups, DARC and Water@Reading, for being so friendly and encouraging. A special thanks to Laura Risley, Riccardo Monfardini, Jeff Da Costa, Devon Francis, Helen Griffiths, and Maureen Wanzala for organising insightful seminars and thought-provoking journal clubs.

Thank you as well to my colleagues at ECMWF, particularly those in the hydrological forecasting team, who have been so understanding as I balanced my PhD alongside full-time work, and who patiently answered all my questions — from the simple to the hypothetical.

I've had the privilege of doing my PhD alongside many admirable early-career researchers, whose enthusiasm for their topics has been an inspiration and whose friendship has been invaluable. My officemates, Mark, Wilson, Daniel, and Linda, deserve a special thank you for stopping the PhD from being a lonely process.

The Meteorology Department is filled with many incredible people from whom I have learnt over the years. Thank you to everyone who has answered or asked a question — especially when that question was "How are you?". In particular, I would like to thank Claire Bulgin and Natalie Harvey, as well as the entire SMPCS WIDE group, whose commitment to improving equality, inclusion, and wellbeing in the workplace is truly inspiring. I would also

like to thank the team of administrators who have made all this possible, including Debbie Turner, Imogen Salt, Wendy Neale, Christine Macfarlane, Nicki Robinson, and Xun Wang.

I thank Becky, Lauren, Sophie, Harshil, Abhi, Lowri, and Ales for making life easier and more fun. Your wisdom and humour were a constant source of strength throughout both the PhD and a global pandemic. Thank you to Hannah and Millie for the Monopoly Deal breaks, gym sessions, and Costa trips.

To Alex, thank you for reminding me that there's more to life than research deadlines, and for agreeing to spend your life with me. Thank you also to the Doyle-Taylor family, who have welcomed me so warmly and unreservedly.

Finally, my immeasurable gratitude goes to Mum, Dad, and Beth for always showing interest in my research — even when I turned a simple idea into a 10-minute tangent. Your unconditional love and unwavering support gave me the perseverance (or, as Dad calls it, 'stubbornness') to finish this thesis.

Authorship of publications

Chapter 3 is an excerpt from the following book chapter. The full chapter is updated from the first edition but the sections provided in Chapter 3 have been entirely written by Gwyneth Matthews. Co-authors provided guidance and review.

- **Chapter 3 (reviewed by editor)**: Matthews, G., Baugh, C., Barnard, C., De Wiart, C. C., Colonese, J., Decremer, D., Grimaldi, S., Hansford, E., Mazzetti, C., O'Regan, K., Pappenberger, F., Ramos, A., Salamon, P., Tasev, D., and Prudhomme, C.: *Chapter 14 - On the Operational Implementation of the European Flood Awareness System (EFAS)*, in Flood Forecasting (Second Edition), edited by T. E. Adams, C. Gangodagamage, and T. C. Pagano, 251–298, Academic Press, 2nd ed., 2025, ISBN 978-0-443-14009-9, https://doi.org/10.1016/B978-0-443-14009-9.00005-5.

Chapters 4, 5, and 7 of this thesis are reproduced from the following publications respectively. Specific author contributions are given for each chapter.

- **Chapter 4 (peer-reviewed)**: Matthews, G., Barnard, C., Cloke, H., Dance, S. L., Jurlina, T., Mazzetti, C., and Prudhomme, C.: *Evaluating the impact of post-processing medium-range ensemble streamflow forecasts from the European Flood Awareness System*, Hydrol. Earth Syst. Sci., 26, 2939–2968, https://doi.org/10.5194/hess-26-2939-2022, 2022.

  - GM designed the study, developed the code, performed the forecast evaluation, interpreted the results, drafted the manuscript and led the manuscript revisions. GM and CB created the post-processed reforecast dataset. TJ provided data and the scripts for the operational post-processing method. HC, SD, and CP provided guidance and review throughout.

- **Chapter 5 (under review)**: Matthews, G., Cloke, H. L., Dance, S. L., and Prudhomme, C.: *Error-correction across gauged and ungauged locations: A data assimilation-inspired approach to post-processing river discharge forecasts*, EGUsphere [preprint], https://doi.org/10.5194/hess-2024-3989, 2025. [under review]

  - GM developed the method, designed the study, developed the code, performed the forecast evaluation and analysis, interpreted the results, drafted the manuscript and led the manuscript revisions. HC, SD, and CP provided supervision, developed the method, and reviewed the manuscript.

- **Chapter 7 (published but not peer-reviewed)**: Matthews, G., Cloke, H. L., Dance, S. L., Hansford, E., Mazzetti, C., and Prudhomme, C.: *Co-Design and Co-Production of Flood Forecast Products: Summary of a Hybrid Workshop*, Bull. Amer. Meteorol. Soc., 104, E1058–E1066, https://doi.org/10.1175/BAMS-D-23-0061.1, 2023.

  - GM organised and facilitated the workshop, completed the ethics review form, collated and interpreted the results, drafted the manuscript and led the manuscript revisions. EH reviewed the workshop material and the manuscript. HC, SD, and CP provided supervision, and reviewed the manuscript.

Chapter 6 is in preparation for submission. The author contributions are as follows:

- GM developed the method, designed the study, developed the code, performed the forecast evaluation and analysis, interpreted the results, drafted the manuscript and led the manuscript revisions. HC, SD, and CP provided supervision and reviewed the manuscript.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Why is this thesis important?

It is difficult to name a part of society that is not impacted by rivers. As well as drinking water and sanitation, rivers contribute to our sports and leisure activities, agricultural and industrial processes, transportation, ecological diversity, and cultural and spiritual inspiration (Anderson et al., 2019; Wang and He, 2022). Accurately forecasting and monitoring the state of our rivers, therefore, has economic, social, political, and environmental benefits (Van Houtven, 2024; Parker et al., 2005). This is particularly true in the instance of extreme events, such as floods, where lives, livelihoods, and infrastructure are at risk (WMO, 2024b; Bubeck et al., 2017; Douben, 2006). Floods occur frequently, accounting for 44% of all recorded natural hazards globally between 2000-2019, affecting 1.65 billion people, and causing $651 billion of economic losses (UNDRR, 2020; Jonkman et al., 2024). In the future the damage caused by floods is expected to increase due to social-economic development and climate change (Rogers et al., 2025). However, the impact of a flood can often be mitigated if the appropriate preparatory actions are taken (Levine et al., 2020). Forecasts play a critical role in disaster mitigation by providing information about upcoming events, ideally, allowing sufficient time for preparatory actions (WMO, 2024a; Golding, 2022; UNDRR, 2015).

Despite substantial development over the past half a century (Todini, 2017), flood forecasts remain imperfect, limiting their ability to support decision-making (WMO, 2024a; Demeritt

et al., 2007). Reducing the uncertainty in forecasts is an ongoing challenge in hydrology (Blöschl et al., 2019) for which several techniques have been developed, accounting for various sources of uncertainty (Li et al., 2017; Camporese and Girotto, 2022; Herrera et al., 2022; Althoff and Rodrigues, 2021). In this thesis, “post-processing" - a statistical technique for correcting forecasts at the end of the forecast modeling chain (Section 2.2)- is investigated. By removing biases and quantifying uncertainty, post-processing aims to transform the forecast to be statistically more representative of real-world conditions (as depicted in Fig.1.1) facilitating decision-making based on the forecast information (Todini, 2017).

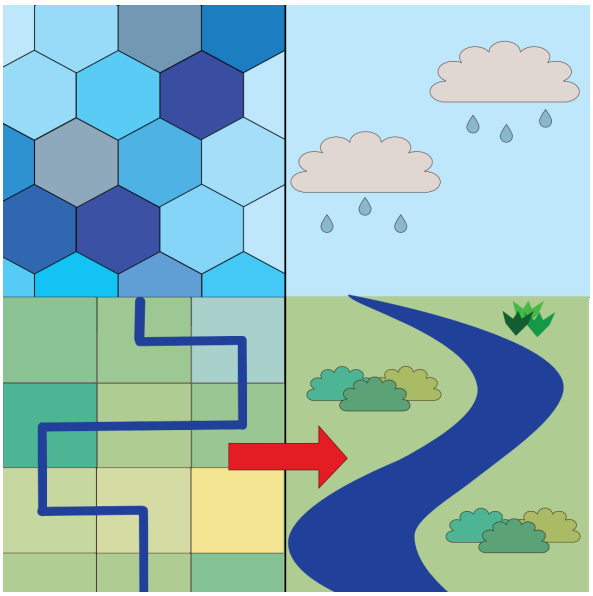

Figure 1.1: Illustration of the goal of post-processing which is to statistically transform the output from the hydrological model to be more like the observations. Colloquially, post-processing transforms the forecast from the model universe (left side of the figure) to the real world (right side of the figure).

Post-processing methods are relatively common in operational flood forecasting systems (e.g., Brown et al., 2014), including in the forecasting system used throughout this thesis, the European Flood Awareness System (EFAS; Smith et al., 2016) of the Copernicus Emergency Management Service (CEMS). Typically, post-processing methods depend on river discharge observations to estimate the characteristics of the forecast errors. River discharge observations are primarily obtained from river gauges (Section 2.1.3). Therefore, a fundamental constraint of many post-processing methods, including that of EFAS, is that they are only applicable at the locations of river gauges. Unfortunately, the global river gauge network is relatively sparse

(Krabbenhoft et al., 2022; Andrews and Grantham, 2024; Hannah et al., 2011), resulting in the majority of the forecast domain being uncorrected. Post-processing methods for ungauged locations have been researched significantly less than their at-gauge counterparts (Li et al., 2017). In general, previously developed post-processing methods for ungauged locations are computationally expensive or cannot adapt to changes in the spatial patterns of forecast errors (Bennett et al., 2022; Skøien et al., 2021). In this thesis, I develop a method for post-processing river discharge forecasts at ungauged locations by leveraging the dynamic and computationally efficient techniques used in data assimilation (Carrassi et al., 2018). Extending post-processing to all locations along a river could enhance decision-makers' confidence in the forecasts and their subsequent actions, ultimately helping to protect lives and livelihoods, even in areas without direct observations.

1.2 Large-scale medium-range river discharge forecasts

1.2.1 Forecasts as part of disaster risk management

Reducing the risk of floods is a cyclic process broadly divided into four overlapping stages: i) preparation for a specific event, ii) responding to the event once it has started, iii) recovering from the event, and iv) mitigating the impact of future events (Madu et al., 2018; Kiptum et al., 2025). Forecasts are a critical tool that contribute primarily to the *preparedness* stage of the disaster risk management cycle (Fig. 1.2; Le Cozannet et al., 2020). A good forecast provides advance knowledge of an upcoming flood event (Stephens and Cloke, 2014). Decision makers are then able to make a more informed choice about which actions will reduce the impact of the flood and enable a rapid response. These actions may involve issuing warnings, mobilizing resources and financial aid, assembling defenses, or evacuating people (Hayes, 2021). Some actions require a minimum lead-time to be effective (Wazed et al., 2021). Medium-range forecasts-those with lead times of 2–15 days-are particularly valuable as they offer a balance between actionable lead times and forecast skill (Bischiniotis et al., 2019; Fakhruddin, 2014; Cloke and Pappenberger, 2009). When integrated into a

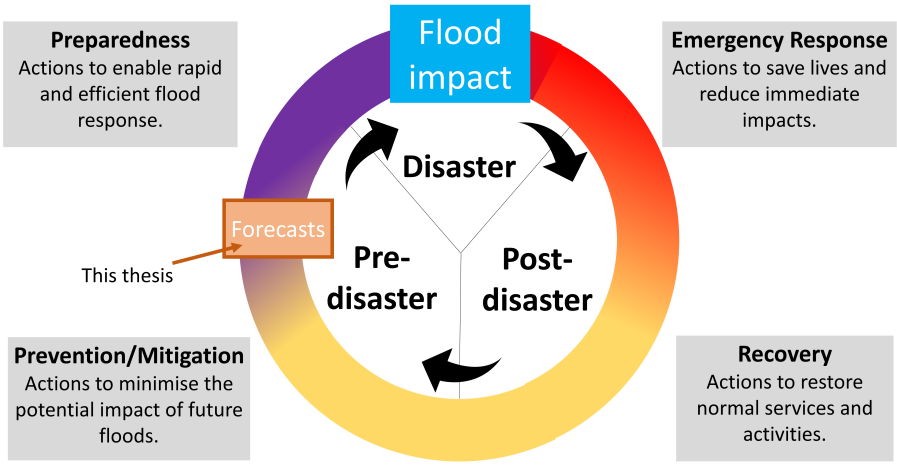

Figure 1.2: Schematic of a simple disaster risk management cycle. Modified from Le Cozannet et al. (2020)

robust Flood Early Warning System (FEWS) forecasts have been shown to help save lives and infrastructure as well as reduce the overall economic burden of floods (UNDRR, 2024; Bank, 2010; Perera et al., 2019; Thielen-del Pozo et al., 2015).

Despite their usefulness, local flood forecasts are not available everywhere often due to limited financial and technical capacity, insufficient data, or fragmented institutional frameworks (UNDRR, 2024). In the past two decades, technological and scientific advances as well as increased availability of global datasets have allowed large-scale flood forecasting systems to become operational (Emerton et al., 2016). A large-scale flood forecasting system refers to a system with broad geographical domains crossing many natural and political borders. There are three primary benefits to large-scale systems (Emerton et al., 2016): 1) they support the United Nations' goal of 'Early Warnings for All' (WMO, 2022) by providing forecasts where local systems are not available, 2) they complement national systems by extending lead times or adding probabilistic components (Thielen et al., 2009), and 3) they provide consistent information across administrative borders, facilitating transboundary coordination and response (Rai et al., 2017). The forecasting system used throughout this thesis, EFAS, is a continental-scale forecasting system designed to provide complementary flood forecast information to organisations responsible for flood protection within their region (Thielen et al., 2009; Smith et al., 2016; Arnal et al., 2019).

1.2.2 A typical large-scale flood forecast modelling chain

The modelling chain of a forecasting system is the set of modules that take the input data and create the forecast of the variable of interest (WMO, 2011). It typically does not include the dissemination of the forecasts to users although this is a crucial step within a FEWS (UNDRR, 2015). Additionally, I focus on forecasts of river discharge but the modelling chain could be extended to predict other variables such as flood inundation, flood impact or hydropower generation (Dottori et al., 2016; Kalas et al., 2019; Chen et al., 2019). A full description of the EFAS modelling chain is given in Chapter 3 but here I focus on the general modelling chain of a large-scale forecasting system.

Three components are required within the modelling chain of a large-scale medium-range flood forecasting system (Emerton et al., 2016; Wu et al., 2020; Cloke and Pappenberger, 2009): 1) meteorological forcings, 2) a set of initial conditions, and 3) a hydrological model to simulate the hydrological processes of runoff generation and routing (Fig. 1.3). Meteorological forcings are typically generated by Numerical Weather Prediction (NWP)

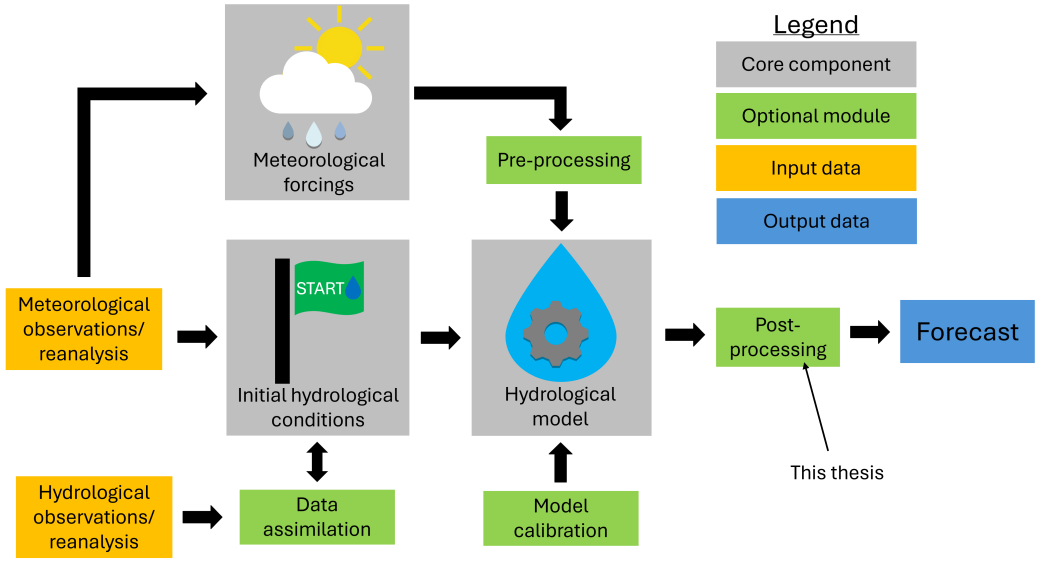

Figure 1.3: Diagram of a large-scale flood forecast modelling chain. Required core components are shown alongside optional modules that could be implemented depending on computational resources and data available. Modified from (Boelee et al., 2019).

systems (Wu et al., 2020; Bartholmes and Todini, 2005). Advances in science and technology have significantly improved forecast complexity and skill (Linsenmeier and Shrader, 2023;

Alley et al., 2019; Stern and Davidson, 2015). Recently, machine learning-based forecasts have become operational, showing skill competitive with traditional NWP systems (Lang et al., 2024; Bi et al., 2023). As will be discussed in Chapter 2 using forecasts as input introduces uncertainty.

To translate meteorological input into river discharge, hydrological models are used to represent runoff generation and river routing processes (WMO, 2011). These models vary in structure, complexity, and data requirements, and are typically classified as physically-based, conceptual, or data-driven (Horton et al., 2022; Knoben et al., 2025). Physically-based models rely on equations derived from our (incomplete) understanding of hydrological processes (Beven, 2012a), while conceptual models apply simplified process representations (Liu et al., 2017). Both typically require parameter calibration (see Section 2.1.2), as not all parameters are directly measurable (e.g., Manning's roughness; García-Pintado et al., 2015). Data-driven models, including machine learning models, define the relationship between the input (e.g., precipitation) and the output variables (e.g., river discharge) via weights or coefficients determined in a training process, without explicitly modelling the hydrological processes (Nearing et al., 2024; Hasan et al., 2024; Remesan and Mathew, 2015). The LISFLOOD hydrological model, that is used in EFAS, adopts a hybrid approach combining conceptual runoff generation with physically-based river routing (van der Knijff et al., 2008). These approaches balance process representation with computational efficiency, making them suitable for large-scale applications.

Hydrological models also differ in spatial representation. Lumped models treat the catchment as a single unit, while distributed models divide it into grid cells, enabling spatial variability (Beven, 2012a). The LISFLOOD is a distributed model that uses a gridded structure, but ignores the impact of lateral flow within the soil (van der Knijff et al., 2008). Representing spatial heterogeneity requires detailed information about catchment properties (Choulga et al., 2024; Conacher, 2002), which has become increasingly feasible due to advances in Earth observation (Emerton et al., 2016).

The initial hydrological conditions provide information about the amount, location, and state of water across the domain at the start time of the forecast (Beven, 2012b). The initial

hydrological conditions are important to be able to predict the response of the catchment to the meteorological forcings. For example, at a simple level, soil that is saturated with water cannot absorb additional water from rainfall leading to greater surface runoff than if water had been able to infiltrate the soil (Davie, 2019). For spatially continuous forecasts, initial hydrological conditions are generally generated using a warm-up of the hydrological model forced by meteorological observations, reanalysis, or short lead-time forecasts (Ceppi et al., 2023).

Additional components can be added to the modelling chain, usually with the aim of reducing or quantifying the uncertainty in the forecasts (Fig. 1.3). These methods are discussed in more detail in Chapter 2 with a focus on post-processing and data assimilation.

1.3 Thesis objectives and research questions

The primary goal of this thesis is to conduct research to improve the actionability of large-scale flood forecasting systems via post-processing and stakeholder engagement. The European Flood Awareness System (EFAS), a distributed continental-scale, medium-range ensemble flood forecasting system is used throughout; however, the methods and conclusions can be generalised to other forecasting systems. There are four objectives addressed in this thesis. The first three objectives consider two post-processing methods: the at-gauge post-processing method used operationally in EFAS and a new post-processing method for ungauged locations. The final objective investigates the usefulness of the current operational EFAS post-processed forecasts, and considers the implications for the operationalisation of the new method for ungauged locations. To aid readability, section dividers are used below to distinguish between the discussions of each objective.

* * *

Post-processing is currently performed operationally at gauged locations in EFAS to account for uncertainty in the meteorological forcings, hydrological model, and initial conditions (see Section 2.1.1). Despite its operational status, the strengths and limitations of the post-processing method, as well as its impact on the skill of the forecasts, have not been explored

previously. Without a robust evaluation of the post-processed forecasts, integrating the forecasts into the decision making process is more difficult. Additionally, without understanding the conditions under which the method performs well or poorly, developers are unable to prioritise scientific developments. The first objective of this thesis is therefore:

Objective 1: To evaluate the skill of the operational EFAS at-gauge post-processing method.

To address *Objective 1*, I compare the post-processed forecasts with river discharge observations at over 500 stations across Europe. The skill of the post-processed forecasts are compared against the skill of the raw (non-postprocessed) forecasts to identify any improvement. In particular, I consider two research questions:

Q1.1. Does the operational post-processing method provide improved forecasts?

Q1.2. What affects the performance of the post-processing method?

Objective 1 is addressed in Chapter 4. The evaluation of the at-gauge post-processed forecasts was used to inform the development and evaluation of the new post-processing method for ungauged locations (*Objective 2* and *Objective 3* below).

* * *

A fundamental limitation of many post-processing methods (including the method used operationally in EFAS) is that they can only be applied at gauged locations (see Section 2.2.2). Post-processing methods for ungauged locations have been developed but typically have three limitations (see Section 2.2.3). First, many methods define a static relationship between gauged and ungauged locations along the river network which does not allow the post-processing method to adapt to the flow situation being forecast. Second, methods that do adapt dynamically are often too computationally expensive for operational use. Finally, the previously developed methods do not work within the current modelling chain of many forecasting systems, requiring an entirely new post-processing method to be implemented. In this thesis I aim to develop a post-processing method for ungauged locations that adapts

to the flow situation, is computationally efficient, and can be implemented as part of any ensemble flood forecasting system. I do this in two steps, corresponding to *Objective 2* and *Objective 3*, respectively.

First, I develop an observation information propagation method using data-assimilation techniques in a post-processing environment. Data assimilation is a state estimation technique that combines a prior model estimate with observations. Data assimilation can be used in flood forecasting systems to improve the estimation of the initial conditions (see Section 2.3). Hydrological data assimilation is much rarer in operational systems that its meteorological counterpart due to data latency, consequences further along the modelling chain, and uncertain benefits at longer lead-times (see Section 2.3.6). However, by applying data assimilation techniques in a post-processing environment I aim to overcome these barriers. The second objective of this thesis is therefore:

Objective 2: To investigate the applicability of data-assimilation in a post-processing environment to propagate observational information from gauged to ungauged locations.

To address *Objective 2*, I adapt the Local Ensemble Transform Kalman Filter (LETKF) for use in a post-processing environment. I assimilate observations at gauged locations and examine the impact at locations where observations are not assimilated. Two research questions are considered:

Q2.1. Can data assimilation techniques be used in a post-processing environment to propagate observation information to ungauged locations?

Q2.2. Are the resulting ensemble predictions of river discharge more skillful than the raw ensemble?

Objective 2 is addressed in Chapter 5. The new observation information propagation method is the first step towards a post-processing method for ungauged locations.

* * *

In the experiments conducted to address *Objective 2*, observations from the forecast period are assimilated to assess the ability of the method to successfully propagate information along the river network. In a true forecasting scenario, no observations are available during the forecast period since, at the time of forecast generation, the forecast period represents the future. There is a need therefore to investigate alternative sources of information that could replace the observations. The third objective of this thesis is

Objective 3: To explore the use of proxy-observations to facilitate the post-processing of ensemble river discharge forecasts at ungauged locations.

To address *Objective 3*, I develop a post-processing method for ungauged locations that uses at-gauge post-processed forecasts as proxy-observations and propagates this information to ungauged locations. To assess the efficacy of the new method I consider two research questions:

Q3.1. Can error corrections derived from at-gauge post-processed forecasts be effectively propagated to ungauged locations using a data-assimilation-based method?

Q3.2. What factors influence the method's ability to improve the forecast skill at ungauged locations?

* * *

Flood forecasts are only of societal benefit if they can be used by decision makers. Stakeholder engagement is a critical, yet often under-emphasized component of the development and operationalisation of flood forecasting systems. Forecast users require forecasts that are not only skillful but also understandable, timely, and useful for decision-making (see Section 2.4). Therefore, it is crucial to consider operational constraints and user requirements during the research phase. Therefore, the fourth objective of the thesis is

Objective 4: To identify potential barriers for the use of the EFAS post-processed forecasts and determine future developmental priorities.

To address *Objective 4*, I designed and organised a hybrid workshop between developers and users of the operational at-gauge EFAS post-processed forecasts. The workshop aimed to answer three research questions:

Q4.1. What is the current usage of the EFAS post-processed forecasts?,

Q4.2. What are the barriers to the use of the EFAS post-processed forecasts?

Q4.3. What developments to the EFAS post-processed forecasts should be prioritised to better satisfy user requirements?

The decision to operationally implement the new post-processing method for ungauged locations (*Objective 3*), as well as the implementation process itself, falls outside the scope of this PhD and is not guaranteed. However, by identifying the limitations of the current operational post-processed forecasts I aim to identify areas where the new method may also face challenges.

* * *

As mentioned, throughout this thesis EFAS is used as the test forecasting system. Conducting research within a large-scale operational system presents both challenges and advantages. On the one hand, output of the research can be more readily used within the operational system, since the constraints and practicalities of the forecasting chain are considered from the start. On the other hand, the operational system continues to evolve during the course of the research, which can create difficulties for evaluation of new developments (e.g., Hasan et al., 2024, ; see discussion in Section 8). The EFAS domain covers a wide range of hydroclimatic regions, enabling post-processing methods to be evaluated across diverse settings, whereas national systems may be more targeted towards specific hydrological processes (e.g., snowmelt). In particular, the relatively large number of river gauges across Europe makes EFAS useful for developing a post-processing method for ungauged locations. Global-scale systems, such as the Global Flood Awareness System (GloFAS; Matthews et al., 2025c), encompass an even broader range of hydroclimatic regions than EFAS, but many of the regions have few

or no river gauges making it more difficult to evaluate post-processing methods for ungauged locations (see 5 and 6). Moreover, because post-processing is already performed at gauged locations within EFAS, much of the necessary workflow and infrastructure for testing new methods is already established. Finally, EFAS has an established and diverse user community, which simplifies the identification of stakeholders and facilitates engagement with end users. These characteristics make EFAS a suitable and representative system for addressing the four objectives of this thesis, although the methods used could be implemented in any hydrological ensemble forecasting system.

1.4 Principal new results

Chapters 4-7 provide the following answers to the thesis objectives outlined in Section 1.3:

1. **Objective 1: The first systematic evaluation of the at-gauge post-processing method for EFAS was conducted.** Using two years of reforecasts, the skill of the post-processed forecast was compared against that of the raw forecasts at 522 stations across Europe. (Q1.1) The assessment finds that the post-processing method corrects the forecasts at most stations with the greatest improvement occurring at short lead-times. (Q1.2) The post-processing method is most beneficial for large, slowly responding catchments with hydrological biases. The method struggles with flashy catchments and those with meteorologically driven errors as well as with more extreme river discharge values. (Impact) The evaluation results have been presented to users allowing them to more effectively use the forecasts. In addition to providing knowledge about the benefits of post-processing the EFAS forecasts (knowledge which previously has not been available to users of the forecasts), this evaluation provides a benchmark on which the other experiments in this thesis can be contextualised.

2. **Objective 2: A new method for propagating observational information along the river network to correct ensemble simulations in a post-processing environment was developed.** The method uses state-augmentation within a LETKF to spread in-

formation from gauged to ungauged locations. Modifications are made to the LETKF to remove the need for additional executions of the computationally expensive hydrological model. (Q2.1) The method's ability to propagate the observational information to ungauged locations is evaluated using a leave-one-out verification strategy for the Rhine-Meuse basin. Overall, the data-assimilation-inspired method was able to effectively propagate observational information from gauged to ungauged locations by using the ensemble covariances of the simulation. (Q2.2) The skill of the ensemble mean was improved at almost all locations and lead-times but when local processes are not captured in the ensemble covariances this correction was less robust. At short lead-times, the ensemble spread was more reliable than for the raw forecasts. At longer lead-times, the spread was not sufficiently adjusted to match the reduction in the error of the mean, leading to under-confident forecasts. (Impact) The ability of this method to produce an improved estimate of the river discharge could aid post-event analysis for ungauged locations. In this thesis, this method is an initial step towards post-processing forecasts at ungauged locations.

3. **Objective 3: A new method for post-processing river discharge forecasts at ungauged locations was developed**. Building on the results from Objective 1 and 2, the new method uses the at-gauge post-processing method to provide proxy-observations during the forecast period, and propagates this information along the river network using the LETKF-based technique. The post-processing method is evaluated for a year of forecasts in the Rhine-Meuse catchment. (Q3.1) The forecasts are generally improved at lead-times up to 5 days and along larger rivers. The skill of the post-processed forecasts varies across the catchment and across the evaluation period. (Q3.2) The skill of the post-processed ensembles at ungauged locations is largely dependent on the accuracy of the proxy-observations. The density of proxy-observations and their location upstream or downstream of the ungauged location also influenced the forecast skill. (Impact) Additional evaluations on different catchments are required before this method can be considered for operational implementation; however, it shows potential

to improve the forecasts at ungauged locations providing users with skillful forecasts across the domain.

4. **Objective 4: Barriers to the use of the operational EFAS post-processed forecasts and developmental priorities were identified.** A workshop was organised between developers and users of the forecasts enabling practical solutions to be identified. (Q4.1) The primary use of the post-processed forecasts was for flood forecasting for lead-times from 1 to 5 days. The post-processed forecasts were always used alongside another meteorological or hydrological forecast product. (Q4.2) Barriers to the use of the EFAS post-processed products included difficulty incorporating the forecasts into local procedures and detailed documentation being hard to find. (Q4.3) Core to many of the identified developmental priorities was the use of locally relevant information. For example, local flood thresholds would allow users to more easily compare with their own systems and incorporate the forecasts into their procedures. Both the forecast visual products and the forecast data itself need to be easily accessible and understandable. (Impact) The workshop enabled a plan for developing the EFAS post-processed forecasts to be outlined. Some of these developments have already been made operational. Additionally, using the outcomes from this workshop I have been able to identify potential roadblocks for any future operational implementation of the new post-processing method for ungauged locations.

1.5 Structure of the thesis

This thesis consists of nine chapters including four chapters which present new research conducted as part of this PhD (Fig. 1.4). The thesis is structured as follows:

- Chapter 2 provides a review of relevant background information including an introduction to the uncertainties in hydrological data, post-processing methods, data assimilation techniques, and a co-production framework.

- Chapter 3 provides a detailed system description of the European Flood Awareness

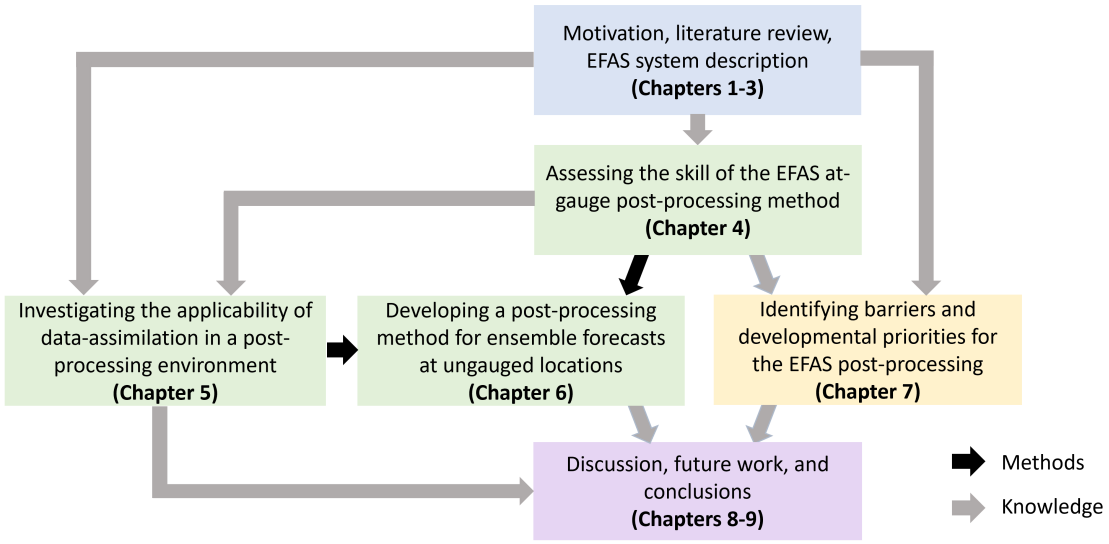

Figure 1.4: Schematic of the structure of this thesis. Colours of the boxes indicate sections with similar objectives: Blue - provide background material; Green - related to the development of a post-processing method; Yellow - related to user requirements; Purple - discussion and summary of the results of this thesis.

System (EFAS) which is the river discharge forecasting system used in all the studies in this thesis. The research studies were conducted alongside the operational system so this chapter also provides context to some of the challenges encountered in the studies. This chapter is an excerpt from Matthews et al. (2025b).

- Chapter 4 presents an evaluation of the EFAS post-processed forecasts at stations across Europe. This was the first skill assessment made available to users of the current operational post-processing method. In addition to investigating the benefits of the post-processing method, we also explore the factors that limit the method's efficacy. This chapter addresses *Objective 1* and is reproduced from Matthews et al. (2022).

- Chapter 5 develops a method for propagating observational information from gauged to ungauged locations using a Local Ensemble Kalman Filter (LETKF) in a post-processing environment. This method allows an ensemble simulation to be corrected at ungauged locations. This chapter addresses *Objective 2* and is reproduced from Matthews et al. (2025a).

- Chapter 6 combines the methods from Chapters 4 and 5 to form a post-processing

method for ensemble forecasts at ungauged locations. The post-processing method for ungauged locations can be applied to any ensemble river discharge forecast. This chapter addresses *Objective 3*.

- Chapter 7 discusses the organisation and outcomes of a workshop held between the producers and the users of the EFAS post-processed forecasts. The workshop identifies developmental priorities for the EFAS post-processed forecasts related to the forecast visualisation, accessibility, training, and post-processing method used. This chapter addresses *Objective 4* and is reproduced from Matthews et al. (2023).

- Chapter 8 discusses the key scientific contributions from each of the research chapters (Chapters 4-7) in relation to the four objectives outlined in Section 1.3. Additionally, the limitations and the impacts of the studies are discussed before potential future work is outlined.

- Chapter 9 highlights the key conclusions of this thesis.

Following each of chapters 2-7 is a 'chapter summary' which briefly summarises the chapter before highlighting the connections with other chapters in this thesis.

Chapter 2

River discharge uncertainty, post-processing, data assimilation, and actionable forecasts

In this chapter, I aim to contextualize the work presented in this thesis. Each of the main research chapters of this thesis contains a specific literature review as they are published works or under review. In Section 2.1, I begin with a discussion of the uncertainties present in river discharge observations and forecasts, and provide a brief description of some common approaches for quantifying and reducing forecast uncertainties. In Section 2.2, I narrow the focus to post-processing methods which I argue is the most effective and efficient technique to statistically improve river discharge forecasts. Section 2.3 introduces data assimilation in more detail with a specific focus on the Local Ensemble Transform Kalman Filter (LETKF), the method used within this thesis. In Section 2.4, I discuss the value of forecasts and the co-production framework. Finally, in Section 2.5, I summarise the the key conclusions from this chapter and highlight where this knowledge is used in future chapters.

I note that Chapter 3 provides more details about the forecast system used throughout this thesis- the Copernicus Emergency Management Service's European Flood Awareness System (EFAS) - so the system is not specifically discussed in Chapter 2.

2.1 Uncertainty in river discharge forecasts and observations

The hydrological community has widely agreed that the quantification and, where possible, reduction of uncertainty is a necessity for the societal benefits of hydrological forecasting to be realized (Blöschl et al., 2019; Juston et al., 2013; Pappenberger and Beven, 2006). Generally, in hydrology uncertainty is categorized into two types: epistemic uncertainty, which arises from lack of knowledge, and aleatory uncertainty, which stems from inherent randomness (McMillan et al., 2018; Gupta and Govindaraju, 2023; Nearing et al., 2016; Di Baldassarre et al., 2016; Beven and Smith, 2015; Montanari, 2007). These uncertainties enter the forecast via several components of the modelling chain. Therefore, I begin this section by outlining the sources of uncertainty in river discharge forecasts (Section 2.1.1). I then introduce techniques that are used to quantify and reduce forecast uncertainty (Section 2.1.2). Two of these techniques, namely post-processing and data assimilation, are key to this thesis and are discussed in greater detail in Sections 2.2 and 2.3, respectively. In Section 2.1.3 we discuss the uncertainty associated with river discharge observations which are key to all forecast uncertainty quantification methods.

2.1.1 Sources of uncertainty in river discharge forecasts

River discharge forecasts contain uncertainty which can make their use in decision making more difficult (Demeritt et al., 2007). Predictive uncertainty is the probability of a forecasted situation occurring conditioned on all knowledge available before the start time of the forecast (Todini, 2009). Knowledge that we may have before the start-time of the forecast includes estimates of the hydrological initial conditions, predictions about the future meteorological situation, past observations, the skill of the hydrological model, and previous forecasts. This uncertainty is introduced to the system throughout the modelling chain as shown in Fig. 2.1 and discussed in more detail below. Whilst the uncertainties are introduced from different sources they interact as they cascade through the forecasting chain making them difficult to

disentangle and account for separately (Zappa et al., 2011; Pappenberger et al., 2005).

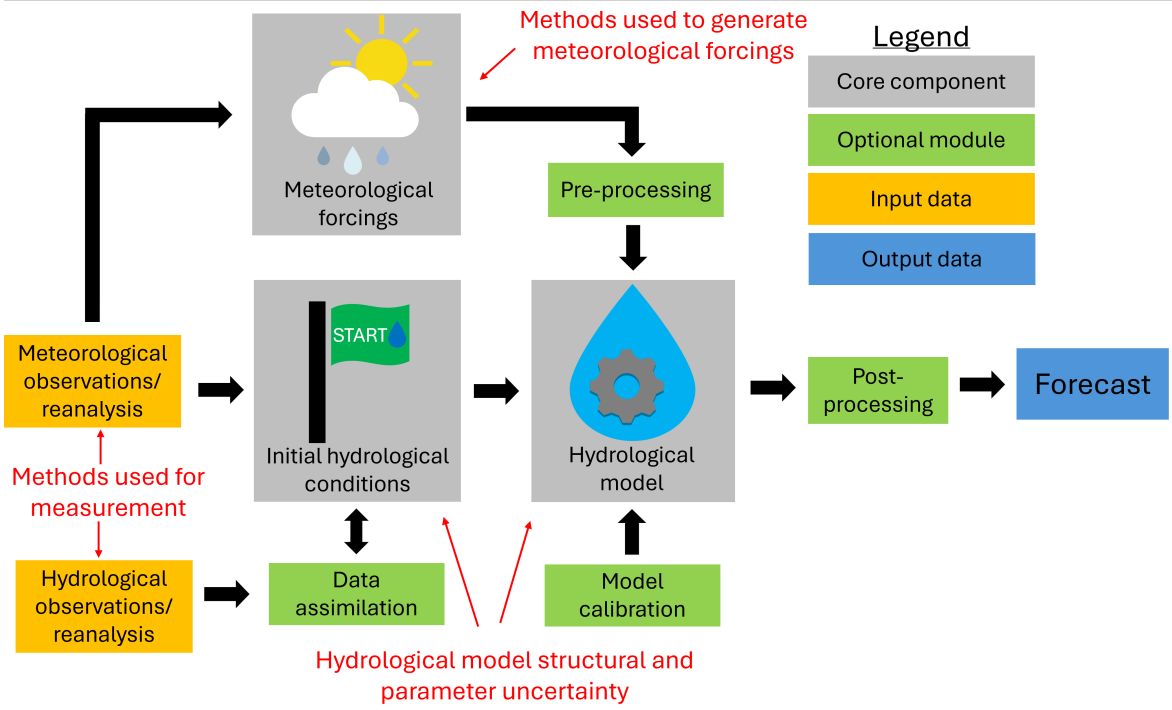

Figure 2.1: Model components and sources of uncertainty for a typical flood forecasting system. Modified from (Boelee et al., 2019).

Meteorological uncertainty or input uncertainty is introduced into the hydrological modelling chain by the meteorological forcings used (Zappa et al., 2010). In this thesis, the river discharge forecasts are generated using the output of a NWP as input; however, forecasts generated using meteorological observations (e.g., radar based nowcasting) will also be impacted by input uncertainty due to uncertainties in the meteorological observations (e.g., Park et al., 2019). Throughout this thesis we treat meteorological uncertainty as a single source of uncertainty, as is often done (e.g., Boelee et al., 2019), but in reality this uncertainty is a combination of uncertainty due to the chaotic nature of the atmosphere and NWP model errors.

Whilst there are many hydrological models (Peel and McMahon, 2020), most are affected by two types of hydrological model uncertainty: i) model structure uncertainty and ii) model parameter uncertainty. For traditional physical models, model structure is related to the hydrological processes represented and their coupling, the equations selected to represent these processes, and the spatial configurations of the model (Moges et al., 2021). Application,

available data, and computing resources, as well as familiarity and system legacy, can influence the choice of model and model configurations (Horton et al., 2022). Model parameter uncertainty is related to the selection of parameter values for the model. Often, there are many parameters (some observable and some implicit) that need to be defined for a model to be used (Arnal et al., 2019). Parametric uncertainty is impacted by the spatial resolution of the model, the forcing data available, the understanding of hydrological processes, and the hydroclimatic region of the domain (Herrera et al., 2022).

Initial condition uncertainty or state uncertainty is due to the initial hydrological conditions on the domain. For single-location forecasts (also called point forecasts), the initial conditions can often be taken as the most recent observation (DeChant, 2014) and are therefore susceptible to the uncertainty in river discharge observations (discussed in Section 2.1.3). For spatially distributed forecasts the initial conditions are often generated by forcing the hydrological model with meteorological observations, reanalyses or forecasts (Harrigan et al., 2020). The initial conditions therefore contain uncertainty from the meteorological forcings of choice and from the hydrological model (Parkes et al., 2013).

These sources of uncertainty are unavoidable but in Section 2.1.2 we discuss some common techniques used to reduce and quantify the uncertainty from each source.

2.1.2 Accounting for uncertainty in river discharge forecasts

Several techniques to reduce and quantify the uncertainty for different sources have been researched. I focus here on techniques that can be used within forecasting systems similar to the EFAS system used in this thesis (i.e., river discharge forecast created by forcing a hydrological model with the output from an ensemble of NWP system; see Chapter 3).

There are two ways of quantifying predictive uncertainty (Boelee et al., 2019): 1) creating an ensemble of realisations showing the different potential future states of the river discharge (Wu et al., 2020; Cloke and Pappenberger, 2009), or 2) statistical methods that can condition components of the model on information available before the start time of the forecast (Panchanthan et al., 2024). Whilst the former technique can quantify the uncertainty, statistical

methods are also capable of reducing the uncertainty by correcting for known errors. Since ensembles can provide uncertainty quantification but not correction it is common to use a statistical method together with a Hydrological Ensemble Prediction System (HEPS; Troin et al., 2021). It is this hybrid approach that will be explored further in this thesis.

2.1.2.1 Ensemble techniques for quantifying uncertainty

An ensemble of potential future states can be generated using multiple sets of meteorological forcings (to account for input uncertainty; Cloke and Pappenberger, 2009), multiple sets of hydrological model parameters (to account for hydrological parameter uncertainty; Mockler et al., 2016), multiple hydrological models (to account for hydrological structural uncertainty; Velazquez et al., 2011), multiple sets of initial hydrological conditions (to account for initial condition uncertainty; Abaza et al., 2014), or a combination of two or more of the above sets (Zappa et al., 2011; Pappenberger et al., 2005; Qu et al., 2017). However, to account for all sources of uncertainty would be very computationally expensive using physically-based or conceptual models (Pappenberger et al., 2005; Leutbecher, 2019; Raynaud and Bouttier, 2017).

Input uncertainty is considered to be the most dominant source of uncertainty in flood forecasting (e.g., Shu et al., 2023; Pappenberger et al., 2005). Therefore, ensemble river discharge forecasts are often created by forcing a single hydrological model with an ensemble of meteorological forecasts (Pappenberger et al., 2019; Emerton et al., 2016; Bao et al., 2011; He et al., 2010). These meteorological forcings may come from an ensemble NWP system or multiple NWP systems (Wu et al., 2020) as is the case for the EFAS system used in this thesis (see Chapter 3). Figure 2.2 shows an example ensemble river discharge forecast hydrograph for a single location created by forcing a hydrological model with an ensemble meteorological forecast.

2.1.2.2 Methods for quantifying and reducing uncertainty

Common statistical methods used to quantify and reduce uncertainty are pre-processing for input uncertainty, hydrological model calibration for parameter uncertainty, and data

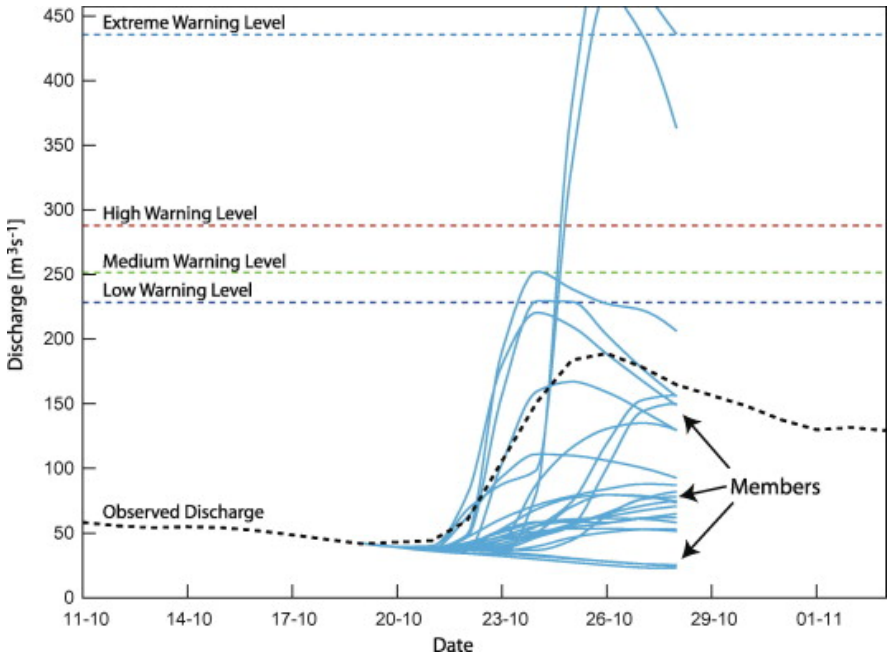

Figure 2.2: An example of an ensemble river discharge forecast generated by forcing a hydrological model with an ensemble weather forecast. Each blue line is a different ensemble member which are all equally likely future scenarios. The dashed black line shows the observed river discharge that occurred. Reproduced from Cloke and Pappenberger (2009) with permission from Elsevier.

assimilation for initial condition uncertainty (Boelee et al., 2019). These statistical methods are implemented in different sections of the modelling chain depending on the source of the uncertainty (Fig. 2.3).

- **Pre-processing** (or post-processing in meteorology) consists of statistically correcting the output from the NWP before using it to force the hydrological model. This has shown to be beneficial in some cases with precipitation being a key variable to correct (Monhart et al., 2019). In other cases pre-processing has led to little difference in the skill of the hydrological forecasts or to a slight degradation (Valdez et al., 2022).

- **Hydrological model calibration** is the process of optimising the model parameters by maximizing the skill of a hydrological simulation relative to observations. Often this results in one set of parameters which are then used to run the hydrological model to create forecasts (e.g., Hirpa et al., 2018). In reality some of the parameters that are often calibrated vary either seasonally or due to changes in the catchment (Blöschl et al., 2015).

- **Data assimilation** is used to improve the estimation of the initial hydrological conditions by combining a prior modelled estimate with observations. Improvements have been seen in forecasts due to data assimilation, primarily in the short lead-times (Bourgin et al., 2014). A key benefit of data assimilation is that river discharge observations are not needed everywhere as data assimilation methods are capable of spreading the observation information to unobserved locations (Johnson et al., 2005). Data assimilation is described in more detail in Section 2.3.

- **Post-processing** is the statistical correction of the river discharge forecasts to quantify uncertainty and reduce biases (Li et al., 2017). Post-processing aims to quantify the predictive uncertainty in the forecast (i.e., the uncertainty in the forecast due to all sources of uncertainty; see Section 2.1.1). Post-processing is described in more detail in Section 2.2.

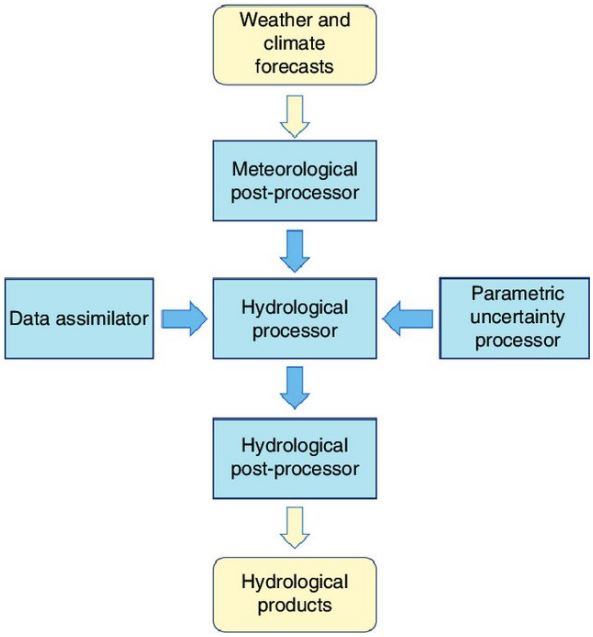

Figure 2.3: Uncertainty analysis components in a flood forecasting system. Reproduced from Li et al. (2017), with permission from John Wiley and Sons.

Several studies have compared different statistical methods to identify which method is most beneficial. Post-processing is consistently found to improve forecast skill more effectively than pre-processing (Kang et al., 2010; Roulin and Vannitsem, 2015; Zalachori et al., 2012; Sharma et al., 2018; Valdez et al., 2022). Improvements due to pre-processing

are limited because errors in the hydrological model are not considered (Sharma et al., 2018). Post-processing is also capable of accounting for errors in model parameters with the potential of replacing the computationally expensive hdyrological model calibration process (Liu et al., 2022; Wu et al., 2018; Roulin and Vannitsem, 2015). While integrating both data assimilation and post-processing often results in the most skillful forecasts, data assimilation had a diminishing effect at longer lead-times (Zhang et al., 2025a; Valdez et al., 2022; Bourgin et al., 2014).

The ability of post-processing to account for multiple sources of uncertainty in a computationally efficient manner has led to it being a key component in hydrological forecasting systems. In Section 2.2 we discuss post-processing in more detail including common methods and ongoing challenges.

2.1.3 Uncertainty in observed river discharge data

River discharge observations are used throughout hydrological research to understand physical processes (Kauffeldt et al., 2013), detect trends in space and time (Harrigan et al., 2018a), and identify appropriate flood warning thresholds (Lucas et al., 2023), to name a few examples. As shown in Section 2.1.2, they are also critical for quantifying and reducing the uncertainty in forecasts. Understanding the uncertainties in these observations is therefore important (McMillan et al., 2022; Rhea et al., 2023).

The observations used in this thesis are traditional in-situ observations, measured by gauging stations at specific locations along the river network. The observations used in this thesis are traditional in-situ observations, measured by gauging stations at specific locations along the river network. In particular, we focus on the uncertainty in continuous in-situ river discharge observations.

River discharge is the volume of water flowing through a river channel at a specific point (Robinson and Ward, 2017). For discrete measurements, the river discharge can be calculated by measuring the velocity of the flow and the cross-sectional area of the river (Herschy, 1993). There are many techniques for measuring these two components with varying complexity,

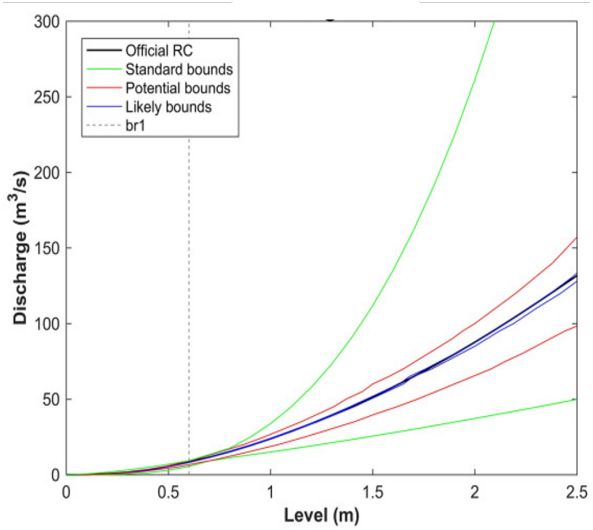

Figure 2.4: Example rating curve and uncertainty bounds for the Agauntza stations on the Oria river in Spain. Uncertainty bounds are calculated in three ways: (1) standard - no hydraulic information is used, (2) potential - basic hydraulic information is used, and (3) likely - detailed and reliable hydraulic information is used. Reproduced from Ocio et al. (2017) with permission from John Wiley and Sons.

cost, and accuracy (Vyas et al., 2024; Muste and Hoitink, 2017). However, continuous real-time river discharge observations are typically measured via the rating-curve method not via the velocity-area measurements (Turnipseed and Sauer, 2010). In the rating curve method, stage (also known as the water level) is measured by the river gauge. A rating curve, which represents the relationship between the stage and the river discharge (Fig. 2.4), is then used to convert the stage to river discharge (Turnipseed and Sauer, 2010).

Broadly, there are three sources of uncertainty in river discharge observations (Turnipseed and Sauer, 2010; Rhea et al., 2023): i) uncertainty in the water level measurement, resulting from factors such as changes in the river bed settlement, waves or surges, vegetation over-growth, and limitations of the sensor (Horner et al., 2022), ii) uncertainty in the rating curve, arising from the non-stationarity of the stage–discharge relationship, errors in the calibration data, parametric assumptions, and extrapolation beyond observed values (Mansanarez et al., 2019), and iii) uncertainty in the recording, transferring, and retrieving of the observations (Viney and Bates, 2004).

The uncertainty in river discharge observations is magnitude-dependent and can vary greatly between locations (e.g., Coxon et al., 2015 find uncertainties range between 10%

and 397%). However, in many studies the range of uncertainty values is often smaller (e.g., 6.2% to 42.8% in Di Baldassarre and Montanari, 2009). Commonly reported magnitudes are 10–20% for medium flows (McMillan et al., 2012). For high and low flows, the percentage of uncertainty is often higher due to the extrapolation of the rating curve required to convert the stage measurement to discharge and limitations of the stage measurements themselves (Horner et al., 2022; Hamilton and Moore, 2012; Tomkins, 2014; Harmel and Smith, 2007).

Estimating the uncertainty in the observations requires information about the river and the gauge itself; information that is often not available at a large-scale (Westerberg and Karlsen, 2024; Coxon et al., 2015). The choice of method used to estimate the uncertainty in river discharge observations can also result in discrepancies in the estimated uncertainty (Kiang et al., 2018; Westerberg et al., 2020). It is therefore difficult to estimate the uncertainty associated with river discharge observations in near real-time at a large-scale. Therefore, it is common to assume either that the observations have no errors or that the uncertainty can be estimated as a constant percentage of the magnitude (Ocio et al., 2017; McMillan et al., 2010a). For example, for the Oder river uncertainties of 10% were estimated for daily average discharge values (Refsgaard et al., 2006). In Chapter 5, I use this percentage of the river discharge value as the standard deviation of the error as 10% falls within the uncertainty ranges of many studies reviewed including those cited above.

2.2 Post-processing of river discharge forecasts

The topic of this thesis is post-processing of medium-range ensemble river discharge forecasts with a focus on developing a post-processing method for ungauged locations. Post-processing was briefly introduced in Section 2.1.2 as a statistical approach for accounting for biases and uncertainty in the output of a hydrological model. In this section, I discuss post-processing in more detail starting with a description of the aims of post-processing in Section 2.2.1. Typically, post-processing at ungauged locations is a two-step process which involves an at-gauge post-processing step followed by the spreading of this correction to ungauged locations. I therefore first discuss previously developed post-processing techniques for gauged locations

in Section 2.2.2 followed by post-processing techniques for ungauged locations in Section 2.2.3.

2.2.1 Introduction to post-processing

The output from a hydrological model contains biases and uncertainties from multiple sources (see Section 2.1.1). Post-processing is a statistical method applied at the end of the forecast modeling chain with the goal of making the forecast more useful to decision makers (Fig. 2.3, Li et al., 2017). The aim of post-processing in general is to reduce and quantify the predictive uncertainty which, as discussed in Section 2.1.1, is the probability of a forecasted situation occurring given all available information at the forecast initiation (Todini, 2009). Let $\mathbf{x}_t \in \mathbb{R}^L$ be the river discharge forecast for a single location with start time t and L timesteps, and let $\mathbf{y}_t \in \mathbb{R}^L$ be the river discharge observations for the same L timesteps. The aim of post-processing is therefore to estimate an improved forecast probability density function, denoted

$$p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}) \tag{2.1}$$

where p indicates the probability density and $\boldsymbol{\theta}$ is a vector of additional input variables that may include, for example, past observations.

There are three techniques that commonly fall under the term “post-processing": bias-correction, error-correction, and uncertainty quantification. Bias-correction corrects systematic errors, such as persistent under- or overestimation of river discharge (Hashino et al., 2007). The aim of bias correction is to minimize the mean forecast error (forecast residual) over time. Unlike bias-correction, which only addresses systematic errors, error-correction also accounts for random errors that vary across forecasts (e.g., underestimation due to river regulation; Bogner and Kalas, 2008). In general, error-correction techniques are more complex than bias-correction techniques and often require real-time observations. For a deterministic point forecast, the error of the forecast for a single lead-time, δ_k, is defined as,

$$\delta_k = y_k - x_k \tag{2.2}$$

where y_k and x_k are the observation and the forecast at the k-th lead-time, respectively. It is often assumed that the observations and forecasts are at the same location and can be directly compared (Li et al., 2018). For an ensemble forecast the error of the forecast may apply to the ensemble mean or to each ensemble member separately depending on the chosen methodology.

Uncertainty quantification focuses on characterizing the likelihood of possible forecast outcomes rather than providing a single corrected value (Krzysztofowicz, 2001). It involves estimating the distribution of the forecast errors or the confidence intervals around predictions (Ramos et al., 2013). Unlike bias- or error-correction, which aim to improve the accuracy of forecasts, uncertainty quantification aims to provide a probabilistic description of forecast confidence by accounting for inherent variability, model limitations, and data quality (Todini, 2009). As they do not reduce errors, uncertainty quantification techniques often follow bias- or error-correction techniques and aim to quantify the remaining uncertainty (Li et al., 2016).

2.2.2 Post-processing methods

Post-processing methods typically require past river discharge observations to estimate the relationship between forecasted and observed river discharge values. As a result, most post-processing approaches have been developed for gauged locations (Li et al., 2017). Over the past few decades, a variety of post-processing techniques have been introduced, which can be broadly categorized into five categories: bias-scaling, regression methods, conditional probability methods, ad-hoc methods, and more recently, machine learning methods.

Bias-scaling techniques are among the most widely used post-processing methods, across all timescales, due to their simplicity (Kolling Neto et al., 2023; Lozano et al., 2025; Tiwari et al., 2022). These methods correct forecast biases by applying a multiplicative or additive adjustment based on the historical relationship between observed and forecasted discharge values. Quantile mapping is commonly used to adjust forecast errors by defining a mapping between the cumulative distribution functions of the hydrological model output and the observations (Tiwari et al., 2022; Kim et al., 2022; Sanchez Lozano et al., 2021; Hashino

et al., 2007). These methods are computationally efficient and easy to implement but do not fully account for non-linear forecast errors.

Regression-based methods improve upon simple bias-scaling by establishing statistical relationships between forecasts and observations (Zhang et al., 2025b; Islam et al., 2023; Tyralis and Papacharalampous, 2023; Hashino et al., 2007). Using multiple linear regression also allows the inclusion of additional predictors such as meteorological variables or recent river discharge observations (Du and Pechlivanidis, 2025; Demargne et al., 2014). River discharge tends to have a long memory (or, mathematically, a high autocorrelation) so the inclusion of recent observations via an autoregressive model has been shown to provide an advantage over simpler methods (Jones et al., 2022; Woldemeskel et al., 2018; Seo et al., 2006). Regression-based techniques effectively correct systematic biases and the additional predictors can allow the output to be tailored to specific forecasting conditions correcting non-systematic errors as well (Bogner and Pappenberger, 2011; Bogner and Kalas, 2008). Several regression-based methods have been developed specifically for ensemble forecasts often adjusting the ensemble mean and standard deviation (e.g., Wood and Schaake, 2008). Of particular interest for this thesis is the Ensemble Model Output Statistics (EMOS; Gneiting et al., 2005), which minimises the error (or maximises the skill depending on the chosen metric) over recent forecasts using multiplicative and additive correction factors to adjust the ensemble mean and standard deviation (Zhong et al., 2020; Siqueira et al., 2021; Bogner et al., 2017). The EMOS method is used in Chapter 4 as part of the at-gauge post-processing method used operationally in the European Flood Awareness System (EFAS). The EMOS method was initially designed for meteorological variables but has shown to be effective at correcting biases and improving the ensemble reliability for river discharge forecasts as well (Hemri et al., 2015; Siqueira et al., 2021).

Conditional probability methods were some of the first hydrological post-processors to quantify uncertainty as well as bias-correct the forecast (Krzysztofowicz and Kelly, 2000; Krzysztofowicz and Maranzano, 2004). They estimate the likelihood of the observed discharge given a forecasted value (Darbandsari and Coulibaly, 2021). The Model Conditional Processor (MCP; Todini, 2008, 2009; Coccia and Todini, 2011) is used in Chapter 4 as

part of the operational post-processing method of the EFAS forecasts. The MCP estimates a multi-Gaussian joint distribution between historic observed and simulated data which is then used to estimate the probability distribution of the observed values during the forecast period based on the forecast values (Romero-Cuellar et al., 2022; Alvisi and Franchini, 2017). Conditional probability methods can effectively correct both bias and spread but require long observational records to produce reliable distributions (Zhao et al., 2011; Coccia and Todini, 2011).

Ad-hoc methods include ensemble dressing which aims to correct the spread of ensemble forecasts. In ensemble dressing, the uncertainty of each ensemble member is assumed to be described by a kernel-distribution which is estimated from previous errors (Pagano et al., 2013; Xu et al., 2020). Ensemble dressing is often used in conjunction with Bayesian Model Averaging which, after dressing each ensemble member, assigns different weights to members based on their historical performance (Raftery et al., 2005; Fraley et al., 2010). Bayesian Model Averaging can also be used to blend a multi-model ensemble of forecasts from different systems (Aminyavari and Saghafian, 2019).

Machine learning methods have been applied to the post-processing of river discharge forecasts for decades (e.g., Shamseldin and O'Connor, 2001). However, the recent curation of large hydrometeorological datasets and the availability of open-source packages has reenergised the discipline. Techniques such as random forests (Du and Pechlivanidis, 2025; Magni et al., 2023), support vector machines (Liu et al., 2022), and long-short term memory networks (Du and Pechlivanidis, 2025; Sharma et al., 2023; Hunt et al., 2022; Frame et al., 2021) have shown promising results in capturing complex, non-linear relationships between forecasts and observations. Machine learning post-processors are highly flexible and can easily incorporate multiple predictors. However, their effectiveness depends on the availability of large training datasets, computational resources, and careful model tuning to avoid overfitting (Hasan et al., 2024; Slater et al., 2024; Shen et al., 2021).

Each of these post-processing methods has distinct advantages and limitations. These are summarised in Table 2.2.2. In practice, the choice of method depends on forecast characteristics, data availability, and operational constraints. Increasingly, hybrid approaches that

combine multiple techniques are being explored to enhance forecast accuracy and reliability (Reggiani et al., 2022). In Chapter 4, we evaluate the at-gauge post-processing method used operationally in the European Flood Awareness System (EFAS). This method uses a hybrid approach combining the MCP and the EMOS methods to account for different sources of uncertainty.

Table 2.1: Advantages and disadvantages of five categories of post-processing methods.

Method Category	**Advantages**	**Disadvantages**
Bias-scaling techniques	Easy to implement; Computationally efficient; Applicable to many systems; Widely used and well-understood.	Do not correct dynamic errors; Do not quantify uncertainty; Data requirements can be large; Assume stationarity.
Regression methods	Can incorporate recent river discharge observations and other predictors; Do not require long time series for calibration; Some account for uncertainty; Flexible and interpretable.	May assume linear relationships; May overfit with many predictors and limited data.
Conditional probability methods	Can incorporate recent river discharge observations; Account for uncertainty; Suitable for probabilistic forecasting.	Require long time series for calibration; May be less interpretable; Assume stationarity.
Ad-hoc methods	Typically tailored to specific systems; Some account for uncertainty; Some require fewer assumptions about statistical distributions.	Some require large datasets of historic forecasts; Not applicable to all systems.
Machine learning methods	Can capture non-linearity in errors; Quick to run (after training); Can handle large numbers of predictors	Require long time series for training; Many methods do not account for uncertainty; May not be interpretable; May overfit without proper tuning.

2.2.3 Post-processing methods for ungauged locations

River discharge gauges are sparsely and inhomogeneously distributed globally (Krabbenhoft et al., 2022). Even on highly observed rivers there are challenges in the maintenance of the gauges and the sharing of the collected data (Hannah et al., 2011; Nelson, 2009). Floods can cause significant impacts in areas without observations, making it crucial to develop accurate and reliable forecasts for ungauged locations. Post-processing river discharge forecasts at ungauged locations is more challenging than at gauged locations—where the process is already complex—because there are no direct observations from which to learn the relationships between forecasted and observed discharge values. It is possible to extend some at-gauge post-processing methods directly for use at ungauged locations. For example, Engeland and Steinsland (2014) applies a conditional-distribution method to post-process forecasts at multiple locations based on the joint distribution of river discharge between the locations. Whilst not requiring recent observations, historic observations are required to define the joint distribution and therefore this approach cannot be used at locations for which observations have never been available.

To overcome this challenge, methods for post-processing at ungauged locations typically rely on post-processing forecasts at gauged locations and then transferring the information to the ungauged locations (e.g., Skøien et al., 2021; Bennett et al., 2022). Therefore, these methods can be conceptualized as consisting of two components: at-gauge post-processing (described in Section 2.2.2) and error propagation, which we focus on in this section. The choice of error propagation technique can directly restrict the selection of at-gauge post-processing methods, as certain transfer techniques are only compatible with specific post-processing approaches. The computational constraints of the forecasting system are also important factors in the selection of error propagation technique.

A common error propagation technique is to match gauged and ungauged locations in a way that allows correction parameters estimated at the gauged location (the "donor") to be applied to an ungauged location (the "recipient"; Fig. 2.5a). This donor-recipient approach is widely used in hydrological model calibration but has also been successfully applied

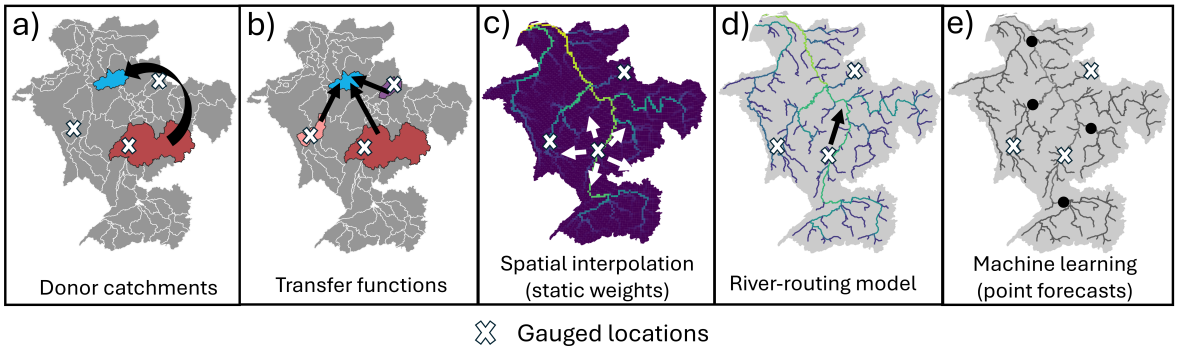

Figure 2.5: Illustration of different methods for transferring information within a post-processing method for ungauged locations.a) Error correction parameters are transferred from a gauged donor catchment (red) to an ungauged catchment (blue). b) Transfer functions are used to estimate error-correction parameters at ungauged locations (blue) based on a physical characteristics of the catchment (red, pink, and purple). c) Spatial interpolation, using static weights (yellow-high, purple-low), of error correction parameters from gauged to ungauged locations. d) Propagation of corrected river discharge along the river network using a river routing model. e) Machine learning models trained at gauged locations and applied to ungauged locations (black dots).

to post-processing ungauged river discharge simulations (Hales et al., 2023; Skoulikaris et al., 2022). By selecting donor gauges based on hydrological connectivity, geographical proximity, and similarity of hydrological characteristics, Hales et al. (2023) improved the skill of GeoGlows (Ashby et al., 2021) simulation across many catchments globally. Another approach involves the use of transfer functions, which typically establishes a regression function between various hydrological characteristics and the correction parameter of interest with the coefficients defined using data from gauged locations (Fig. 2.5b). While transfer functions have proven effective for model calibration (e.g., Samaniego et al., 2010), their application in post-processing introduces additional uncertainty, which can reduce the amount of skill gained (Farmer et al., 2018). Typically, these error propagation techniques are used in conjunction with bias-correction methods rather than error correction or uncertainty quantification methods.

Rather than assigning a parameter value based on similarity with gauged locations, topological kriging (or 'top-kriging') is a spatial interpolation method which assigns parameter values based on previously defined covariances between locations (Skøien et al., 2006, ; Fig. 2.5c). Top-kriging estimates the parameter value for an ungauged location as a weighted average of nearby gauged locations taking into account the river network and the upstream

area of both the gauged and ungauged locations. Skøien et al. (2021) used top-kriging to spread the bias and spread correction parameters of an at-gauge EMOS method to ungauged locations across the whole of Europe. The results show an improvement in the skill of the forecasts at ungauged locations. However, the evaluation focuses on correction of errors due to the meteorological forcings and ignores the errors due to the hydrological model.

An alternative to transferring correction parameters from gauged to ungauged locations is to propagate the post-processed river discharge downstream from gauged locations using a river routing model (Bennett et al., 2022, ; Fig. 2.5d). This approach has the advantage of not requiring an explicit definition of hydrological similarity between locations or the estimation of their covariances. However, errors from the river routing model may offset improvements made at the gauged location, particularly as the distance from the gauge increases. Additionally, running the river routing model a second time adds computational costs, which may be prohibitive for large-scale continental or global forecasting systems.

As discussed in Section 2.2.2, machine learning (ML) methods have increasingly been applied as post-processors of forecasts from physically based models. In general, when near real-time river discharge observations are provided as input variables to an ML model, the forecast skill improves (Nearing et al., 2021). However, at ungauged locations, it is not possible to provide such observations, often resulting in a significant decrease in the added benefit of the ML method (Nearing et al., 2024). Unlike other error propagation approaches, which explicitly separate at-gauge post-processing and error propagation, ML-based post-processing learns error behaviors from gauged locations during an offline training procedure and then assumes the model is applicable to both gauged and ungauged locations (Fig. 2.5e; Konapala et al., 2020; Won et al., 2023; Choi and Kim, 2025). Most ML-based hydrological forecasting studies have focused on pure ML systems (e.g., Kratzert et al., 2018). However, ML-based post-processing (where a physically-based hydrological forecast is an input variable to the ML model) has been shown to outperform purely ML-generated forecasts, particularly at ungauged locations (Xiao et al., 2025; Choi and Kim, 2025; Sharma et al., 2023).

In Chapter 5 we present a data-assimilation-inspired method for post-processing river

discharge forecasts at unguaged locations.

2.3 Data Assimilation

This thesis uses data assimilation techniques which are modified for use in a post-processing environment. Data assimilation is a technique often applied in environmental sciences for state estimation (e.g., estimation of the river discharge) at ungauged locations, and therefore techniques used to overcome challenges in hydrological data assimilation may be useful for overcoming challenges in hydrological post-processing. In this section I introduce some of the key concepts of data assimilation (Section 2.3.1), and some data assimilation techniques within the Kalman filter family of methods in particular (Section 2.3.2). I also motivate our choice of data assimilation technique, the Local Ensemble Transform Kalman Filter (LETKF), and provide equations for its implementation (Sections 2.3.3 and 2.3.4). Finally, I describe some considerations for the use of data assimilation in hydrological forecasting systems (Section 2.3.5).

2.3.1 Introduction to data assimilation

Data assimilation is a mathematical framework that combines two sources of data to produce a more accurate estimate of the state of the system along with an estimate of the associated uncertainty (Asch et al., 2016; Nichols, 2010). In environmental disciplines, including meteorology, oceanography, and hydrology, the two primary sources of data are: 1) modelled data, in the form of a short-term forecast referred to as the *prior state* (or background state), and 2) observations, which are direct or indirect measurements of the system (Carrassi et al., 2018; Reichle, 2008). Both sources of data have advantages and limitations. For instance, the prior state provides a spatially consistent estimate of the system, with values for each variable for each grid-box of the domain but may contain model errors. In contrast, observations are tied directly to the true state of the system but are only available at specific locations and times. Combining the prior state and the observations, data assimilation exploits the advantages of both data sources. The resulting state estimate, called the analysis, is often

used as the initial conditions for a forecast or as part of a long-term record, called a reanalysis.

Various data assimilation techniques have been developed to meet specific needs of different systems, data characteristics, and computational constraints (e.g., variational methods (Bannister, 2017); Kalman filters (Sun et al., 2016); and particle filters (Fearnhead and Künsch, 2018)). This thesis uses the Local Ensemble Transform Kalman Filter (LETKF; Hunt et al., 2007), a sequential data assimilation technique within the Kalman filter family of methods. As will be described further in in Sections 2.3.2-2.3.5, the LETKF is computationally efficient, scalable, and can handle nonlinearities in the hydrological processes, making it an appropriate choice for large-scale hydrological systems. In this section, I introduce some concepts that are common across many data assimilation techniques; however, I use a Kalman filter lens to describe their importance.

A data assimilation technique typically has two steps:

1. An **update step** which combines the prior state and observations to create the analysis state.

2. A **propagation step** which evolves the state of the system, and the associated uncertainty, between timesteps using a model of the system.

The update step in *sequential* data assimilation techniques, such as the LETKF, combines the prior state and the observations at a single analysis time whilst *continuous* data assimilation techniques use values from across an assimilation window spanning multiple time steps (Fairbairn, 2009). Regardless of the data assimilation framework, the update step in data assimilation requires the prior state and the observations to be represented within the same vector space to enable a meaningful comparison. The prior state is transformed from the state space, $\mathbb{R}^n$, to the observation space, $\mathbb{R}^p$, using an observation operator (or observation function), typically denoted by $\mathbf{H} \in \mathbb{R}^{p \times n}$ (or $\mathbf{h}$), such that

$$\mathbf{y}_k^x = \mathbf{H}\mathbf{x}_k \tag{2.3}$$

where $\mathbf{y}_k^x \in \mathbb{R}^p$ is the model equivalent of the observations such that they should be directly

comparable. The form of the operator depends on the observation type and location, the dimension of the prior state, and whether the variables of the prior state are directly observable (Cooper et al., 2019). If the observed variable is also a model state variable, the observation operator extracts the locations in the model domain (e.g., grid-boxes) corresponding to the observation locations (e.g., Li et al., 2018). In cases where the observed variables and the model state variables differ, more complex observation operators are required (e.g., Emery et al., 2018). Typically, this transformation from state space to model space is imperfect resulting in added uncertainty referred to as representation errors (Janjić et al., 2018).

Once both data sources are represented in the same space, they can be compared and combined. In many methods, including the LETKF, the analysis state is computed as a weighted average of the prior state and the observations (Khodarahmi and Maihami, 2023). The weights are determined by the uncertainties associated with each source which are often assumed to be unbiased and Gaussian-distributed (Carrassi et al., 2018). These uncertainties are typically quantified via error covariance matrices, although this is not common across all data assimilation methods (e.g., particle filters). An error covariance matrix is a symmetric, positive-definite matrix (Golub and Van Loan, 2013) that represents the magnitude of uncertainties for each variable (variances; diagonal elements) and the relationships between uncertainties of different variables (covariances; non-diagonal elements). Many data assimilation techniques require error covariance matrices for both the prior state and the observations, the structures of which are critical to the efficacy of the data assimilation process (e.g., Fowler et al., 2018a; Bannister, 2008a,b). Section 2.3.3 describes how the error covariance matrix of the prior state is estimated using an ensemble of states within ensemble data assimilation techniques, such as the LETKF. The observation errors are often assumed to be uncorrelated such that the observation error covariance matrix is diagonal. This simplification is common in hydrology (e.g., Li et al., 2012) although may not always be appropriate, particularly as satellite observations become more common in hydrology (Schumacher et al., 2016).

In addition to determining the weighting of the observations and prior state within the analysis state, the error covariance matrices enable both observed and non-observed variables

to be updated (Bannister, 2008a; Sawada et al., 2018). This ability is the reason why I investigate the use of data assimilation techniques for post-processing at ungauged locations in this thesis and is discussed in more detail in Section 2.3.2. Non-observed variables may be observable variables (e.g., river discharge) at ungauged locations or parameters that are not easily measured (e.g., the Manning's roughness coefficient).

2.3.2 Kalman filter

The Kalman filter (Kalman, 1960) is a sequential technique that under certain conditions, namely linear dynamics and Gaussian errors, provides a minimum variance estimate of the state of the system. These strict requirements are rarely satisfied in environmental systems, leading to the development of several descendants of the Kalman filter which attempt to allow these requirements to be relaxed (Wang et al., 2023). As discussed in Section 2.3.1, the Kalman filter consists of two steps: an *update step* and a *propagation step*. To aid in describing these steps I use a simple example of the Lorenz 1963 model (Lorenz, 1963) which has three variables (Fig. 2.6).

2.3.2.1 Update Step

The aim of the update step is to produce an improved estimate of the system state, called the analysis, and its associated uncertainty. To achieve this, the update step combines the prior state, $\mathbf{x}_k^f \in \mathbb{R}^n$, and the observations, $\mathbf{y}_k \in \mathbb{R}^{p_k}$. It is assumed in this thesis, that the observations are related to the true state of the system, $\mathbf{x}_k^{true}$, as

$$\mathbf{y}_k = \mathbf{H}\mathbf{x}_k^{true} + \boldsymbol{\epsilon}_k \tag{2.4}$$

where $\boldsymbol{\epsilon}_k$ is a vector of unbiased Gaussian noise with covariance matrix $\mathbf{R}_k \in \mathbb{R}^{p_k \times p_k}$ and $\mathbf{H}_k \in \mathbb{R}^{p_k \times n}$ is a linear observation operator (described in Section 2.3.1). The result of the update step is a weighted mean between the modelled data and the observed data with the influence of each determined by their respective uncertainties (represented by their error covariance matrices; Emery et al., 2020). The weighting is controlled by the Kalman gain

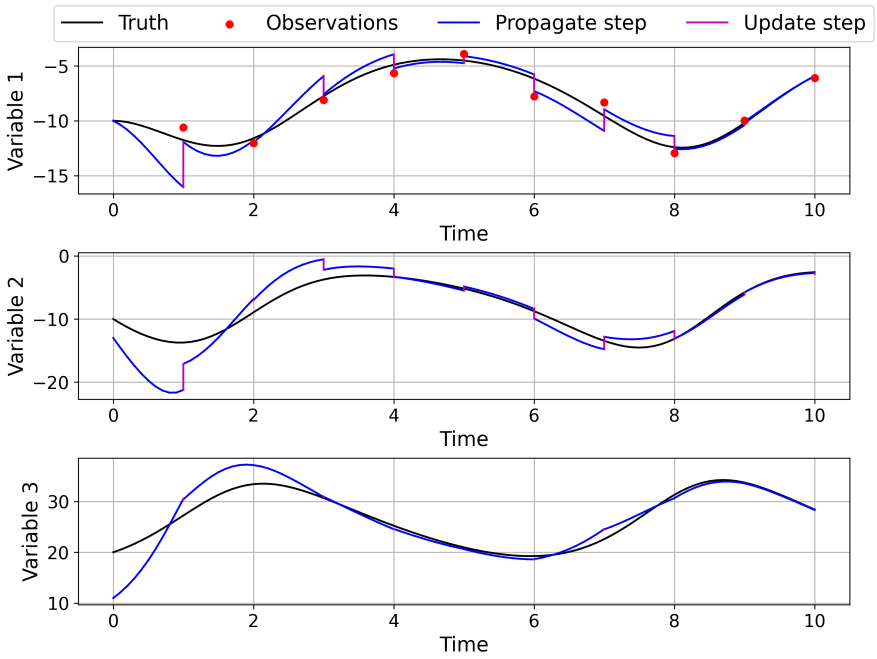

Figure 2.6: Demonstration of the application of a sequential data assimilation method. Only observations for Variable 1 are assimilated, but all variables are updated at each analysis time. All variables are then propagated to the next timestep using the Lorenz 1963 3-variable model. Figure produced using open source code (DARC, 2025).

matrix, $\mathbf{K}_k \in \mathbb{R}^{n \times p_k}$, which is defined as

$$\mathbf{K}_k = \mathbf{P}_k^f \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k^T + \mathbf{R}_k)^{-1} \tag{2.5}$$

where $\mathbf{P}_k^f \in \mathbb{R}^{n \times n}$ is the error covariance matrix of the state before the update step and $\mathbf{R}_k \in \mathbb{R}^{p \times p}$ is the error covariance matrix of the observations (Grewal and Andrews, 2014). The analysis state is then defined as

$$\mathbf{x}_k^a = \mathbf{x}_k^f + \mathbf{K}_k (\mathbf{y}_k - \mathbf{H}_k \mathbf{x}_k^f) \tag{2.6}$$

where the superscripts f and a indicate the state before and after the update step. The analysis error covariance matrix, which represents the uncertainty in the analysis state, is defined as

$$\mathbf{P}_k^a = (\mathbf{I}_k - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_k^f \tag{2.7}$$

where $\mathbf{I}_k \in \mathbb{R}^{n \times n}$ is the identity matrix and all other variables are as defined above.

In Fig. 2.6, the observations of *Variable 1* and the prior state are combined resulting in an analysis state that is closer to the truth (purple lines show the update step). In this example, it is only *Variable 1* that is observed (red dots) but *Variable 2*, and sometimes *Variable 3*, are also updated (note the jumps at each timestep for the middle panel). This is because the error cross-covariances allow the observations of *Variable 1* to influence the analysis state of *Variable 2* and *3*. *Variable 3* is the least impacted and this is due to a smaller cross-covariance with *Variable 1*. This ability to correct non-observed variables is a key benefit of data assimilation, and is why I investigate the use of data assimilation techniques to post-process river discharge forecasts at ungauged locations (Chapters 5 and 6).

These update functions (Eq. (2.6) and Eq. (2.7)) are fundamental to each member of the Kalman filter family of data assimilation methods, including the Ensemble Transform Kalman Filter (ETKF) described in Section 2.3.3.

2.3.2.2 Propagation Step

The propagation step evolves the analysis state, $\mathbf{x}^a \in \mathbb{R}^n$, and the analysis error covariance matrix, $\mathbf{P}^a \in \mathbb{R}^{n \times n}$, of the system between timesteps for which observations are available for assimilation. This is done via a linear model, $\mathbf{M} \in \mathbb{R}^{n \times n}$, such that the state propagates as

$$\mathbf{x}^f_{k+1} = \mathbf{M}_{k \to k+1} \mathbf{x}^a_k \tag{2.8}$$

and the error covariance matrix propagates as,

$$\mathbf{P}^f_{k+1} = \mathbf{M}_{k \to k+1} \mathbf{P}^a_k \mathbf{M}^T_{k \to k+1} \tag{2.9}$$

where the superscripts f and a indicate the prior and analysis, respectively, and the superscript T indicates the transpose of the matrix. In Fig. 2.6 the propagation is shown for each variable by the blue lines.

In Chapters 5 and 6 of this thesis, the propagation step is modified for use with precomputed states.

2.3.3 Ensemble Transform Kalman Filter (ETKF)

The linearity requirement in the propagation step of the Kalman filter (see Section 2.3.2.2) limits its use within environmental systems which are often highly non-linear. To overcome this limitation, several flavours of Kalman filter have been derived (e.g., Extended Kalman Filter (Wishner et al., 1969); Unscented Kalman Filter (Wan and Van Der Merwe, 2000); Ensemble Kalman Filter (Evensen, 1994); Ensemble Square-root Kalman Filters (Tippett et al., 2003)). Ensemble-based Kalman filters are commonly used in the environmental sciences due to their ability to account for non-linearity in a computationally efficient manner (Lahoz and Schneider, 2014). Figure 2.7 illustrates an ensemble Kalman filter framework. In ensemble-based Kalman filters, the full (usually) non-linear model can be used in the propagation step (Eq. (2.8)) to propagate an ensemble of state vectors between analysis times. At each timestep the ensemble members are treated as samples from the distribution of the prior state. The prior state is estimated as the ensemble mean defined as

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)} \in \mathbb{R}^n \tag{2.10}$$

where $\mathbf{x}^{(i)}$ is the i-th ensemble member, N is the ensemble size, and n is the dimension of the system. The state error covariance matrix is estimated from the ensemble covariance matrix

$$\mathbf{P} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}^{(i)} - \overline{\mathbf{x}})(\mathbf{x}^{(i)} - \overline{\mathbf{x}})^T. \tag{2.11}$$

Estimating the error covariance matrix using the ensemble members rather than explicitly propagating the full matrix between timesteps allows the use of a non-linear model without the need for linearization. The computational savings made by not directly propagating the $n \times n$ matrix also outweigh the increase in computational resources needed to propagate N state vectors between timesteps for modest ensemble sizes of the order of those used in operational ensemble forecasting systems (10-100 members; Ott et al., 2004). However, the computation of the error covariance matrix can still be prohibitively expensive for large systems (Hamill, 2002).

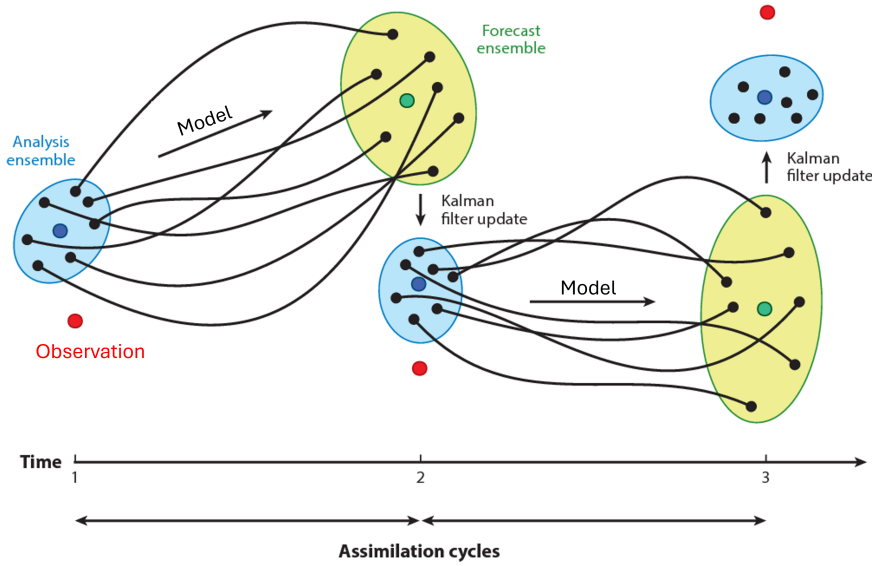

Figure 2.7: Schematic diagram of the ensemble Kalman filter procedure. During the forecast step, an ensemble of simulations (black) is integrated by the nonlinear model until data are available. The forecast ensemble provides an estimate of the model error covariance, which is used in a Kalman filter update step to provide an analysis and a new analysis ensemble. Two assimilation cycles are shown. Generalised for application to any system from Edwards et al. (2015) with permission from Annual Reviews.

Ensemble square-root filters are a type of ensemble-based Kalman filter that only require the computation of the ensemble perturbation matrix, $\mathbf{X} \in \mathbb{R}^{n \times N}$, rather than the full error covariance matrix, making them more computationally efficient (Tippett et al., 2003). The ensemble perturbation matrix is defined such that the i-th column represents the i-th ensemble member's departure from the ensemble mean,

$$\mathbf{X} = \frac{1}{\sqrt{N-1}} \left(\mathbf{x}^{(1)} - \overline{\mathbf{x}} \quad \mathbf{x}^{(2)} - \overline{\mathbf{x}} \quad \cdots \quad \mathbf{x}^{(N)} - \overline{\mathbf{x}} \right). \tag{2.12}$$

The perturbation matrix is a square-root of the ensemble covariance matrix such that

$$\mathbf{P} = \mathbf{X}\mathbf{X}^T. \tag{2.13}$$

The specific ensemble-based Kalman filter used in this thesis is the Local Ensemble Transform Kalman Filter (LETKF; Hunt et al., 2007) which is a localised variant of the Ensemble Transform Kalman Filter (ETKF; Bishop et al., 2001) which in turn is a type

of Ensemble square-root filter. In the ETKF, the update step is performed partially in the N-dimensional ensemble space rather than the n dimensional state space (where typically $n >> N$) by defining the Kalman filter in terms of a so-called square root transform matrix, $\mathbf{T} \in \mathbb{R}^{N \times N}$ (see Eqs. (2.17) - (2.20)). This further reduces the computational demands of the matrix operations due to the smaller matrix sizes. The update equations of the ETKF are

$$\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^f + \mathbf{K}(\mathbf{y} - \mathbf{H}\overline{\mathbf{x}}), \tag{2.14}$$

and

$$\mathbf{X}^a = \mathbf{X}^f \mathbf{T} \tag{2.15}$$

where the superscripts f and a indicate before and after the update step, $\overline{\mathbf{x}}$ and $\mathbf{X}$ are the ensemble mean and perturbation matrix, respectively, $\mathbf{K}$ is the Kalman gain matrix, $\mathbf{y}$ is the observation vector, and $\mathbf{H}$ is the observation operator. Although the Kalman gain serves the same purpose as in the original Kalman filter, weighting the contributions of the prior state and observations based on their respective uncertainties (Eq. (2.5)), its calculation is performed differently, as will be shown below.

Ensemble square-root filters, such as the ETKF, do not require the addition of stochastic noise to the observations as is done in the EnKF (Houtekamer and Mitchell, 1998; Burgers et al., 1998; Tippett et al., 2003). Instead, the ensemble members are transformed such that the analysis error covariance matrix satisfies Eq. (2.7),

$$\mathbf{P}^a = \mathbf{X}^a \mathbf{X}^{aT} = \left(\mathbf{I}^{n \times n} - \mathbf{K}\mathbf{H}\right) \mathbf{X}^f \mathbf{X}^{fT}. \tag{2.16}$$

Although the choice of the matrix $\mathbf{T}$ is non-unique, to preserve the ensemble mean and ensure an unbiased state estimation it must satisfy,

$$\mathbf{T}\mathbf{T}^T = (\mathbf{I} - \mathbf{Y}^T \mathbf{D}^{-1} \mathbf{Y}) \tag{2.17}$$

where $\mathbf{Y} = \mathbf{H}\mathbf{X}^f$ is the model-observation ensemble perturbation matrix and $\mathbf{D}$ satisfies

$$\mathbf{D} = (\mathbf{Y}\mathbf{Y}^T + \mathbf{R}) \tag{2.18}$$

where $\mathbf{R}$ is the error covariance matrix of the observations (Livings et al., 2008; Nerger et al., 2012). In order to ensure numerical stability, it is advised to normalise the model-observation ensemble perturbation by multiplying with the square-root of the observation error covariance matrix (Livings et al., 2008). The elements of the resulting normalised matrix, $\tilde{\mathbf{Y}} = \mathbf{Y}\mathbf{R}^{1/2}$, are of similar orders resulting in a more well-conditioned matrix. Using this nomalisation, and the Sherman-Morrison-Woodbury formula (Sherman and Morrison, 1950) to manipulate Eq. (2.18), this condition on $\mathbf{T}$ can be expressed as

$$\mathbf{T}\mathbf{T}^T = (\mathbf{I} + \tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}})^{-1}. \tag{2.19}$$

Finally, following from this constraint, and the definition given in Eq. (2.5), the Kalman gain can be expressed as

$$\mathbf{K} = \mathbf{X}^f(\mathbf{T}\mathbf{T}^T)\tilde{\mathbf{Y}}^T\mathbf{R}^{-1/2}. \tag{2.20}$$

Despite its deterministic nature, the ETKF can still suffer due to sampling errors caused by small ensemble size in the form of spurious correlations (Ehrendorfer, 2007). The LETKF uses localisation to minimise the effect of the spurious correlations as will be discussed in the next section.

2.3.4 Localisation and the Local Ensemble Transform Kalman Filter (LETKF)

In this section, I provide the necessary equations for the implementation of the LETKF, the derivation of which is available for the interested reader in Hunt et al. (2007). As the propagation step has already been discussed in Section 2.3.2, here I keep the discussion to the update step only. Computational resources restrict possible ensemble sizes to be

much smaller than the size of the system state (i.e., $n >> N$). The ensemble can therefore only span N dimensions of the n-dimensional state space potentially leading to an incorrect estimation of the state error covariances, represented by the perturbation matrix in the ETKF (Ehrendorfer, 2007). Misspecification of the covariances can lead to spurious correlations between variables (Evensen, 2009). In the update step these spurious correlations can result in observations incorrectly impacting state variables with which they have no physical connection. Localisation is a technique used to restrict the impact of observations to variables for which the correlations are presumed to be trustworthy (Anderson, 2012; Petrie and Dance, 2010; Ott et al., 2004; Hamill, 2002).

In the Local Ensemble Transform Kalman Filter (LETKF, Hunt et al., 2007), localisation is applied in the observation space by multiplying the observation error covariance matrix, $\mathbf{R} \in \mathbb{R}^{p \times p}$, by a localisation matrix, $\boldsymbol{\rho} \in \mathbb{R}^{p \times p}$, such that

$$\mathbf{R}_{local}^{-1} = \boldsymbol{\rho} \circ \mathbf{R}^{-1} \tag{2.21}$$

where $\mathbf{R}_{local} \in \mathbb{R}^{p_k \times p_k}$ is the localised observation-error covariance matrix, and the symbol $\circ$ indicates an element-wise matrix multiplication called the Schur product (Golub and Van Loan, 2013). The LETKF updates each variable of the system state separately and the localisation matrix will differ for each variable. Each element of the localisation matrix is a weight that, if less than one, increases the effective uncertainty of an observation. A weight of zero enforces an infinite effective uncertainty, and thus the observation has no impact in the update equations (Hunt et al., 2007). For in-situ observations, the weights are often determined by the distance from the observation (Miyoshi et al., 2007).

Once the localisation weights have been determined and the observation-error covariance matrix has been weighted, the element of the ensemble mean and the row of the perturbation matrix corresponding to the variable being updated are identified. They are then modified using Eqs. (2.14) and (2.15) (Hunt et al., 2007). The matrix $\mathbf{T}$ is calculated using an

eigendecomposition of the constraint given in Eq. (2.19), such that

$$\mathbf{T}\mathbf{T}^T = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T \tag{2.22}$$

where each column of $\mathbf{V}$ represents an eigenvector and $\mathbf{\Sigma}$ is the corresponding diagonal matrix of eigenvalues. The matrix $\mathbf{T}$ can therefore be calculated as

$$\mathbf{T} = \mathbf{V}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{V}^T. \tag{2.23}$$

The i-th analysis ensemble member is then calculated as

$$\mathbf{x}^{(i)a} = \overline{\mathbf{x}}^a + \mathbf{X}^a[:, i] \tag{2.24}$$

where $\mathbf{X}^a[:, i]$ is the i-th column of the analysis perturbation matrix. The LETKF is often used in hydrology as will be discussed in Section 2.3.5

2.3.5 LETKF and hydrology

Data assimilation in hydrological sciences is newer than the application of the mathematical technique in the atmospheric sciences (e.g., Liu and Gupta, 2007; Sun et al., 2015). However, ensemble Kalman filter techniques are often used in hydrological forecasting systems due to their efficiency and ability to account for non-linear behaviours (Sun et al., 2016). The LETKF is commonly chosen for larger domains such as regional or global applications (e.g., Revel et al., 2023, 2021; Li et al., 2020). Previous studies have focused on different hydrological variables including snow cover (e.g., Micheletty et al., 2022), soil moisture (e.g., Jadidoleslam et al., 2021; Baugh et al., 2020), flood extent (e.g., Dasgupta et al., 2021; Nguyen et al., 2022), water level (e.g., Zijl et al., 2015), and river discharge (e.g., Wang et al., 2018; Oubanas et al., 2018). For hydrological variables there are some additional considerations for the application of data assimilation.

For river discharge in particular the choice of localisation function is important. Atmo-

spheric variables often show spatial correlations with neighbouring locations in all directions (Jang and Matteson, 2023). On the other hand the spatial correlations of hydrological variables are often directed by preferential flow which, at the catchment scale, is along the river network (Blöschl, 2022). Locations that are geographically close but in different catchments (i.e., the water drains into rivers and/or body of water) may show limited correlation whereas locations on the same river stretch will be more strongly correlated. Several data assimilation studies have shown better results when applying localisation along the river network rather than using Euclidean distance (El Gharamti et al., 2021a; Emery et al., 2018; Revel et al., 2019; García-Pintado et al., 2015; Khaniya et al., 2022). Figure 2.8 shows how an along the river network localisation function may look spatially for different localisation length scales. Notably, 1) the length of river for which the localisation weight is non-zero is not constant due to the different upstream drainage areas, and 2) observations cannot impact areas outside of the catchment. However, despite the acceptance that the distance should be calculated along the river network the treatment of upstream and downstream locations varies between studies (García-Pintado et al., 2013, 2015; Revel et al., 2019; Mason et al., 2020; El Gharamti et al., 2021a). Additionally, previous studies use a tuning process to define the localisation length scale or attempt to define a physically informed length scale determined by lag-times of the river discharge (Revel et al., 2019). This means the localisation length scale is not easily transferable between applications.

The LETKF estimates the error covariance matrix from the ensemble members and assumes a Gaussian distribution of the errors (Emery et al., 2020). In general river discharge has a non-Gaussian distribution and, particularly during precipitation events, can demonstrate tipping point behaviour often leading to the ensemble members showing a skewed distribution (Blöschl, 2022, see for example Fig. 2.2). Therefore, the error covariance matrix estimated from a skewed ensemble will break the Gaussian error assumption of the Kalman filter making the update sub-optimal. For larger river discharge values this skewedness can also lead to an underestimation in the uncertainty of the prior state. River discharge distributions can be transformed to be more Gaussian like via, for example, the log-transform or the Normal Quantile Transform (used and described in Chapter 4), which has been shown to improve

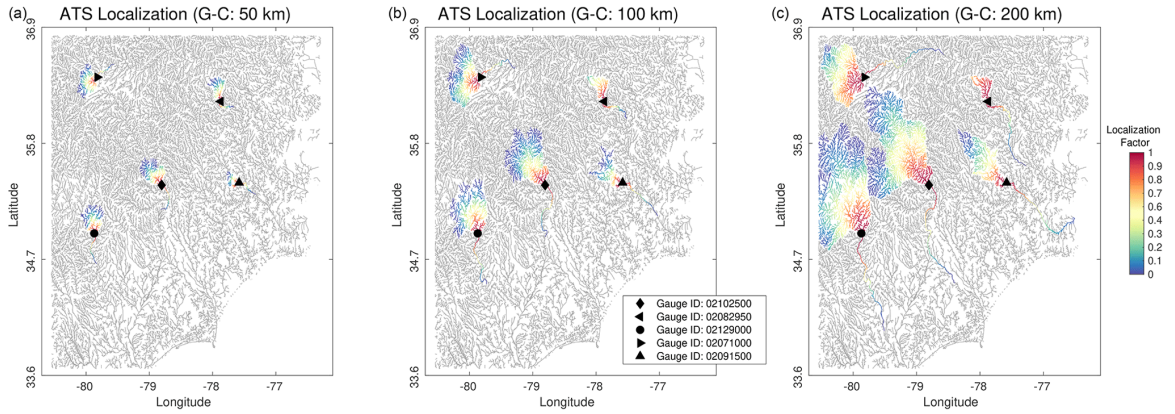

Figure 2.8: Demonstration of the along-the-stream (ATS) localization function in the model domain for three different effective localization radii: 50 km (a), 100 km (b), and (c) 200 km. The colour shows the localization weighting for five different stream gauges based on the along-the-stream distance and the Gaspari–Cohn correlation function. Reproduced from (El Gharamti et al., 2021a) under the CC BY 4.0 license.

the analysis state in previous studies (Clark et al., 2008; Nguyen et al., 2023; Musuuza et al., 2020). Alternatively, inflation techniques can be used to ensure that filter divergence does not occur due to non-Gaussian errors. El Gharamti et al. (2021a) show that posterior inflation, where the ensemble spread is inflated after the update step, is particularly suitable for accounting for non-Gaussian errors. However, they also found that the error due to the hydrological model dominates over the error due to non-Gaussianity. Prior-inflation, where the ensemble spread is inflated before the update step, was shown to be more useful for accounting for model error. In Chapter 5 I use prior-inflation to mitigate filter divergence.

Joint estimation of the state and model parameters is commonly performed in hydrological data assimilation due to the limited knowledge regarding the bathymetry and morphology of rivers around the world (Pathiraja et al., 2016; Pedinotti et al., 2014). State augmentation is a technique that allows for the simultaneous estimation of the system state and model parameters by augmenting the model parameters of interest to the end of the state vector (Evensen et al., 2022; Smith et al., 2013; Dee, 2005). A more detailed explanation of state augmentation is given in Chapter 5, where the technique is used to estimate the errors at ungauged locations.

One of the key attractions of Kalman Filters, and data assimilation techniques in general, is the ability to adapt to different observation types. Data assimilation methods can use any

type of observations for which an observation operator can be defined and the observation uncertainty can be estimated. This is beneficial as, due to the sparse distribution of in-situ river discharge observations (Krabbenhoft et al., 2022) the use of non-traditional observations such as crowd-sourced observations and Earth Observation measurements is increasing within the hydrological community (e.g., Mazzoleni et al., 2017; Vandaele et al., 2021; Musuuza et al., 2023).

2.3.6 Operational considerations for hydrological data assimilation

Despite several studies demonstrating the benefits of data assimilation for hydrological forecasting, the operational implementation of hydrological data assimilation is difficult (Wu et al., 2020). This is partially due to a lack of real-time observations (WMO, 2024a). Whereas significant investment has been made to facilitate the reliable exchange of meteorological data in real-time (WMO, 2023), less investment has been made for hydrological data. Whilst attempts have been made, they are typically for very few stations on the major rivers or for historic data (e.g., Maurer, 2005; Färber et al., 2024; Kratzert et al., 2023). Particularly at the large-scale where large regions of the domain may by unobserved, the lack of reliable observations is one reason why data assimilation is not commonly implemented.

The implementation of a data assimilation approach would also require significant changes to the forecasting system. First, the forecasting schedule would need to be altered to allow additional time for the data assimilation module to run. As many large-scale systems already have a tight schedule to deliver timely forecasts (e.g., Matthews et al., 2025c) this may not be feasible. Second, many distributed large-scale systems use a model climatology to calculate flood thresholds for each grid-box, from which the extremeness of the river discharge can be determined (Emerton et al., 2016). The addition of a data assimilation module could change the forecast distribution such that it is no longer compatible with the thresholds. In post-processing, this is often overcome by having a separate 'post-processing product' along side the raw (non-postprocessed) forecasts (e.g., Matthews et al., 2025b). Since data assimilation impacts the raw forecasts it would not be practical to have two forecast products (i.e., one

with and one without data assimilation). Additionally, the observation impact time (i.e., the length of time for which data-assimilation improves the forecast skill) has been shown to be limited to short lead-times (see Section 2.1.2; Valdez et al., 2022). However, in this thesis I am interested in improving medium-range forecasts.

On the other hand, many post-processing methods do not require near real-time observations (Section 2.2.2), the output from the hydrological model is not intrinsically changed, and post-processing has been shown to improve forecasts at longer lead-times than data assimilation (Valdez et al., 2022). Therefore, in this thesis I combine the strengths of data assimilation and post-processing to improve the skill of the forecasts at ungauged locations. Specifically, I use data assimilation techniques, which can spread information from gauged to ungauged locations, to spread the correction estimated using a post-processing method. The application of data assimilation to a post-processing environment (i.e., with a precomputed forecast) requires some modification and assumptions to be made (see Chapter 5) as the hydrological model cannot be used between timesteps, as is typically done in data assimilation.

2.4 Creating Actionable Forecasts

River discharge forecasts are a critical tool to support disaster preparedness and water resource management (Adams et al., 2024). However, numerous studies have highlighted a 'usability gap' between the production of forecasts and their actual use in decision-making contexts (McNie, 2007). In this thesis, decision-making typically refers to the choice around whether flood mitigation actions should be taken and what actions are most appropriate. For example, a forecast user may need to decide if flood defenses need to be erected or if a town should be evacuated. Bridging the usability gap requires technical advances in hydrological and meteorological modelling *and* consideration of the forecast design, dissemination, and interpretation. In this section I discuss what makes a 'good' forecast (Section 2.4.1), the use of a co-production as a framework for making forecasts more useful (Section 2.4.2), and the challenges of making useful large-scale forecasts (Section 2.4.3).

2.4.1 Forecast skill and forecast value

Forecast "skill" traditionally refers to the statistical accuracy of a prediction compared to a reference (also known as a benchmark), such as climatology or persistence (Pappenberger et al., 2015b). Traditionally, a change to a forecasting system is deemed beneficial if a skill metric of choice is improved (Murphy, 1993; Bauer et al., 2015). There are many metrics that aim to quantify the skill of the forecast depending on what factor is most important (e.g., timing, magnitude, etc.; Cloke and Pappenberger, 2008). In hydrology common metrics include the Nash Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970), the Kling-Gupta Efficiency (KGE; Gupta et al., 2009; Kling et al., 2012), the Brier score (Brier, 1950), the Continuous Ranked Probability Skill Score (CRPS; Hersbach, 2000), contingency tables scores (Gold et al., 2020), and many more (Jolliffe and Stephenson, 2012). Additionally, graphical measures of skill such as the Relative Operating Characteristics (ROC) diagram (Kharin and Zwiers, 2003) and the rank histogram (Hamill, 2001) are commonly used. I use many of these metrics in this thesis (see Chapters 4, 5, and 6). These metrics are useful at identifying patterns in forecast behaviour and identifying situations when forecasts are most accurate. However, the magnitude of many of these metrics may not directly reflect the usefulness of a forecast (Giuliani et al., 2020). In contrast, forecast value depends on the extent to which information leads to better decisions and outcomes (Katz and Murphy, 1997). A forecast might be technically accurate but go unused if it is not trusted, untimely, or hard to interpret. Forecast value is highly contextual and emerges from the interaction between forecast characteristics, the decision environment, and the user's capacity to act (Lemos and Rood, 2010). This disconnect between skill and value has motivated a shift in emphasis from purely improving forecast accuracy to improving forecast usefulness. This involves understanding user needs, the context in which decisions are made, and barriers to forecast use (Lemos et al., 2012; Golding, 2022).

2.4.2 Co-production framework

There has been a growing interest in *co-production* within the hydrological sciences to ensure the products are relevant, usable, and actionable (e.g., Singletary and Sterle, 2020; Grainger et al., 2021; Jagannathan et al., 2021; Kliskey et al., 2023). Co-production is the joint development of knowledge, services, or systems by all relevant stakeholders, including researchers, software engineers, policymakers, data providers, and users of the forecast (Carter et al., 2019). Rather than treating forecast users as simply recipients of data, in co-production, users are partners in the creation of the forecasts (Kats, 2024). The definition of co-production is broad encompassing a range of viewpoints (Bremer and Meisch, 2017). For clarity, here we discuss the *normative* definition of co-production (e.g., 'How can developers and users work together to create better river discharge forecasts?') rather than the *descriptive* definition (e.g., 'How do the relationships between developers and users influence the design of river discharge forecasts?'). More specifically, in this thesis I take the *normative iterative interaction* definition of co-production as the "iterative interaction of science providers and users along an interdisciplinary research process designed to produce more usable ... information" (Bremer and Meisch, 2017).

To be effective the co-production process must begin by identifying all relevant stakeholders and building partnerships between the people and organisations (Visman et al., 2022). The co-production process consists of several iterative stages, including the co-exploration of needs, co-design of solutions, co-delivery of products, and co-evaluation of outcomes (Fig. 2.9; Carter et al., 2019). These stages help align scientific outputs with user priorities whilst maintaining technical feasibility and emphasising the usability of forecasts as the key objective (Meadow et al., 2015). In the co-exploration phase, all stakeholders jointly identify barriers and constraints, as well as information needs (Daniels et al., 2020). Solutions are then developed in the co-design phase. This could involve, for example, defining forecast specifications, designing visualisations, or developing a prototype (Hirons et al., 2021). The selected developments are then jointly implemented in the co-delivery phase. Depending on the aim of the project this may include establishing data acquisition or dissemination sys-

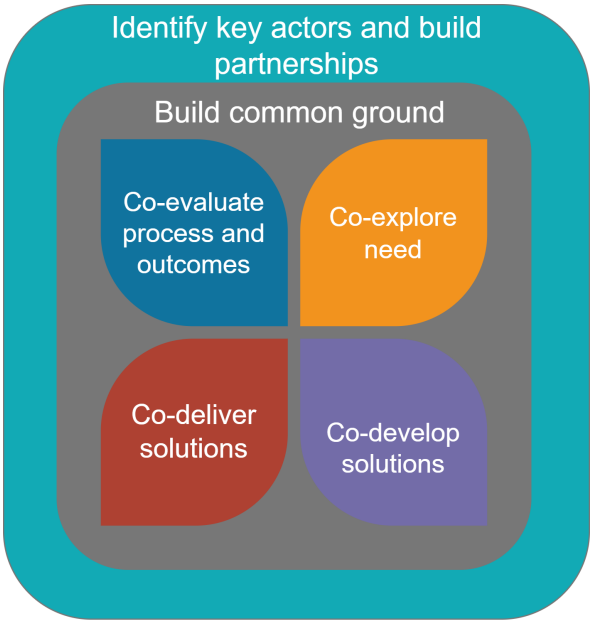

Figure 2.9: Core components of the co-production process. Modified from Carter et al. (2019).

tems, integrating forecasts into decision-making procedures, or conducting training sessions (Cantone et al., 2023). In the co-evaluation phase, all stakeholders evaluate the co-production process (to enable lessons to be learnt in the next cycle) as well as the usability of the forecasts (Gebhardt and Kuhlicke, 2024; Hyytinen et al., 2019). Both aims of the co-evaluation phase can be achieved via workshops, questionnaires or interviews with stakeholders. Forecasts created using the co-production process are more likely to contain the information most relevant to the user, as well as being more accessible in terms of format, forecast time frame, and language (Howarth et al., 2019; Arnott et al., 2020; Vincent et al., 2018; Speight et al., 2025).

In hydrology, co-production is often conducted within climate adaptation projects (e.g., Pretorius et al., 2019) or for the development of local forecasting systems (e.g., Fujisaki-Manome et al., 2022). Alternatively, some studies have used co-production successfully to develop early action triggers that use the output of large-scale forecasting system. For example, forecast-based financing programmes, which aim to release funds before an event occurs, are usually established via a co-production project (e.g., Robbins et al., 2022; Coughlan de Perez et al., 2015). However, fewer studies have explored how co-production principles can be applied within existing, large-scale operational forecasting systems, where institu-

tional complexity and scale pose challenges to stakeholder engagement and customization (e.g., Cantone et al., 2023; Visman et al., 2022). In this thesis, I attempt to incorporate the co-production practices into the development process of the post-processed forecasts of an established large-scale forecasting system (Chapter 7).

2.4.3 Challenges of co-production for large-scale systems

Co-production has shown promise at the local level but its implementation within large-scale systems presents several challenges (Landström et al., 2024; Lemos et al., 2018). One of the primary difficulties lies in the number and diversity of stakeholders involved (Vincent et al., 2018). Stakeholders may have conflicting information needs—for example, a flood forecaster may require different information from a reservoir manager. Whilst conflicting needs are a challenge for local co-production, with more stakeholders it becomes even more challenging to address the needs of each individual or organisation (Grainger et al., 2021). Large-scale forecast systems are typically designed for broad consistency and coverage (e.g., Matthews et al., 2025c), whereas co-production is intended to be flexible and decision-process driven (Vincent et al., 2018).

In transnational or multi-sectoral contexts, these challenges become even more pronounced. Engaging meaningfully with the full range of stakeholders is difficult when decision contexts, institutional priorities, and technical capacities differ significantly across countries or sectors (Page and Dilling, 2019; Lemos et al., 2018). Communication also becomes more complex, particularly when multiple languages, cultural perspectives, and backgrounds must be navigated (Vincent et al., 2021; Chambers et al., 2021).

Logistical challenges also increase with system scale. Organising inclusive meetings — especially in-person—requires more coordination, time, and financial resources (Arnott et al., 2020). As more actors are involved, existing inequalities, such as those related to gender, language, or institutional power, may not be identified or may even be exacerbated (Chambers et al., 2021). Without sufficient resources and long-term commitment, partnerships may remain superficial, weakening the co-production process and discouraging full stakeholder

engagement (Arnott et al., 2020). Moreover, many user communities may lack the capacity to fully participate unless appropriately supported (van der Graaf et al., 2021).

Large organisations may also resist the iterative and often nonlinear nature of co-production. Institutional structures, strict mandates, short funding cycles, or performance metrics focused on outputs rather than engagement, can impact the efficacy of co-production (Arnott et al., 2020; Käyhkö et al., 2025). Additionally, co-production requires a shift in professional norms for scientists and forecasters, requiring them to become facilitators (Lemos et al., 2018). This transition will require training and institutional support (Arnott et al., 2020).

In Chapter 7, I use co-production practices to identify barriers to the use of the EFAS post-processed forecasts, and to design solutions to overcome these barriers. EFAS is a large-scale flood forecasting system and as such many of the challenges discussed in this section were encountered. I discuss these challenges further in Chapter 8.

2.5 Summary of Chapter 2

In this chapter, I have discussed the sources of uncertainty present in river discharge data, the aims and common methods of post-processing river discharge forecasts, data assimilation for hydrological applications, and co-production in the context of a large-scale forecasting system. *Objectives 1-3* of this thesis are regarding the evaluation and development of post-processing methods. For *Objective 1*, I will conduct an evaluation of the operational at-gauge post-processing method used in EFAS (Chapter 4). This post-processing method uses a combination of the Ensemble Model Output Statistics (EMOS) method and the Model Conditional Processor (MCP), which are both described in Section 2.2.2. The Kalman Filter, described in Section 2.3.2, is also used in the operational at-gauge post-processing method to combine the output of the EMOS and MCP methods. In Section 2.2.3, I review some of the post-processing methods that have been developed for ungauged locations. These methods tend to assume static relationships between the errors at gauged and ungauged locations, or are computationally expensive. Therefore, *Objectives 2* and *3* are related to the development

of a new post-processing method for ungauged locations (Chapters 5 and 6). To address both objectives, I use a data assimilation method, namely the Local Ensemble Transform Kalman Filter (LETKF), to propagate error information from gauged to ungauged locations. The LETKF is described in Section 2.3.4 and some of the considerations necessary for its application to hydrological systems are discussed in Section 2.3.5. *Objective 4* relates to understanding the barriers to using the EFAS post-processed forecasts. Section 2.4.2 motivates the organisation of a co-production workshop between users and developers of the forecasts in order to address this objective (Chapter 7). Section 2.4.3 discusses some of the challenges of co-production at a large-scale. These challenges are discussed in Chapter 8.

Chapter 3

European Flood Awareness System (EFAS)

This chapter provides a system description of the European Flood Awareness System (EFAS) which is used throughout this thesis. The operational forecasting chain used to create the EFAS medium-range forecasts used in this thesis, is described along with the input data required. The research presented in this thesis was conducted alongside the EFAS operational system to ensure relevance to real-world forecasting workflows. Therefore, this chapter also provides context to the research in terms of the motivation to establish the system, the organisational structure, the computational constraints of the operational system, the catalogue of forecast products (including the post-processed forecasts focused on in this thesis), the forecast dissemination, and user engagement processes. These factors were considered throughout the research presented in this thesis. This chapter is an excerpt from:

G. Matthews, C. Baugh, C. Barnard, C. C. De Wiart, J. Colonese, D. Decremer, S. Grimaldi, E. Hansford, C. Mazzetti, K. O'Regan, F. Pappenberger, A. Ramos, P. Salamon, D. Tasev, and C. Prudhomme. Chapter 14 - On the operational implementation of the european flood awareness system (EFAS). In T. E. Adams, C. Gangodagamage, and T. C. Pagano, editors, Flood Forecasting (Second Edition), pages 251–298. Academic Press, second edition, 2025. ISBN 978-0-443-14009-9. doi: https://doi.org/10.1016/B978-0-443-14009-9.00005-5.

The book chapter was submitted in April 2024; however, no major changes have been made between then and the submission of this thesis. I have removed sections which do not discuss

the medium-range forecasts at all. I note that this chapter was written after Chapters 4 and 7 were published. Chapter 4 is referenced as Matthews et al. (2022) and the post-processed forecast products discussed in Section 3.6.2 are the new products resulting from the workshop discussed in Chapter 7.

In Section 3.11 I summarise the chapter and highlight the key sections for the remainder of the thesis.

Abstract

Within Europe severe flood events are often transboundary requiring collaboration between several national and regional authorities who are responsible for water and flood management. In these situations, inconsistent or erroneous information may arise, for example, from incomplete communication between authorities or different forecasts resulting in diverging assessments of the flood event. This chapter discusses the European Flood Awareness System (EFAS) which operates on a pan-European scale to provide coherent flood forecasts and related information at a range of timescales. EFAS serves as an independent reference forecast for most of the hydrological services responsible for flood forecasting in Europe as well as the European Civil Protection. The forecasts produced by EFAS are discussed along with the systematic approach used to ensure the information is beneficial to users. This approach includes an organisational structure that allows experts to focus on their specializations, a modelling chain that takes advantage of the latest developments in hydrometeorological and computational sciences, and an increased focus on user engagement. Since the system is constantly evolving, a snapshot of EFAS as of April 2024 is provided. Alongside an overview of the managerial and technical aspects of the system, case studies are used to illustrate the effectiveness of this system in providing early warning of the potential for flooding to authorities.

3.1 Introduction

In Europe more than 40 rivers cross at least one national border with the Danube, which is shared by 18 countries, being the most transnational river in Europe. This means that in the event of a flood different authorities involved in water resource management, civil protection, and the organisation of aid must, ideally, communicate, share data and information, and take concerted actions to reduce the impact of the flooding along the course of the river. In these situations, flood risk management can become challenging as inconsistent information may arise from incomplete communication between authorities, differing results from different forecasting models and subsequent assessment of the ongoing and forecasted flood event, or simply misunderstandings due to language barriers. Inconsistent information can introduce uncertainties and errors into the assessment of the ongoing and upcoming situation leading to incoherent or uncoordinated decision making and actions across the chain of responsibilities, which can be counterproductive to reducing the impacts of the flood event (Demeritt et al., 2007; Zeitoun et al., 2013).

To avoid discrepancies in information content, clear communication channels and agreed protocols for exchange of data and information are necessary and many countries have agreed bi-lateral protocols accordingly (Krengel et al., 2018). However, even for countries with bi-lateral agreements different models and forecasting systems often exist for different countries or even administrative units (Alfieri et al., 2012). This can lead to a lack of consistent information that is available to all parties involved which can make assessing the situation complicated, particularly for those not covered by bi-lateral agreements with upstream countries or those responsible for the management of European aid.

Significant flooding across Europe in the early 2000s highlighted the need for improvements in flood risk and crisis management. Post event analysis lead the European Commission to start, amongst other important initiatives, the development of the European Flood Awareness System, originally named European Flood Alert System (EFAS; Bartholmes et al., 2009; Burek et al., 2011; Commission, 2002; Thielen et al., 2009) and based on the initial research activities of the European Flood Forecasting System project (EFFS De Roo et al., 2003;

Commission, 2005; Gouweleeuw et al., 2005; Pappenberger et al., 2005). The objectives of EFAS are to provide pan-European medium-range streamflow forecasts and early warning information, in particular for large transnational river basins, in direct support to the national forecasting services, as well as harmonized information on possible high-impact flooding to the Emergency Response Coordination Centre (ERCC) of the European Commission. In the case of major flood events, EFAS contributes to the better protection of the citizens, the environment, property, and cultural heritage.

From 2003 to 2012, EFAS was developed and tested at the Joint Research Centre (JRC), the European Commission's in house science service, in close collaboration with national hydrological and meteorological services across Europe, the European Civil Protection through the ERCC, and other research institutes (Buizza et al., 2009; Cloke et al., 2009; Kalas et al., 2008; Pappenberger et al., 2008, 2011, 2012; Ramos et al., 2009, 2007; Raynaud et al., 2015; Younis et al., 2008). The European Commission's Communication "Towards Stronger European Union Disaster Response" adopted and endorsed by the Council in 2010 (Commission, 2010) underpins the importance of strengthening concerted actions for natural disasters including floods, which are amongst the costliest natural disasters in the EU. Partially in response to this, EFAS became part of the European Union's Copernicus Emergency Management Service (CEMS) in 2011 and in 2012 it was transferred from research to operational service. Figure 3.1 shows some key developments in the history of EFAS.

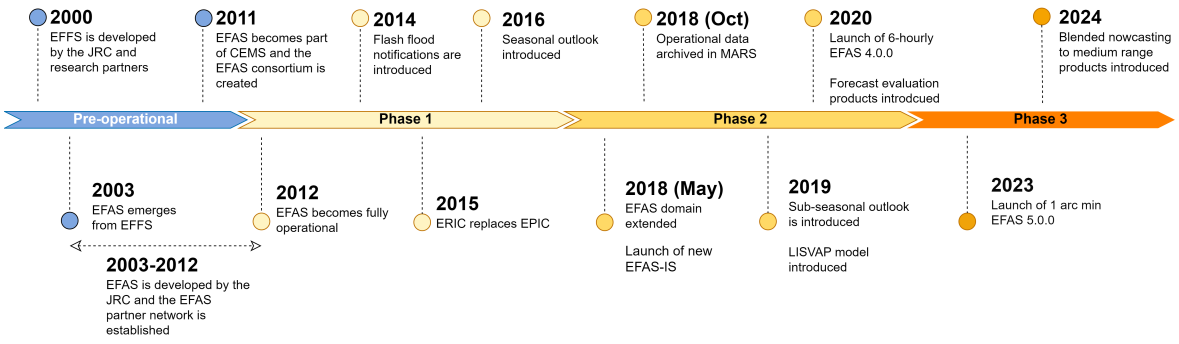

Figure 3.1: Timeline of EFAS developmental and operational history.

The importance of pan-European early warning systems in complementing national information systems was further highlighted in 2013 with the decision on a Union Civil Protection

Mechanism (UCPM), where it is stated that the European Commission "shall contribute to the development and better integration of transnational detection and early warning and alert systems of European interest in order to enable a rapid response, and to promote the interlinkage between national early warning and alert systems and their linkage to the ERCC and the CECIS" (Commission, 2013). CEMS, including its early warning systems component for better emergency management, was endorsed in Regulation (Commission, 2014). As a result, over the past decade EFAS has become increasingly integrated into national and European flood risk management (e.g., the Danube Flood Risk Management Plan ICPDR, 2021).

EFAS has also evolved substantially over the past 10 years, incorporating new scientific and computational advances, and responding to user needs. For example, since the first operational forecasts, the collection of EFAS forecast products has grown to now include 46 products updated in real-time and 14 additional products to help interpretation. The medium-range flood forecasts are still a key product within EFAS, however, products expanding the forecast horizon have been added. In 2014, the first flash flood notifications were introduced using the European Precipitation Index based on Climatology (EPIC; Alfieri and Thielen, 2015). These were replaced by the European Rainfall Index based on Climatology (ERIC; Raynaud et al., 2015) in 2016, providing forecasts up to 5 days. The forecast horizon was extended to longer lead-times via the seasonal (Arnal et al., 2018) and sub-seasonal (Wetterhall and di Giuseppe, 2018) outlooks which were introduced in 2016 and 2019, respectively. From 2017 onwards forecast products based on flood impact are also included in the EFAS portfolio (Alfieri et al., 2014b; Dottori et al., 2017). Finally, since 2020, forecast evaluation products are available alongside the forecasts to aid users in making informed decisions. Throughout the years, funding from the EU, for example via the UCPM, has aided the inclusion of new experimental products, such as the flash flood prototype, TAMIR, and its predecessor, ERICHA (Berenguer et al., 2011; Park et al., 2019), into the EFAS workflow.

In parallel to forecast products, the range of services provided by EFAS has also expanded. In 2018 the EFAS-Information System (EFAS-IS) was updated to make it easier for users to navigate and to allow feedback to be collected. The raw hydrological forecast data (the forecasts before post-processing has been performed to create the visual forecast products)

have been made available to end-users via the Common Data Store Engine (CDS-E) and dedicated Data Store (DS) since 2019, with datasets containing multidecadal simulations and reforecasts added more recently. These datasets are also archived in the ECMWF Meteorological Archival and Retrieval System (MARS) since 2018.

Developments to EFAS are largely guided by feedback from users as well as scientific and technological advances, including improved hydrological modelling, more efficient algorithms, and advances in computing and IT infrastructure. In 2018, the EFAS domain was extended to cover all of mainland Europe (see Fig. 3.2). Another major upgrade to the system was released in 2020 which, among other improvements, increased the temporal resolution of the medium-range forecasts from daily to 6-hourly. In 2023, EFAS version 5.0 was released increasing the spatial resolution of the EFAS forecasts from 5 km to 1 arc min (approximately 1.3 km). The system domain also increased (see Fig. 3.2).

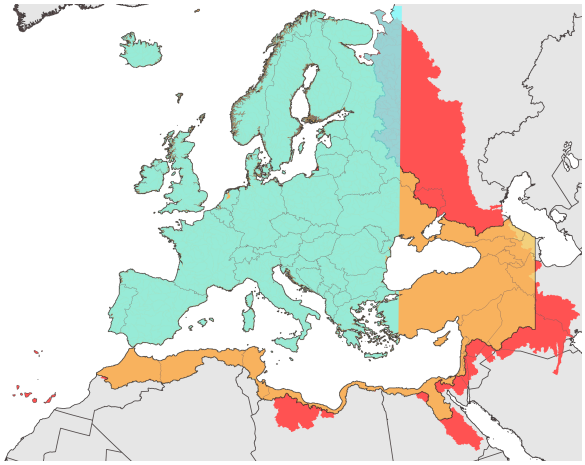

Figure 3.2: Map showing the regions of the EFAS domain modelled by LISFLOOD until Spring 2018 (cyan), until September 2023 (orange), and as of April 2024 (red).

As of April 2024, 180 hydrological and civil protection services in Europe are part of the EFAS network. To ensure that EFAS does not interfere in the 'one voice warning mandate' postulated by the World Meteorological Organisation, EFAS flash flood and medium range forecast products are not publicly available in real-time. Instead, national and EU authorities mandated to inform or act on ongoing or upcoming flood situations can get access to EFAS after having signed a 'conditions of access' contract. This regulates the dissemination of EFAS information between the EFAS centres (see Section 3.2) and the partner organisations.

However, the sub-seasonal and seasonal forecasts, and the initial conditions of the hydrological model are immediately available to everyone.

In this chapter the status of the operational EFAS system as of April 2024 is outlined. The chapter starts with an organisational overview of EFAS in Section 3.2 before proceeding to outline the workflow used to create the forecasts and forecast derived products. In Section 3.3 the required input data are described followed by the model components of the forecasting chain in Section 3.4. The computational infrastructure used for generating the forecasts is described in Section 3.5 along with the monitoring of the system that is performed 24/7 to ensure timely forecasts. Following this the forecast products and their dissemination are described in Section 3.6 and Section 3.7, respectively. Section 3.9 describes the user engagement within EFAS and the ways in which users help to shape the development of the system. Some concluding remarks are made in Section 3.10.

An issue that is often experienced when writing about operational systems is that the system will have evolved by the time the writing process is completed. Whilst this chapter describes the status of the operational EFAS system as of April 2024, as highlighted above, the system is always evolving. The background and development of EFAS is not covered in detail, and readers are referred to Thielen et al. (2009), and citations therein, for a description of the early developments.

3.2 Organisational Structure of EFAS

The overall objective of EFAS is to provide complimentary information to support preparatory measures for major floods particularly in large transnational rivers. To achieve this aim EFAS generates hydrological forecasts for a range of lead-times for the whole of Europe. These forecasts and a collection of derived forecast products are made available to EFAS full and third-party partners, national and regional authorities mandated to provide flood forecasts and flood management in their respective regions. Additionally, research organisations can apply to become an EFAS research partner allowing them restricted access to the EFAS forecast products in real-time. The organisational structure put in place to facilitate the running of

EFAS is outlined in Fig. 3.3.

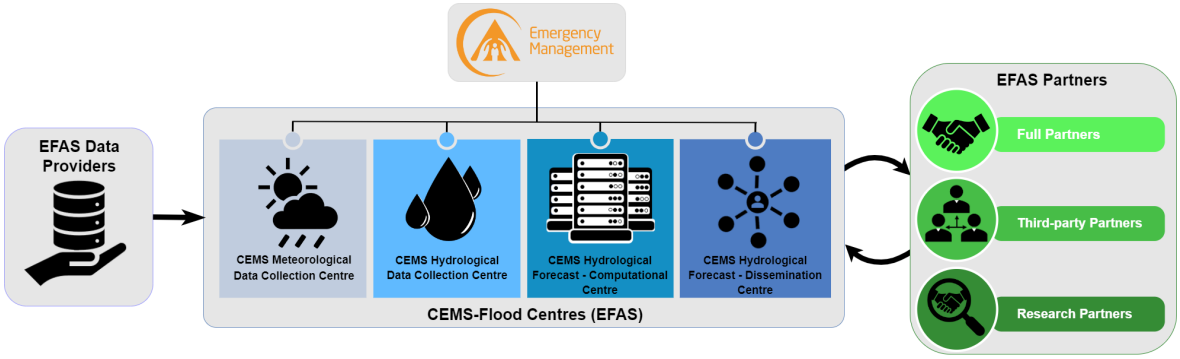

Figure 3.3: Overview of the organisational structure of EFAS and its relationship with organisations that provide data to EFAS and the EFAS partners.

As a component of CEMS, the technical and administrative management of EFAS is performed by the Joint Research Council (JRC) of the European Commission but the operational EFAS services have been outsourced to four centres. As of April 2024, following an open tendering process for contracts for the third phase of EFAS (2021-2027), the following aspects of the system operations have been issued:

- CEMS Hydrological Data Collection Centre: Soologic Technological Solution S.L. collects historic and real-time river discharge and water level data.
- CEMS Meteorological Data Collection Centre: A consortium of KISTERS AG and Deutscher Wetterdienst collects historic and real-time observed meteorological data.
- CEMS Hydrological Forecast Centre – Computation: The European Centre for Medium Range Weather Forecasts (ECMWF) collates numerical weather predictions, generates the forecast products and operates the EFAS Information System web platform and data access systems.
- CEMS Hydrological Forecast Centre – Analytics and Dissemination: A consortium of the Swedish Meteorological and Hydrological Institute (SMHI), the Slovak Hydrometeorological Institute (SHMU) and the Rijkswaterstaat Waterdienst (RWS, the Netherlands) analyses the results on a daily basis, assesses the situation, and disseminates information to the EFAS partners and to the European Commission.

The division of work between four centres (henceforth referred to as the EFAS centres) was designed to harvest the diverse skills within the European meteorological and hydrological communities by allowing institutions to focus on their areas of expertise. Close collaboration between the EFAS centres has proven essential for the smooth running of the operational services, the continued development of the system, and the delivery of the highest quality service addressing user-needs. The CEMS hydrological forecast centre – analytics and dissemination are responsible for issuing notifications to EFAS partners if a set of predefined criteria are satisfied. However, the publication of flood warnings is performed by the appropriate EFAS full and third-party partners, so that the distribution of EFAS information is executed by authorities who are experts in the field of flood forecasting as well as mandated to communicate such information with civil protection agencies. This ensures the necessary competence to understand the complexity of legal issues associated with flood forecasting and civil protection within the countries. This also guarantees that at no point does the EU system interfere with the national one voice warning principle thus allowing trust to be built between the different partners involved. Although some forecast products such as the sub-seasonal and seasonal outlooks are available in real-time to the public, most forecast products are only openly available after a 30-day embargo. The full EFAS notifications sent to EFAS partners are only ever available to the respective EFAS partners but a list of the sent EFAS notifications is included in the EFAS bulletin issued bi-monthly.

3.3 Input Data

EFAS requires hydrometeorological observations, meteorological forecasts, geo-spatial and exposure data, and hydrological reference data, in order to create hydrological forecasts. This section describes these datasets, outlines the data acquisition process, and the use of the data within the modelling chain (see Section 3.4).

3.3.1 Hydro-meteorological observations

Hydrological and meteorological data from in-situ observations are needed to calculate the initial hydro-meteorological conditions and to aid the creation of the forecast products. Meteorological and hydrological national services, and river basin authorities across Europe provide near real-time and historic data to EFAS. A complete list of data providers is provided at `https://www.efas.eu/en/share-your-data-efas`. For EFAS, the CEMS meteorological and hydrological data collection centres manage the existing network of providers of observed data. The EFAS centres can also contact potential providers and negotiate standard data license agreements between the provider and the Copernicus services. Data are collected on a 24/7 basis. The CEMS hydrological data collection centre supplies quality-checked real-time and historic in-situ hydrological observed data collected from EFAS data providers. Real-time data are used in the generation of post-processed forecast products while the historic data are also used for model calibration. As of April 2024, river discharge data are collected for 4256 stations with over 2230 stations providing near real-time data (see Fig. 3.4). The hydrological data is then transferred to the CEMS hydrological forecast centre – computational via a web API.

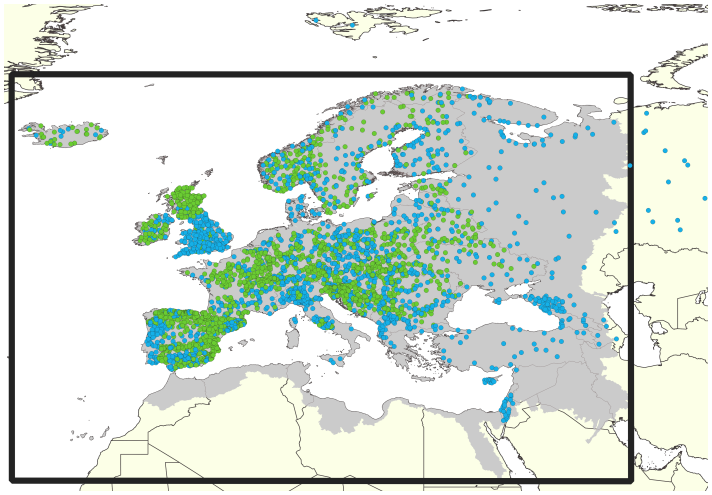

Figure 3.4: Coverage of river gauge stations providing river discharge observations to EFAS. Stations that provide historic data are shown in blue whilst stations that also provide near real-time data are in green (data from the CEMS hydrological data collection centre). Grey shaded area shows the region modelled by LISFLOOD. The black box is the EFAS domain.

The CEMS meteorological data collection centre collates several variables from gauges

including precipitation, temperature, and wind speed, although not all variables are collected from all stations. Fig. 3.5 shows all gauges that provided near real-time data for one or more meteorological variables on the 13 April 2024. Each day the collected meteorological data are quality checked for outliers and interpolated to the regular EFAS domain grid. Observations are resampled to 6-hourly and then transferred to the CEMS hydrological forecast centre – computational. Historic meteorological data are used in the calibration of LISFLOOD and to create a multidecadal simulation (see Section 3.3.4).

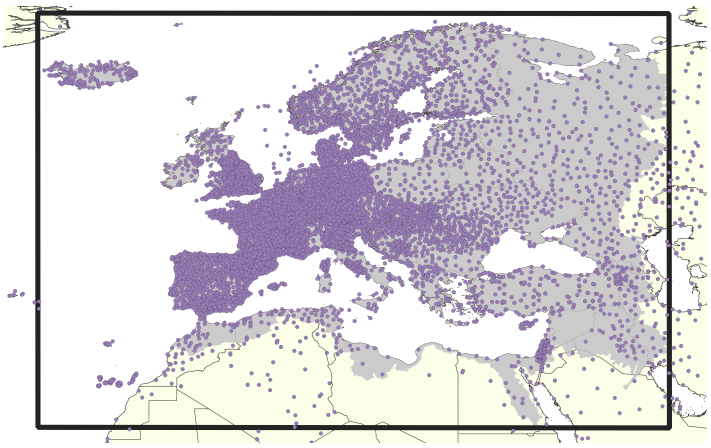

Figure 3.5: Coverage of EFAS real-time meteorological gauging stations. Stations shown are those that provided real-time meteorological observations on the 13 April 2024 (data from the CEMS meteorological data collection centre). Grey shaded area shows the region modelled by LISFLOOD. The black box is the EFAS domain.

Alongside the in-situ observations HSAF (`http://hsaf.meteoam.it/`) satellite-derived soil moisture and snow coverage products are also obtained for visualisation purposes. This data is acquired through the ECMWF data archival system or the relevant satellite data providers.

Additionally, rainfall data for the whole of Europe at a high spatiotemporal resolution is required for some products. The TAMIR flash flood forecasts, which indicate regions that are susceptible to flash flooding due to heavy rainfall, use radar composites from the radar rainfall network of the EUMETNET project OPERA (Park et al., 2024; Saltikoff et al., 2019). The radar composites have a spatial resolution of 2 km and a temporal resolution of 15 minutes. These composites are acquired by the CEMS hydrological forecast centre – computational from the ECMWF Production Data Store (ECPDS).

3.3.2 Meteorological Forecasts

To capture some of the uncertainty in the weather predictions, many of the EFAS forecasts are created using ensemble meteorological forecasts. All other EFAS forecasts are created using the output from a number of numerical weather prediction (NWP) systems (see Table 3.1) capable of providing the required forcings for the LISFLOOD hydrological model (see Section 3.3.4). Different forecast horizons and products have different requirements. For example, the ERIC surface runoff forecasts require forcings with as high a spatial resolution as possible, and an associated reforecast dataset such as that produced by the COSMO-LEPS ensemble forecast system for most of the EFAS domain. In contrast, the EFAS medium-range forecasts are based on a multi-model approach and requires meteorological forcings from different NWP systems with the aim of capturing the uncertainty due to the NWP model structures. Currently, forcings are used from four meteorological forecasts, provided by three forecasting centres (the European Centre for Medium-range Weather Forecast, the German Weather Service and the Consortium for Small-scale modelling). The seasonal and sub-seasonal hydrological outlooks are generated by forcing LISFLOOD with the seasonal (SEAS5; Johnson et al., 2019) and the extended-range (ENS-Extended; by utilising the first 50 ensemble members; Vitart et al., 2019) forecasts from ECMWF, which have the necessarily long forecast horizons.

As of April 2024, the four NWP systems used across the range of EFAS products are:

- The Integrated Forecasting System (IFS) of the European Centre for Medium Range Weather Forecasting (ECMWF). Cycle 48r1 (ECMWF, 2023) has been operational from the 27th of June 2023. This version of the IFS is used to create the HRES, ENS, ENS-Extended forecasts.

- The SEAS5 forecasts provided by ECMWF are created using cycle 43r1 of ECMWF's IFS which is coupled with the Nucleus for European Modelling of the Ocean (NEMO) model version 3.4 (Maass, 2021).

- The Icosahedral Nonhydrostatic (ICON) model of the German Weather Service (Reinert et al., 2021; Zängl et al., 2015). Version 2.6.6-nwp0 of the ICON model was released

Table 3.1: Summary details of the meteorological forcings used in generating EFAS forecasts as of April 2024. Brackets indicate the subset of the meteorological forecasts used in EFAS.

Product type	**Meteorological forcing**	**Provider**	**Maximum lead-time**	**Number of members**	**Spatial resolution**
Flash flood - TAMIR	OPERA radar composites[a]	EUMETNET	NA	NA	~2 km
Flash flood - TAMIR	ENS	ECMWF	15 days (5 days)	51	~9 km
Flash flood - ERIC	COSMO-LEPS	COSMO consortium	5.5 days	20	~7 km
Medium range	ICON (Global)[b]	DWD	7 days	1	~13 km
Medium range	ICON-EU[b]	DWD	5 days	1	~6.5 km
Medium range	HRES	ECMWF	10 days	1	~9 km
Medium range	ENS	ECMWF	15 days	51	~9 km
Medium range	COSMO-LEPS	COSMO consortium	5.5 days	20	~7 km
Sub-seasonal	ENS-Extended	ECMWF	46 days	101 (51)	~36 km
Seasonal	SEAS5	ECMWF	7 months	51	~36 km

[a] OPERA radar composites are meteorological observations rather than forecasts.

[b] DWD deterministic forecasts use the regional ICON-EU model for the first 3 days, then the global ICON model for days 4-7 of the forecasts.
NA: Not applicable.

into operations in September 2023. More information about previous versions can be found in Reinert et al. (2021). The ICON model is used for both the ICON global forecasts and the ICON-EU forecasts which is created using a nested model at a higher resolution within the global model.

- The Limited-area Ensemble Prediction System (LEPS) of the Consortium for Small-scale Modelling (COSMO; Montani et al., 2003, 2016), but only for the EFAS spatial domain it covers (see Fig. 3.6). Version 6 of the COSMO model was released in December 2021.

The NWP systems used in EFAS are continuously evolving. Updates to the NWP systems are not usually considered as updates to the EFAS system unless they require a change to the

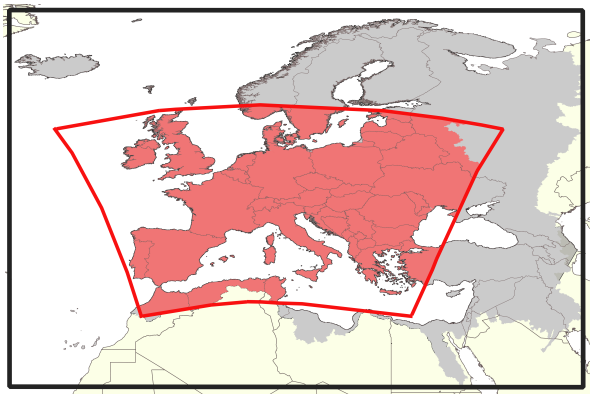

Figure 3.6: Domain boundaries for EFAS (black box) and COSMO-LEPS (red). The shaded (red and grey) region is the area modelled by LISFLOOD.

workflow of EFAS (see Section 3.4.1).

3.3.3 Geospatial Data

The EFAS hydrological forecasts are created using a range of methods which represent the transfer of water from the atmosphere to rivers. These methods are described in detail in Section 4. However, to represent the hydrological processes as accurately as possible, information about the physical characteristics of the land is required. This information is generated from several local, European, and global datasets, and mapped to the spatial resolution of the EFAS domain. Currently 91 geospatial data maps, each representing a single land-surface variable, are used within EFAS across the full range of forecast products. The variables include information on the topography (e.g., digital elevation model), the rivers (e.g., location of lakes and reservoirs, channel width, Manning's roughness), the land (e.g., soil depth and texture, and vegetation), water demand (e.g., irrigation), and social-economic data (e.g., population density, land-use, and infrastructure).

The geospatial data is considered invariant; therefore, the maps are updated infrequently. However, due to the upcoming increase in the spatial resolution of the EFAS domain, the maps have been updated to use the most recent datasets available at a spatial resolution of 1 arc min. For more information about the methods and datasets used in this process please see Choulga et al. (2024).

3.3.4 Hydrological Reference Data

The method by which the EFAS hydrological forecasts are created is described in Section 3.4. However, to convert hydrological forecasts into flood and flood impact forecast products, some form of hydrological reference data is necessary. For most EFAS forecast products the reference data are flood thresholds used to define floods of different severities. This section describes the different hydrological reference data used in the production of the EFAS forecast products.

For most EFAS forecast products the reference data is needed at every grid point and therefore cannot be derived from observations. Instead, for the medium-range forecasts and the seasonal and sub-seasonal outlooks, the thresholds are calculated from a multidecadal hydrological simulation (henceforth referred to as the EFAS reference simulation; Mazzetti et al., 2020a) created by forcing the LISFLOOD hydrological model (see Section 3.4.3) with meteorological observations (see Section 3.3.1). From the EFAS reference simulation the 1.5-, 2-, 5-, and 20-year river discharge return period thresholds are calculated and used to generate the medium-range forecast products (see Section 3.6.1).

Each time a change to the LISFLOOD hydrological model requires a new calibration to be performed, a new EFAS reference simulation is produced, and new associated flood thresholds are calculated. The benefit of deriving flood thresholds from a modelled timeseries is their availability throughout the domain and their consistency across national and administrative borders. Using a reference simulation that is as long as possible (at least 30 years for the EFAS products) also allows for a more rigorous estimation of the thresholds than if river discharge observations, for which there may only be a couple of years, were used. Additionally, since the same hydrological model is used to generate both the flood thresholds and the forecast, the impact of any biases in the model is reduced.

3.4 Modelling Chains

3.4.1 Overview

At its onset, EFAS only produced medium-range (i.e., up to 15 days) forecasts but since its founding the range of forecasts has expanded to also include short-range flash flood forecasts up to 5 days ahead (TAMIR and ERIC), and sub-seasonal and seasonal outlooks which predict up to 6 and 8 weeks ahead, respectively. This section briefly outlines the models and methods used to create each forecast category, with Fig. 3.7 illustrating the basic modelling chain used to create all EFAS forecast products.

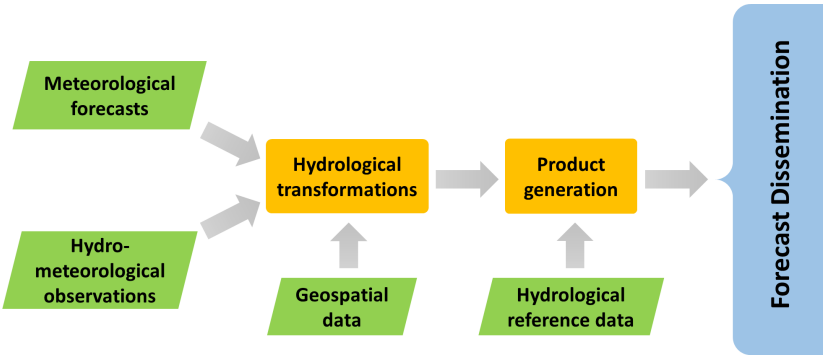

Figure 3.7: Simplified schematic of the modelling chain used to create all EFAS forecast products. Hydrological transformation and product generation steps (yellow) are discussed in Sections 3.4 and 3.6, respectively. Input data (green) are discussed in Section 3. The forecast dissemination step (blue) is discussed in Section 3.7.

This section focuses on the methods used to represent the transfer of water from the atmosphere to rivers, which we call hydrological transformations. These methods range in complexity from a simple catchment aggregation used to create the TAMIR nowcasts to a physically based distributed hydrological model used to produce forecasts for longer lead-times. Since the product generation step is intrinsically linked to the specific forecast product, a discussion of this step is delayed until Section 3.6 where the corresponding forecast products are introduced. The required input data, including the meteorological forecasts, hydro-meteorological observations, geospatial data, and hydrological reference data are described in Section 3.3. The models in the EFAS modelling chain are constantly being developed and upgraded; therefore, Section 3.4.2 provides more details on the versioning scheme used

within EFAS and outlines what is required before an upgrade can be made operational.

3.4.2 System Versions

The EFAS system is constantly evolving, both in terms of the modelling system and the derived products, as shown by the numerous improvements made over the past decade. To make it possible for the EFAS users to understand and adapt to changes that might affect their work, all changes are documented in a freely accessible web-platform (hence forth referred to as the EFAS Wiki, `https://confluence.ecmwf.int/x/PwLMDw`). If major changes are expected to the hydrological outputs, new datasets (EFAS reference simulation, reforecasts, and real-time forecasts) are made available to users ahead of the release so that they can adapt their practices if necessary (from EFAS version 5). Such a versioning system makes changes in the characteristics and skill of the forecasts over time more easily attributable to a specific update.

There are three types of system updates which are categorised in terms of the expected impact on the forecast products and summarised in Table 3.2. Each system update is labelled by a version number consisting of three digits referring to each type of update, respectively. Information on all changes to the system are documented on the dedicated EFAS Wiki and users can receive clarification via the user support team (see Section 3.9). Changes within the numerical weather prediction systems used to generate the forecasts are not explicit in the EFAS versioning system.

Table 3.2: Types of system updates and the required steps for them to become operational.

Update type	**Description of impact**	**Example scenarios**	**Release process**
Major	Characterized by a change in the hydrological model output. Change will affect how the forecasts are interpreted by the CEMS hydrological forecast centre – analytics and dissemination or the ERCC.	Calibration of hydrological model, major changes to the physics of the hydrological model (e.g., increase in resolution).	Communicated to CEMS hydrological forecast centre – analytics and dissemination at least 3 months before implementation and to EFAS Partners at least 4 weeks prior. Includes a period of testing and pre-release for feedback.
Minor	Major technical change or minor change in the output.	New layer, new product, change of format, or update of thresholds.	Communicated at least 1 week before implementation.
Patch	Bug fix or update of information in the system. May affect post-processed layers, but no impact on the model output.	Bug fix, update of layer names, or minor technical updates.	Immediately released into production.

3.4.3 Distributed Hydrological Model (LISFLOOD)

LISFLOOD is a distributed, physically based hydrological model originally developed by the Joint Research Centre (JRC) of the European Commission (Burek et al., 2013; De Roo et al., 2001, 2000; van der Knijff et al., 2008) and co-developed with the CEMS hydrological forecast centre – computational to optimise its use in an operational context. LISFLOOD has been designed for the modelling of rainfall-runoff processes in large and transnational catchments, for a variety of applications, including flood simulation and forecasting; water resources assessment (drought forecast); analysis of the impacts of land use changes, river regulation measures, and other water management plans; assessment of the impact of climate change on the terrestrial water cycle. The numerical simulation is driven by meteorological forcing

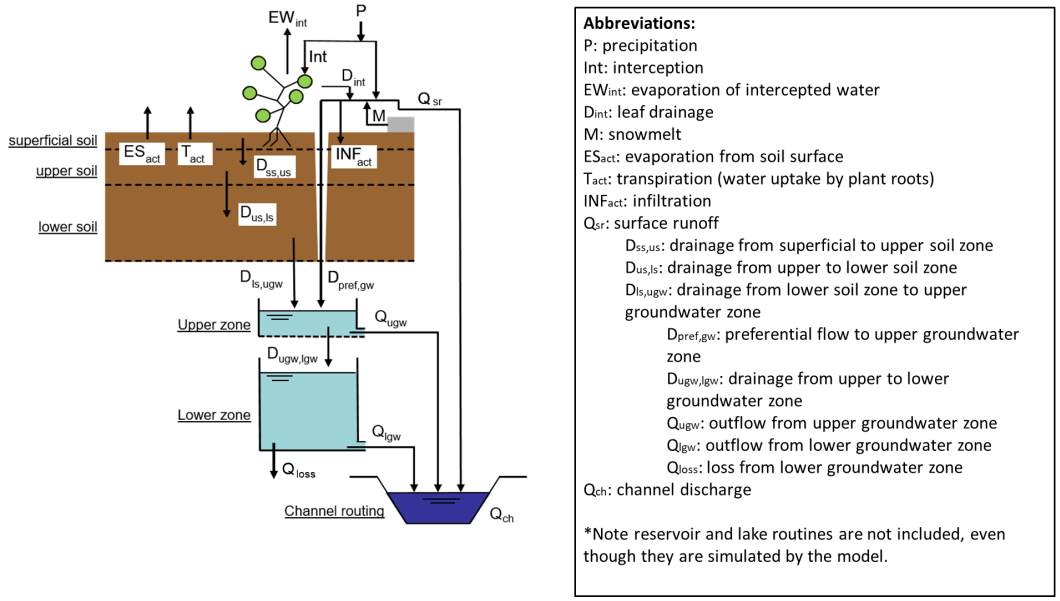

Figure 3.8: Schematic of the LISFLOOD hydrological model for a non-reservoir/lake grid-point (modified old Fig. 11.4).

data (precipitation, temperature, and potential evapotranspiration). A set of raster maps, describing the terrain morphology, soil properties, land cover and land use features, and water demand, enables the modelling of runoff processes in different climates and socio-economic contexts. LISFLOOD solves the water balance at every time step and for each grid cell. More specifically, as shown in Fig. 3.8, the model accounts for the main hydrological processes as well as for the main anthropogenic impacts on the water cycle: partition of precipitation into rain and snow, evapotranspiration and evaporation, snow melting, soil freezing, canopy

interception of rain, water infiltration into the soil and preferential flow, redistribution of soil moisture within the soil profile, groundwater storage and base flow, overland flow, lakes, dams, irrigation and other human water uses (e.g. domestic consumption), flow in the rivers and in the floodplains.

3.4.3.1 Initial conditions

To create a forecast, LISFLOOD requires initial hydrological condition which provide information about the water stored in the catchments and rivers at the start of the forecast. In EFAS the initial conditions would ideally be generated by forcing LISFLOOD with gridded meteorological observations up to the forecast date. However, there is a delay in the availability of the meteorological observations hence LISFLOOD can only be driven by these observations up to 18 hours prior to the forecast start time for the 00 UTC cycle and 30 hours prior to the forecast start time for the 12 UTC. In a process called the 'fill-up', the ECMWF and DWD deterministic forecasts valid for the gap between the last available observational data and the forecast start time are used to drive LISFLOOD and generate the initial conditions for the forecast. This procedure is shown in Fig. 3.9.

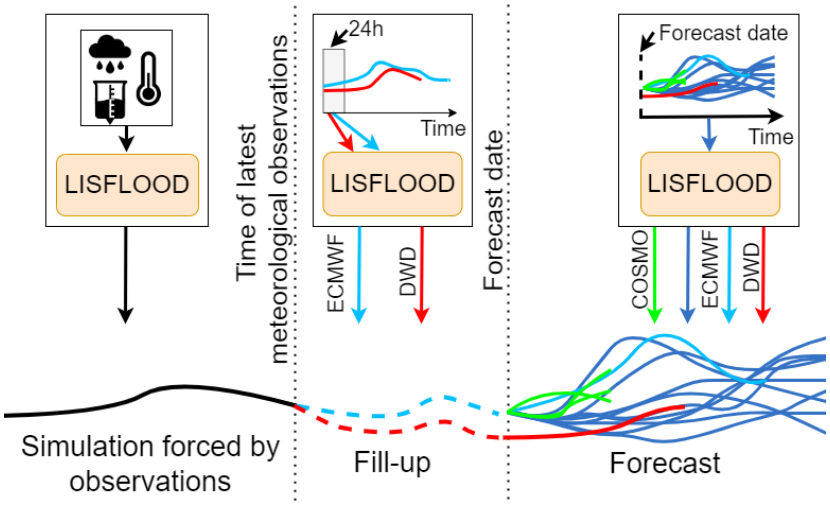

Figure 3.9: Schematic of the fill-up process for the medium-range multi-model ensemble forecasts.

3.4.3.2 Model Calibration

Some processes represented in LISFLOOD require optimal parameters to be identified via a calibration exercise. In EFAS this exercise results in the production of 14 European-wide parameter maps. The approach used in EFAS uses an Evolutionary Algorithm (EA) to select the most likely set of parameters by maximising the modified Kling-Gupta Efficiency metric (KGE'; Gupta et al., 2009; Kling et al., 2012) for the LISFLOOD output river discharge given realistic parameter constraints. Historic river discharge data, collected by the CEMS hydrological data collection centre, are used in the calibration. Since EFAS version 4, when the temporal resolution of the EFAS reference simulation and medium-range forecasts increased to 6-hourly timesteps, the calibration is conducted using 6-hourly observations as much as possible. Where only daily river discharge observations are available the simulated river discharge is aggregated to daily timesteps within the calibration procedure (Mazzetti et al., 2021a).

Figure 3.10 shows the KGE' and the correlation coefficient associated with the river discharge time series generated with the calibrated LISFLOOD model used in EFAS version 5. The metrics were calculated by comparison with the corresponding 6-hourly observational records between 1993-2021 (median record length 17 years). The median KGE' across all stations is 0.694 with no notable difference in the score between the stations calibrated with 6-hourly observations and those calibrated with daily observations (not shown). Time series produced with the calibrated LISFLOOD model also have a high correlation with the 6-hourly observation records particularly in Central Europe. Notwithstanding the overall good agreement between the observed and simulated flow statistics, large discrepancies do occur at a small number of stations, particularly in the Iberian Peninsula. Deviations from the observations may be attributed to errors in the meteorological forcings, the spatial interpolation of meteorological data, as well as to shortcomings in the hydrological model, the geospatial input data (see Section 3.3.3), and the calibration of its parameters. Some of the differences may also be due to man-made modifications of flow regimes present in many catchments, but which are not fully accounted for in the LISFLOOD hydrological model.

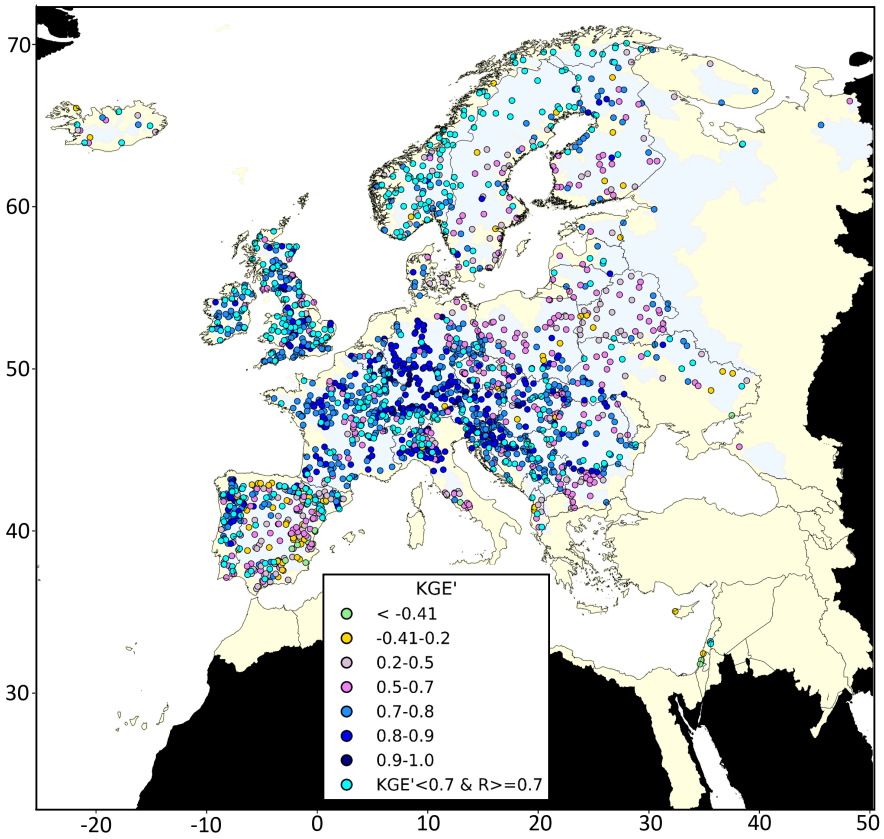

Figure 3.10: The modified KGE associated with the reference simulation of EFAS version 5.0 at a 6-hourly timestep (also called EFAS historical data in the DS). Stations with a low modified KGE but a correlation over 0.7 are coloured cyan as a high correlation is a key requirement for flood forecasts.

3.5 Generating Forecasts

The generation of forecasts is the responsibility of the CEMS hydrological forecast centre – computational. The procedure can be subdivided into five main tasks: (i) collating all the necessary forcing and input data; (ii) running the hydrological models and algorithms, (iii) statistical analysis and preparation of the results for visualisation, (iv) uploading the derived forecast products to the EFAS Information System, and (v) archiving the forecasts on the ECMWF Meteorological Achieve Retrieval System. Section 3.5.1 outlines the hardware used to perform these tasks whilst Section 3.5.2 provides details on the scheduling and software used for their execution. Section 3.5.3 outlines the steps taken to monitor forecast generation to make sure the forecasts are not delayed.

3.5.1 Overview

In October 2022 ECMWF, which houses the CEMS hydrological forecast centre – computational, upgraded its supercomputer infrastructure. The new HPC system in Bologna, Italy, consists of four Atos BullSequana XH2000 clusters and replaced the two Cray XC40 clusters that ran on-site in Reading, UK. The Atos clusters are all self-sufficient but inter-connected, increasing operational resistance should one cluster fail or require maintenance. All four clusters are connected to the high-performance clusters discussed below. Each cluster comprises of 2,032 nodes each with two AMD Epyc Rome 64 processors. Of these nodes 1,920 are compute nodes used for parallel jobs and have 256 GB of memory, and 112 are 'GPIL' nodes for general purpose and interactive workloads with 512 GiB of memory. Other nodes have special functions, such as managing the system, running the job scheduler (see Section 3.5.2), and connecting to the storage. The storage of the new system uses the Lustre parallel file system and is provided by DataDirect Networks (DDN) EXAScaler appliances. The fast access storage will provide more than 90 Petabytes in total split between dedicated storage for operational time-critical processes and research.

ECMWF's new computing facility in Bologna comprises two redundant 15 Gbps links to GARR the Italian Research and Education Network. The connection is to provide data rates of approximately 40Gbps. GARR has high speed connections to the rest of the Internet, especially to the GÉANT network, which provides a high-speed backbone between most research networks within Europe and the US. The clusters have been installed in two different computer halls in the data centre. All areas of the computer hall are equipped with an inert gas fire suppression system. The data centre site is fed through two separate redundant power cables. The internal power distribution infrastructure is also redundant with Diesel rotary UPS units (located in a separate building) providing emergency power.

3.5.2 Scheduling and Execution

EFAS forecasts are run through the ecFlow software (`https://github.com/ecmwf/ecflow`), a workflow manager that enables users to run many jobs (1442 in the case of EFAS) with

dependencies on each other and on time in a controlled environment. The ecFlow software provides an abstraction of workflows and task executions through the concept of a "suite". A suite is a collection of tasks, grouped into "families" to provide a graphical view of the system. These tasks are written as independent pieces that the suite is responsible for submitting at a specific time and in a specified order. Operational suites are designed so that they can be moved from one computing cluster (see Section 3.5.1) to another, by the simple change of a variable in the suite. Similarly, the workflow can be moved to use different storage systems in case of hardware failure. The ecFlow software provides a reasonable tolerance to hardware and software failures, combined with good restart capabilities. The scheduling of tasks considers the dependencies between them as well as date and time dependencies. This makes ecFlow particularly suited for use in EFAS where tasks require sequential evaluation yet must be performed simultaneously to ensure timely delivery of the forecasts; for example, running LISFLOOD to generate forecasts must occur after the initial conditions are determined but each set of meteorological forecasts can be evaluated simultaneously.

EFAS operational services are currently run using 9 suites. Some of these suites run the hydrological models and generate the products, others archive the data, etc. The workflows of the different EFAS forecast products may take different paths through the EFAS suites but all follow the overarching structure shown in Fig. 3.11 with the operational suite being the main suite that performs the hydrological transformation and the product generation.

Separate but similar ‘test’ and ‘experimental’ suites are used to test developments to the EFAS system before they are put into production which both streamlines the research to operational pipeline and allows EFAS partners to experience changes to the system before they are officially released (see Section 3.4.2).

3.5.3 Monitoring

The entire chain of EFAS computations as well as the underlying hardware and software infrastructure are monitored at all times to ensure uninterrupted availability and timely delivery of the EFAS products.

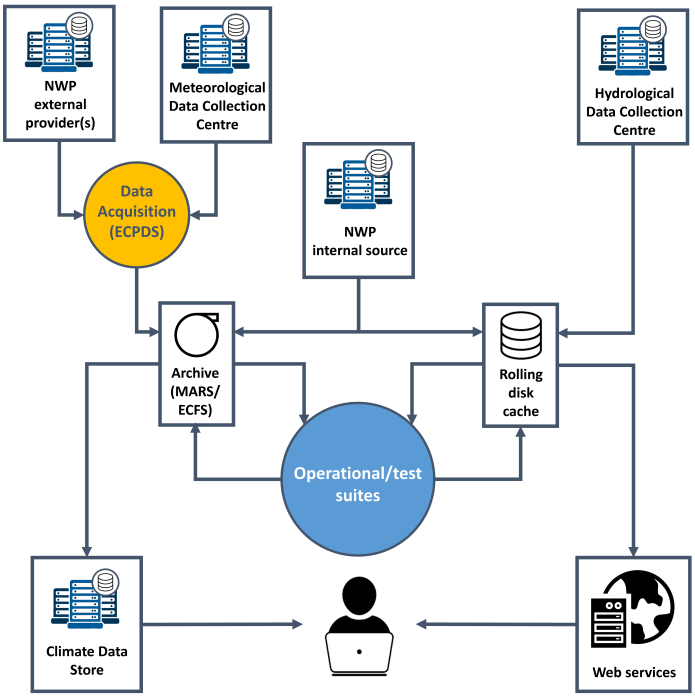

Figure 3.11: Schematic of the ecFlow suite used to generate the EFAS forecasts.

The core monitoring services, and primary support are provided by a dedicated team of operators who are always available at the ECMWF premises. This core team follows established procedures to rectify issues themselves or forward the issue to the secondary support staff (Table 3.3). The secondary support is provided by specialized teams of experts on 24/7/365 call-out duty with remote access to ECMWF IT infrastructure. Finally, tertiary support is provided by in-house and third-party technical and scientific experts.

Table 3.3: Summary of key monitoring procedures for the EFAS operational suite.

Service	**Trigger**	**Action**
Data acquisition	ecFlow monitoring interface alerts the operators	Operator informs the data provider.
EFAS computations	Failure/delay signaled by ecFlow or abnormal state detected by OpsView	Operator follows a recovery procedure or calls an analyst.
EFAS web interface	Abnormal state detected by OpsView, or an email/phone call from a user.	Operator follows a recovery procedure or calls an analyst.

Operators on duty have several mechanisms at their disposal to monitor the activity of the EFAS system. The first one is built into the ecFlow workflow managing software – its

graphical user interface visualises the progress of computations and instantly alerts operators on any failures.

The second mechanism is a dedicated subset of ecFlow jobs which are executed at specified times and check if EFAS computations have reached the expected stage. Additionally, the state of the EFAS system and underlying infrastructure (computational clusters, web servers and network) is monitored by the OpsView service (`http://www.opsview.com`) which is a monitoring and alerting tool for servers, switches, applications and services. The acquisition of input data from external providers is monitored via the web interface built into the ECMWF Product Dissemination System (ECPDS) data acquisition system which, for example, raises an alarm if no new data has been received for a prolonged period. In Table 3.3 key parts of the EFAS operational suite are listed along with the trigger used to identify a failure and the action taken by the operators following a failure. If the failure is likely to result in late, incomplete, or incorrect products, the JRC and the CEMS hydrological forecast centre – analytics and dissemination are informed.

3.6 Forecast Products

Several forecast products are derived from the hydrological forecasts created as described in Section 3.4. The form of these products is one of the most dynamic parts of the system with their evolution driven by user requirements (Pappenberger et al., 2012; Ramos et al., 2009; De Roo et al., 2011; Wetterhall et al., 2013). For brevity, only some key forecast products are described in this section, but a full list is available at `https://confluence.ecmwf.int/display/CEMS/EFAS+products`. In addition to the forecast products, auxiliary products are provided designed as complementary information to aid users in their decision making. These include products describing the initial hydrometeorological conditions of the forecasts (updated for every forecast), summarising the skill of the forecasts (updated for every major cycle) or showing the geospatial data underlying the modelling system (updated occasionally, see Section 3.3.3).

3.6.1 Reporting Points (Medium range)

The medium-range reporting points were one of the first forecast products in the EFAS portfolio, although they have been updated several times since 2012 in response to user feedback, for example in May 2022. The aim of the reporting points is to highlight areas forecasted to have high flows in the next 10 days. The 'high flow' criteria are defined by three river discharge thresholds (2-, 5-, and 20- year return periods) which are determined using the EFAS reference simulation (see Section 3.3.4). The probability of each of these thresholds being exceeded is calculated for every grid-box in the EFAS domain and shown as a range of products (typically split by threshold return period, forecast horizon, or forecast model). However, to avoid potential floods being missed in a swarm of information, in-depth details are only provided at specific locations, called 'Reporting Points'. Two types of reporting points are defined in EFAS: static and dynamic. Static reporting points are river gauge locations where historical and/or near real-time river discharge data are available. Static reporting points allow users to make a direct comparison with observations regardless of whether a threshold has been exceeded; this is particularly important for those who want to monitor the evolution of the river at a particular point of interest. Dynamic reporting points are locations for which no hydrological data is shared with EFAS and are forecast-dependent so may change from one model run to the next depending on the expected future flow conditions. Dynamic reporting points allow floods to be forecast anywhere and are defined as outlined in Fig. 3.12. The reporting point criteria valid from EFAS version 4.4 is available at `https://confluence.ecmwf.int/display/CEMS/EFAS+Reporting+Points`.

The reporting points are shown on the EFAS-IS map viewer as points on a queryable map with a pop-out window available for each point containing more information. Fig. 3.13 shows examples of the reporting points. Static points for which no flood signal meeting the EFAS reporting point criteria is forecast are shown as a grey square unless the post-processed hydrograph product (see Section 3.6.2) is available in which case the point is indicated by a blue square (see Fig. 3.13A). Yellow, red, and purple squares indicate points where the flow meets the reporting point criteria for the EFAS 2-year, 5-year and 20-year return periods,

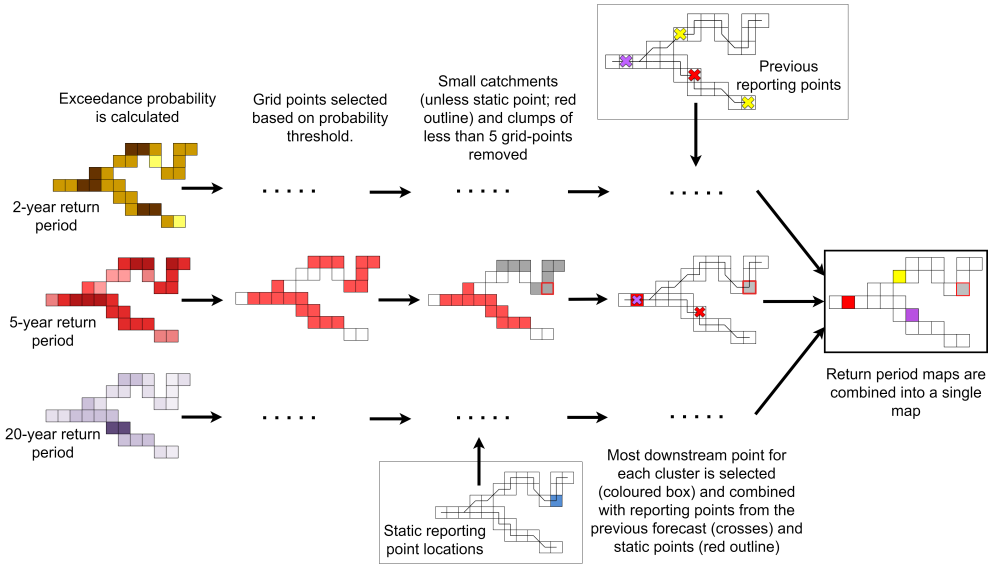

Figure 3.12: Schematic of the algorithm used to define dynamic and static reporting points as for EFAS version 5.0 medium-range forecasts. For each of the 2-, 5-, and 20-year return period thresholds the probability of exceedance is calculated. Reporting points are then calculated using these exceedance probabilities, persistence criteria based on the reporting points of the previous forecast, and upstream area thresholds.

respectively (see Fig. 3.13B). A dark outline indicates the threshold exceedance is expected within the next 48h (see Fig. 3.13B). If zoomed in on a region (as in Fig. 3.13), labels beside the point show the forecast maximum exceedance probabilities and the tendency of the total probability over the last 2 forecasts (arrows). A coloured outline to a blue or grey square indicates that the flow is forecast to exceed the respective threshold, but the upstream area is too small to meet the reporting point criteria (see Fig. 3.13A).

3.6.2 Post-processed hydrograph

The forecasts created by EFAS (or any forecasting system to date) are imperfect. As described in Section 3.3.4 most thresholds used within EFAS are calculated using the climatology of the hydrological model. Among other benefits this removes the need to consider any consistent bias in the hydrological model compared to observations. However, the predicted river discharge value (and not just the relative extremeness) is also of interest to some users of the forecasts. The post-processed hydrograph products are created by post-processing the ensemble forecast to minimise errors in the timing, volume, and the magnitude of the peak

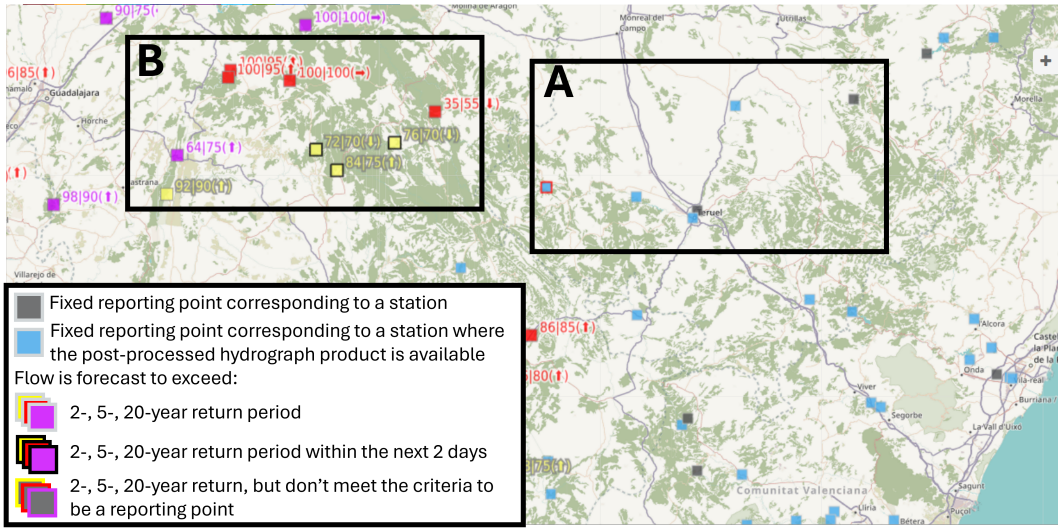

Figure 3.13: Reporting points for the EFAS medium-range streamflow forecast from the 26th of March 2024. Squares indicate the location of reporting points. Text beside reporting points indicate probability of exceedance indicated by ECMWF-ENS and COSMO-LEPS forecasts, and the tendency over the past three forecasts. Boxes label A and B show examples discussed in the text.

when compared to observations, and to quantify the total uncertainty in the forecast due to meteorological forcings, initial conditions, and hydrological model.

The method used in EFAS estimates distributions which account for the errors and uncertainty due to the meteorological forcings and the hydrological model separately using an Ensemble Model Output Statistics approach (EMOS; Gneiting et al., 2005), and a Model Conditional Processor approach (MCP; Todini, 2008), respectively. The two distributions are then combined using a Kalman filter. Full details of the method are available in (Matthews et al., 2022). Post-processing is done at locations with at least 2 years of historic river discharge observations, and where near real-time observations are available (see Fig. 3.4). These locations are identified by a blue square on the EFAS map viewer as described in Section 3.6.1.

The visualisation of the post-processed hydrograph was improved in March 2024 based on feedback from the EFAS Partners. An example of the post-processed hydrograph from the test system for EFAS version 4.1 (released operationally on 21 March 2024) is shown in Fig. 3.14. The shaded blue region shows the probability distribution with darkest blue showing the median. In this example hydrograph, three thresholds are shown by the horizontal

lines. Hydrographs for all post-processed stations show the mean river discharge threshold (Mean(Q)) and the mean annual maximum river discharge threshold (AM(Q)) which are both derived from the available historical observed time-series (and not the EFAS climatology) which may vary in length between stations. Additionally, since EFAS version 5, EFAS partners can provide up to four threshold levels (TL) to be included in the hydrograph shown by pink horizontal lines. The forecast probabilities of the river discharge exceeding the thresholds are shown as bar charts (see Fig. 3.14).

An analysis performed using the EFAS version 4 reforecasts, showed that the post-processed forecasts skilfully predicted the observed river discharge at most stations particularly those impacted by river regulation or slow hydrological processes such as snowmelt (Matthews et al., 2022).

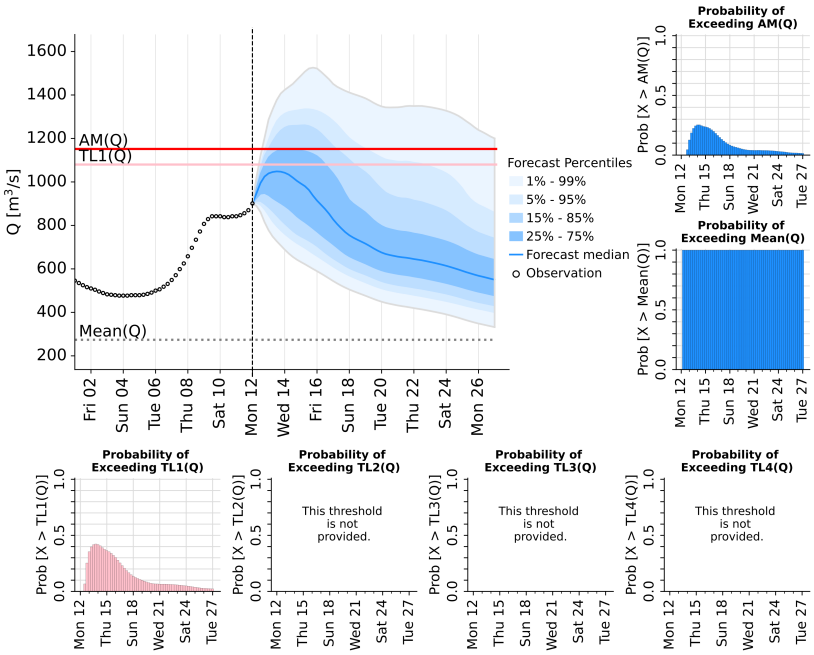

Figure 3.14: Post-processed hydrograph product from the EFAS version 4.1 test system from the 12 February 2024 for the Wittenberg / Lutherstadt station on the River Elbe, Germany. The main panel (left) shows the probability distribution of the post-processed forecast for lead-times up to 15 days. Black dots show recent observed river discharge values. The two panels on the right show the probability of exceedance of the mean annual maximum river discharge (AM(Q)) threshold (upper panel) and the mean river discharge (Mean(Q)) threshold (lower panel) for each lead-time. The four panels below the hydrograph show the probability of exceedance of up to four thresholds provided by EFAS Partners.

3.7 Forecast Dissemination

Dissemination of EFAS forecasts products to end users is carried out in two ways. The first is through the use of a web-based interface: the EFAS Information System (EFAS-IS). It is possible to access the EFAS-IS anonymously, but only for a limited number of layers and forecasts cannot be viewed in real-time but only with a delay of 30 days. The EFAS full and third-party partners have unrestricted access to the EFAS-IS in real-time. The second is for the CEMS hydrological forecast centre – analytics and dissemination to pro-actively contact EFAS full and third-party partners to notify them of a flood signal within their region.

The raw forecast data (i.e., the output from LISFLOOD) are also available to EFAS partners (including full, third-party, and some research partners) through several access points, and to the public after a 30-day embargo period. These include a dedicated Data Store powered by the Common Data Store Engine, a dedicated ftp server, and different web-services.

3.7.1 EFAS-IS

The EFAS Information System (EFAS-IS or sometimes colloquially the EFAS Portal; `https://www.efas.eu/`) is a Rich Internet Application providing the same level of interactivity and responsiveness as desktop applications (see Fig. 3.15). It was carefully designed, alongside the forecast products, with the aims of users in mind. The EFAS-IS allows control and management of the content within the web portal based on user specific roles and permits various workflows in a collaborative environment. It grants authorised users the ability to contribute to and share information and helps improve communication by allowing EFAS users to raise queries with the EFAS centres.

The EFAS-IS is loosely split into two sections. The first section is entirely open access and contains general information about EFAS as well as news bulletins, and monthly summaries of recent flood events and the hydrometeorological situation across Europe. In this section users can find links to the EFAS wiki and recent webinars (see Section 3.9) in addition to official reports and peer-reviewed journal articles about different components of EFAS.

Figure 3.15: Homepage of the EFAS-IS as visible for a registered EFAS user.

Visitors to the EFAS-IS can also find information about collaborating with EFAS. The second section of the EFAS-IS, including real-time access to the map viewer and notifications which are also sent to EFAS partners, is restricted to EFAS partners (including full, third-party, and some research partners), the JRC, and EFAS centres.

3.8 EFAS Map Viewer

The EFAS map viewer is arguably one of the easiest ways for users to access and view the forecast products described in Section 3.6. The map viewer is available to anyone however only EFAS partners can view the forecast products in real-time. There is a delay of 30 days before the forecast products are open access to ensure the one-voice policy for flood forecast warnings is not violated. Additionally, notifications, which are also viewable via the map viewer, are always restricted to EFAS full and third-party partners. All forecast products can be accessed for any forecast since they were introduced into the operational system. For example, the sub-seasonal outlooks were introduced in December 2019 and users can access all the sub-seasonal outlooks produced since.

Forecast products are organised into tabs on the map viewer such as the Flood Summary tab that is shown in Fig. 3.16. The hydrological tab allows users to view the seasonal and sub-seasonal outlooks and additional medium-range forecasts products, and the meteorological tab allows users to see the precipitation forecast products. Other tabs include information to

aid users with interpreting the forecast products. For example, static data that may impact the skill of the forecast, such as the drainage network and the location of modelled reservoirs and lakes, are available. Administrative borders and large rivers can also be overlaid to allow users to identify events that may impact them. The evaluation tab allows users to see the skill of the forecasts alongside the forecast themselves and a notifications layer allows EFAS partners to see both active and deactivated notifications.

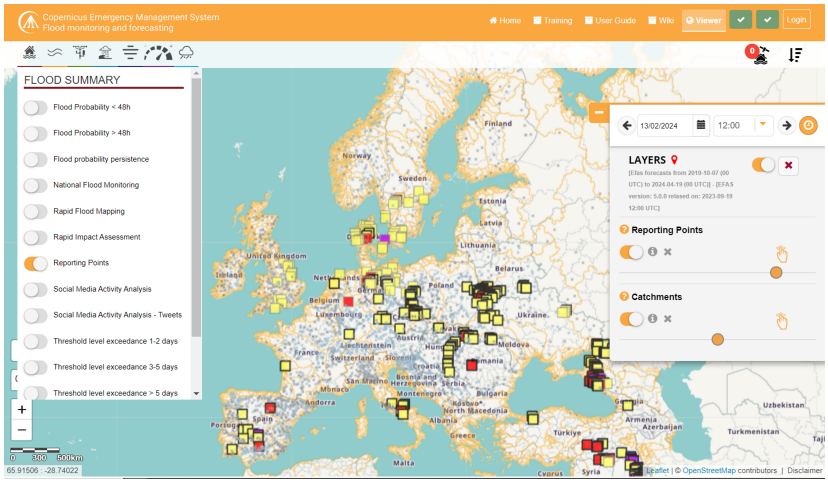

Figure 3.16: Screenshot of the EFAS map viewer showing the tabbed layout and menu for selection of products to be visualised as seen by an EFAS full partner. Reporting point forecast layer and the catchments static layer are selected.

The operational status of the EFAS forecasting system and whether partners can expect a delay or a degradation to the forecasts is shown via the status icons (green boxes in the top right of Fig. 3.16) for ERICHA (cloud) and all other forecast products (tick). A pop-out window also provides partners with a logbook of current and recent issues that may have impacted the forecasts. A key development to the map viewer in recent years is the extension of the feedback capabilities. For each notification issued, partners can say whether the event occurred and how accurate the forecast was in terms of timing and location. Partners can also indicate if an event was missed completely by the system. The information is archived and made available to EFAS centres for evaluation purposes.

3.9 User Engagement

EFAS aims to provide flood forecasts to its partners to aid them in their decision making. Therefore, developments to the EFAS system are largely guided by feedback from users. This section outlines the ways in which the EFAS partners engage with the EFAS centres whether that be to provide feedback or to solve a query regarding the forecasts.

3.9.1 Meetings, Webinars, and Outreach

EFAS runs different outreach activities each year including events specifically designed for EFAS partners which are organised by the EFAS centres and the JRC, and more general conferences related to flood forecasting. A mixture of face-to-face, online, and hybrid meetings are held to maximise attendance whilst fostering an interactive network.

3.9.1.1 Meetings, workshops, and trainings

The EFAS centres proactively engage with the EFAS partners through regular meetings, workshops, and training events. In addition to an EFAS annual meeting, delegates from the EFAS centres attend scientific workshops and conferences to engage in the scientific and operational field of hydrological and flood forecasting. EFAS holds training events, open to new and current partners, to highlight updates in products, methods, and to aid partners in accessing the forecasts products and data. Training events are coordinated by the CEMS hydrological forecast centre – analytics and dissemination with contribution from the JRC and other EFAS centres when relevant. The JRC and EFAS centres also contribute to international workshops on hydrological modelling and forecasting, including the highly successful Joint Virtual Workshop on "Connecting global to local hydrological modelling and forecasting: scientific advances and challenges", which attracted over one thousand attendees in 2020 (Dasgupta et al., 2022). EFAS uses these events to promote its products and services, as well as scientific advances.

3.9.1.2 Webinars

The EFAS centres design and deliver webinars on an annual basis or more frequently depending on the complexity and number of updates within the year. Webinars are a very effective way to reach out to a wide community as well as acting as training material for more expert audiences. The webinars have two aims: firstly, to describe the science behind the products; and secondly to demonstrate their use in specific case studies, each designed for targeted audiences and user groups. Webinars produced by EFAS engage a live audience during the hosted event and are also post-produced as short videos to act as training materials for those unable to attend the live event. Examples of webinars and tutorials produced for EFAS users can be found in the webinars section of the EFAS website and on the CEMS dedicated YouTube channel (`https://www.youtube.com/channel/UCIb8QBCOoDoKlo4JKxrse0w`, see Fig. 3.17).

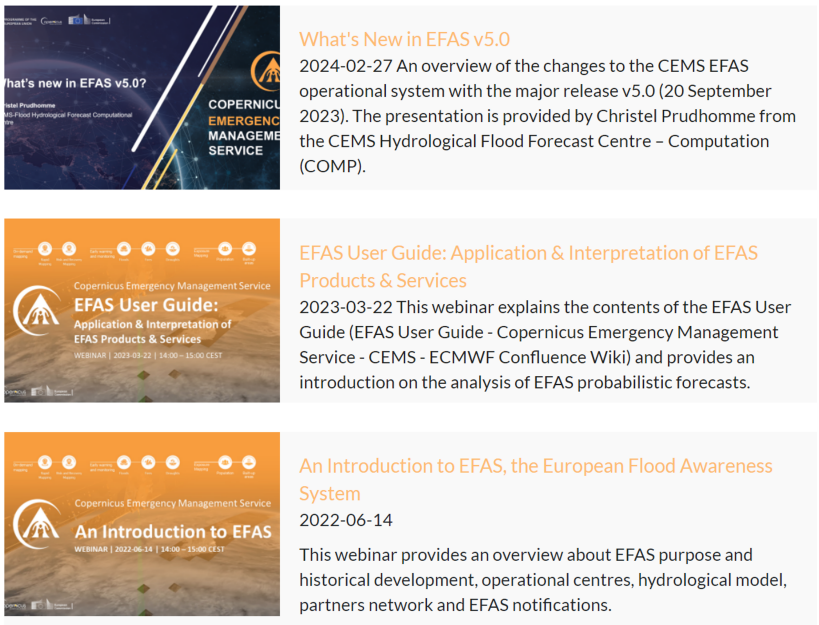

Figure 3.17: Example of webinars produced by ECMWF to communicate EFAS version 4 characteristics to EFAS partners, with recordings made available on the EFAS website and YouTube.

3.10 Conclusions

Following the devastating, trans-national floods affecting many countries in Europe in 2002, a pan-European flood forecasting system was developed to enhance the European Union's

capabilities for flood preparedness and coordination of aid. The resulting European Flood Awareness System (EFAS) has been a fully operational system since 2012, as part of the European Commission's Copernicus Emergency Management Services, providing early flood warnings across Europe to eligible national entities. The system has grown from a research experiment used to provide forecast information on an ad-hoc basis to a complex operational system in which the hydrological model forms a small part of a sophisticated forecasting and dissemination chain for a range of forecast horizons. This chapter represents a snapshot of the EFAS forecasting system in April 2024.

The EFAS operational forecasting systems can be divided into six major function blocks:

a Data acquisition, which includes the acquisition of all static and dynamic data used to operate the EFAS system including weather forecasts, geospatial data, and hydrological and meteorological observations.

b Model components, which includes different hydrological methods with a range of complexities and calibrations approaches.

c Computational infrastructure, which relates to the underlying hardware infrastructure and the way the workflow of the forecasting system is managed.

d Forecast and auxiliary products, which includes all products produced as part of the forecast including flood notifications, hydrographs, and supporting information such as precipitation maps.

e Dissemination, which deals will all aspects of disseminating the forecasts and corresponding forecast products and as such includes the web site and data distribution services.

f Performance monitoring, which includes the monitoring of the technical system performance and reliability as well as statistical skill of the forecasts.

The EFAS system is continuously evolving to adapt to the needs of its users. In its decade of operational service some of the key changes to EFAS include 1) broadening its domain which

now covers catchments between Iceland and Azerbaijan, and Scandinavia and the Maghreb, 2) improving the representation of hydrological processes by developing the hydrological models used, and increasing the temporal and spatial resolution of the forecasts, and 3) expanding the forecast horizon to include flash flood forecasts for the next 6 hours up to seasonal outlooks for the next 8 weeks. The aim of EFAS is to provide partner organisations with complementary forecast information to support their decision making. As of April 2024, EFAS has 83 full partners, 43 third-party partners, and 2 research partners. Partners can access the forecast products via a dedicated web portal, and the forecast data and companion datasets are available via several data access points. Forecasts and corresponding products are available to the public after a 30-day embargo period. Data access (real-time and historic data) has been designed with the users' objectives in mind whether that is direct use of the forecast products in decision making procedures, the incorporation of the data into downstream forecasting chains, or research studies. A versioning system has been developed to allow users to confidently navigate the expanding EFAS system. EFAS also provides data to other Copernicus services: for example, EFAS data are included in the multi-model forecast of the Copernicus Climate Change Service (C3S), the Copernicus State of the Climate report, and the CEMS European Drought Observatory.

Early flood warning systems are intended to reduce the negative impacts of floods by improving preparedness. It has been demonstrated that an early flood warning systems such as EFAS provides an immense monetary benefit (about 400 Euros for every 1 Euro invested; Pappenberger et al., 2015a). Over the years EFAS has successfully predicted several large floods aiding users in saving both money and lives. To continue doing so in the future EFAS must continue to develop in line with users' needs and balance these requirements with scientific and operational demands.

3.11 Summary of Chapter 3

This chapter provides the context in which the research in this thesis was conducted. The history of the system and the organisational structure is briefly discussed before the hydro-

logical modelling chain and medium-range river discharge forecasts products are described. Additionally, the supporting computational infrastructure required to create and disseminate the forecasts is presented. Finally, the user engagement processes are described.

The methods evaluated and developed in this thesis can be applied within any ensemble river discharge forecasting system. I use the EFAS medium-range forecasts throughout this thesis to maintain consistency between studies. This allows direct comparisons of results with regards to Chapters 5 and 6, and allows the benefits of different evaluation techniques to be compared with regards to Chapters 4 and 7.

Chapter 4

Skill assessment of the at-gauge post-processed forecasts of the European Flood Awareness System (EFAS)

This chapter addresses the first objective of the thesis: *To evaluate the skill of the operational EFAS at-gauge post-processing method.*

A forecast evaluation is conducted across 522 stations to assess the skill of the post-processed forecasts, and to identify factors that may influence the ability of the post-processing method to correct forecast errors. The content of the chapter is reproduced from:

G. Matthews, C. Barnard, H. Cloke, S. L. Dance, T. Jurlina, C. Mazzetti, and C. Prudhomme. *Evaluating the impact of post-processing medium-range ensemble streamflow forecasts from the European Flood Awareness System.* Hydrology and Earth System Sciences, 26(11):2939–2968, 2022

In Section 4.7 we summarise the key results of the chapter and identify connections with other chapters in this thesis.

Abstract

Streamflow forecasts provide vital information to aid emergency response preparedness and disaster risk reduction. Medium-range forecasts are created by forcing a hydrological model with output from numerical weather prediction systems. Uncertainties are unavoidably introduced throughout the system and can reduce the skill of the streamflow forecasts. Post-processing is a method used to quantify and reduce the overall uncertainties in order to improve the usefulness of the forecasts. The post-processing method that is used within the operational European Flood Awareness System is based on the Model Conditional Processor and the Ensemble Model Output Statistics method. Using 2-years of reforecasts with daily timesteps this method is evaluated for 522 stations across Europe. Post-processing was found to increase the skill of the forecasts at the majority of stations both in terms of the accuracy of the forecast median and the reliability of the forecast probability distribution. This improvement is seen at all lead-times (up to 15 days) but is largest at short lead-times. The greatest improvement was seen in low-lying, large catchments with long response times, whereas for catchments at high elevation and with very short response times the forecasts often failed to capture the magnitude of peak flows. Additionally, the quality and length of the observational time-series used in the offline calibration of the method were found to be important. This evaluation of the post-processing method, and specifically the new information provided on characteristics that affect the performance of the method, will aid end-users to make more informed decisions. It also highlights the potential issues that may be encountered when developing new post-processing methods.

4.1 Introduction

Preparedness for floods is greatly improved through the use of streamflow forecasts resulting in less damage and fewer fatalities (Field et al., 2012; Pappenberger et al., 2015a). The European Flood Awareness System (EFAS), part of the European Commission's Copernicus Emergency Management Service, supports local authorities by providing continental-scale medium-range streamflow forecasts up to 15 days ahead (Thielen et al., 2009; Smith et al., 2016).

These streamflow forecasts are produced by driving a hydrological model with an ensemble of meteorological forecasts from multiple numerical weather prediction (NWP) systems including two NWP ensembles and two deterministic NWP forecasts (Smith et al., 2016). However, the streamflow forecasts are subject to uncertainties that decrease their skill and limit their usefulness for end-users (Roundy et al., 2019; Thiboult et al., 2017; Pappenberger and Beven, 2006). These uncertainties are introduced throughout the system and are often categorised as *meteorological uncertainties* (or input uncertainties) which propagate to the streamflow forecasts from the NWP systems, and *hydrological uncertainties* which account for all other sources of uncertainty including those from the initial hydrological conditions and errors in the hydrological model (Krzysztofowicz, 1999). It should be noted that throughout the paper meteorological uncertainties refers to the uncertainty in the streamflow forecasts that is due to the meteorological forcings and not the uncertainty in the meteorological forecasts themselves. These differ as the meteorological variables are usually aggregated by the catchment system (Pappenberger et al., 2011). According to Krzysztofowicz (1999) and Todini (2008), a reliable forecast will include the total predictive uncertainty which is the probability of a future event occurring conditioned on all the information available when the forecast is produced.

Several approaches have been developed to reduce hydrological forecast errors and account for the predictive uncertainty. Improvements to the NWP systems used to force the hydrological model have been shown to reduce the uncertainty in the streamflow forecasts (Dance et al., 2019; Flack et al., 2019; Haiden et al., 2021). Additionally, the use of ensemble NWP systems to represent the uncertainty due to the chaotic nature of the atmosphere is becoming increasingly common and the use of multiple NWP systems can account for model parameter and structural errors in the meteorological forecasts (Wu et al., 2020; Cloke and Pappenberger, 2009). Regardless of whether deterministic or ensemble NWP systems are used, pre-processing of the meteorological input can reduce biases and uncertainties often present in the forecasts (Verkade et al., 2013; Crochemore et al., 2016; Gneiting, 2014). Data assimilation schemes can be used to improve accuracy in the initial hydrological conditions (e.g. Liu et al., 2012; Mason et al., 2020) and calibration of the hydrological model can reduce

model parameter uncertainties (Kan et al., 2019). To represent the hydrological uncertainties using an ensemble, similarly to the meteorological uncertainties, would require creating an ensemble of initial hydrological conditions and using several sets of model parameters or potentially using multiple hydrological models (Georgakakos et al., 2004; Klein et al., 2020). Operationally this is usually prohibited by computational and temporal constraints particularly if an ensemble of meteorological forcings is already included. An alternative, relatively quick and computationally inexpensive approach is to post-process the streamflow forecasts.

Post-processing the streamflow forecast allows all uncertainties to be accounted for. Over the past few decades several techniques have been proposed. These techniques can be split into two approaches: (1) methods accounting for the meteorological and hydrological uncertainties separately and (2) lumped approaches which calculate the total combined uncertainty of the forecast. One of the first examples of the former approach was the Bayesian forecasting system which was applied to deterministic forecasts and consists of the Hydrological Uncertainty Processor (HUP; Krzysztofowicz, 1999; Krzysztofowicz and Kelly, 2000; Krzysztofowicz and Herr, 2001; Krzysztofowicz and Maranzano, 2004) and an Input Uncertainty Processor (IUP; Krzysztofowicz, 1999). The development of the Bayesian Ensemble Uncertainty Processor (Reggiani et al., 2009), an extension of the HUP for application in ensemble prediction systems, attempts to remove the need for the IUP by assuming the meteorological ensemble fully represents the input uncertainty. However, as streamflow forecasts are often under-spread this assumption is not always appropriate. The Model Conditional Processor (MCP) first presented in Todini (2008) also uses a conditional distribution-based approach by defining the joint distribution between the model output and the observations using a multi-variate Gaussian distribution. The MCP has the capacity to determine the total combined uncertainty if the joint distribution is defined between the observations and the forecasts of the operational system. To define this joint distribution a large set of historic forecasts is required which is not always available as operational systems are upgraded regularly. Therefore, often it is used to account for the hydrological uncertainty only (as it is in this paper, see Section 4.3). However, the method is attractive as it can be efficiently extended to allow for multivariate, multi-model, and ensemble forecasts (Coccia, 2011; Coccia

and Todini, 2011; Todini, 2013; Todini et al., 2015). The method discussed in this study is partially motivated by the Multi-Temporal Model Conditional Processor (MT-MCP; Coccia, 2011) which extends the original MCP method for application to multiple lead-times simultaneously.

Many regression-based methods have been developed to post-process streamflow forecasts because of their relatively simple structure (e.g. quantile regression (Weerts et al., 2011), indicator cokriging (Brown and Seo, 2010, 2013), and the General Linear Model Post-Processsor (Zhao et al., 2011)). The Ensemble Model Output Statistics (EMOS; Gneiting et al., 2005) method adjusts the mean and variance of an ensemble forecast using linear functions of the ensemble members and the ensemble spread respectively (Gneiting et al., 2005; Hemri et al., 2015). This allows variations in ensemble spread to be used when estimating the predictive uncertainty. The strong autocorrelation in time observed in hydrological timeseries lends itself to the use of autoregressive error-models (e.g. Seo et al., 2006; Bogner and Kalas, 2008; Schaeybroeck and Vannitsem, 2011) although some of these methods do not account for uncertainty and instead try to correct errors in the trajectory of the forecast. These methods should therefore be used alongside a separate method which attempts to quantify the uncertainty. On the other hand, kernel-based (or "dressing") methods define a kernel to represent the uncertainty which is superimposed over the forecast or over every member for an ensemble forecast (Pagano et al., 2013; Verkade et al., 2017; Boucher et al., 2015; Shrestha et al., 2011). Depending on the approach used to define the kernel, this technique can account for the hydrological uncertainties or the total uncertainty but often requires a bias-correction method to be applied to the forecast beforehand (Pagano et al., 2013).

All the methods mentioned above, and many more that have not been mentioned (see Li et al., 2017, for a more comprehensive review), have been shown to be effective at improving the skill of forecasts in one or a few catchments. The Hydrological Ensemble Prediction Experiment (HEPEX; Schaake et al., 2007) post-processing intercomparison experiment resulted in comparisons between the different techniques (van Andel et al., 2013; Brown et al., 2013) but still relatively few studies have evaluated the performance of post-processing methods across many different catchments. Some exceptions include studies comparing

the performance of post-processing techniques for limited numbers of basins in the USA (Brown and Seo (2013), 9 basins; Ye et al. (2014), 12 basins; and Alizadeh et al. (2020), 139 basins), and recently, Siqueira et al. (2021) evaluated two post-processing methods at 488 stations across South America. Skøien et al. (2021) compared variations of the EMOS method at the 678 stations across Europe and investigated the forecast features that indicated when post-processing was beneficial. However, as post-processing is incorporated into more large-scale, multi-catchment flood forecasting systems, such as the EFAS, there is a greater need to understand which catchment characteristics, as well as which forecast features, can affect the post-processing. In this paper, the operational post-processing method of the EFAS is evaluated at 522 stations to investigate how the performance of the post-processing method varies across the domain.

The EFAS domain covers hundreds of catchments across several hydroclimatic regions with different catchment characteristics. The raw forecasts (i.e. forecasts that have not undergone post-processing) have varying levels of skill across these catchments (Alfieri et al., 2014a), and are regularly evaluated in order to identify possible areas of improvement and to allow end-users to understand the quality of the forecasts. At the locations of river gauge stations, where near real-time and historic river discharge observations are available, the raw forecasts are post-processed using a post-processing method which is motivated by the MCP and EMOS techniques. However, the post-processed forecasts do not currently undergo regular evaluation. This study aims to assess the post-processing method used within the EFAS. Additionally, new information is provided about the effect that characteristics of the catchments and properties of the forecasting system have on the performance of the post-processing method. Specifically, the paper will address the following questions:

- Does the post-processing method provide improved forecasts?
- What affects the performance of the post-processing method?

The remainder of the paper is set out as follows. In Section 4.2 we briefly describe the EFAS system used to produce forecasts operationally. In Section 4.3 we introduce the post-processing method being evaluated and explain in detail how the post-processed

forecasts are created. In Section 4.4, the evaluation strategy is described. This includes an explanation of the criteria used to select stations, details of the reforecasts used in this evaluation, and a description of the evaluation metrics considered. We separate the results section (Section 4.5) into two main sub-sections. In Section 4.5.1 we assess the affect of post-processing on different features of the forecast such as the forecast median and the timing of the peak. In Section 4.5.2 we investigate how the benefits of post-processing vary due to different catchment characteristics such as response time and elevation. Finally, in Section 4.6 we state our conclusion that post-processing improves the skill of the streamflow forecasts for most catchments and highlight the main factors affecting the performance of the post-processing method.

4.2 European Flood Awareness System (EFAS)

The focus of this paper is the evaluation of the post-processing method used operationally to create the product referred to as the 'real-time hydrograph' (see Fig. 4.4). In this section, we describe the production of the (raw) EFAS medium-range ensemble forecasts that are inputs for the post-processing method described in Section 4.3. The EFAS system has recently been updated, therefore reforecasts are used in this study allowing for a larger number of forecasts to be evaluated. Reforecasts are forecasts for past dates created using a forecasting system as close to the operational system as possible (Hamill et al., 2006; Harrigan et al., 2020). However, there are differences between the reforecasts and the operational system due to limited computational resources and data latency in the operational system. Therefore, we also highlight the differences between the evaluated reforecasts and the operational forecasts.

Version 4 of EFAS (operational October 2020) uses the LISFLOOD hydrological model at the increased temporal resolution of 6 hours and a spatial resolution of 5 km (Mazzetti et al., 2021b). LISFLOOD is a GIS-based spatially distributed gridded rainfall-runoff-routing model specifically designed to replicate the hydrological processes of large catchments (van der Knijff et al., 2008; De Roo et al., 2000). At each timestep LISFLOOD calculates the discharge as the average over the previous 6 hours for each grid-box in the EFAS domain.

For EFAS 4 the model calibration of LISFLOOD was performed using a mixture of daily and six hourly observations where available for the period 1990-2017 (Mazzetti et al., 2021a; Mazzetti and Harrigan, 2020). The reforecasts used in this evaluation are created using the same hydrological model.

Operationally, the medium-range ensemble forecasts are generated twice daily at 00 UTC and 12 UTC with a maximum lead-time of 15 days (Smith et al., 2016). The forecasts are created by forcing LISFLOOD with the precipitation, temperature, and potential evaporation outputs from four NWP systems (Smith et al., 2016; EFAS, 2020): two deterministic forecasts and two ensemble forecasts. For further information about the NWP systems see EFAS (2020). The reforecasts used in this study are generated twice weekly at 00 UTC on Mondays and Thursdays by forcing LISFLOOD with reforecasts from the European Centre of Medium-Range Weather Forecasts (ECMWF) ensemble system which have 11 ensemble members.

The hydrological initial conditions for the streamflow forecasts are determined by forcing LISFLOOD with meteorological observations to create a simulation henceforth referred to as the *water balance simulation*. The water balance simulation would provide the starting point of the forecast in terms of water storage within the catchment and discharge in the river. However, there is an operational time delay in receiving the meteorological observations. Therefore, the deterministic meteorological forecasts are used to drive the LISFLOOD model for the time period between the last available meteorological observation and the initial timestep of the forecast in a process called the 'fill-up'. For the reforecasts, all necessary meteorological observations are available so there is no need for the fill-up process.

4.3 Post-processing method

This section describes the post-processing method evaluated. Post-processing is performed at stations for which near real-time and historic river discharge observations are available. The method is motivated by the Multi-Temporal Model Conditional Processor (MT-MCP; Coccia, 2011) and Ensemble Model Output Statistics (EMOS; Gneiting et al., 2005) which are used to quantify the hydrological and meteorological uncertainties, respectively. The

Kalman filter is then used to combine these uncertainties. Since these methods assume Gaussianity, the Normal Quantile Transform (NQT) is used to transform the discharge values from physical space to standard Normal space. As with many post-processing methods, an offline calibration is required to define a so-called *station model*. In Section 4.3.1 some notation is introduced. Details on the post-processing method are given in Sects. 4.3.2 to 4.3.4. Figure 4.1 outlines the structure of the method. A discussion of the input data is postponed until Section 4.4.2.

4.3.1 Notation

In this section notation and definitions used throughout the paper are introduced. The aim of post-processing is to correct the errors and account for the uncertainty that may be present in a forecast. As described in Section 4.2, the EFAS produces ensemble streamflow forecasts for the whole of Europe on a 5 km grid with 6-hourly timesteps. However, post-processing is performed at daily timesteps and only at stations for which near real-time and historic river discharge observations are available. Therefore, the discharge values corresponding to the grid-boxes representing the locations of the stations are extracted and temporally aggregated to daily timesteps. This creates a separate streamflow forecast for each station and it is these single station forecasts that are henceforth referred to as the *raw forecasts*. The post-processing method evaluated in this paper is applied separately at each station creating a corresponding *post-processed forecast* for each raw forecast.

The input data shown in Fig. 4.1a is the input data required for the post-processing of a single raw forecast (i.e. for one station). As shown, the input data can be separated into three time periods. These time periods are henceforth referred to (from left to right in Fig. 4.1a) as the *historic period*, the *recent period*, and the *forecast period*. The length of the historic period, denoted p, varies between stations depending on the length of the historic observational record available. However, a minimum of 2 years of observations since 1991 is required for the offline calibration. For a forecast produced at time t, the recent period has q timesteps and extends from time $t - q + 1$ to time t. The forecast period extends from time

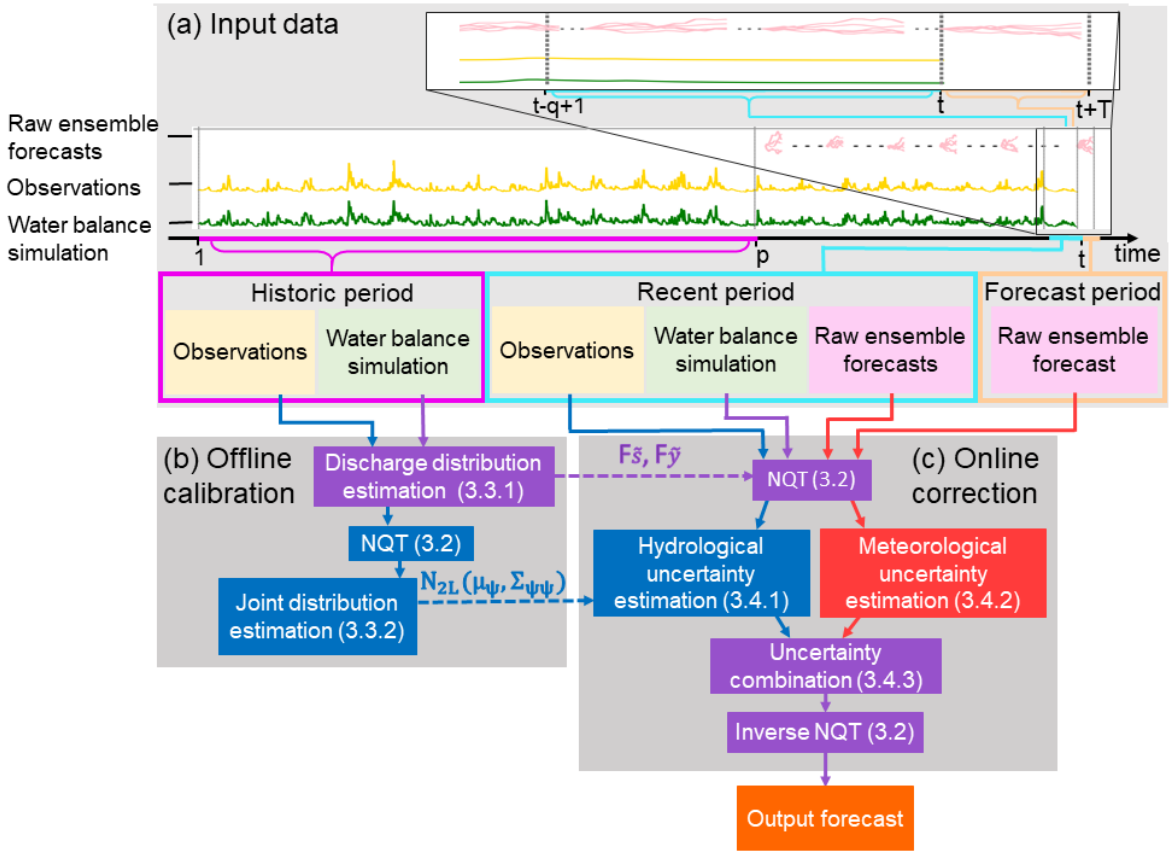

Figure 4.1: Flow chart describing the post-processing method at a station. (a) Input data are separated by time period (historic period: fuchsia, recent period: cyan, forecast period: peach) and by data type (observations: yellow, water balance simulations: green, raw ensemble forecasts: pink). The top time-series is a magnification of the bottom time-series for the period $t-q+1$ to $t+T$. The historic period has length p. For a forecast produced at time t, the recent period starts at time $t-q+1$ and the forecast period ends at time $t+T$. (b) Offline calibration steps. (c) Online correction steps. NQT is the Normal Quantile Transform. Blue and red arrows and boxes show the data and methods used to account for the hydrological uncertainty and the meteorological uncertainty, respectively. Data and methods used to account for both the hydrological and meteorological uncertainties are shown in purple. Dashed arrows show data stored in the station model such as the cumulative distribution functions of the water balance simulation and observations, denoted $F_{\tilde{s}}$, and $F_{\tilde{y}}$, respectively, and the joint distribution between the water balance simulation and observations, denoted $N_{2L}(\boldsymbol{\mu}_\psi, \boldsymbol{\Sigma}_{\psi\psi})$. Section numbers given in parentheses contain more details.

$t+1$ to time $t+T$ for a forecast with a maximum lead-time of T timesteps. The length of the recent period and the forecast period combined is $L = q+T$. For convenience, we introduce a *timestep notation* of the form $t_i : t_j$ to represent all timesteps between time t_i and time t_j i.e. $t_i{:}\ t_j$ means $t_i, t_i+1, t_i+2, \ldots t_j-1, t_j$.

The raw ensemble forecast that is post-processed is the only data available in the forecast

period. This forecast is produced at time t, has M ensemble members, and a maximum lead-time of T timesteps. The full ensemble forecast is represented by a matrix, denoted $\tilde{\mathbf{x}}_t(t+1:t+T) \in \mathbb{R}^{T\times M}$, where each column corresponds to an ensemble member and contains a vector of discharge values for each timestep in the forecast period. Throughout the paper, the tilde notation indicates that the discharge values are in physical space whereas discharge values without the tilde are in the standard Normal space (see Section 4.3.2). The subscript t indicates the forecast production time, and the range of timesteps for which discharge values are available is shown using the timestep notation. The raw ensemble forecasts from the recent period are denoted using similar notation such that, for example, the forecast produced at $t-q+1$ is denoted $\tilde{\mathbf{x}}_{t-q+1}(t-q+2:t-q+1+T) \in \mathbb{R}^{T\times M}$. All forecasts are from the same forecasting system and so all have M ensemble members and maximum lead-times of T timesteps.

The time-series of observations for a single station is denoted by the vector $\tilde{\mathbf{y}}$ where each element represents a daily discharge observation. The observations in the historic period are used in the offline calibration (see Fig. 4.1b and Section 4.3.3) and are denoted $\tilde{\mathbf{y}}(1:p) \in \mathbb{R}^p$ where the timestep notation is used to show the range of timesteps for which observations are available. This vector is the same for all forecasts for this station as the station model is not updated between forecasts. The observations in the recent period (the q timesteps up to the production time of the forecast) are used in the online correction (see Fig. 4.1c and Section 4.3.4) and are denoted $\tilde{\mathbf{y}}(t-q+1:t) \in \mathbb{R}^q$. Since $\tilde{\mathbf{y}}(t-q+1:t)$ is a function of t the observations in this vector are different for each forecast production time.

Similarly, the time-series of the water balance simulation, denoted by the vector $\tilde{\mathbf{s}}$, is used in both the offline calibration and the online correction. Each element of the vector represents a daily water balance simulation value calculated by forcing LISFLOOD with meteorological observations (see Section 4.2). The water balance simulation values from the historic period, $\tilde{\mathbf{s}}(1:p)$, are selected to correspond to the timesteps of the p observations from the same period. The water balance simulation values from the recent period are denoted $\tilde{\mathbf{s}}(t-q+1:t)$ and are dependent on the forecast production time, t.

4.3.2 Normal Quantile Transform (NQT)

The methods used in this post-processing method utilise the properties of the Gaussian distribution but discharge values usually have highly skewed non-Gaussian distributions (Hemri, 2018). Therefore, the NQT is used to transform the discharge data to the standard Normal distribution which has a mean of zero and a variance of 1, denoted $N(0, 1)$. The NQT is applied separately to all input data (observed, simulated, and forecast) for a given station, therefore, it is defined here for any scalar discharge value $\tilde{\eta}$.

The NQT defines a one-to-one map between the quantiles of the Cumulative Distribution Function (CDF) of the discharge distribution in physical space, $F_{\tilde{\eta}}(\tilde{\eta})$, and the CDF of the standard Normal distribution, $Q(\eta)$. The scalar function $F_{\tilde{\eta}}$ is dependent on whether $\tilde{\eta}$ represents a modelled discharge value (simulated or forecast) or an observed discharge value. The calculation of the discharge distributions and their subsequent CDFs are described in Section 4.3.3.1. The NQT transforms each scalar discharge value such that

$$\eta = Q^{-1}\left(F_{\tilde{\eta}}\left(\tilde{\eta}\right)\right). \tag{4.1}$$

After the forecast values have been adjusted by the post-processing method, the inverse NQT,

$$\tilde{\eta} = F_{\tilde{\eta}}^{-1}\left(Q\left(\eta\right)\right), \tag{4.2}$$

is applied to transform the discharge values from the standard Normal space back to the physical space (see Fig. 4.1c).

4.3.3 Offline calibration

The offline calibration (see Fig. 4.1b) has two main aims: to determine the distributions of the observed, $\tilde{\mathbf{y}}$, and simulated, $\tilde{\mathbf{s}}$, discharge values at a station, and to define the joint distribution between the transformed observations, $\mathbf{y}$, and the transformed water balance simulation, $\mathbf{s}$. These distributions are then stored in the station model for use in the online post-processing step (shown by dashed lines in Fig. 4.1). The input data required for the offline calibration is

an historic record of observations for the station, denoted by the vector $\tilde{\mathbf{y}}(1:p) \in \mathbb{R}^p$, and, for the same period, an historic time-series of the water balance simulation for the grid-box representing the location of the station, denoted by the vector $\tilde{\mathbf{s}}(1:p) \in \mathbb{R}^p$. The length of these vectors, p, is equal to the number of data points in the historic records and varies between stations. A minimum of 2 years historic data is required to guarantee that $p >> L$ (see Section 4.3.1).

4.3.3.1 Discharge distribution approximation

The NQT requires the Cumulative Distribution Function (CDF) of the observed and simulated discharge values in physical space, denoted $F_{\tilde{y}}$ and $F_{\tilde{s}}$ respectively, to be defined. This section describes the approach used to estimate these functions. First, the discharge density distributions are estimated using the observations, $\tilde{\mathbf{y}}(1:p) \in \mathbb{R}^p$, and the water balance simulation values, $\tilde{\mathbf{s}}(1:p) \in \mathbb{R}^p$, from the historic period. These historic time-series are often only a few years long and therefore may not represent the full discharge distribution due to the relative rarity of larger discharge values. To avoid the issues that short time-series commonly cause in the inverse NQT (discussed in Bogner et al., 2012), rather than using the empirical distribution as was done in the original MCP method (Todini, 2008), an approximation of the discharge distribution is determined using a method similar to that presented in MacDonald et al. (2011). The approximation method applies kernel density estimation to the bulk of the distribution (Węglarczyk, 2018) and fits a Generalised type II Pareto Distribution (GPD) to the upper tail (Kleiber and Kotz, 2003) to create a composite distribution (see Fig. 4.2). The GPD is an extreme value distribution that is fully defined by three parameters: the location parameter a, the scale parameter b, and the shape parameter c. Within this composite distribution the location parameter also serves as the breakpoint which separates the kernel density and the GPD, and is shown in Fig. 4.2. The parameters of the GPD are determined using the concentrated likelihood method (see steps 4-6 below). The concentrated likelihood method allows the maximum likelihood estimates of multiple parameters to be determined by first expressing one parameter in terms of the others (Takeshi, 1985). The time-series of discharge values, $\tilde{\boldsymbol{\eta}}(1:p) \in \mathbb{R}^p$, is used here to described the

distribution approximation which is implemented as follows:

1. All values in the time-series, $\tilde{\boldsymbol{\eta}}$, are sorted into descending order with $\tilde{\eta}_1$ denoting the largest value in the time-series, $\tilde{\eta}_2$ denoting the second largest value and so on.

2. A Gaussian kernel is centered at each data point such that

$$K_i(x) = \frac{1}{\sigma_{\tilde{\eta}}\sqrt{2\pi}} e^{-(x-\tilde{\eta}_i)^2/2\sigma_{\tilde{\eta}}^2} \tag{4.3}$$

   where K_i is the kernel centered at $\tilde{\eta}_i$, and $\sigma_{\tilde{\eta}}$ is the Silverman's "rule of thumb" bandwidth (Silverman, 1984). The bandwidth is calculated using the in-built R function *bw.nrd0* (Team, 2019; Venables and Ripley, 2002) and all values in the time-series, $\tilde{\boldsymbol{\eta}}$.

3. The kernel density is estimated using a leave-one-out approach such that the density at $\tilde{\eta}_j$ is

$$P(\tilde{\eta}_j) = \frac{1}{p-1} \sum_{i \neq j} K_i(\tilde{\eta}_j). \tag{4.4}$$

   This makes sure the density is not over-fitted to any individual data point.

4. To guarantee data points in the tail, the largest 10 values are always assumed to be in the upper tail of the distribution (within the GPD) and the next 990 values (i.e. $\tilde{\eta}_{11}$ to $\tilde{\eta}_{1000}$) are each tried as the location parameter, a, of the GPD. If there are less than 1000 data points (i.e. $p < 1000$) then all data points are tried as the location parameter.

5. For each test value of a:

   i The scale parameter, b, is determined analytically by the constraints that the density distribution must be equal at the breakpoint for both the GPD and the KDE distributions, and the integral of the full density distribution function with respect to discharge must be equal to 1.

   ii The shape parameter, c, is determined numerically by finding the maximum likelihood estimate, given the values of a and b, within the limits of $-1 \leq c \geq \frac{b}{\tilde{\eta}_1}$ (de Zea Bermudez and Kotz, 2010). The upper limit guarantees the upper bound

of the distribution is greater than the maximum value in the time-series, $\tilde{\eta}_1$, and the lower limit constrains the number of values considered to reduce the computational time required.

For stations with $p > 1000$, this produces 990 sets of parameters.

6. The full distribution is the combination of the KDE and GPD weighted by their contribution to the total density, $F_{\tilde{\eta}}(a)$ and $1 - F_{\tilde{\eta}}(a)$, respectively (MacDonald et al., 2011). The likelihood function for the full distribution is used to determine the maximum likelihood estimate of the location parameter, a, given the values of b and c that were calculated in step 5 for each possible value of a. This results in the most likely set of parameters (a_{ML}, b_{ML}, c_{ML}) to define the GPD fitted to the upper tail of the distribution.

The six steps outlined above are applied separately to both the simulated time-series, $\tilde{\mathbf{s}}(1:p)$, and the observed time-series, $\tilde{\mathbf{y}}(1:p)$. Figure 4.2 illustrates the approximation method for the simulated discharge distribution for a single station.

Once the variables that define the discharge density distribution, namely $\sigma_{\tilde{\eta}}$, a_{ML}, b_{ML}, and c_{ML}, have been determined, the Cumulative Distribution Function (CDF) can be calculated analytically for both the observed and simulated discharge distributions. All input data (for both the online and offline parts of the method) must be transformed to the standard Normal space using the NQT. However, it is too computationally expensive to calculate the analytical CDF for each data point. To increase the computational efficiency of the NQT, the KDE parts of the CDFs are approximated as piecewise linear functions. Each data point in the historic time-series is considered a knot (a boundary point between pieces of the piecewise function). The CDF values at the mid-points between knots are approximated using linear interpolation. If the approximated and analytical CDFs differ by more than 1×10^{-5} then the mid-points are added as additional knots. The process is repeated until the approximated CDF is accurate to within 1×10^{-5}. Ensuring that the CDF for any discharge value can be determined using linear interpolation makes the application of the NQT more efficient.

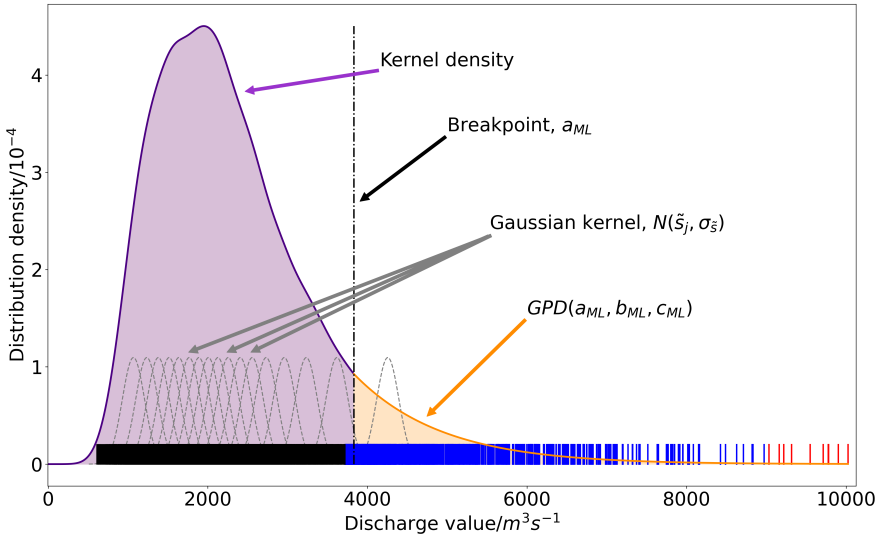

Figure 4.2: Schematic of the distribution approximation method. All data points are shown by the short solid lines. The largest 10 data points are red (always in the upper tail), the next 990 largest data points are blue (tried as the location parameter), and the remaining data points are black. Gaussian kernels (grey dashed lines) are used to calculate the kernel density (purple line). For clarity, only the kernels centered at every 500^{th} data point are plotted. The upper tail is fitted with a Generalised type II Pareto distribution (orange line). The breakpoint (dot-dashed black line) defines the separation between the two distributions. The integral of the density distribution function with respect to discharge (the sum of the purple and orange shaded areas) equals 1.

4.3.3.2 Joint distribution estimation

This section describes the calculation of the joint distribution used in the online hydrological uncertainty estimation (see Section 4.3.4.1). First, the discharge distributions defined in Section 4.3.3.1 are used within the NQT to transform the historic observations and water balance simulation to the standard Normal space (see Fig. 4.1b). This allows the joint distribution to be calculated as a multivariate Gaussian distribution. The joint distribution is defined between the observations and water balance simulation values at L timesteps which, as noted in Section 4.3.1, is equal to the length of the recent period (q timesteps) and forecast period (T timesteps) combined. The L timesteps are defined relative to a timestep k such that the joint distribution is a $2L$-dimensional distribution that describes the relationship between the observations, $\mathbf{y}(k-q+1:k+T)$, and the water balance simulation values, $\mathbf{s}(k-q+1:k+T)$. To ease notation we introduce the vector $\phi(t_i:t_j)$, here defined generally for arbitrary timesteps, which includes the observed and simulated discharge values for all

timesteps between timestep t_i and timestep t_j, such that

$$\boldsymbol{\phi}(t_i : t_j) = \begin{pmatrix} \mathbf{y}(t_i : t_j) \\ \mathbf{s}(t_i : t_j) \end{pmatrix}. \tag{4.5}$$

Following on from Eq. (4.5), we define the vector $\boldsymbol{\psi} \in \mathbb{R}^{2L}$

$$\boldsymbol{\psi}(k-q+1 : k+T) = \begin{pmatrix} \boldsymbol{\phi}(k-q+1 : k) \\ \boldsymbol{\phi}(k+1 : k+T) \end{pmatrix} = \begin{pmatrix} \mathbf{y}(k-q+1 : k) \\ \mathbf{s}(k-q+1 : k) \\ \mathbf{y}(k+1 : k+T) \\ \mathbf{s}(k+1 : k+T) \end{pmatrix} \in \mathbb{R}^{2L}. \tag{4.6}$$

The splitting of the observed and simulated variables into two distinct time periods is discussed below. The joint distribution can now be defined in terms of $\boldsymbol{\psi}(k-q+1 : k+T)$.

The joint distribution is denoted $N_{2L}(\boldsymbol{\mu}_{\psi}(k-q+1 : k+T), \boldsymbol{\Sigma}_{\psi\psi}(k-q+1 : k+T, k-q+1 : k+T))$ where the subscript $2L$ indicates its dimensions and the subscript $\boldsymbol{\psi}$ indicates that the distribution is for both observed and simulated variables. The distribution is fully defined by its mean, $\boldsymbol{\mu}_{\psi}(k-q+1 : k+T) \in \mathbb{R}^{2L}$, and covariance matrix, $\boldsymbol{\Sigma}_{\psi\psi}(k-q+1 : k+T, k-q+1 : k+T) \in \mathbb{R}^{2L\times 2L}$. Since both the observed and simulated historic timeseries have been transformed into the standard Normal space the mean discharge value is zero for both distributions and therefore the mean vector is defined as $\boldsymbol{\mu}_{\psi}(k-q+1 : k+T) = \mathbf{0}$. The covariance matrix of the joint distribution is calculated as

$$\boldsymbol{\Sigma}_{\psi\psi}(k-q+1 : k+T, k-q+1 : k+T) = \frac{1}{p-L} \sum_{k=q}^{p-T} \boldsymbol{\psi}(k-q+1 : k+T)\boldsymbol{\psi}(k-q+1 : k+T)^T \in \mathbb{R}^{2L\times 2L}, \tag{4.7}$$

where $\boldsymbol{\psi}(k-q+1 : k+T)$ is defined as in Eq. (4.6) for each timestep, k, in the historic period. Since many stations have short time-series the impact of the seasonal cycle on the joint distribution is not considered. Additionally, any spurious correlations resulting from these short time-series are not currently treated.

To ensure that the covariance matrix, $\boldsymbol{\Sigma}_{\psi\psi}(k-q+1:k+T, k-q+1:k+T)$, is positive definite the minimum eigenvalue method is used (Tabeart et al., 2020). The covariance matrix is decomposed into the eigenvalues and eigenvectors. A minimum eigenvalue threshold is set as $1 \times 10^{-7}\lambda_1$, where λ_1 is the largest eigenvalue. All eigenvalues below this threshold are set to the threshold. The matrix is then reconstructed and scaled to match the variance of the original covariance matrix.

As mentioned, the joint distribution is used in the estimation of the hydrological uncertainty in the online part of the post-processing method (see Section 4.3.4.1). If the joint-distribution is defined such that k is equal to the production time of a forecast then timesteps $k-q+1$ to k correspond to the recent period and timesteps $k+1$ to $k+T$ correspond to the forecast period. Therefore, the joint distribution can be used to condition the unknown observations and water balance simulation values in the forecast period on the known observations and water balance simulation values from the recent period. Here, we introduce notation that is used to split the joint distribution into the variables corresponding to each of these two periods. First, the mean vector is split by timestep (as in Eq. (4.6)) such that

$$\boldsymbol{\mu}_{\psi}(k-q+1:k+T) = \begin{pmatrix} \boldsymbol{\mu}_{\phi}(k-q+1:k) \\ \boldsymbol{\mu}_{\phi}(k+1:k+T) \end{pmatrix}, \tag{4.8}$$

where the subscript ϕ indicates the distribution is for the observed and simulated variables for a single time period, following the structure shown in Eq. (4.5), rather than for both time periods as indicated by the subscript ψ. The covariance matrix can be expressed as

$$\boldsymbol{\Sigma}_{\psi\psi}(k-q+1:k+T, k-q+1:k+T) = \begin{pmatrix} \boldsymbol{\Sigma}_{\phi\phi}(k-q+1:k, k-q+1:k) & \boldsymbol{\Sigma}_{\phi\phi}(k-q+1:k, k+1:k+T) \\ \boldsymbol{\Sigma}_{\phi\phi}(k+1:k+T, k-q+1:k) & \boldsymbol{\Sigma}_{\phi\phi}(k+1:k+T, k+1:k+T) \end{pmatrix} \tag{4.9}$$

where $\boldsymbol{\Sigma}_{\phi\phi}(k-q+1:k, k-q+1:k)$ and $\boldsymbol{\Sigma}_{\phi\phi}(k+1:k+T, k+1:k+T)$ are the covariance matrices for variables in the recent and forecast periods, respectively, and $\boldsymbol{\Sigma}_{\phi\phi}(k-q+1:k, k+1:k+T)$ and $\boldsymbol{\Sigma}_{\phi\phi}(k+1:k+T, k-q+1:k)$ represent the

cross-covariance matrices of variables in both time periods.

These submatrices can be further decomposed into the components referring to the observed and the simulated variables such that, for example,

$$\boldsymbol{\Sigma}_{\phi\phi}(k+1:k+T, k+1:k+T) = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{yy}}(k+1:k+T, k+1:k+T) & \boldsymbol{\Sigma}_{\mathbf{ys}}(k+1:k+T, k+1:k+T) \\ \boldsymbol{\Sigma}_{\mathbf{sy}}(k+1:k+T, k+1:k+T) & \boldsymbol{\Sigma}_{\mathbf{ss}}(k+1:k+T, k+1:k+T) \end{pmatrix}, \quad (4.10)$$

where the subscripts **y** and **s** indicate that the distribution refers to the observed and simulated variables, respectively (in contrast to the subscript ϕ which indicates both observed and simulated variables are included). The mean vector can also be split in this way such that

$$\boldsymbol{\mu}_{\phi}(k+1:k+T) = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{y}}(k+1:k+T) \\ \boldsymbol{\mu}_{\mathbf{s}}(k+1:k+T) \end{pmatrix}. \quad (4.11)$$

4.3.4 Online correction

This section describes the online correction part of the post-processing method (see Fig. 4.1c). The online correction quantifies and combines the hydrological and meteorological uncertainties for a specific forecast to produce the final probabilistic forecast. This forecast is produced at time t and has a maximum lead-time of T days, $\tilde{\mathbf{x}}_t(t+1:t+T) \in \mathbb{R}^{M \times T}$ (see Section 4.3.1 for a description of the notation). As shown in Fig. 4.1, as well as the current forecast produced at time t, the online correction requires the following input data from the recent period:

- observations for the station, $\tilde{\mathbf{y}}(t-q+1:t) \in \mathbb{R}^q$
- the water balance simulation for the grid-box containing the station's location, $\tilde{\mathbf{s}}(t-q+1:t) \in \mathbb{R}^q$
- a set of ensemble streamflow forecasts (from the same system as the forecast $\tilde{\mathbf{x}}_t$) for the grid-box containing the station's location, $\{\tilde{\mathbf{x}}_{t-q+1}, \tilde{\mathbf{x}}_{t-q+2}, \ldots, \tilde{\mathbf{x}}_{t-1}\}$.

Previous work used tuning experiments to determine that a recent period of length 40 days (i.e. $q = 40$) was most appropriate (Paul Smith, personal communication, September 2020) . All the input data is transformed to the standard Normal space using the NQT (see Eq. (4.1)) and the CDFs determined in the offline calibration (see Section 4.3.3) and stored in the station model, $F_{\tilde{y}}$ and $F_{\tilde{s}}$. The observations are transformed using $F_{\tilde{y}}$, and the water balance simulation and forecasts are transformed using $F_{\tilde{s}}$. The following sections provide more detail on the methods used to account for the uncertainties and are performed within the standard Normal space. For simplicity, it is assumed that all data is available and there are no data latency issues such that the most recent observation available is $\tilde{y}(t)$ for the timestep when the forecast is produced. In practice, some observations from the recent period may not be available, and additionally the operational system does have data latency of approximately 1 day.

4.3.4.1 Hydrological uncertainties

The hydrological uncertainty is quantified using a MCP method which uses the discharge values from the recent period and the joint distribution, $N_{2L}(\boldsymbol{\mu}_{\psi}, \boldsymbol{\Sigma}_{\psi\psi})$, defined in the offline calibration (see Section 4.3.3.2). The joint distribution defines the relationship between the observations and water balance simulation across $L = q + T$ timesteps. The hydrological uncertainty is estimated by conditioning the unknown observations and water balance simulation values in the forecast period on the known observed and simulated discharge values from the recent period using the joint distribution. First, the station observations and water balance simulations from the recent period are combined into a single vector, $\boldsymbol{\phi}(t-q+1:t)$, as defined in Eq. (4.5).

In Section 4.3.3.2, the L timesteps of the joint distribution were defined relative to a timestep k. Here, k is set equal to the production time of the forecast, t, such that the timesteps from $t - q + 1$ to t correspond to the recent period and the timesteps from $t + 1$ to $t + T$ correspond to the forecast period. Thus, the mean vector of the joint distribution can be

expressed, as discussed in Section 4.3.3.2, as

$$\boldsymbol{\mu}_{\psi}(t-q+1:t+T) = \begin{pmatrix} \boldsymbol{\mu}_{\phi}(t-q+1:t) \\ \boldsymbol{\mu}_{\phi}(t+1:t+T) \end{pmatrix} \tag{4.12}$$

where $\boldsymbol{\mu}_{\phi}(t-q+1:t)$ represents the mean of the variables (both observations and water balance simulation) in the recent period, for which we have known values, $\phi(t-q+1:t)$, and $\boldsymbol{\mu}_{\phi}(t+1:t+T)$ represents the mean of the variables in the forecast period, which we are required to predict.

The sub-matrices of the covariance matrix of the joint distribution that were defined in Eq. (4.10) are also positioned relative to timestep t, such that,

$$\boldsymbol{\Sigma}_{\psi\psi}(t-q+1:t+T, t-q+1:t+T) = \begin{pmatrix} \boldsymbol{\Sigma}_{\phi\phi}(t-q+1:t, t-q+1:t) & \boldsymbol{\Sigma}_{\phi\phi}(t-q+1:t, t+1:t+T) \\ \boldsymbol{\Sigma}_{\phi\phi}(t+1:t+T, t-q+1:t) & \boldsymbol{\Sigma}_{\phi\phi}(t+1:t+T, t+1:t+T) \end{pmatrix}. \tag{4.13}$$

By positioning the joint distribution in this way, $\boldsymbol{\mu}_{\phi}(t+1:t+T) \in \mathbb{R}^{2T}$ and the sub-matrix $\boldsymbol{\Sigma}_{\phi\phi}(t+1:t+T, t+1:t+T) \in \mathbb{R}^{2T\times 2T}$ create a climatological forecast for the observations and water balance simulation in the standard Normal space. It is this climatological forecast that is conditioned on the discharge values from the recent period.

The conditional distribution of the unknown discharge values in the forecast period conditioned on the known discharge values in the recent period, denoted $N_{2T}(\widehat{\boldsymbol{\mu}}_{\phi}(t+1:t+T), \widehat{\boldsymbol{\Sigma}}_{\phi\phi}(t+1:t+T, t+1:t+T))$, is calculated using the properties of a multivariate Gaussian joint distribution (Dey and Rao, 2006) such that

$$\widehat{\boldsymbol{\mu}}_{\phi}(t+1:t+T) = \boldsymbol{\mu}_{\phi}(t+1:t+T) + \boldsymbol{\Sigma}_{\phi\phi}(t+1:t+T, t-q+1:t)\boldsymbol{\Sigma}_{\phi\phi}(t-q+1:t, t-q+1:t)^{-1} \times \Big(\phi(t-q+1:t) - \boldsymbol{\mu}_{\phi}(t-q+1:t)\Big) \tag{4.14}$$

and

$$\widehat{\boldsymbol{\Sigma}}_{\phi\phi}(t+1:t+T, t+1:t+T) = \boldsymbol{\Sigma}_{\phi\phi}(t+1:t+T, t+1:t+T) - \boldsymbol{\Sigma}_{\phi\phi}(t+1:t+T, t-q+1:t) \times \boldsymbol{\Sigma}^{-1}_{\phi\phi}(t-q+1:t, t-q+1:t)\boldsymbol{\Sigma}_{\phi\phi}(t-q+1:t, t+1:t+T). \quad (4.15)$$

where the hat notation indicates it is conditioned on the discharge values from the recent period.

The resulting predicted distribution, $N_{2T}(\widehat{\boldsymbol{\mu}}_{\phi}(t+1:t+T), \widehat{\boldsymbol{\Sigma}}_{\phi\phi}(t+1:t+T, t+1:t+T))$ is referred to as the *hydrological uncertainty distribution* and can be partitioned into two T-dimensional forecasts; one for the water balance simulation and one for the unknown observations in the forecast period such that,

$$\begin{bmatrix} \mathbf{y}(t+1:t+T) \\ \mathbf{s}(t+1:t+T) \end{bmatrix} \sim N_{2T}\left(\begin{bmatrix} \widehat{\boldsymbol{\mu}}_{\mathbf{y}}(t+1:t+T) \\ \widehat{\boldsymbol{\mu}}_{\mathbf{s}}(t+1:t+T) \end{bmatrix}, \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_{\mathbf{yy}}(t+1:t+T, t+1:t+T) & \widehat{\boldsymbol{\Sigma}}_{\mathbf{ys}}(t+1:t+T, t+1:t+T) \\ \widehat{\boldsymbol{\Sigma}}_{\mathbf{sy}}(t+1:t+T, t+1:t+T) & \widehat{\boldsymbol{\Sigma}}_{\mathbf{ss}}(t+1:t+T, t+1:t+T) \end{bmatrix}\right). \quad (4.16)$$

The subscripts **y** and **s** indicate that the distribution refers to the observed and simulated variables, respectively.

4.3.4.2 Meteorological uncertainty

This section describes the part of the online correction that estimates the meteorological uncertainty in the forecast of interest. As stated at the beginning of Section 4.3.4, the forecast of interest and the input data from the recent period are transformed into standard Normal space. The full transformed forecast, denoted by the forecast matrix $\mathbf{x}_t(t+1:t+T) \in \mathbb{R}^{T\times M}$ where each column represents an ensemble member (see Section 4.3.1), has ensemble mean $\overline{\mathbf{x}}_{\mathbf{t}}(t+1:t+T) \in \mathbb{R}^T$. The i-th component of $\overline{\mathbf{x}}_{\mathbf{t}}(t+1:t+T)$ represents the ensemble mean

discharge at the i-th lead-time and is calculated as

$$\overline{\mathbf{x}}_{\mathbf{t}}(t+1:t+T)[i] = \frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_t(t+1:t+T)[i,m] \tag{4.17}$$

The auto-covariance matrix of the forecast, $\mathbf{\Gamma}_t(t+1:t+T, t+1:t+T) \in \mathbb{R}^{T\times T}$, is calculated such that the element corresponding to the i-th row and j-th column is given by

$$\begin{aligned}\mathbf{\Gamma}_t(t+1:t+T, t+1:t+T)[i,j] = \\ \frac{1}{M-1}\sum_{m=1}^{M}(\mathbf{x}_t(t+1:t+T)[i,m] - \overline{\mathbf{x}}_{\mathbf{t}}(t+1:t+T)[i]) \\ \times(\mathbf{x}_t(t+1:t+T)[j,m] - \overline{\mathbf{x}}_{\mathbf{t}}(t+1:t+T)[j])^T.\end{aligned} \tag{4.18}$$

The uncertainty that propagates through from the meteorological forcings is partially captured by the spread of the ensemble streamflow forecast. However, these forecasts are often under-spread particularly at shorter lead-times. The Ensemble Model Output Statistics method (EMOS; Gneiting et al., 2005) is used here to correct the spread only. Biases from the hydrological model are ignored in this section as the same hydrological model is used to create the water balance simulation and the forecasts. It is assumed that there is no bias in the meteorological forcings relative to the meteorological observations that are used to produce the water balance simulation (see Section 4.2) and that each ensemble member is equally likely. These assumptions allow the value of the water balance simulation at any time k to be expressed as

$$s(k) = \overline{\mathrm{x}}_l(k) + \epsilon \tag{4.19}$$

where $\overline{\mathrm{x}}_i(k)$ is the ensemble mean for the timestep k of a forecast produced at time l (where $l+1 <= k <= l+T$), and ϵ is an unbiased Gaussian error. The value of the ensemble mean at timestep k, $\overline{\mathrm{x}}_l(k)$, is therefore a random variable from the distribution $N(s(k), \sigma_\epsilon^2)$.

The variance of ϵ, σ_ϵ^2, should equal the expected value of the spread of the forecast, $E[\mathbf{\Gamma}_t]$. However, this is not always satisfied. To correct the spread, a set of forecasts from the recent period are used to estimate two spread correction parameters. The corrected

covariance matrix, $\mathbf{\Gamma}_t^c(t+1:t+T, t+1:t+T) \in \mathbb{R}^{T\times T}$, is then calculated, using these spread correction parameters, such that

$$\mathbf{\Gamma}_t^c(t+1:t+T, t+1:t+T) = \zeta\left(\delta\mathbf{I} + \mathbf{\Gamma_t}(t+1:t+T, t+1:t+T)\right) \tag{4.20}$$

where $\mathbf{I}$ is the identity matrix, and ζ and δ are the scalar spread correction parameters to be determined.

The ensemble mean at each lead-time and the auto-covariance matrices are calculated for each of the forecasts from the recent period after they have been transformed to the standard Normal space (not including the forecast produced at time t that is being corrected). Using the concentrated likelihood method (Takeshi, 1985) the spread correction parameters are defined as the maximum likelihoood estimates, ζ_{ML} and δ_{ML}, for the likelihood function

$$L(\zeta, \delta|\{\mathbf{x_{t-q+1}}, \ldots, \mathbf{x_{t-1}}\}) = \prod_{k=t-q+1}^{t-1} \frac{1}{\sqrt{2\pi\zeta(\delta\mathbf{I} + \mathbf{\Gamma}_k)}} \exp\left(-\frac{1}{2\zeta(\delta\mathbf{I} + \mathbf{\Gamma}_k)}(\overline{\mathbf{x}}_k - \mathbf{s})^2\right) \tag{4.21}$$

where we have used a shorthand notation for clarity, such that $\overline{\mathbf{x}}_k = \overline{\mathbf{x}}_k(k+1:k+T)$, $\mathbf{\Gamma}_k = \mathbf{\Gamma}_k(k+1:k+T, k+1:k+T)$, and $\mathbf{s} = \mathbf{s}(k+1:k+T)$ as defined above.

The current forecast, $\mathbf{x_t}(t+1:t+T)$, is spread-corrected to account for the meteorological uncertainty by applying the parameters, ζ_{ML} and δ_{ML}, as described in Eq. (4.20). This resultant distribution is referred to as the *meteorological uncertainty distribution* and provides a prediction of the water balance simulation in the forecast period, such that

$$\mathbf{s}(t+1:t+T) \sim N(\overline{\mathbf{x}}_\mathbf{t}(t+1:t+T), \mathbf{\Gamma}_t^c(t+1:t+T, t+1:t+T)). \tag{4.22}$$

4.3.4.3 Combining uncertainties

The update step equations of the Kalman Filter (Kalman, 1960) are used to combine the hydrological and meteorological uncertainties to produce the final probabilistic forecast. The hydrological uncertainty distribution, defined in Eq. (4.16) and denoted $N_{2T}(\widehat{\boldsymbol{\mu}}_\phi(t+1:t+T), \widehat{\mathbf{\Sigma}}_{\phi\phi}(t+1:t+T, t+1:t+T))$, is a predicted distribution for the water balance simulation

and the observations during the forecast period. The meteorological uncertainty distribution, defined in Eq. (4.22) and denoted $N(\overline{\mathbf{x}}_\mathbf{t}(t+1:t+T), \mathbf{\Gamma}_t^c(t+1:t+T, t+1:t+T))$, is a predicted distribution for the water balance simulation in the forecast period. The predictions of the distribution of the water balance are compared within the Kalman filter. In order to extract the water balance simulation part of the hydrological uncertainty distribution we define the matrix "observation operator" $\mathbf{H}$ such that

$$\widehat{\boldsymbol{\mu}}_\mathbf{s}(t+1:t+T) = \mathbf{H}\widehat{\boldsymbol{\mu}}_\psi(t+1:t+T) = \mathbf{H}\begin{pmatrix} \widehat{\boldsymbol{\mu}}_\mathbf{y}(t+1:t+T) \\ \widehat{\boldsymbol{\mu}}_\mathbf{s}(t+1:t+T) \end{pmatrix} \in \mathbb{R}^T \tag{4.23}$$

where the subscripts $\mathbf{y}$ and $\mathbf{s}$ denote the observed and water balance simulation variables, respectively.

The update step of the Kalman filter is applied to produce a probabilistic forecast in the standard Normal space containing information about both the meteorological and hydrological uncertainties. The distribution of this forecast is denoted $N_{2T}(\widehat{\boldsymbol{\mu}}^a_\psi(t+1:t+T), \widehat{\mathbf{\Sigma}}^a_{\psi\psi}(t+1:t+T, t+1:t+T))$, where the superscript a signifies the Kalman filter has been applied. The mean, $\widehat{\boldsymbol{\mu}}^a_\psi(t+1:t+T))$, is calculated as

$$\widehat{\boldsymbol{\mu}}^a_\psi(t+1:t+T) = \widehat{\boldsymbol{\mu}}_\psi(t+1:t+T) + \mathbf{K}(\overline{\mathbf{x}}_\mathbf{t}(t+1:t+T) - \mathbf{H}\widehat{\boldsymbol{\mu}}_\psi(t+1:t+T)) \tag{4.24}$$

where $\mathbf{K}$ is the Kalman gain matrix, defined as

$$\mathbf{K} = \widehat{\mathbf{\Sigma}}_{\psi\psi}(t+1:t+T, t+1:t+T)\mathbf{H}^\mathbf{T}(\mathbf{H}\widehat{\mathbf{\Sigma}}_{\psi\psi}(t+1:t+T, t+1:t+T)\mathbf{H}^\mathbf{T} \tag{4.25}$$

$$+\mathbf{\Gamma}_t^c(t+1:t+T, t+1:t+T))^{-1}, \tag{4.26}$$

and $\mathbf{H}$ is the matrix observation operator defined above. The auto-covariance matrix is calculated as

$$\widehat{\mathbf{\Sigma}}^a_{\psi\psi}(t+1:t+T, t+1:t+T) = (\mathbf{I} - \mathbf{K}\mathbf{H})\widehat{\mathbf{\Sigma}}_{\psi\psi}(t+1:t+T, t+1:t+T) \tag{4.27}$$

where $\mathbf{I}$ is the identity matrix and all other symbols are as before. The distribution produced by combining these two sources of uncertainty, $N_{2T}(\widehat{\boldsymbol{\mu}}^a_{\psi}(t+1:t+T), \widehat{\boldsymbol{\Sigma}}^a_{\psi\psi}(t+1:t+T, t+1:t+T))$, is for both the unknown observations and the water balance simulations variables in the forecast period. This distribution is partitioned into two T-dimensional forecasts which are in the standard Normal space such that

$$\begin{bmatrix} \mathbf{y}(t+1:t+T) \\ \mathbf{s}(t+1:t+T) \end{bmatrix} \sim N_{2T}\left(\begin{bmatrix} \widehat{\boldsymbol{\mu}}^a_{\mathbf{y}}(t+1:t+T) \\ \widehat{\boldsymbol{\mu}}^a_{\mathbf{s}}(t+1:t+T) \end{bmatrix}, \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}^a_{\mathbf{yy}}(t+1:t+T, t+1:t+T) & \widehat{\boldsymbol{\Sigma}}^a_{\mathbf{ys}}(t+1:t+T, t+1:t+T) \\ \widehat{\boldsymbol{\Sigma}}^a_{\mathbf{sy}}(t+1:t+T, t+1:t+T) & \widehat{\boldsymbol{\Sigma}}^a_{\mathbf{ss}}(t+1:t+T, t+1:t+T) \end{bmatrix}\right) \tag{4.28}$$

where the subscripts $\mathbf{y}$ and $\mathbf{s}$ denote the observed and water balance simulation variables, respectively.

The T-dimensional distribution corresponding to the predicted distribution of the unknown observations in the forecast period, $N_T(\widehat{\boldsymbol{\mu}}^a_{\mathbf{y}}(t+1:t+T), \widehat{\boldsymbol{\Sigma}}^a_{\mathbf{yy}}(t+1:t+T, t+1:t+T))$, is transformed back into physical space using the inverse NQT, defined in Eq. (4.2), and the CDF of the observed discharge distribution, $F_{\tilde{y}}$. This forecast is then used to produce the 'real-time hydrograph' (see Fig. 4.4 for an example of this forecast product).

4.4 Evaluation strategy

4.4.1 Station selection

To maintain similarity with the operational system, the station models used in this evaluation are those calibrated for use in the operational post-processing. To avoid an unfair evaluation, station models must have been calibrated using observations from before the evaluation period. An evaluation period of approximately 2-years (from 1 January 2017 to 14 January 2019) was chosen to balance the length of the evaluation period with the number of stations evaluated. Of the 1200 stations post-processed operationally, 610 stations have calibration

time-series with no overlap with the evaluation period. Additionally, stations were required to have at least 95% of the daily observations for the evaluation period reducing the number of stations to 525. A further three stations were removed after a final quality control inspection (see Section 4.4.2.2 for details of the observations and the quality control system used). The locations of the 522 stations are shown in Fig. 4.3. The marker colour shows the CRPS of the raw ensemble forecast for a lead-time of 6 days. The spatial pattern of these CRPS values are discussed in Section 4.5.1.4. Although all 522 stations are evaluated, specific stations (labelled in Fig. 4.3) are used to illustrate key results (see Section 4.5.2).

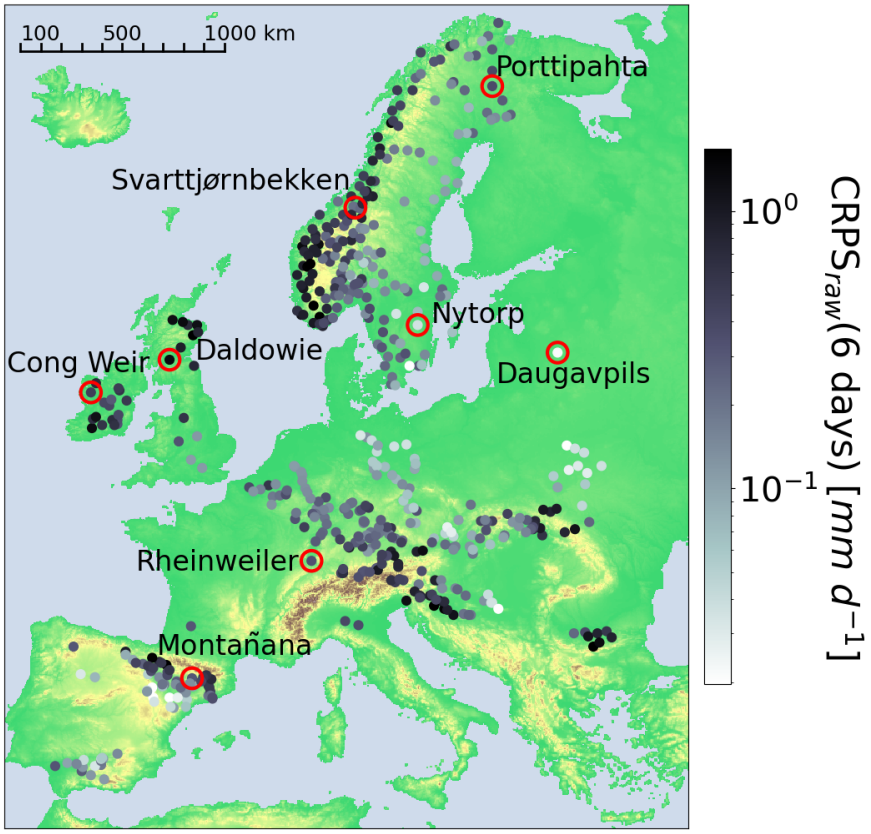

Figure 4.3: Map showing the locations of the 522 stations evaluated. The marker colour shows the Continuous Ranked Probability Score (see Section 4.4.3.4) for the raw forecast at a lead-time of 6 days on a log scale. Perfect score: CRPS=0. Stations used as examples in Section 4.5 are labelled and highlighted by the red circles.

4.4.2 Data

4.4.2.1 Reforecasts

The reforecasts used in this study are a subset of the EFAS 4.0 reforecast dataset (Barnard et al., 2020a). This dataset contains twice-weekly reforecasts for dates that correspond to each Monday and Thursday in 2019. For example, 3 January 2019 is a Thursday, so the dataset contains reforecasts for 3 January for every year from 1999 to 2018. The chosen evaluation period (see Section 4.4.1) includes 208 reforecasts. The raw forecasts were used as input for the post-processing method. Using twice-weekly reforecasts, rather than daily, reduces the temporal correlations between forecasts and therefore limits the dependence of the results on the autocorrelation of the river discharge (Pappenberger et al., 2011). However, this means any single event cannot be included in the evaluation for all lead-times. For example, an event that occurs on a Saturday will not be included within the evaluation of the forecasts at a lead-time of 1 day which can only be a Tuesday or a Friday. Where necessary the evaluation metrics were combined over several lead-times (see Sects. 4.4.3.2 and 4.4.3.3). Additionally, fewer reforecasts were available to estimate the EMOS parameters in the meteorological uncertainty estimation (see Section 4.3.4.2). Whereas operationally daily forecasts for each day of the recent period are available, here only two reforecasts are available for each week of the recent period. This reduces the number of forecasts used to estimate the EMOS parameters from 40 to 11. We did not extend the recent period to maintain consistency with the operational system and to avoid introducing errors due to any seasonal variation in the EMOS parameters.

The reforecasts, and the operational forecasts (see Section 4.2), have a 6-hourly timestep. However, currently post-processing is performed at daily timesteps. Therefore, the reforecasts were aggregated to daily timesteps with a maximum lead-time of $T = 15$ days.

4.4.2.2 Observations

All discharge observations were provided by local and national authorities and collected by the Hydrological Data Collection Centre of the Copernicus Emergency Management Service

and are the observations used operationally. The operational quality control process was applied to remove incorrect observations before they were used in this study (McMillan et al., 2012). Additionally, simple visual checks were performed to account for any computational errors introduced after the operational quality checks. Average daily discharge observations were used in three parts of the study. For each station, an historic time-series was used in the calibration of the station model (see Section 4.3.3). The length of the historic time-series, denoted p in Section 4.3.1, varies in length between stations. However, a minimum of 2-years of observational data between 1 January 1990 and 1 January 2017 is required. It should be noted that there is an overlap between the observations used for the calibration of the station models and the observations used for the calibration of the LISFLOOD hydrological model. For each reforecast, records of near real-time observations from the $q = 40$ days prior to the forecast time were used as the observations in the recent period (see Section 4.3.4.1). Observations from the evaluation period were used as the truth values in the evaluation (see Section 4.4.3).

4.4.2.3 Water balance simulation

The EFAS 4.0 simulation (Mazzetti et al., 2020a) was used as the water balance simulation for dates between 1 January 1990 and 14 January 2019. As described in Section 4.2, the water balance simulation is created by driving LISFLOOD with gridded meteorological observations. This dataset provides simulations for the whole of the EFAS domain. The values for the grid-boxes representing the locations of the stations were extracted creating a simulated time-series for each station. These time-series were aggregated from 6-hourly timesteps to daily timesteps (00 UTC to 00 UTC) and were used in three ways in this study. The water balance values for dates corresponding to the available observations in the historic period were used to calibrate the station model (see Section 4.3.3). For dates within the recent period for each reforecast, the water balance values were used in the post-processing (see Section 4.3.4.1). Finally, the water balance values corresponding to the 15 day lead-time of each reforecast was used to estimate the average meteorological error of each station (see Section 4.5.2.1).

4.4.3 Evaluation metrics

The evaluation of the post-processing method is performed by comparing the skill of the raw forecasts with the corresponding post-processed forecasts. Since the aim of the post-processing is to create a more accurate representation of the observation probability distribution all metrics use observations as the “truth" values. As mentioned in Section 4.2, the output from the post-processing method evaluated here is expressed operationally in the ‘real-time hydrograph’ product, an example of which is shown in Fig. 4.4. Therefore, the evaluation will consider four main features of forecast hydrographs.

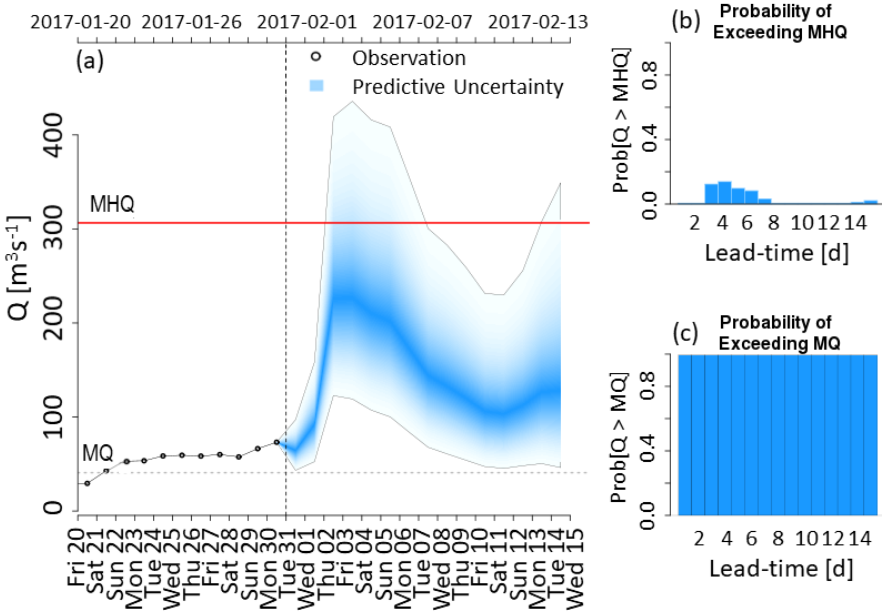

Figure 4.4: Example of the ‘real-time hydrograph’ product for the station in Brehy, Slovakia on 31 January 2017. (a) Probability distribution of the post-processed forecast. The darkest shade of blue indicates the forecast median (50th percentile) with each consecutive shade indicating a percentile difference such that the extent of the total predictive uncertainty is shown by the shaded region. Solid grey lines indicate the upper (99th percentile) and the lower (1st percentile) bounds of the forecast probability distribution. The red line shows the mean annual maximum (MHQ) threshold, and the dashed grey line shows the mean flow (MQ) threshold. Black circles represent observations positioned at the centre of the timestep over which they are calculated. (b) Bar chart showing the probability of the discharge exceeding the MHQ threshold at each lead-time. (c) Bar chart showing the probability of the discharge exceeding the MQ threshold at each lead-time.

4.4.3.1 Forecast median

In the real-time hydrograph the darkest shade of blue indicates the forecast median making it the easiest and most obvious single-valued summary of the full probabilistic forecast for end-users. The ensemble median of the raw forecasts is used in this evaluation because operationally the ensemble forecasts are often represented by boxplots where the median at each timestep is shown.

The skill of the forecast median is evaluated using the modified Kling-Gupta Efficiency score (KGE', Kling et al., 2012; Gupta et al., 2009). The forecast median is determined for the post-processed forecasts by extracting the 50th percentile of the probability distribution at each lead-time. For the raw forecasts the ensemble members are sorted by discharge value and the middle (i.e. 6th) member is chosen. This is done separately for each lead-time so the overall trajectory may not follow any single member. The forecast median is denoted x to distinguish it from the full forecast, $\mathbf{x_t}(t+1:t+T)$. The KGE' is calculated as,

$$KGE' = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \tag{4.29}$$

with

$$\beta = \frac{\overline{x}}{\overline{y}}, \tag{4.30}$$

and

$$\gamma = \frac{\sigma_x/\overline{x}}{\sigma_y/\overline{y}}, \tag{4.31}$$

where r is the Pearson's correlation coefficient, $\overline{x}$ and $\overline{y}$ are the mean values of the forecast median and the observations, respectively, and σ_x and σ_y are their standard deviations. The correlation, r, measures the linear relationship between the forecast median and the observations indicating the ability of the forecasts to describe the temporal fluctuations in the observations. The bias ratio, β, indicates if the forecast consistently under or over-predicts the observations. The variability ratio, γ, measures how well the forecast can capture the variability of the discharge magnitude. The KGE' is calculated separately for each lead-time. The KGE' ranges from $-\infty$ to 1, r ranges from -1 to 1, and both β and γ range from $-\infty$ to

∞. A perfect score for the KGE' and each of the components is 1.

4.4.3.2 Peak discharge

The timing of the peak discharge is an important variable of flood forecasts. The Peak-Time Error (PTE) is used to evaluate the effect of post-processing on the timing of the peak within the forecast. The PTE requires a single-valued forecast trajectory. For the reasons stated in Section 4.4.3.1, the PTE is calculated using the forecast median, x. Peaks are defined as the maximum forecast value and the PTE is calculated for forecasts where this peak exceeds the 90th percentile discharge threshold of the station. This threshold is calculated using the full observational record for the station. The PTE is calculated as,

$$PTE = t_n^x - t_n^y \tag{4.32}$$

where t_n^x is the timestep of the maximum of the forecast median for the n-th forecast and t_n^y is the timestep of the maximum observed value in the same forecast period. A perfect score is $PTE = 0$. A negative PTE value indicates the peak is forecast too early and a positive PTE value indicates the peak is forecast too late. As the maximum lead-time is 15 days, the maximum value of the PTE is 14 days and the minimum value is -14 days.

4.4.3.3 Threshold exceedance

Two discharge thresholds are shown in the real-time hydrograph: the mean discharge (MQ) and the mean annual maximum discharge (MHQ). Both thresholds are determined using the observations from the historic period. For the post-processed forecasts, the probability of exceedance of the MQ threshold, $PoE(MQ)$, is calculated such that

$$PoE(MQ) = 1 - F_{\tilde{x}}(MQ) \tag{4.33}$$

where $F_{\tilde{x}}(MQ)$ is the value of the forecast CDF at the MQ threshold. The CDF is assumed to be linear between any two percentiles. The same method is applied for the MHQ threshold.

For the ensemble forecast, each ensemble member above the threshold contributes one eleventh to the probability of the threshold being exceeded. The probability of the threshold being exceeded is calculated separately for each lead-time.

The Relative Operating Characteristic (ROC) score and ROC diagram (Mason and Graham, 1999) are used to evaluate the potential usefulness of the forecasts with respect to these two thresholds. The ROC diagram shows the probability of detection vs the false alarm rate for alert trigger thresholds from 0.05 to 0.95 in increments of 0.1. The ROC score is the area below this curve with a ROC score of less than 0.5 indicating a forecast with less skill than a climatological forecast. As discharge values of above the MHQ threshold are rare, all stations are combined and lead-times are combined into 3 groups; 1-5 days, 6-10 days, and 11-15 days. Since the reforecasts are only produced on Monday and Thursdays, an event that occurs on a Saturday can only be forecasted at lead-times of 2, 5, 9, and 12 days. Using 5-day groupings of lead-times guarantees that each group is evaluated against each event at least once but allows the usefulness of the forecasts to be compared at different lead-times. A perfect forecasting system would have a ROC score of 1.

Reliability diagrams are used to evaluate the reliability of the forecast in predicting the exceedance of the two thresholds. Reliability diagrams show the observed frequency vs the forecast probability for bins of width 0.1 from 0.05 to 0.95. A perfectly reliable forecast would follow the one-to-one diagonal on a reliability diagram. The same combination of stations and lead-times is used as with the ROC diagrams.

4.4.3.4 Full probability distribution

A commonly used metric to evaluate overall performance of a probabilistic or ensemble forecast is the Continuous Ranked Probability Score (CRPS, Hersbach, 2000). The CRPS measures the difference between the CDF of the forecast and that of the observation and is defined as

$$CRPS\left(F_{\tilde{x}}, y\right) = \int_{-\infty}^{\infty} \left(F_{\tilde{x}}\left(\tilde{\eta}\right) - \theta\left(\tilde{\eta} - y\right)\right)^2 d\tilde{\eta} \tag{4.34}$$

where $F_{\tilde{x}}$ represents the CDF of the forecast and $\theta(\tilde{\eta} - y)$ is the step function (Abramowitz and Stegun, 1972), defined such that

$$\theta(\tilde{\eta}) = \begin{cases} 0 & \tilde{\eta} < 0 \\ 1 & \tilde{\eta} \geq 0 \end{cases} \tag{4.35}$$

and represents the CDF of the observation, y. The post-processed forecasts are defined via their percentiles, therefore by assuming the CDF is linear between percentiles the CRPS can be calculated directly. The empirical CDF of the raw forecasts, defined via point statistics, is used and the CRPS is calculated using a computationally efficient form (Jordan et al., 2019, Equation 3). It should be noted that the error in the calculation of the CRPS for the raw ensemble forecasts is likely to be large compared to that of the post-processed forecasts because of the limited number of ensemble members (Zamo and Naveau, 2018). However, as this evaluation is of the post-processing method no corrections to account for the ensemble size are made (e.g. Ferro et al., 2008) since the impact of the post-processing would be difficult to differentiate from that of the CRPS correction. The CRPS ranges from a perfect score of 0 to ∞.

4.4.3.5 Comparison

For some of the metrics described in Sects. 4.4.3.1-4.4.3.4, the impact of post-processing is shown using the respective skill score, SS, with the raw forecast as the benchmark,

$$SS = \frac{S_{pp} - S_{raw}}{S_{perf} - S_{raw}} \tag{4.36}$$

where S_{pp} and S_{raw} are the scores for the post-processed forecast and the raw forecast, respectively, and S_{perf} is the value of the score for a perfect forecast. The skill score gives the fraction of the gain in skill required for the raw forecast to become a perfect forecast that is provided by the post-processing. A value $SS < 0$ means the forecast has been degraded by the post-processing, a value of $SS > 0$ indicates that the forecast has been improved by the post-processing, and a value of $SS = 1$ means that the post-processed forecast is perfect.

Henceforth, the skill score for a metric is denoted by adding 'SS' to the metric name.

4.5 Results and Discussion

4.5.1 Performance of the post-processing method

This section focuses on the overall impact of post-processing at all 522 of the evaluated stations across the EFAS domain and aims to address the research question: *Does the post-processing method provide improved forecasts?*

4.5.1.1 Forecast median

The modified Kling-Gupta Efficiency Skill Score (KGESS) is used to evaluate the impact of post-processing on the forecast median (see Section 4.4.3.1). Figure 4.5a shows the KGESS for all stations at every other lead-time such that each boxen plot (also known as letter-value plots, Hofmann et al., 2017) contains 522 values, one for each station. For each lead-time the central black line shows the median KGESS value. The inner box (the widest box) represents the interquartile range and contains 50% of the data points. Each subsequent layer of boxes splits the remaining data points in half such that the second layer of boxes are bounded by the 12.5th and 87.5th percentiles and contains 25% of the data points. The outliers represent a total of 2% of the most extreme data points. The lower panels of Fig. 4.5 show the three components of the KGE' (b: correlation, c: bias ratio, d: variability ratio) for lead-times of 3, 6, 10, and 15 days for all stations for both the raw forecasts (orange) and the post-processed forecasts (purple). The chosen lead-times are representative of the results.

Figure 4.5a shows that most stations have positive KGESS values at all lead-times indicating that post-processing increases the skill of the forecast median. However, the magnitude of this improvement decreases at longer lead-times with most of the reduction occurring in the first 7 days. The proportion of stations for which post-processing degrades the forecast median increases with lead-time However, the lowest KGESS values become less extreme (i.e. not as negative). This increase in the KGESS of the most degraded stations is due to a

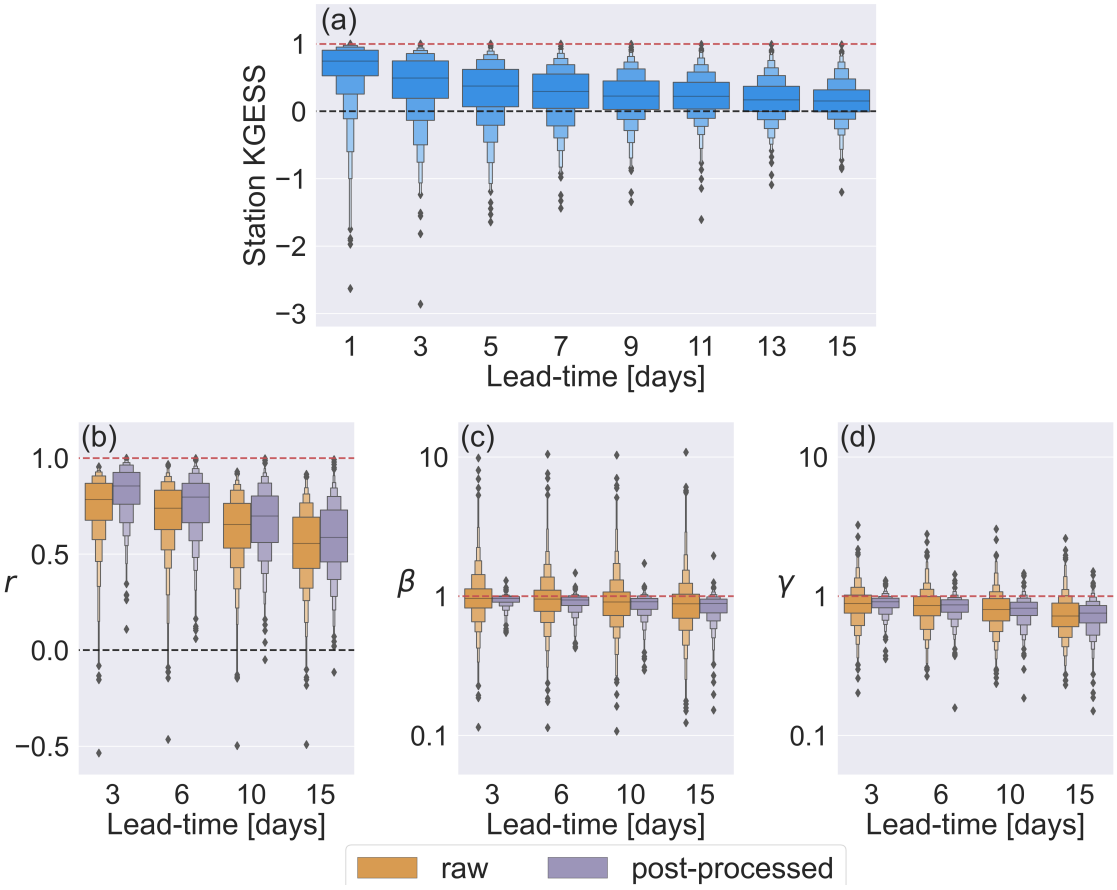

Figure 4.5: Comparison of the raw and post-processed forecast medians. (a) The Kling-Gupta Efficiency Skill Score (KGESS) for the forecast medians at all 522 stations for every other lead-time. Red dashed line shows the perfect score of KGESS = 1. Black dashed lines show KGESS value of 0. KGESS > 0 indicates the skill of the forecast median is improved by post-processing. KGESS < 0 indicates the skill of the forecast median is degraded by post-processing. The three components of the KGE': (b) Correlation component, r. Black dashed line shows r = 0. (c) Bias ratio component, β. (d). Variability ratio component, γ. Red dashed lines show the perfect scores of 1 for all components. Both (c) and (d) have logarithmic y-axes.

decrease at longer lead-times in the skill of the raw forecast (used as the benchmark for the skill score) rather than an increase in the skill of the post-processed forecasts. This shows that the effect of naïve skill on the results should be considered; however, as the aim is to evaluate the impact of post-processing, it is appropriate to use the raw forecasts as the benchmark (Pappenberger et al., 2015b).

Figure 4.5b shows that post-processing improves the correlation between the forecast median and the observations for most stations, particularly at short lead-times The impact of post-processing on the correlation component of the KGE' varies greatly between stations.

Notably, the flashiness of the catchment and whether or not the river is regulated can affect the performance of the post-processing (see Section 4.5.2.2). Additionally, the quality and length of the calibration time-series also have an effect (see Section 4.5.2.3).

Figure 4.5c shows the bias ratio, β, which indicates if on average the forecasts over or under-estimate the discharge at a station. In the hydrological uncertainty estimation part of the online correction (see Section 4.3.4.1) the mean of the hydrological uncertainty distribution is calculated in Eq. (4.14) as the mean flow of the observed time-series from the historic period (term 1) plus an amount dependent on the discharge values in the recent period (term 2). Therefore, assuming the mean flow does not change between the calibration (historic) and evaluation periods, any consistent biases in the hydrological model climatology should be corrected.

Figure 4.5c shows the variability ratio, γ, which indicates if the forecast median is able to capture the variability of the flow. In general, the post-processing method does reduces the bias in the forecast median. For raw forecasts, the β-values range from approximately 10 (an over-estimation by an order of magnitude) to 0.1 (an under-estimation of an order of magnitude) For the post-processed forecasts the β-values are more tightly clustered around the perfect value of $\beta = 1$. The largest improvements to the β-values are for stations where the flow is under-estimated by the raw forecasts. Some stations with raw β-values of greater than 1 are over-corrected such that the post-processed forecasts have β-values of less than 1. This is supported by the similarity of the median β-values for the raw and post-processed forecasts despite the decrease in the range of values. For stations where the over-estimation by the raw forecast is relatively small the over-correction can result in the post-processed forecasts being more biased than the raw forecasts. The over-correction is generally due to the under-estimation of high flows (see discussion on the third component of the KGE', the variability ratio) which results in an under-estimation of the average flow and hence a β-value of less than 1.

There is a small decrease in the β-values at longer lead-times for both the raw and post-processed forecasts. This is primarily caused by an increase in the under-estimation of high flows at longer lead-times as the skill of the forecast decreases. However, for some stations

the drift in β-values at longer lead-times is also caused by nonstationarity of the discharge distribution. A change in the discharge distribution from that of the calibration period means the hydrological uncertainty is calculated using an inaccurate climatological mean (term 1 of Eq. (4.14)) . The impact of the discharge values from the recent period (term 2 of Eq. (4.14)) decreases with lead-time because the autocorrelation weakens. Therefore, any errors in the climatological forecast are more pronounced at longer lead-times.

Figure 4.5d shows that the variability of the flow tends to be under-estimated by the raw forecast (γ less than 1). The under-estimation is because the magnitudes of the peaks relative to the mean flow are not predicted accurately particularly at longer lead-times. This decrease in γ-values at longer lead-times is also visible for the post-processed forecasts. However, at all lead-times most stations show an improvement after post-processing (i.e. have a value of γ closer to 1). Stations where the raw forecast over-estimates the variability (γ above 1) are more likely to have the variability corrected by post-processing particularly at longer lead-times.

The two factors impacting the ability of the post-processed forecasts to capture the variability of the flow are 1) the level of indication of the upcoming flow by the discharge values in the recent period, and 2) the spread of the raw forecast. In the Kalman filter when the hydrological uncertainty distribution and the meteorological uncertainty distribution are combined (see Section 4.3.4.3) the weighting of each distribution is dependent on their relative spreads. The spread of the hydrological uncertainty is impacted by the discharge values in the recent period. Due to the skewedness of discharge distributions, the climatological forecasts tend to have a low probability of high flows. If the recent discharge values show no indication of an upcoming high flow (i.e. no increase in discharge), the low probability of high flows is reenforced . This decreases the spread of the hydrological uncertainty distribution and increases its weight within in the Kalman filter.

The meteorological uncertainty distribution is the spread corrected raw forecast and includes the variability due to the meteorological forcings. For floods with meteorological drivers, if the magnitude of the peaks is under-predicted by the raw forecasts then the post-processed forecasts are also likely to under-predict the magnitude of the peaks. Alternatively,

if the raw forecast is unconfident in the prediction of a peak (e.g. only a couple of members predict a peak) then it may not have a sufficient impact within the Kalman filter and the post-processed forecast may not predict the peak regardless of the accuracy of the ensemble members that do predict the peak. The impact of the spread correction is discussed further in Section 4.5.2.1.

The ensemble mean is another commonly used single-valued summary of an ensemble forecast (Gneiting, 2011). Although the comparison presented here uses the ensemble median we also show the three components of the KGE' for the ensemble mean in Fig. 1 of the supplementary material. The ensemble means (see Fig. 1b of the supplementary material) do not show the general drift in β-values with increasing lead-time that is discussed above for both the ensemble median and post-processed forecasts. However, the range of β-values is similarly large for both the ensemble median and the ensemble mean. In terms of the correlation coefficient and the variability ratio the ensemble mean performs similarly or worse than the ensemble median (see Fig. 1a and 1c of the supplementary material, respectively).

4.5.1.2 Timing of the peak discharge

To evaluate the impact of post-processing on the ability of the forecast to predict the timing of the peak flow accurately the Peak-Time Error (PTE, see Section 4.4.3.2) is used. The aim of this assessment is to see how well the forecast is able to identify the time within the forecast period with highest flow and therefore greatest hazard. A PTE of less than 0 indicates the peak is predicted too early whereas a PTE of greater than 0 indicates the peak was predicted too late. Figure 4.6 shows the distribution of the PTE values for both the post-processed and raw forecasts for all forecasts where the maximum forecast value exceeds the 90th percentile . The forecasts are split into three categories dependent on the lead-time at which the forecast maximum occurs. Therefore, the distributions shown in each panel are truncated at different values of PTE. For example, if the forecast maximum occurs on a lead-time between 1 and 5 days, it can at most be predicted 5 days early.

Approximately 40% of the forecast medians of the raw forecasts have no error in the timing of the peak for peaks that occur within lead-times of 1 to 5 days. This drops to 37%

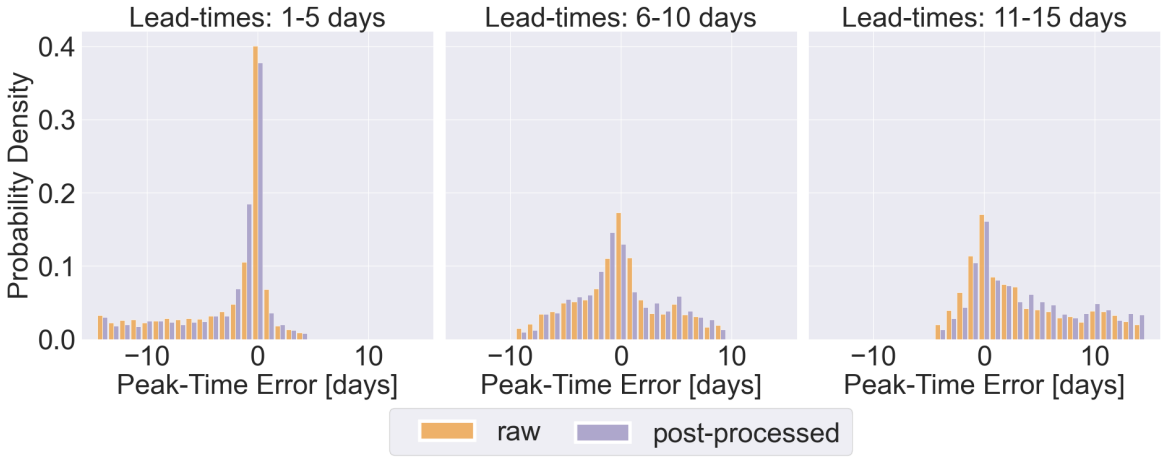

Figure 4.6: Histograms showing the probability distribution of Peak-Time Errors for all forecasts where the maximum observation is above the 90% percentile for the station (26807 forecasts) for raw forecasts (orange) and post-processed forecasts (purple). (a) Maximum observations occurs on lead-times of 1 to 5 days. (b) Maximum observations occurs on lead-times of 6 to 10 days. (c) Maximum observations occurs on lead-times of 10 to 15 days.

for post-processed forecasts. Both sets of forecasts have approximately 60% of forecasts with timing errors of 1 day or less. However, the post-processed forecasts are more likely to predict the peak too early. For maximum forecast values occurring on lead-times of 6 to 10 days, the post-processed forecasts still tend to predict peaks earlier than the raw forecasts. However, for maximum forecast values occurring on lead-times of 11 to 15 days the post-processed forecasts are more likely to predict the peaks several days too late. This suggests that for floods forecast at longer lead-times by the post-processing forecasts should be considered carefully.

Overall the impact of post-processing is small but tends towards the early prediction of the peak flow for short lead-times and late peak predictions for longer lead-times. However, there are three main limitations with this analysis. The first is that both sets of forecasts are probabilistic and therefore the median may not provide an adequate summary of the forecast. Secondly, the evaluation here is forecast-based rather than peak-based in that the focus is the timing of the highest discharge value in the forecast within the forecast period and not the lead-time at which a specific peak is predicted accurately. This was intentional as the twice-weekly production of the reforecasts means that a specific peak does not occur at each lead-time. Finally, the combination of forecasts at all stations means the relationship between

the runoff generating mechanisms and the PTE cannot be assessed.

4.5.1.3 Threshold exceedance

The Relative Operating Characteristic (ROC) diagrams for the mean flow (MQ) and the mean annual maximum flow (MHQ) thresholds (see Section 4.4.3.3) are shown in Fig. 4.7. The diagrams show the probability of detection against the false alarm rate for varying decision thresholds. The forecast period is split into three lead-time groups: 1-5 days, 6-10 days, and 11-15 days (see Section 4.4.3.3). The ROC scores for the MQ and MHQ thresholds are given in Table 4.1 for each lead-time group for the raw and post-processed forecasts, along with the corresponding skill scores (ROCSS). Both the raw and post-processed forecasts have ROC scores greater than 0.5 showing that they are more skilful than a climatological forecast.

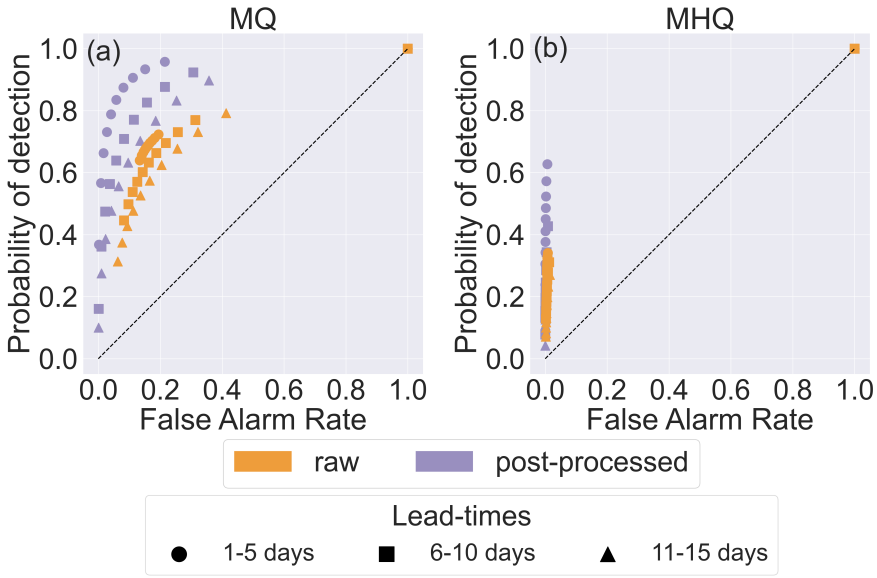

Figure 4.7: Relative Operating Characteristic diagrams for (a) the MQ threshold (118,888 observations above MQ), and (b) the MHQ threshold (2783 observations above MHQ). All stations are combined and groupings of lead-times are used (see Section 4.4.3.3.)

The spread of the raw forecasts is small at short lead-times. This is shown by the overlapping of the points in Fig. 4.7a for lead-times of 1-5 days (orange circles). The similarity of the points indicates that the decision thresholds are usually triggered simultaneously and therefore that the forecast distribution is narrow. The spread of the forecast increases with lead-time as the ensemble of meteorological forcings increases the uncertainty in the forecasts.

	MQ			MHQ		
Lead-time	ROC_{raw}	ROC_{pp}	ROCSS	ROC_{raw}	ROC_{pp}	ROCSS
1-5 days	0.78	0.96	0.87	0.68	0.83	0.48
6-10 days	0.78	0.91	0.56	0.68	0.74	0.20
11-15 days	0.76	0.87	0.45	0.67	0.69	0.08

Table 4.1: Relative Operating Characteristic Scores (ROCS) and corresponding skill scores (ROCSS) for the raw and post-processed (pp) forecasts for lead-times of 1-5 days, 6-10 days, and 10-15 days for the mean flow threshold (MQ) and the mean annual maximum threshold (MHQ).

Although the skill of the forecast median decreases with lead-time (see Section 4.5.1.1), the introduction of the meteorological uncertainty means the usefulness of the raw forecasts is similar for lead-times of 1-5 days and 6-10 days. This is shown by similarity of the ROC scores for these lead-time groups for the raw forecast.

Post-processing also accounts for the hydrological uncertainty allowing for a more complete representation of the total predictive uncertainty. In addition, as shown in Fig. 4.5c, post-processing bias corrects the forecast relatively well at short lead-times. The combination of spread and bias correction leads to an increase in the probability of detection for all but the highest decision thresholds and a decrease in the false alarm rate for almost all decision thresholds and lead-times. The added reliability gained from post-processing decreases with lead-time. The ROCSS for lead-times of 1-5 days at the MQ level is 0.8 but is only 0.45 for lead-times of 11-15 days.

The ROC diagram for the MHQ threshold (Fig. 4.7b) shows that the raw forecasts tend to cautiously predict high flows with the forecast much more likely to miss a flood than to issue a false alarm even for the lowest decision threshold. There is less improvement from post-processing than for the MQ threshold with the ROCSS for the MHQ threshold only reaching 0.48 for 1-5 days lead-time. For the MHQ threshold, the post-processing increases the probability of detection and decreases the false alarm rate at short lead-times. At longer lead-times the false alarm rate is still decreased by post-processing, but the probability of detection is also decreased for the largest decision thresholds. This reluctance to forecast larger probabilities also occurs with the MQ threshold and is due to the interaction between the hydrological and meteorological uncertainty in the Kalman filter discussed in Section

4.5.1.1.

Figure 4.8 shows reliability diagrams for the MQ and MHQ thresholds. For the MQ threshold (Fig. 4.8a) the raw forecasts are over-confident leading to under-estimation of low probabilities and over-estimation of high probabilities. The post-processed forecasts are more reliable but also tend to under-estimate low probabilities. The raw forecasts increase in reliability with lead-time whereas the reliability of the post-processed forecasts decreases. This is also true for the MHQ threshold.

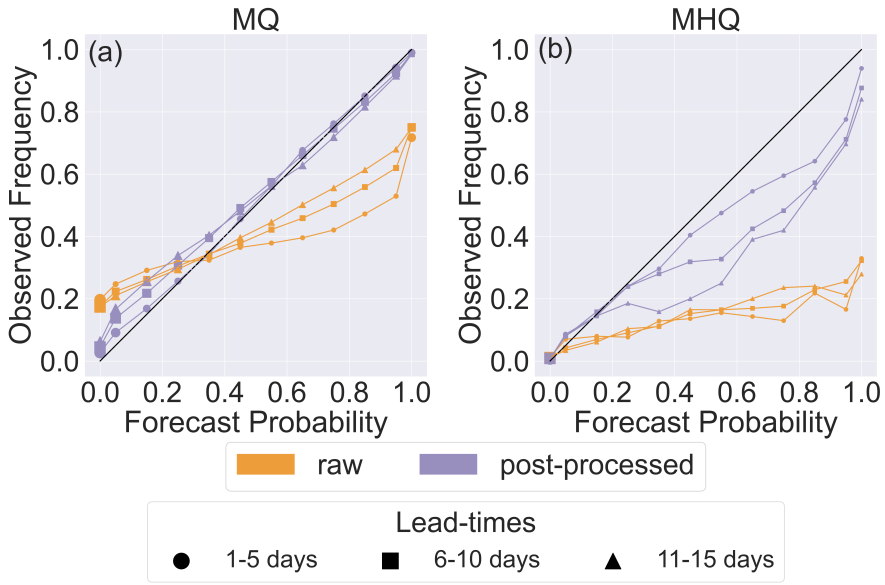

Figure 4.8: Reliability diagrams for (a) the mean flow threshold (MQ), and (b) the mean annual maximum flow (MHQ). All stations are combined and groupings of lead-times are used (see Section 4.4.3.3.)

Both sets of forecasts are consistently below the diagonal in the MHQ reliability diagram (Fig. 4.8b) indicating unconditional biases. However, the post-processed forecasts have smaller biases consistent with the results discussed in Section 4.5.1.1. In addition, the raw forecast shows relatively poor resolution with events occurring at approximately the same frequency regardless of the forecast probability.

The distribution of forecasts (shown by marker size) is more uniform for the post-processed forecasts particularly at shorter lead-times. Since the ensemble reforecasts evaluated have 11 members and the operational forecasts have 73 members, the distribution for operational raw forecasts is expected to be slightly more even as the additional members allow for greater

gradation in the probability distribution. The distribution of forecasts is skewed towards low probabilities showing similarly to the ROC diagrams (Fig. 4.7) that both sets of forecasts tend to cautiously forecast flows exceeding the MHQ threshold.

4.5.1.4 Overall skill

The Continuous Ranked Probability Skill Score (CRPSS) is used to evaluate the impact of post-processing on the overall skill of the probability distribution of the forecasts. Figure 4.9 shows the CRPSS for each station at lead-times of 3, 6, 10, and 15 days. Stations that are degraded by post-processing (CRPSS < 0) are circled in red. Stations that show a large increase in skill after post-processing (CRPSS > 0.9) are circled in cyan.

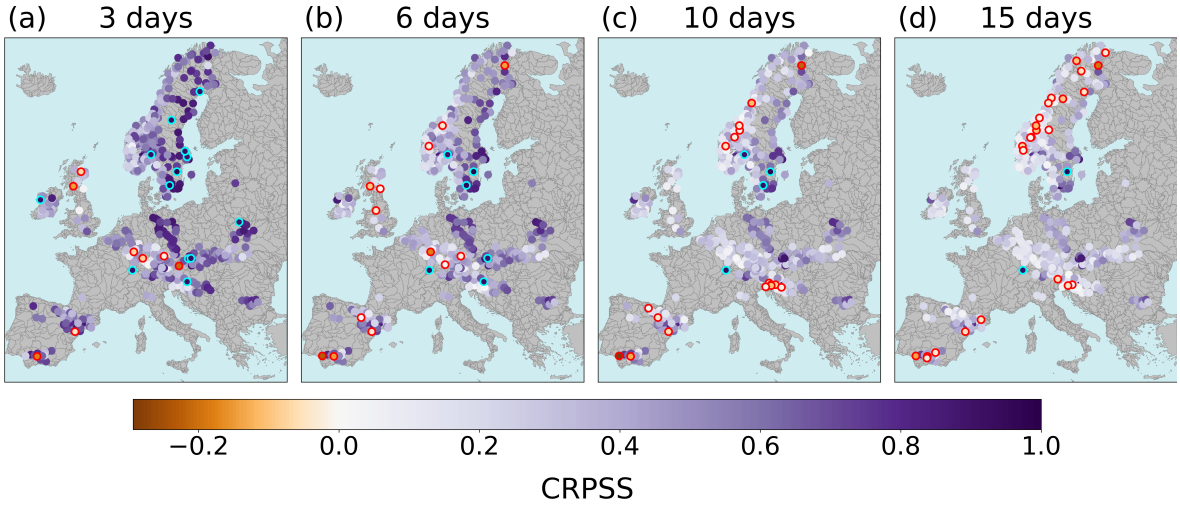

Figure 4.9: The Continuous Ranked Probability Skill Score (CRPSS) for all 522 stations for lead-times of 3, 6, 10, and 15 days. CRPSS values below 0 indicate the forecast probability distribution is on average less skillful after post-processing and values above 0 indicate added skill after post-processing. Markers are outlined in red if the CRPSS is below 0 and in cyan if the CRPSS is above 0.9.

As was seen with the KGESS for the forecast median, there is a decrease in the improvement offered by post-processing at longer lead-times. This can be seen in Fig. 4.9, by the gradual change from dark purple to light purple/white values for panels (a) to (d). It is also shown in the increase of red circles and the decrease of cyan circles. Approximately 55% of stations have a CRPSS of above 0.5 at a lead-time of 3 days and this decreases to 10% by a lead-time of 15 days. At a lead-time of 3 days, 8 stations are degraded by the post-processing and 13 stations have a CRPSS greater than 0.9. By a lead-time of 15 days these change

to 24 degraded stations and only 2 stations with CRPSS values greater than 0.9. Many of the stations that are improved significantly have large hydrological biases. For example, one of the most improved stations at a lead-time of 15 days is in Rheinweiler, Germany (see Fig. 4.3) which has a large bias in the hydrological model output due to limitations in the representation of the drainage network in the model domain. The post-processing method can account for these biases (see Section 4.5.1.1) resulting in CRPSS values greater than 0.9 at all lead-times.

The lack of clustering of the stations with CRPSS values above 0.9 suggest that the magnitudes of the largest corrections are due to station dependent characteristic. On the other hand, the degraded stations at a lead-time of 3 days appear to cluster in three loose regions. In all three regions the degradation is due to high short-duration peaks being captured better by the raw forecasts than the post-processed forecasts. At longer lead-times the Spanish catchments are still degraded but the Scottish stations are not. As discussed in Section 4.5.1.1 for the lowest KGESS values, this is due to a decrease in the skill of the raw forecasts. The degraded stations at lead-times of 10 and 15 days cluster in Spain, around the Kjolen Mountains, and in the Sava catchment. The poorly post-processed forecasts in the Sava catchment are downstream of a reservoir the impact of which is discussed in Section 4.5.2.2.

Comparing the CRPSS values in Fig. 4.9 with the raw CRPS values shown in Fig. 4.3 shows similarities in the spatial pattern of the raw forecast skill and the spatial pattern of the magnitude of improvement due to post-processing. In general, stations with low CRPS scores (high skill) for the raw forecasts are improved the most by post-processing. For example, the west coast of the Scandinavian Peninsula has a lower raw skill in general and the level of improvement is also lower than that of the east coast. However, there are some anomalies to this pattern. For example, the station in Cong Weir, Ireland has a relatively low raw forecast skill compared to surrounding catchments due to regulation of the streamflow but has a high CRPSS value at all lead-times. Additionally, whilst stations on the River Rhine and the River Oder have similar raw CRPS values the River Oder is improved more by post-processing. This suggests that post-processing is more effective at dealing with certain types of error and

therefore that the benefit of post-processing is catchment dependent. This is discussed in Section 4.5.2.

As mentioned, many of the stations with CRPSS values below 0 at short lead-times are degraded due to peak flows being better predicted by the raw forecasts. Therefore, the skill of the forecast at different flow levels is evaluated. Figure 4.10 shows the distribution of CRPSS values for all stations evaluated over the 4 quartiles of discharge (Q1 lower quartile to Q4 upper quartile) such that each boxenplot contains 522 CRPSS values, one for each station evaluated over approximately 52 forecasts. Only lead-times of 3, 6, 10, and 15 days are shown but these lead-times are representative of the results at similar lead-times.

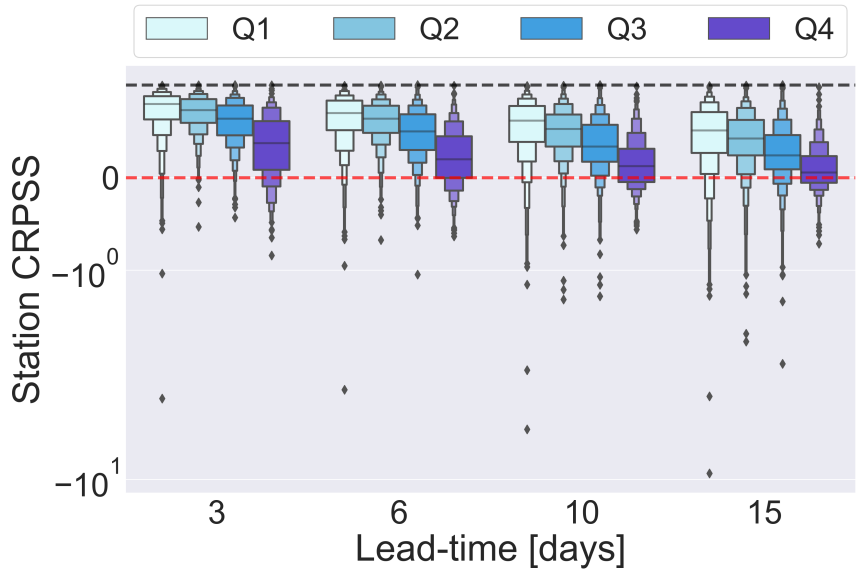

Figure 4.10: The Continuous Ranked Probability Skill Score (CRPSS) for all 522 stations calculated over the forecasts (approximately 52 forecasts) with flow values in the lowest quartile (Q1) to the highest quartile (Q4). CRPSS values below 0 indicate the forecast probability distribution is on average less skillful after post-processing and values above 0 indicate added skill after post-processing.A log-scale is used on the y-axis.

The improvement for all 4 quartiles decrease with lead-time as has been seen previously in Fig. 4.5 and Fig. 4.9. The improvement from post-processing is smaller for higher flows. However, the majority of stations are still improved for these high flows with over 60% of stations being improved for discharge values in Q4 at a lead-time of 15 days. The high flows are often under-predicted by both sets of forecasts. As discussed in Section 4.5.1.1, the ability of the post-processed forecasts to capture the magnitude of peaks is often determined by the

relative spread of the hydrological and meteorological uncertainty distributions. Although Q4 is the category with the greatest number of degraded stations (CRPSS < 0), some stations are degraded more (have a lower CRPSS value) for discharge values in Q1. This is mainly due to the larger proportional errors for lower flows.

4.5.2 What impacts the performance of the post-processing method?

In the previous section the impact of post-processing was shown to vary greatly between stations. The following sections investigate the factors that influence the effect of the post-processing method. The CRPSS is used in this analysis as it provides an assessment of the improvement or degradation to the overall skill of the probabilistic forecast.

To aid the discussion of the key results some stations are highlighted. See Fig. 4.3 for the locations of the stations. Figure 4.11 shows the observed time-series (solid black line) for half the evaluation period (1 October 2017 to 30 September 2018) for six example stations; (a) Daldowie, Scotland. (b) Nytorp, Sweden. (c) Svarttjørnbekken, Norway. (d) Daugavpils, Latvia. (e) Porttipahta, Finland. (f) Montañana, Spain. The forecast median of the raw forecasts (oranges) and the post-processed forecasts (purples) are also plotted for lead-times of 3 days (circles), 6 days (crosses), and 15 days (triangles). These stations are discussed throughout Section 4.5.2 and were chosen as they allow some of the impacts of the post-processing to be visualised. Table 4.2 summaries the key results that each of the example stations highlight and all results are summarised in Section 4.6.

4.5.2.1 Type of uncertainty

This section looks at how meteorological and hydrological uncertainties affect the performance of the post-processing method. As mentioned in Section 4.1, the term meteorological uncertainties is used to refer to the uncertainty in the streamflow forecasts due to the error and uncertainty in the meteorological forcings, and not the error in the meteorological forecasts themselves. The magnitude of meteorological uncertainty is represented here by the CRPS of the raw ensemble forecast at each lead-time respectively. To remove the uncertainty due

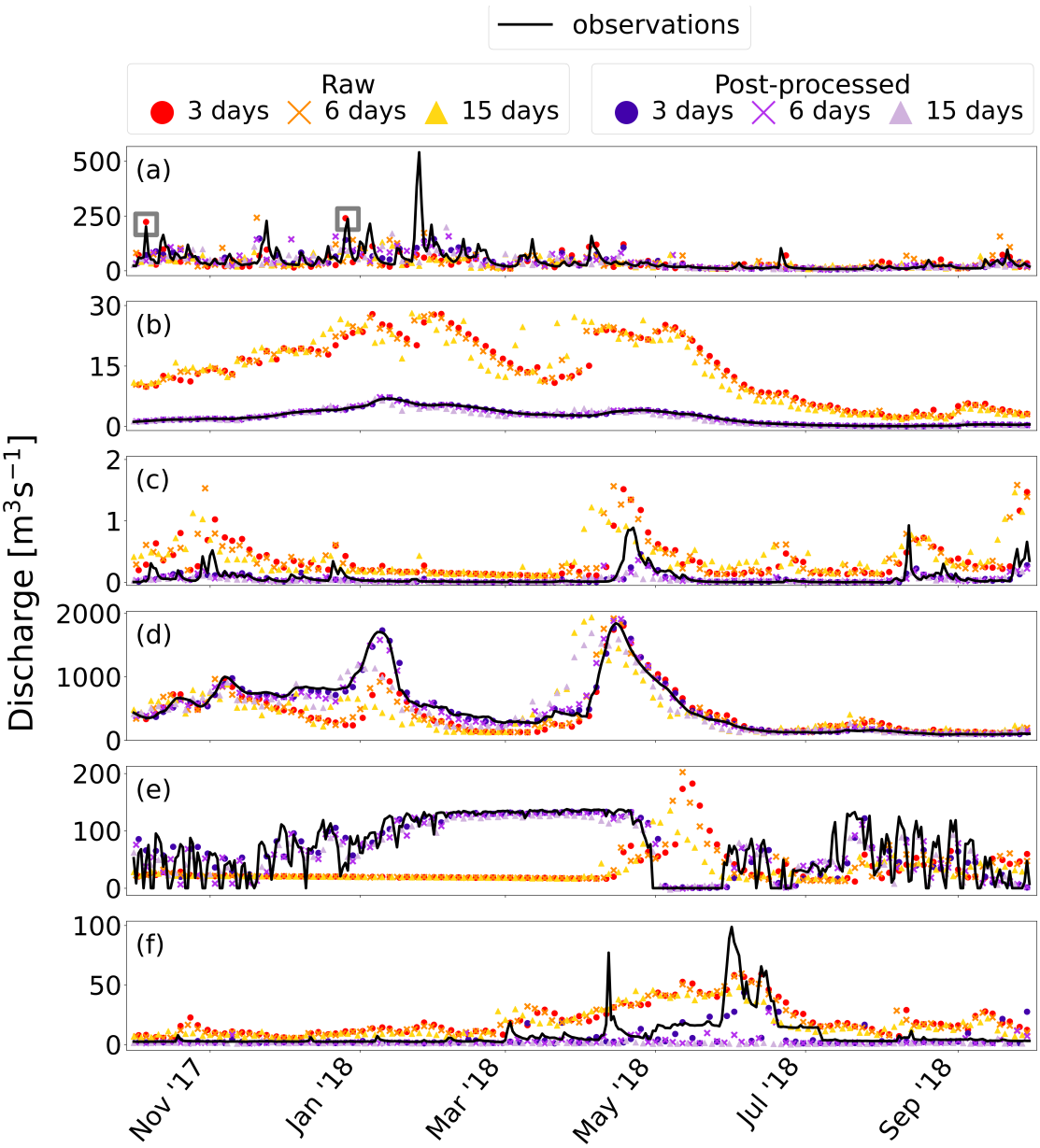

Figure 4.11: Observation time-series for one year of the evaluation period from October 2017 to October 2018 for 6 example stations. The forecast medians of the raw and post-processed forecasts are shown for lead-times of 3, 6, and 15 days. (a) Daldowie, Scotland. (b) Nytorp, Sweden. (c) Svarttjørnbekken, Norway. (d) Daugavpils, Latvia. (e) Porttipahta, Finland. (f) Montañana, Spain.

to the hydrological model, the water balance simulation is used as the “truth" value in the calculation of the CRPS, replacing the value of the observation, y, in Eq. (4.34). As both the forecast and the water balance simulation are produced using the same hydrological model, and the water balance simulation provides the initial conditions for the reforecasts, the only remaining uncertainty is from the forcings. The errors of the meteorological observations

Panel	Station	Description of key results	Section
(a)	Daldowie, Scotland	• Meteorological errors are not corrected as well as hydrological errors. • Poor post-processing of peaks for flashy catchments.	4.5.2.1, 4.5.2.2
(b)	Nytorp, Sweden	• Large biases due to limitations of the drainage network are well corrected.	4.5.2.1, 4.5.2.2
(c)	Svarttjørnbekken, Norway	• Post-processing is beneficial for stations where the hydrological model is uncalibrated.	4.5.2.2
(d)	Daugavpils, Latvia	• Slowly responding catchments benefit from post-processing the most. • Post-processing can account for poor modelling of slow hydrological processes such as snowmelt.	4.5.2.2
(e)	Porttipahta, Finland	• Regulated catchments benefit from post-processing.	4.5.2.2
(f)	Montañana, Spain	• The quality of the calibration time-series is more important than the length of the time-series.	4.5.2.3

Table 4.2: Key results and the section that provide more information for each of the six stations used as examples and for which time-series are shown in Fig. 4.11

used to create the water balance simulation are considered negligible compared to those of the meteorological forecasts. The magnitude of the hydrological uncertainty is represented by the CRPS of the water balance simulation, with the observations used as the “truth" values, at each lead-time respectively. As both these values are deterministic the CRPS is equivalent to the absolute error between the two values. Both metrics, for the meteorological and hydrological uncertainties, are averaged over all 208 forecasts for each station. So that the errors are comparable between catchments they are calculated in terms of specific discharge (mmd^{-1}) instead of discharge (m^3s^{-1}).

Figure 4.12 shows density plots of the CRPSS values for all stations vs the hydrological

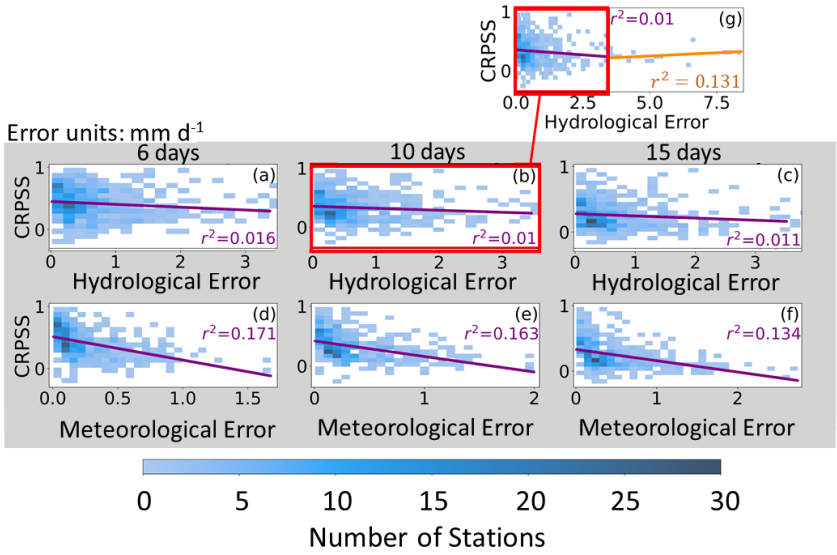

Figure 4.12: Density plots showing the station CRPSS for lead-times of 6 days (a and d), 10 days (b and e), and 15 days (c and f), against hydrological error (a-c) and meteorological error (d-f). The largest 15 hydrological errors are excluded from panels a-f. See Section 4.5.2.1 for an explanation of the metrics used to represent the hydrological and meteorological errors. Purple lines show the line-of-best-fit calculated using linear regression and the associated r^2 are given within each panel. (g) The CRPSS against hydrological error including the 15 largest hydrological errors for a lead-time of 10 days. The orange line shows the line-of-best-fit for the station with large hydrological errors.

errors (a-c) and meteorological errors (d-f) for lead-times of 6, 10, and 15 days. A lead-time of 3 days is not shown here as the meteorological forcings have often not had a significant effect on the forecasts resulting in a small distribution of meteorological errors across stations. However, the relationships discussed below are present at all lead-times. The 15 stations with the largest hydrological errors at each lead-time have been removed from the main analysis because these stations show a different pattern as shown in Fig. 4.12g and discussed below.

The purple lines in Fig. 4.12 show the least-squares regression line of best fit for the relationship between the CRPSS vs the hydrological and meteorological errors. In general, an increase in either the hydrological or meteorological uncertainties, decreases the improvement due to post-processing. However, this relationship is much stronger for the meteorological errors ($r > 0.13$ compared to $r \approx 0.01$ for hydrological errors) which suggests that hydrological errors are better corrected by the post-processing method. The EMOS method is used to correct the spread of the raw forecast to account for the meteorological uncertainty (see Section 4.3.4.2) but no bias correction is performed as is sometimes done

(e.g., Skøien et al., 2021; Gneiting et al., 2005; Hemri, 2018; Zhong et al., 2020). Whereas both bias and spread correction are performed for the hydrological uncertainties. In Section 4.5.1.4 it was noted that the raw forecasts for the Rhine and the Oder catchments have similar skill but the Oder was improved more by post-processing. It was found (not shown) that this is because the errors in the raw forecast of the Rhine were mainly meteorological but those of the Oder were mainly hydrological.

Although the r^2 values are small some trends are observed in their variation with lead-time. The relationship between the meteorological errors and the CRPSS value is slightly stronger at shorter lead-times. This is partly because the EMOS spread correction parameters are lead-time invariant. The spread of the raw forecast tends to be small at short lead-times, because all ensemble members have the same initial conditions, but increases as the differing meteorological forcings propagate through the catchment system. Skøien et al. (2021) found that the value of the variance inflation factor (ζ in Eq. (4.20) of this paper) decreases with increasing lead-time even becoming less than 1 (a reduction in spread) for lead-times greater than 8 days (see top left panel of Fig. 8. in Skøien et al. (2021)). This alters the structure of the forecast spread increasing the uncertainty at shorter lead-times and decreasing the uncertainty at longer lead-times. However, here the spread at all lead-times is multiplied by a constant value such that the spread retains its original structure. Therefore, at shorter lead-times the meteorological forcings are more influential within the Kalman filter than at longer lead-times . On the one hand, if the raw forecast is skilled at short lead-times then this greater influence is beneficial and may, for example, allow the post-processed forecast to predict an upcoming peak. On the other hand, any large errors contained in the raw forecasts propagate through to the post-processed forecasts. For example, the largest peak in the time-series for the station in Daldowie, Scotland (see Fig. 4.11a) is not predicted by the raw forecast; therefore, no information about the upcoming, precipitation-driven peak is provided to the post-processed forecast. Using a lead-time dependent EMOS method may allow for better use of the information provided by meteorological forcings.

Alternatively, the hydrological uncertainty distribution may have a greater weight within the Kalman filter. Some peaks at the Daldowie station in winter 2017/2018 are forecast

accurately by the raw forecast median (grey boxes in Fig. 4.11a) but are not forecast by the post-processed forecast. This suggests that the hydrological uncertainty distribution is most impactful in the Kalman filter. The observations in the recent period often don't indicate an upcoming flood resulting in a hydrological uncertainty distribution which confidently, but incorrectly, predicts a low flow. The confidence of the hydrological uncertainty distribution results in the information of the upcoming flow provided by the meteorological uncertainty distribution being ignored. This ignoring of the meteorological information is also the reason for the poorly post-processed forecasts for some stations in Spain (see Fig. 4.9) which have very low hydrological variability except for rare large peaks. Since extreme precipitation can be an important runoff generating mechanism in this region (Berghuijs et al., 2019), post-processed forecasts for these catchments should be used cautiously particularly when the raw forecasts predict a flood.

For the hydrological errors the r^2 values decrease for lead-times of 1 day to approximately 6 days (not shown) and for lead-times longer than 6 days the r^2 values remain at approximately 0.01. This suggests that forecast dependent errors due to the initial conditions and the interaction of the meteorological forcings in the hydrological model are corrected at shorter lead-times, but at longer lead-times the correction is mainly to consistent hydrological model errors.

The 15 stations with the largest hydrological uncertainties show a small increase in average CRPSS with increasing hydrological uncertainties. This trend is visualised by the orange line in Fig. 4.12g but the limited number of data points makes the calculation irresolute. The relationship is only shown here for a lead-time of 10 days but is present at all longer lead-times. Most of the hydrological uncertainty in these cases is caused by large consistent biases rather than forecast dependent errors. For example, the station in Nytorp, Sweden has a large bias in the raw forecasts (see Fig. 4.11b). As discussed in 4.5.1.1 the post-processing method is able to correct for consistent biases resulting in post-processed forecasts that much more closely follow the observations as shown in Fig. 4.11b and higher CRPSS values when the bias of the raw forecasts is larger.

4.5.2.2 Catchment characteristics

The catchments within the EFAS domain vary greatly in terms of size, location, and flow regime. This section discusses catchment characteristics that impact the performance of the post-processing method namely: upstream area, response time, elevation, and regulation. In Fig. 4.13, box-and-whisker plots are used to show the distribution of the CRPSS values for all stations at every other timestep with the whiskers extending to the 5th and 95th percentiles. The stations are split into categories depending on (a) the size of the upstream area, (b) the time of concentration, and (c) the elevation. Values for these characteristics are extracted from static LISFLOOD maps used operationally.

Figure 4.13a shows that in general large catchments (larger than 5000 km^2) are improved more by post-processing than medium (between 1000 km^2 and 5000 km^2) and small (less than 1000 km^2) catchments, particularly at short lead-times. The relationship between medium and small catchments is less consistent. At short lead-times the median CRPSS value for small catchments is higher than for medium catchments but for longer lead-times the converse is true. However, it was found that by removing stations with an upstream area smaller than 500 km^2 (henceforth referred to as very small catchments) from the analysis the remaining small stations (with upstream areas between 500 km^2 and 1000 km^2) are in general improved less by post-processing than medium catchments at all lead-times. This results in a single trend: that in general post-processing improves forecasts more for larger catchments. A partial reason for this is that smaller catchments are impacted more by spatiotemporal errors in the meteorological forcings than larger catchments (Pappenberger et al., 2011) and, as discussed in Section 4.5.2.1, meteorological errors are difficult to correct.

There are two reasons why very small catchments must be removed to clearly identify the trend between upstream area and CRPSS. Firstly, most stations with upstream areas (provided by local authorities) smaller that 500 km^2 were not included in the calibration of LISFLOOD for EFAS 4 (Mazzetti et al., 2021b). The uncalibrated model has varying skill between catchments with some very small catchments having large hydrological errors. As discussed in Section 4.5.2.1 hydrological errors are well-corrected by post-processing,

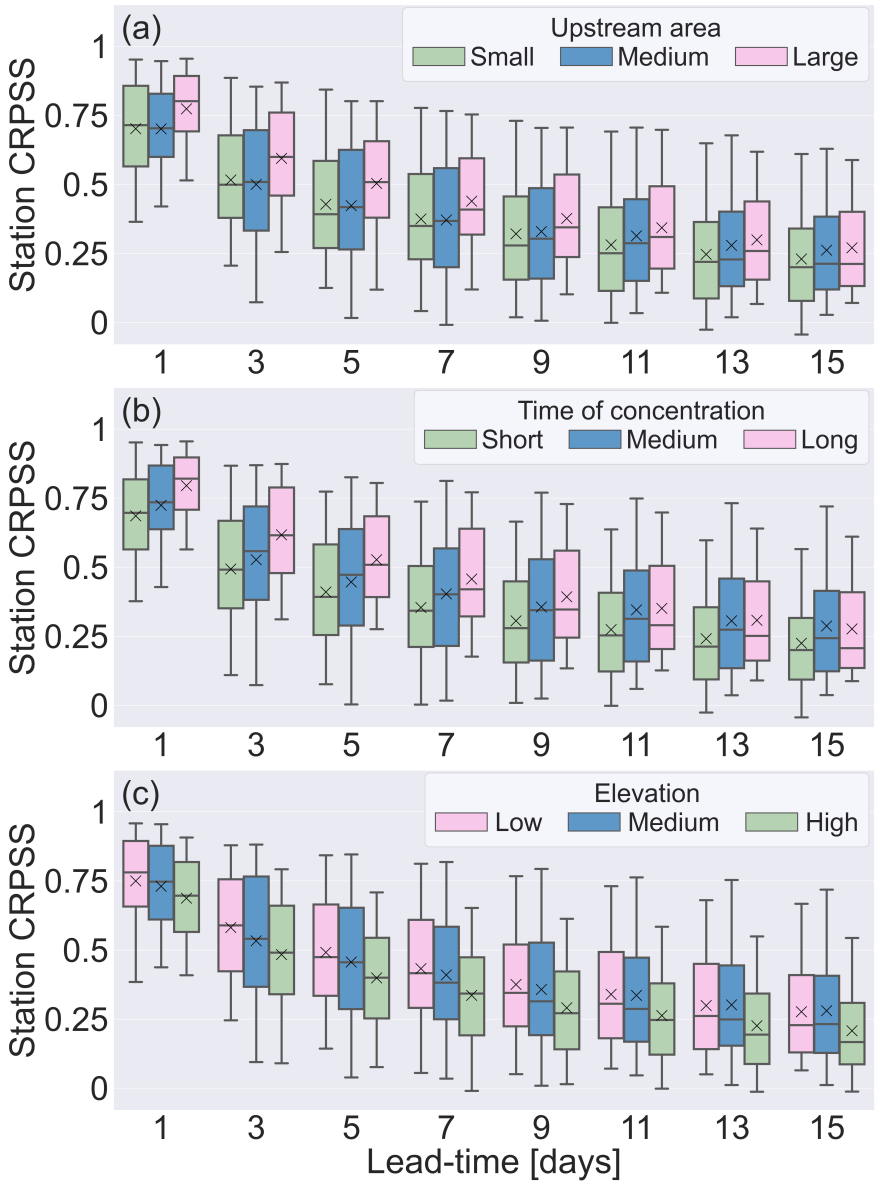

Figure 4.13: The CRPSS for all 522 stations at every other lead-time with stations categorised by their catchment characteristics. (a) Upstream area. Small catchments: less than 1000 km^2 (165 stations), Medium catchments: between 1000 km^2 and 5000 km^2 (204 stations), Large catchments: larger than 5000 km^2 (153 stations). (b) Time of concentration. Fast response catchments: less than 24 hours (253 stations), Medium response catchments: between 24 and 48 hours (144 stations), Slow response catchments: more than 48 hours (126 stations). (c) Elevation Low-elevation catchments: less than 150 m (178 stations), Medium-elevation catchments: between 150 m and 400 m (168 stations), High-elevation catchments: more than 400 m (177 stations)

therefore resulting in larger CRPSS values for some very small uncalibrated catchments than for larger calibrated catchments. Secondly, the minimum area increment of the LISFLOOD static map used to categories the stations is the area of one grid-box, 25 km^2. Therefore,

the upstream areas are multiples of 25 km^2 and thus may not represent the real size of the catchment which could lead to large hydrological biases. For example, the station in Svarttjørnbekken, Norway has a catchment area provided by local authorities of 3.4 km^2 and was therefore not included in the calibration. Additionally, in LISFLOOD its upstream area is rounded to 25 km^2 (over 7 times the size of the catchment). Figure 4.11c shows that these issues result in an over-estimation of the variability of the flow and a consistent bias in the raw forecast even at low flows. Both issues are corrected by post-processing.

In Fig. 4.13b the time of concentration is used to represent the catchment response time. Stations are split into fast response catchments (response times of less than less than 24 hours), moderate response catchments (between 24 and 48 hours), and slow response catchments (more than 48 hours). At short lead-times, slowly responding catchments outperform medium and fast responding catchments. Since large catchments tend to have slower responses, this suggests response time is partly responsible for the greater improvement experienced by large catchments. Slower responses result in stronger autocorrelations therefore the recent observation are more informative about the state of the river during the forecast period. This is shown by comparing the time-series of the Daugavpils station (Fig. 4.11d) which has a time of concentration of approximately 195 hours with that of the Daldowie station (Fig. 4.11a) which has a time of concentration of 27 hours. The Daugavpils station has a slow response with peaks lasting two months (longer than the length of the recent period) whereas the Daldowie station responds faster with peaks only lasting a week at most (shorter than the length of the forecast period). As such the post-processing method can correct forecasts much better for the Daugavpils station. It should be noted that most stations still benefit from being post-processed even at lead-times larger than their time of concentration. This is useful as operationally there is a delay in the availability of the meteorological observations used to create the water balance simulation whereas here it is assumed that all observations up to the production time of the reforecast are available. Therefore, these results suggest that although the CRPSS may be smaller there is still an operational benefit to post-processing.

In Fig. 4.13c catchments are categorised by the height of the station above sea level: low-elevation catchments (less than 150 m), medium-elevation catchments (between 150 m

and 400 m), and high-elevation catchments (more than 400 m). At all lead-times catchments at higher elevations are improved less than lower-lying catchments. This is partly due to mountainous catchments tending to have faster response times. Additionally, precipitation forecasts in mountainous regions can be biased due to insufficient resolutions to represent the orography in the NWP systems (Lavers et al., 2021; Haiden et al., 2014, 2021). Alfieri et al. (2014a) found that when compared to the water balance simulation (i.e. equivalent to the metric for meteorological error used here) the raw ensemble forecasts are negatively biased in mountainous regions due to an under-estimation of the precipitation. The effect of station elevation on the performance of the post-processing method explains the cluster of degraded stations around the Kjolen Mountains (see Fig. 4.9).

The regulation of rivers via reservoirs and lakes is difficult to model. Raw forecasts for many regulated catchments were found to have negative correlation with the observations. In this study, a station is considered regulated if it is within 3 grid-boxes downstream of a reservoir or lake in the LISFLOOD domain or if data providers have reported that the station is on a regulated stretch of the river. Figure 4.14 shows the CRPSS values of the 42 regulated stations (black lines) and the distribution of the CRPSS values of the unregulated stations (green distribution) for lead-times of 3, 6, 10, and 15 days. The distribution for the unregulated stations is estimated using kernel density estimation with the dashed line showing the median value and the dotted lines showing the interquartile range. The mean CRPSS values are indicated by crosses of the respective colours.

At all lead-times, the CRPSS values of most regulated stations are above the median of the unregulated stations. Additionally, the mean CRPSS value of the regulated stations is at least 0.1 higher than that of the unregulated stations for all lead-times longer than 1 day. The improvement due to post-processing at regulated stations is dependent on whether the reservoir is in the same state during the recent and forecast periods and hence whether the discharge values from the recent period provide useful information about the state of the reservoir. At longer lead-times it becomes more likely that the reservoir will have changed state and therefore that the information provided by the recent discharge values is not useful. However, if the reservoir is in the same state then the magnitude of the improvement from

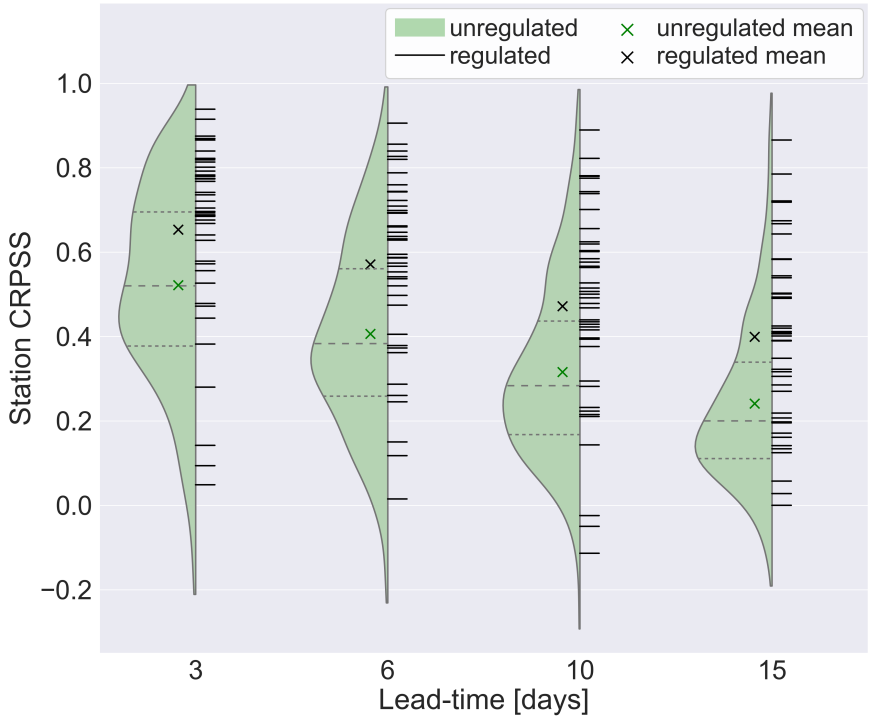

Figure 4.14: Violin plot of the CPRSS values for the 480 unregulated stations (green distribution) and the 42 regulated catchments (black lines) at lead-times of 3, 6, 10, and 15 days.

post-processing can be large. For example, the Porttipahta station in Finland is located at the Porttipahta reservoir and its time-series is shown in Fig. 4.11e. In May 2018 the discharge is 0 $m^3 s^{-1}$ for approximately a month. The raw forecast does not capture this decrease in discharge, but the post-processed forecast median is very accurate even at longer lead-times. However, at the start and end of this zero-flow period the post-processed forecasts do not perform as well for a lead-time of 15 days (purple triangles) because the reservoir has changed state since the forecast production time. It is thought that small but regular regulation is partly responsible for the cluster of degraded stations on the River Sava shown for a lead-time of 10 days in Fig. 4.9c. Three of the degraded stations in this cluster are regulated and are the three regulated stations with the lowest CRPSS values at all lead-times shown in Fig. 4.14.

It is interesting to consider whether other hydrological processes that are difficult to model can be accounted for by post-processing. For example, the peak in the winter and spring in the Daugavpils catchment (see Fig. 4.11d) is largely dominated by snow and ice melt (Skute et al., 2008) which are difficult processes to model (Alfieri et al., 2014a). Figure 4.11d shows that the raw forecasts do not predict the magnitude of the peak in late January

but the post-processed forecasts, which are conditioned on recent observations that indicate the increase in discharge due to snowmelt, do accurately predict the peak. Similar results were seen in other catchments with snow dominated regimes. Although the identification of dominating runoff generating mechanisms for all catchments and seasons is beyond the scope of this study, the results presented in this section suggest that post-processing can correct for errors introduced by the imperfect modelling of slow hydrological processes.

4.5.2.3 Calibration time-series

The length of the time-series used to calibrate the station model varies between stations. The maximum length is dictated by the water balance simulation which is available from 1 January 1990. However, many stations have shorter time-series due to the availability of observations. Figure 4.15 shows the CRPSS values for each lead-time with stations split by the length of their calibration time-series into unequally sized categories (see caption): very short time-series (up to 15 years), short time-series (between 15 and 20 years), medium time-series (between 20 and 25 years), and long time-series (over 25 years). These categories were chosen to investigate the impact of the length of the calibration time-series whilst keeping the number of stations in each category as large as possible. These initial comments ignore the very short time-series (green) which are discussed in more detail below.

At short lead-times long time-series in general lead to more improvement by post-processing than shorter time-series. Longer time-series allow the joint distribution between the observations and the water balance simulation to be more rigorously defined allowing a more accurate conditioning of the forecast on the discharge values from the recent period. For lead-times greater than 7 days the CRPSS distributions for all categories are similar. As discussed in Section 4.5.2.1, post-processing corrects forecast specific errors at short lead-times but at longer lead-times it is mainly consistent errors to the climatology that are corrected. The similarity of the CRPSS distributions suggests that short time-series are sufficient to capture these consistent errors. This is also shown by the relatively good performance of stations with very short time-series. Although a full sensitivity analysis is beyond the scope of this study, these results suggest that very short time-series can be used, if necessary, to correct for

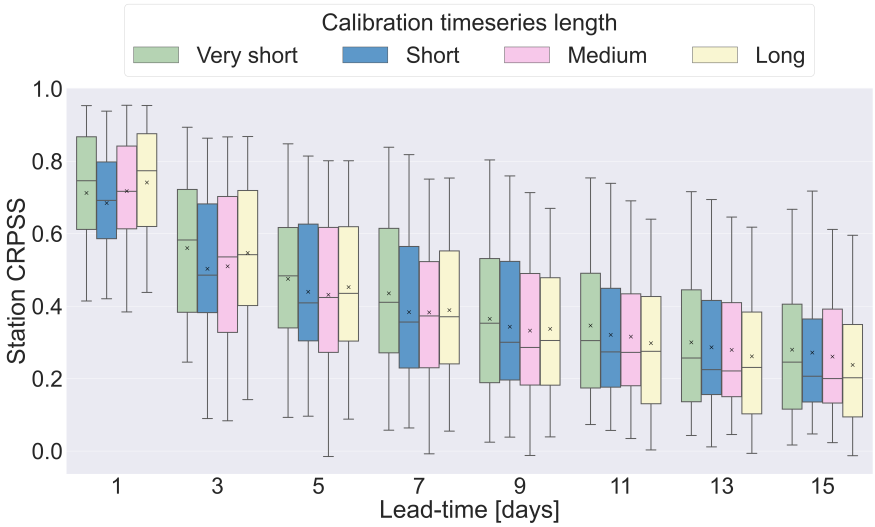

Figure 4.15: The CRPSS for all 522 stations at every other lead-time with stations categorised by the length of their calibration time-series. Very short time-series: less than 15 years (63 stations), Short time-series: 15 to 20 years (93 stations), Medium time-series: 20 to 25 years (119 stations), Long time-series: over 25 years (247).

consistent biases, although longer time-series are preferable. However, care should be taken when forecasting high flows since a short timeseries will not allow for a robust calculation of the upper tail of the discharge distribution (see Section 4.3.3.1) which will likely cause errors in the forecast probability distribution (Bogner et al., 2012).

In general, shorter time-series tend to be more recent and so benefit from improved river gauging technology and also because non-stationarity between the calibration and evaluation period is less likely to be an issue. The station in Montañana (shown in Fig. 4.11f) is an example of a station where a period of poor quality observations in the calibration time-series impact the calibration resulting in a large jump in the CDF of the observed discharge distribution as highlighted by a red circle in Fig. 4.16b. This CDF is used in the NQT and the large jump results in non-smooth forecast probability distributions. Additionally, these errors were found to impact the estimation of the joint distribution which resulted in a decrease in the correlation coefficient after post-processing. Removing the erroneous observations improved the discharge estimations suggesting that the priority should be to use the best quality data available even if the resultant calibration time-series is shorter.

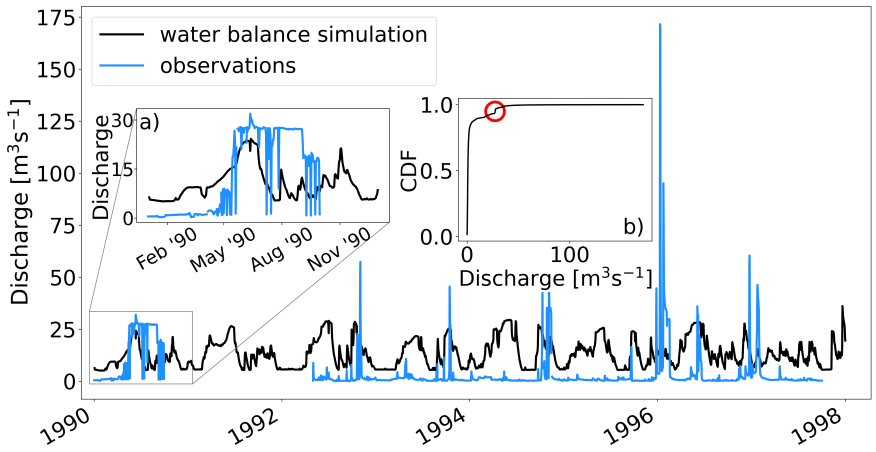

Figure 4.16: Observations (blue) and water-balance simulation (black) time-series used in the calibration of the station model for the station in Montañana. a) Section of the calibration time-series with errors in the observations. b) the Cumulative Distribution Function (CDF) of the observed discharge distribution calculated during the calibration. Red circle indicates a jump in the CDF due to the section of the time-series shown in a).

4.6 Conclusions

Post-processing is a computationally efficient method of quantifying uncertainty and correcting errors in streamflow forecasts. Uncertainties enter the system from multiple sources including the meteorological forcings from numerical weather prediction systems (here referred to as meteorological uncertainties), and the initial hydrological conditions and hydrological model (here referred to as hydrological uncertainties). The post-processing method used operationally in the European Flood Awareness System (EFAS) uses a method motivated by the Ensemble Model Output Statistics (Gneiting et al., 2005) method to account for the meteorological uncertainty and the Multi-Temporal Model Conditional Processor (Coccia, 2011) to account for the hydrological uncertainty. The EFAS domain includes catchments of varying characteristics for which the same post-processing method is used. In this paper we used reforecasts to investigate the added skill gained by post-processing and how these improvements vary across the domain. This study aimed to answer two research questions.

First, does the post-processing method provide improved forecasts? Our results show that for the majority of stations the post-processing improves the skill of the forecast with median Continuous Ranked Probability Skill Scores (CRPSS) of between 0.74 and 0.2 at all lead-times. This improvement is greatest at shorter lead-times of up to 5 days but post-processing

is still beneficial up to the maximum lead-time of 15 days. The bias and spread correction provided by the post-processing increases the reliability of the forecasts and increased the number of correctly forecast flood events without increasing the number of false alarms. However, the post-processed forecasts also led to the flood peak often being forecast too early by approximately a day. Although, forecasts for floods events at most stations did benefit from post-processing the greatest improvements were to forecasts for normal flow conditions.

Second, what affects the performance of the post-processing method? Several factors were found to impact the performance of the post-processing method at a station. The post-processing method is more easily able to correct hydrological errors than meteorological errors. This is mainly because no bias-correction is performed for the meteorological errors whereas hydrological errors are bias corrected by conditioning the forecast on the recent observations. Therefore, stations where the errors were primarily due to hydrological errors were improved more. As the hydrological errors tend to be larger than the meteorological errors this is beneficial; however, more research is required to fully account for biases due to the meteorological forcings as well.

The post-processing method was found to easily account for consistent hydrological biases that were often due to limitations in the model representation of the drainage network. However, the correction of forecast specific errors (due to initial conditions and meteorological forcings) was largely determined by the response time of the catchment. Therefore, the greatest improvement was seen in catchments larger than 5000 km^2 and catchments less than 100 m above sea level as these catchments tended to have longer response times. Additionally, post-processing was able to correct for errors due to difficult to model hydrological processes, such as regulation and snowmelt, when recent observations contained relevant information about the discharge.

The use of long historic observational time-series for the offline calibration is beneficial particularly for correcting forecast specific errors. However, time-series shorter than 15 years were found to be sufficient for correcting consistent errors in the model climatology even at a lead-time of 15 days. The quality of the observations in the historic time-series is important and errors in the time-series degraded the performance of the post-processing method and

limit the usefulness of the forecasts.

These results highlight the importance of post-processing within the forecasting chain of large-scale flood forecasting systems. They also provide a benchmark for end-users of the EFAS forecasts and show the situations when the post-processed forecasts can provide more accurate information that the raw forecasts. These results also highlight possible areas of improvement within the EFAS and the factors that must be considered when designing and implementing a post-processing method for large-scale forecasting systems.

4.7 Summary of Chapter 4

In this chapter, I present an evaluation of the at-gauge post-processing method that is operational in the European Flood Awareness System (EFAS) to address *Objective 1*. This evaluation is the first skill assessment of the current operational post-processing method. (Q1.1.) In general, the forecasts are improved by post-processing across the domain, particularly at short lead-times. (Q1.2) I identify some situations where the post-processing method struggles, such as in quickly responding catchments, and when the errors are primarily caused by the meteorological forcings. Longer calibration timeseries are beneficial at shorter lead-times, but forecasts are improved even when the available timeseries is short.

A key limitation of the operational post-processing method is that it can only be applied at gauged locations. In EFAS version 4.0, this leaves 99.5% of the domain uncorrected. However, there are clear spatial patterns in the improvement gained by post-processing the forecasts (e.g., Fig. 4.9) suggesting that the errors at locations connected along the river network are correlated. Chapters 5 and 6 present the development of a post-processing method for ungauged locations based on the forecast covariances between locations (*Objectives 2* and *3*). The at-gauge post-processing method evaluated in this chapter is used in Chapter 6 to provide proxy-observations during the forecast period. These predictions are then propagated to ungauged locations using the LETKF-based method that will be presented in Chapter 5. Key results from the above evaluation that are relevant for these chapters are:

- The improvement from the at-gauge post-processing method is greatest at short lead-

times.

- Large rivers are corrected more than smaller rivers in general.
- High river discharge events are predicted with less accuracy than normal flows.
- Errors due to the meteorological forcings are less effectively corrected than those due to the hydrological model or initial conditions. The Rhine catchment is largely dominated by errors due to the meteorological forcings.

The results of this evaluation have been presented to users to aid forecast interpretation, and to help identify development objectives. In Chapter 7, I present the outcomes from a workshop focused on identifying the barriers to the use of the EFAS post-processed forecasts, and prioritising developments (*Objective 4*). The results from this chapter were included in a presentation at the start of the workshop that aimed to provide all participants with a baseline knowledge of the forecasts.

Overall, the evaluation presented in this chapter highlights the benefit of post-processing, providing useful knowledge of the forecast skill to developers and users. This information has also guided the development of the post-processing method for ungauged locations that is presented in this thesis.

Chapter 5

Data assimilation for ensemble river discharge hindcast post-processing

This chapter addresses the second objective of the thesis: *To investigate the applicability of data-assimilation in a post-processing environment to propagate observational information from gauged to ungauged locations.*

A new method for propagating observation information from gauged to ungauged locations is presented. The method is based on common data assimilation techniques: state augmentation and the Local Ensemble Transform Kalman Filter (defined in Section 2.3.4). However, these techniques are modified for application within a post-processing environment with pre-computed simulations. Error vectors, which represent the deviation of a river discharge ensemble member from the true river discharge value at each grid-box, are estimated at every timestep. The ensemble members are then corrected using the corresponding error vector. The new method is evaluated for the Rhine-Meuse catchment using a leave-one-out cross-verification strategy to evaluate the skill at ungauged locations. The content of the chapter is reproduced from:

Matthews, G., Cloke, H. L., Dance, S. L., and Prudhomme, C.: *Error-correction across gauged and ungauged locations: A data assimilation-inspired approach to post-processing river discharge forecasts*, EGUsphere [preprint], `https://doi.org/10.5194/hess-2024-`

3989, 2025. [accepted]

In Section 5.10, I summarise the key results of the chapter and identify connections with other chapters in this thesis.

Abstract

Forecasting river discharge is essential for disaster risk reduction and water resource management, but forecasts of future river state often contain errors. Post-processing reduces forecast errors but is usually only applied at the locations of river gauges, leaving the majority of the river network uncorrected. Here, we present a data-assimilation-inspired method for error-correcting ensemble simulations across gauged and ungauged locations in a post-processing step. Our new method employs state augmentation within the framework of the Localised Ensemble Transform Kalman Filter (LETKF) to estimate an error vector for each ensemble member. The LETKF uses ensemble error covariances to spread observational information from gauged to ungauged locations in a dynamic and computationally efficent manner. To improve the efficiency of the LETKF we define new localisation, covariance inflation, and initial ensemble generation techniques that can be easily transferred between modelling systems and river catchments. We implement and evaluate our new error-correction method for the entire Rhine-Meuse catchment using forecasts from the Copernicus Emergency Management Service's European Flood Awareness System (EFAS). The resulting river discharge ensembles are error-corrected at every grid box but remain spatially and temporally consistent. The skill is evaluated at 89 proxy-ungauged locations to assess the ability of the method to spread the correction along the river network. The skill of the ensemble mean is improved at almost all locations including stations both up- and downstream of the assimilated observations. Whilst the ensemble spread is improved at short lead-times, at longer lead-times the ensemble spread is too large leading to an underconfident ensemble. In summary, our method successfully propagates error information along the river network, enabling error correction at ungauged locations. This technique can be used for improved post-event analysis and can be developed further to post-process operational forecasts providing more accurate knowledge about the

future states of rivers.

5.1 Introduction

River discharge forecasts are essential tools for taking effective preparatory actions for disaster mitigation and water resource planning (UNDRR, 2015). However, despite the increased sophistication of forecasting systems over the past few decades, river discharge forecasts still contain uncertainty (Boelee et al., 2019). The uncertainty is introduced at several stages of the forecasting system including the meteorological forcings, the initial conditions, and the hydrological model structure and parameters (Valdez et al., 2022). Ensemble river discharge forecasts typically aim to account for the meteorological uncertainty by forcing a hydrological model with many meteorological forcings (Cloke and Pappenberger, 2009; Wu et al., 2020). However, ensemble forecasts can still contain biases and errors in the representation of uncertainty. Different methods for correcting these errors have been developed including pre-processing of the meteorological forcings, calibration of the hydrological model, improving the initial conditions using data assimilation, and post-processing of the river discharge forecast (Bourdin et al., 2012). Of these approaches post-processing is often considered the most computationally efficient and its ability to correct for multiple sources of errors simultaneously is appealing.

In meteorological forecasting, post-processing at non-observed locations is common (see Vannitsem et al., 2021). However, hydrological forecasting requires consideration of the spatial heterogeneity introduced by the river network (e.g., Li et al., 2017; Woldemeskel et al., 2018; Ye et al., 2014; Xu et al., 2019; Liu et al., 2022; Lee and Ahn, 2024) making hydrological post-processing at ungauged locations a difficult challenge. The global river gauge network is sparse (Krabbenhoft et al., 2022), and even in regions where gauges exist, river discharge data are often not widely shared (Lavers et al., 2019; Hannah et al., 2011). Therefore, the development of post-processing techniques for ungauged locations is essential. However, current techniques are generally too computationally expensive for operational river flow forecasting applications (Emerton et al., 2016). For example, defining a joint

distribution between the river discharge at multiple locations would allow forecasts to be conditioned on observations available at specific locations (Engeland and Steinsland, 2014). However, for large-scale distributed systems and multiple lead-times the size of the joint distribution quickly becomes too large. Alternatively, error-correction can be performed at a gauged location and the results interpolated to ungauged locations. One such method used to interpolate error-correction parameters is top-kriging (Pugliese et al., 2018; Skøien et al., 2021). Top-kriging takes into account the river network but the relationship between errors at different locations is assumed static regardless of the hydrometeorological situation (Skøien et al., 2016, 2006). Another option is to use a river routing model to propagate error-corrected river discharge forecasts between gauged locations (Bennett et al., 2022). Whilst this approach maintains spatial consistency between locations, the additional run of the model could be computationally expensive for an operational application.

The aim of this paper is to present and evaluate a novel technique for spreading observation information from gauged to ungauged locations in a computationally efficient and temporally varying manner. The new method is based on data assimilation techniques. Data assimilation is a mathematical technique that combines modelled predictions and observations to produce a better estimate of the true state of the river (Nichols, 2003, 2010). Data assimilation is often used to improve the initial conditions of forecasts (Valdez et al., 2022). However, in this paper we modify the techniques to apply them in a post-processing environment such that additional, computationally expensive, executions of the hydrological model are not required. The error correction method proposed in this study is based on state augmentation (Dee, 2005) and the Local Ensemble Transform Kalman Filter (LETKF, Hunt et al., 2007). State augmentation is a technique that allows the estimation of the state and parameters/biases of a system simultaneously, and is often used for online bias-estimation in data assimilation (Ridler et al., 2018; Gharamti and Hoteit, 2014; Smith et al., 2013, 2009; Martin et al., 2002). The LETKF is part of the Kalman filter family of methods and uses an ensemble of model states to estimate the state error covariances. Due to their computational efficiency and ability to handle non-linear dynamics without an adjoint model, ensemble Kalman filters are common data assimilation methods in hydrological research (Rouzies et al., 2024; Li et al.,

2023; Mason et al., 2020; Ridler et al., 2018; Khaki et al., 2017; Xie and Zhang, 2010; Clark et al., 2008).

Whilst many studies have shown the benefits of data assimilation for hydrological forecasting (Tanguy et al., 2025; Valdez et al., 2022; Piazzi et al., 2021), the process is rare in operational systems (Pechlivanidis et al., 2025), particularly in large-scale systems (Wu et al., 2020). This limited uptake is partly due to data latency issues (WMO, 2024a), time constraints, and the potential impact on the interpretation of the forecasts (e.g., thresholds based on model climatology may no longer be consistent; Emerton et al., 2016). Additionally, the benefit of data assimilation at longer lead-times is uncertain (e.g., Valdez et al., 2022). In this paper, we leverage key advantages of data assimilation—such as the ability to propagate observational information to ungauged locations—within a post-processing framework that is more readily integrated into operational systems.

The proposed method aims to improve the skill of the ensemble mean and the reliability of the ensemble spread by adjusting each ensemble member, as will be discussed in more detail in Section 5.2. However, it is equally, if not more, important that the ensembles are spatially and temporally consistent in order to aid with decision making (Bennett et al., 2022). This is particularly important for large scale systems that provide forecasts across administrative boundaries, such as the Copernicus Emergency Management Service's (CEMS) European Flood Awareness System (EFAS) used in this study (Matthews et al., 2025b). The specific research questions to be addressed in this study are therefore,

1. Can data assimilation techniques be used in a post-processing environment to propagate observation information to ungauged locations?

2. Are the resulting ensemble predictions of river discharge more skillful than the raw ensemble?

This paper is organised as follows. In Section 5.2 we define the errors which we aim to correct and introduce some terminology and notation. In Section 5.3 we describe the data assimilation techniques used within this study. In Section 5.4 we outline the proposed error-correction method and detail how the ensemble is corrected. Section 5.5 provides some

additional components of the method that improve the efficacy of the method but which can be adjusted to suit the data availability of any system and/or domain. Section 5.6.4 outlines the strategy used to evaluate the efficacy of the proposed method. Section 5.7 presents the results, first investigating the impact of assimilating the observations, and then assessing the skill of the error-corrected ensembles. In Section 5.8 we discuss key features of the proposed method and their impact on the error-corrected ensembles. In Section 5.9 we conclude that the proposed method successfully improves the skill of the ensemble mean, and highlight priorities for future developments.

Please note that throughout the paper 'hindcast ensemble' refers to the ensembles of river discharge that we are error-correcting. In this paper, these ensembles are past operational EFAS forecasts (see Section 5.6.1). However, when we perform the error-correction we use observations that are available within the forecast (hindcast) period. Observations are not available during the forecast period in an operational system, since these timesteps are in the future. Therefore, we refer to these river discharge ensembles as hindcasts to indicate that the ensembles are not valid forecasts.

5.2 Ensemble error-correction framework

Here, we define the errors which we aim to correct and provide some notation that is used throughout the paper. Where possible we follow the standard data assimilation notation provided in Ide et al. (1997). Let the true state of the system at time k be defined as $\mathbf{x}_k^{true} \in \mathbb{R}^n$, where each element represents the true river discharge in one of the n grid boxes in the domain of interest. Hydrological forecasts generally estimate the true state of the system using a modelled state, denoted $\mathbf{x}_k$, where the lack of superscript 'true' indicates it is a modelled estimate. In this study, the hydrological ensemble forecasts consist of N potential realizations of future river discharge, referred to as ensemble members. We define the ensemble river discharge hindcasts as

$$\left\{\mathbf{x}_k : \mathbf{x}_k^{(i)}, \text{for } i = 1, 2, \ldots, N \text{ and } k = 0, 1, \ldots, L\right\}. \tag{5.1}$$

where the superscript (i) indicates the i-th member of the ensemble, N is the ensemble size, the timestep k refers to the lead-time of the hindcast, and L is the maximum lead-time. The ensemble mean is defined as

$$\overline{\mathbf{x}}_k = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_k^{(i)} \in \mathbb{R}^n. \tag{5.2}$$

The ensemble perturbation matrix is defined as

$$\mathbf{X}_k = \begin{pmatrix} \mathbf{x}_k^{(1)} - \overline{\mathbf{x}}_k & \mathbf{x}_k^{(2)} - \overline{\mathbf{x}}_k & \cdots & \mathbf{x}_k^{(N)} - \overline{\mathbf{x}}_k \end{pmatrix} \in \mathbb{R}^{n\times N} \tag{5.3}$$

where the i-th column represents the i-th ensemble member's departure from the ensemble mean at lead-time k. The perturbation matrix contains information about the spread of the ensemble and the spatial structure of the deviations of each ensemble member from the mean. From the definition of the perturbation matrix, the ensemble covariance matrix is defined as

$$\mathbf{P}_k = \frac{1}{N-1}\mathbf{X}_k\mathbf{X}_k^T \in \mathbb{R}^{n\times n}. \tag{5.4}$$

where the superscript T indicates the matrix transpose.

In this paper, we propose a method to spread an error-correction from gauged locations to every grid box in the domain. The proposed method estimates an additive error vector for each hindcast ensemble member at each timestep. Each element of the error vector represents the error associated with a single grid box in the domain. Collectively, these error vectors form an ensemble defined as,

$$\left\{\mathbf{b}_k^{(i)} \in \mathbb{R}^n \text{ for } i = 1, 2, \ldots, N\right\} \tag{5.5}$$

where N is the same ensemble size as the river discharge hindcast, n is the number of grid-boxes in the hindcast domain, and k is the timestep. The error ensemble mean, $\overline{\mathbf{b}}_k$, and the ensemble perturbation matrix, $\mathbf{B}_k$, are calculated by substituting $\mathbf{b}_k^{(i)}$ in place of $\mathbf{x}_k^{(i)}$ in Eqs. (5.2) and (5.3), respectively. We assume there is an additive relationship between each hindcast ensemble member and the corresponding error vector such that the i-th error-

corrected ensemble member, $\mathbf{x}_k^{new,(i)}$, is defined as

$$\mathbf{x}_k^{new,(i)} = \mathbf{x}_k^{(i)} + \mathbf{b}_k^{(i)} \in \mathbb{R}^n. \tag{5.6}$$

The estimation of the error ensemble at each timestep is described in Section 5.4.

To aid with the estimation of the error vectors, we assume that at each timestep the system is observed at p_k river discharge gauges. We assume the observation vector, $\mathbf{y}_k \in \mathbb{R}^{p_k}$, is related to the true state of the system as

$$\mathbf{y}_k = \mathbf{H}_k(\mathbf{x}_k^{true}) + \boldsymbol{\epsilon}_k \tag{5.7}$$

where $\boldsymbol{\epsilon}_k \in \mathbb{R}^{p_k}$ is a vector of unbiased Gaussian noise with covariance matrix $\mathbf{R}_k \in \mathbb{R}^{p_k \times p_k}$, and $\mathbf{H}_k \in \mathbb{R}^{p_k \times n}$ is the linear observation operator. The observation operator maps the variables from the state space to observation space. In this study, the observation operator selects the grid boxes within the modelled drainage network that correspond to the locations of the river gauges.

5.3 Data Assimilation

As discussed in Section 5.1, the proposed method is based on common data assimilation techniques: state augmentation and the Local Ensemble Transform Kalman Filter (LETKF). In this section, we provide an overview of these techniques and introduce the necessary equations. In Section 5.4, we adapt and apply these methods in a non-standard way due to their application in a post-processing environment.

5.3.1 State augmentation

State augmentation is a technique used for online bias-correction in data assimilation that allows the simultaneous estimation of the system state and biases. An augmented state is defined by appending the biases to the state vector, allowing both to be updated by the data

assimilation method. In this study, the i-th member of the augmented ensemble is defined as

$$\mathbf{w}_k^{(i)} = \begin{pmatrix} \mathbf{x}_k^{(i)} \\ \mathbf{b}_k^{(i)} \end{pmatrix} \in \mathbb{R}^{2n}. \tag{5.8}$$

where $\mathbf{x}^{(i)} \in \mathbb{R}^n$ and $\mathbf{b}^{(i)} \in \mathbb{R}^n$ are the i-th hindcast and error ensemble members, respectively. The augmented ensemble mean and perturbation matrix are given by

$$\overline{\mathbf{w}}_k = \begin{pmatrix} \overline{\mathbf{x}}_k \\ \overline{\mathbf{b}}_k \end{pmatrix} \in \mathbb{R}^{2n} \quad \text{and} \quad \mathbf{W}_k = \begin{pmatrix} \mathbf{X}_k \\ \mathbf{B}_k \end{pmatrix} \in \mathbb{R}^{2n \times N} \tag{5.9}$$

where $\overline{\mathbf{x}}$ and $\overline{\mathbf{b}}$ are the ensemble means of the hindcast and error ensembles, respectively, and $\mathbf{X}$ and $\mathbf{B}$ are the perturbation matrices of the hindcast and error ensembles, respectively.

In this study, state augmentation is used within an LETKF (described in Section 5.3.2) and it is therefore necessary to define the evolution of the augmented states between timesteps. The evolution of the hindcast and error ensembles determines the evolution of the augmented states. The hindcasts used in this study were generated using the LISFLOOD hydrological model, which is used in the EFAS operational system (van der Knijff et al., 2008). As the true evolution of the error vectors at all grid-boxes is unknown, we assume a simple persistence model, such that $\mathbf{b}_k^{(i)} = \mathbf{b}_{k-1}^{(i)}$. This is a common assumption used in state augmentation (Pauwels et al., 2020; Ridler et al., 2018; Rasmussen et al., 2016; Martin, 2001). Based on the independent evolution of the hindcast and error ensembles, and the additive relationship between their members (Eq. (5.6)), we define the propagation of the augmented ensemble members as

$$\mathbf{w}_k^{(i)} = \begin{pmatrix} \mathbf{M}_{k-1} & \mathbf{I}_{k-1} \\ \mathbf{0}_{k-1} & \mathbf{I}_{k-1} \end{pmatrix} \begin{pmatrix} \mathbf{x}_{k-1}^{(i)} \\ \mathbf{b}_{k-1}^{(i)} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_k^{(i)} + \mathbf{b}_{k-1}^{(i)} \\ \mathbf{b}_{k-1}^{(i)} \end{pmatrix}. \tag{5.10}$$

where $\mathbf{M}_{k-1} \in \mathbb{R}^{n \times n}$ is a linear evolution operator representing the LISFLOOD hydrological model and $\mathbf{I}_{k-1} \in \mathbb{R}^{n \times n}$ is the identity matrix. Since we use precomputed hindcast ensembles the propagation of the hindcast ensemble members requires no additional computation. The full non-linear LISFLOOD hydrological model is also used without the need to define a linear

approximation.

5.3.2 Local Ensemble Transform Kalman Filter (LETKF)

The Local Ensemble Transform Kalman Filter (LETKF; Hunt et al., 2007) updates the mean state and the perturbation matrix of an ensemble by combining the modelled and observed data. As a sequential data assimilation method, the LETKF consists of a *propagation step* (also known as a *forecast step*) and an *update step* (also known as an *analysis step*) that are cycled. In this method, we use the LETKF to update the ensemble of error vectors at each hindcast timestep for which observations are available. However, we modify the propagation step to use precomputed hindcasts. The propagation step evolves the augmented states forward in time from time $k-1$ to k, as described in Eq. (5.10). Rather than evolve the hindcast ensemble explicitly (which would require the hydrological model), we use the precomputed hindcast at timestep k.

The update step of the LETKF calculates the state of the system at timestep k by combining the modelled augmented states and observations. Both data are weighted by their respective uncertainties, represented by their covariance matrices. As the LETKF is a well documented method we only provide the key update equations. For more detailed derivations, we direct the reader to Hunt et al. (2007) and Livings et al. (2008). To apply the LETKF to the augmented ensemble, we create a *model-observation ensemble* with an ensemble mean, $\overline{\mathbf{y}}_k^{\mathbf{x}}$, defined as

$$\overline{\mathbf{y}}_k^{\mathbf{x}} = (\mathbf{H}_k \;\; \mathbf{0})\,\overline{\mathbf{w}}_k = \mathbf{H}_k\overline{\mathbf{x}}_k + \mathbf{H}_k\overline{\mathbf{b}}_{k-1} \tag{5.11}$$

where $\mathbf{H}_k \in \mathbb{R}^{p_k \times 2n}$ is the observation operator defined in Eq. (5.7). The LETKF can then update the augmented ensemble mean, $\overline{\mathbf{w}}_k$, such that,

$$\overline{\mathbf{w}}_k^a = \overline{\mathbf{w}}_k^f + \begin{pmatrix} \mathbf{K}_{\mathbf{x}_k} \\ \mathbf{K}_{\mathbf{b}_k} \end{pmatrix} (\mathbf{y}_k - \overline{\mathbf{y}}_k^{\mathbf{x}f}), \tag{5.12}$$

where the superscripts f and a indicate the state before and after the update step, respectively; $\mathbf{K}_{\mathbf{x}_k} \in \mathbb{R}^{n \times p}$ and $\mathbf{K}_{\mathbf{b}_k} \in \mathbb{R}^{n \times p}$ are the components of the Kalman gain matrix acting on the

hindcast ensemble and the error ensemble respectively; and $\mathbf{y}_k \in \mathbb{R}^p$ is the observation vector defined in Eq. (5.7). The difference between the observations and the model state in observation space (i.e., $\mathbf{y}_k - \overline{\mathbf{y}}_k^{\mathbf{x}f}$) is called the innovation vector. The Kalman gain matrix determines the impact of the innovation vector in the update step. The respective uncertainties of the prior modelled state and the observations determine their weight within the LETKF. Large observation uncertainties reduce the Kalman gain, while large uncertainties in the prior state increase the Kalman gain. Both the hindcast and the error components of the Kalman gain are functions of the covariance matrix of the augmented ensemble (see Appendix A.1; Bell et al., 2004). The covariance matrix describes the state error covariances between grid-boxes allowing the Kalman gain to spread the observation information to ungauged locations. To update the error component specifically, it is the cross-covariances between the error component and the hindcast component that control the spread of the observation information to ungauged locations (see Eq. (9) in Bell et al., 2004). This ability to spread the observational information is key to the error-correction method presented in this study.

The LETKF updates the augmented ensemble perturbation matrix, $\mathbf{W}_k$, such that,

$$\mathbf{W}_k^a = \mathbf{W}_k^f \mathbf{T}_k \tag{5.13}$$

where $\mathbf{T}_k \in \mathbb{R}^{N \times N}$ is the square root transform matrix (Livings et al., 2008). The square root transform matrix is derived using the Kalman gain matrix which gives the weighting between the modelled state and the observations (Livings et al., 2008). Using an eigenvector decomposition, the square root transform matrix rescales and rotates the ensemble members such that the updated perturbation matrix represents the uncertainty in the updated ensemble mean. The square root transform matrix allows the covariance matrix of the ensemble to be updated without the need for the covariances to be explicitly calculated which can be computationally expensive (Bishop et al., 2001; Hunt et al., 2007). These update equations are used to update the error component only, as will be discussed in Section 5.4.1.

5.4 Spatially consistent error-correction method for river discharge

In this section, we describe how we use the data assimilation techniques discussed in Section 5.3, to post-process the hindcasts across the domain, including at ungauged locations (Fig. 5.1). The correction is applied in a post-processing environment, avoiding the need for additional executions of the hydrological model which can be computationally expensive. In Section 5.4.1, we describe how the error ensemble is updated at every timestep. In Section 5.4.2, we describe how the updated error ensemble is used to error-correct the hindcast ensemble. Specific experimental design choices are discussed in Section 5.5.

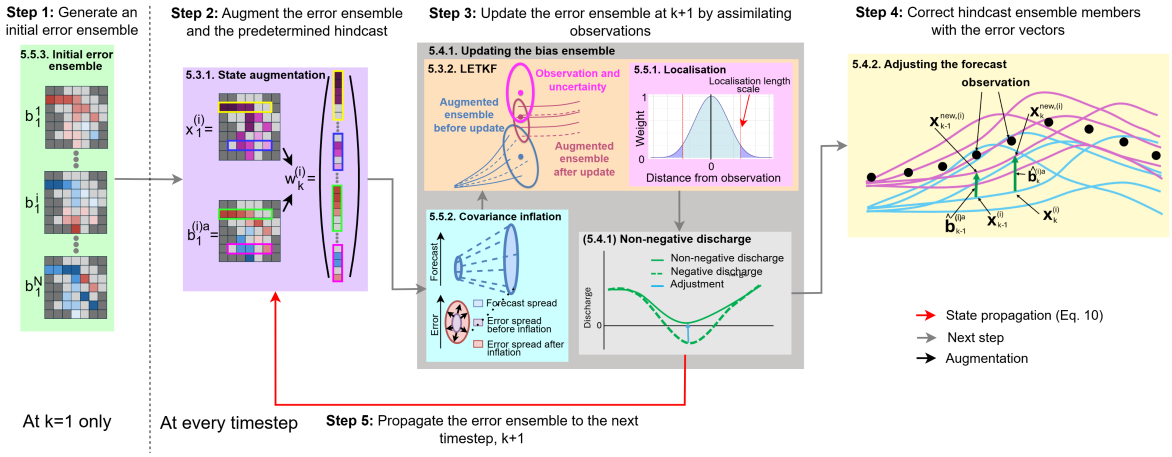

Figure 5.1: Schematic of the new error-correction method for gauged and ungauged locations. Coloured boxes indicate different components of the method. An initial error ensemble is created for timestep k=1 (green box). Then, the error ensemble is augmented to the hindcast ensemble (purple box). At each timestep the covariance of the augmented ensemble is inflated (cyan box) before being updated using the LETKF, which uses localisation to improve the results of the update (collectively the orange box). The updated error ensemble is adjusted to ensure non-negative discharge values (light grey box) before being used to error-correct the hindcast (yellow box). The non-negative error ensemble is propagated to the next timestep (red arrow). More details are provided for each component in the section indicated in the top left corner of the corresponding box.

5.4.1 Updating the error ensemble

At each timestep the error ensemble is updated to estimate the optimal set of error vectors to correct the hindcast at that timestep. The update is performed using the LETKF defined in

Section 5.3.2. Using the definition of the augmented state in Eq. (5.8) the update equations for the error ensemble only are,

$$\overline{\mathbf{b}}_k^a = \overline{\mathbf{b}}_k^f + \mathbf{K}_{\mathbf{b}_k}(\mathbf{y}_k - \overline{\mathbf{y}}_k^{\mathbf{x}}) \tag{5.14}$$

and

$$\mathbf{B}_k^a = \mathbf{B}_k^f \mathbf{T}_k. \tag{5.15}$$

As the hindcast component is not explicitly evolved, we assume that the raw hindcast is a good approximation for the hindcast analysis state if the component were to be updated. This allows the substitution of the precomputed hindcast in place of the propagated state at the next timestep. Thus, the updated mean of the augmented ensemble can be defined as

$$\overline{\mathbf{w}}_k^a = \begin{pmatrix} \overline{\mathbf{x}}_k \\ \overline{\mathbf{b}}_k^a \end{pmatrix} \in \mathbb{R}^{2n}. \tag{5.16}$$

where $\overline{\mathbf{x}}_k$ is the ensemble mean of the raw hindcast ensemble and $\overline{\mathbf{b}}_k^a$ is the updated error ensemble mean (Eq. (5.14)). The perturbation matrix of the updated augmented ensemble follows a similar pattern such that

$$\mathbf{W}_k^a = \begin{pmatrix} \mathbf{X}_k \\ \mathbf{B}_k^a \end{pmatrix}. \tag{5.17}$$

where $\mathbf{X}_k$ is the ensemble perturbation matrix of the raw hindcast ensemble and $\mathbf{B}_k^a$ is the updated error ensemble perturbation matrix (Eq. (5.15)). The assumptions made in Eqs. (5.16) and (5.17) make our system sub-optimal from a data assimilation perspective but are necessary to avoid rerunning the hydrological model. Importantly, we aim to estimate the error of the precomputed model output at each lead time. Therefore, while the lack of state evolution makes the hindcast component update sub-optimal, the update of the error ensemble remains mathematically consistent. In this study, we provide proof-of-concept that the resulting error ensemble improves the skill of the hindcast (see Section 6.4).

The Kalman filter is not constrained to enforce non-negativity of the analysis state, and therefore, could lead to negative discharge values for some grid boxes if the cross-covariances are incorrectly defined. We enforce non-negativity by further adjusting the error ensemble members after the LETKF update step. For any ensemble member and grid box where the sum of the hindcast discharge and the updated error is negative, we modify the error value so that the total becomes a small positive value, mitigating the potential for instabilities caused by zero-values. This small positive value is sampled from a Gaussian distribution with a mean of zero and a standard deviation equal to 10% of the standard deviation of the updated error ensemble at the grid-box of interest.

The updated positive-definite augmented states are propagated to the next timestep as defined in Eq. (5.10). The updated positive-definite augmented states are also used to error-correct the hindcast (Section 5.4.2).

5.4.2 Adjusting the forecast

After the error component of the augmented state has been updated using Eqs. (5.14) and (5.15), and non-negativity has been enforced (Section 5.4.1), the error ensemble members are added to the respective hindcast ensemble members such that

$$\mathbf{x}_k^{new,(i)} = \mathbf{x}_k^{(i)} + \hat{\mathbf{b}}_k^{(i)a} \tag{5.18}$$

where $\mathbf{x}_k^{new,(i)}$ and $\mathbf{x}_k^{(i)}$ are the i-th ensemble members of the error-corrected and raw hindcast ensembles, respectively, and $\hat{\mathbf{b}}_k^{(i)a}$ is the error vector associated with the i-th error ensemble member where the caret indicates a non-negativity check has been applied. Consequently, the error-corrected hindcast ensemble mean and perturbation matrix are given by

$$\overline{\mathbf{x}}_k^{new} = \overline{\mathbf{x}}_k + \overline{\hat{\mathbf{b}}}_k^{a} \tag{5.19}$$

and

$$\mathbf{X}_k^{new} = \mathbf{X}_k + \hat{\mathbf{B}}_k^{a}. \tag{5.20}$$

This update results in an additive spread correction matrix, $\mathbf{\Gamma}_k$, with the form

$$\mathbf{\Gamma}_k = \mathbf{X}_k \hat{\mathbf{B}}_k^{aT} + \hat{\mathbf{B}}_k^a \mathbf{X}_k^T + \hat{\mathbf{B}}_k^a \hat{\mathbf{B}}_k^{aT} \tag{5.21}$$

where $\mathbf{X}_k$ and $\hat{\mathbf{B}}_k^a$ are the perturbation matrices of the raw hindcast and error ensembles, respectively, and the superscript T indicates the matrix transpose (Section 5.2 in Martin, 2001). Whilst $\mathbf{\Gamma}_k$ can be negative it should be noted that the covariance matrix of the corrected ensemble, $\mathbf{X}_k^{new} \mathbf{X}_k^{new,T}$, is positive-definite by definition.

5.5 Experimental implementation

In Section 5.4 we presented a new method of spreading observation information to ungauged locations in a post-processing environment based on common data assimilation techniques. In this section, we describe three key components of the method—localisation, covariance inflation, and the generation of the initial error ensemble—which are crucial for its performance but can be implemented in various ways.

5.5.1 Localisation

Localisation is used to reduce the effect of spurious correlations which can arise due to sampling errors caused by the small ensemble size (Hamill et al., 2001; Hunt et al., 2007). The LETKF uses observation localisation which reduces the impact of observations by multiplying the inverse of the observation-error covariance matrix by a *localisation matrix*, $\boldsymbol{\rho} \in \mathbb{R}^{p_k \times p_k}$, such that

$$\mathbf{R}^{-1} = \boldsymbol{\rho} \circ \mathbf{R}_{nl}^{-1} \tag{5.22}$$

where $\mathbf{R} \in \mathbb{R}^{p_k \times p_k}$ is the localised observation-error covariance matrix used in the LETKF (Section 5.3.2), $\mathbf{R}_{nl} \in \mathbb{R}^{p_k \times p_k}$ is the non-localised observation-error covariance matrix, and the symbol $\circ$ indicates the Schur product (also known as the Hadamard product) which is an element-wise matrix multiplication (Golub and Van Loan, 2013). We assume that $\mathbf{R}_{nl}$ and, by definition, $\mathbf{R}$ are diagonal matrices. In this study we use distance-based localisation so the

impact of the multiplication described in Eq. (5.22) is to increase the effective uncertainty of distant observations and thus decrease their impact on the analysis state. The impact of the localisation on the spatial extent of the analysis increments is demonstrated in Section 5.7.1.

The localisation matrix is defined using the Gaspari-Cohn function which has a parameter called the *localisation length scale* (Appendix A.2; Gaspari and Cohn, 1999). The Gaspari-Cohn function smoothly decreases the weights assigned to an observation as the distance from the observation location increases, starting from a value of 1 at the observation location and reaching 0 for distances greater than twice the localisation length scale (pink box, Fig. 5.1). In this study, the distance is calculated along the river network which has been shown to improve the analysis for fluvial applications (García-Pintado et al., 2015; El Gharamti et al., 2021b; Khaniya et al., 2022). The distance between a grid-box and the location of an observation is calculated using the local drainage direction map and the channel length of the hydrological model (Choulga et al., 2023). As the distance is defined along the river network, observations cannot impact grid-boxes in a different drainage basin.

Sensitivity experiments conducted during the development of this method found that the optimal length scale varied by location, lead-time, and tuning metric of choice, but overall, the differences were small for length scales from 65 km to 786 km (not shown). Therefore, we propose instead for the localisation length scale to be defined as the maximum distance between any grid point and its closest observation. This 1) ensures that all grid boxes are updated in the update step of the LETKF reducing the potential for discontinuities in the analysis state, 2) can adapt to changes in the availability of observations, and 3) can be applied to different domains and hydrological model configurations without requiring a tuning experiment.

5.5.2 Covariance inflation

Small ensemble sizes can cause underestimation of the ensemble spread, reducing the impact of the observations on the analysis (Furrer and Bengtsson, 2007). Additionally, we assume the error ensemble is constant between timesteps which, while simplifying implementation,

could introduce model errors into the ensemble (Evensen et al., 2022). To ameliorate these issues, various covariance inflation techniques are often used (Duc et al., 2020; Scheffler et al., 2022). We implement a heuristic covariance inflation method inspired by the *relaxation-to-prior perturbations* technique (Zhang et al., 2004; Kotsuki et al., 2017). However, as we are working within a post-processing context, we adapt the method for use with predefined ensembles (i.e., without evolving the inflated perturbations between timesteps).

We blend the prior perturbation matrix at $k + 1$ with an estimated perturbation matrix $\mathbf{W}^{est}_{k+1}$ similar to the use of a climatological covariance matrix in Valler et al. (2019). The resulting perturbation matrix is given by

$$\mathbf{W}^{inf}_{k+1} = (1 - \alpha) \begin{pmatrix} \mathbf{M}_k & \mathbf{I}_k \\ \mathbf{0}_k & \mathbf{I}_k \end{pmatrix} \mathbf{W}^{a}_{k} + \alpha \mathbf{W}^{est}_{k+1} \tag{5.23}$$

where α is an inflation parameter to be defined (and the definition of the matrices $\mathbf{M}_k$ and $\mathbf{I}_k$ are given in Section 5.3.1). This blending of matrices introduces both additive and multiplicative inflation. We define $\mathbf{W}^{est}_{k+1}$ as

$$\mathbf{W}^{est}_{k+1} = \begin{pmatrix} \mathbf{X}^{est}_{k+1} + \mathbf{B}^{est}_{k+1} \\ \mathbf{B}^{est}_{k+1} \end{pmatrix}. \tag{5.24}$$

where $\mathbf{X}^{est}_{k+1}$ and $\mathbf{B}^{est}_{k+1}$ can be estimated separately. When substituted into Eq. (5.23), this form of $\mathbf{W}^{est}_{k+1}$ maintains consistency between the hindcast and error components of the augmented ensemble.

During development, it was found that the estimated matrices must have spatial structures consistent with the river network and be forecast and lead-time-dependent. For simplicity, and as the raw hindcast perturbations satisfy these requirements, we set both $\mathbf{X}^{est}_{k+1}$ and $\mathbf{B}^{est}_{k+1}$ equal to the raw hindcast perturbation matrix (Dee, 2005; Martin et al., 2002). In future studies, the estimated perturbation matrices could be defined using alternative models to evolve the analysis perturbation matrix between timesteps or be climatological matrices (Valler et al., 2019).

The inflation parameter α_k controls the weighting between the prior and estimated matrices. To account for changing uncertainty across lead-times and forecasts, we define α_k using a smoothed estimate of the relative change in hindcast ensemble variance

$$\alpha_k = \frac{1}{k} \sum_{l=k-2}^{l=k} max\left\{ \frac{\left|Tr(\mathbf{P}_l) - Tr(\mathbf{P}_{l+1})\right|}{Tr(\mathbf{P}_l)}, 1 \right\} \tag{5.25}$$

where k is the current timestep and $\mathrm{Tr}(\mathbf{P}_l)$ is the trace of the raw hindcast covariance matrix at timestep l. A maximum value of 1 is set to avoid instabilities, particularly at short lead-times where the change in variance between timesteps can be large. The average over the past three timesteps is taken to ensure that α is smoothly changing between timesteps, again to avoid instabilities. This approach of estimating α was selected after sensitivity testing (not shown for brevity), which indicated that the inflation factor must be both lead-time dependent and forecast dependent. While α is not spatially varying, it is applied to perturbation matrices with spatial structures consistent with the river network, ensuring physically plausible ensemble perturbations.

5.5.3 Initialising the error ensemble for the first timestep

We must define an initial error ensemble to perform the state augmentation at the first timestep. In a forecast post-processing environment there is no "warm-up" period in which a state of equilibrium can be reached, and therefore the initial error ensemble must be physically plausible. Here, the initial error ensemble is defined using three sets of river discharge data: in-situ observations, simulations created by forcing a hydrological model with meteorological observations, and the ensemble mean and ensemble perturbation matrix of a single lead-time from a previous hindcast. A single ensemble is generated for the full EFAS domain, from which the elements associated with the domain of interest (in this study the Rhine-Meuse catchment) are extracted.

The estimation has two main steps: estimating the mean error and generating the perturbations around that mean. The ensemble mean is intended to capture biases in the hydrological model at the initial time. It is computed as follows:

1. **Calculate the errors at gauged locations:** For each river gauge location, we calculate the average relative error between observed and simulated river discharge over the past 10 days. To limit the influence of outliers or representation errors, these errors are capped at ± 100

2. **Interpolate the errors to ungauged locations:** Using inverse distance weighting, we interpolate the errors from gauged to ungauged locations. The value at each grid-box is a weighted average of relative errors from the 100 nearest stations, with closer stations given more influence (Lu and Wong, 2008). All available stations, including those outside the catchment of interest, are used in this calculation to capture spatial variability.

3. **Impose the river network structure:** The interpolated error field is then multiplied by the simulated river discharge values at each grid point. This enforces the spatial structure of the river network, ensuring errors are proportional to the size of the river.

Since the true error covariance is unknown, we assume a reasonable estimate can be derived from a previous river discharge ensemble forecast as follows:

1. **Calculate the ensemble statistics:** We calculate the ensemble mean and perturbation matrix from the second lead-time of a hindcast issued two days prior. This choice avoids unrealistically low spread often seen at the first lead-time due to a single set of initial conditions.

2. **Inflate the covariance matrix:** The perturbation matrix is adjusted by calculating the error of the hindcast ensemble mean at each grid-box relative to a simulation forced by meteorological observations. This provides a set of scaling factors used to inflate the perturbation matrix. To avoid underestimating uncertainty, we impose a minimum threshold on the resulting standard deviation of 10% of the local simulated river discharge.

The resulting error ensemble mean and perturbations define the initial ensemble, which is then updated using the LETKF with state augmentation, as described in Section 5.4.1.

5.6 Evaluation strategy

5.6.1 European Flood Awareness System (EFAS)

The hindcasts used in this study were produced by the European Flood Awareness System (EFAS) as operational forecasts (Barnard et al., 2020b). EFAS is part of the Early Warning component of the European Commission's Copernicus Emergency Management Service (CEMS), and aims to provide complementary forecast information to hydro-meteorological services throughout Europe (Matthews et al., 2025b). EFAS streamflow forecasts are produced by forcing a calibrated hydrological model, LISFLOOD (De Roo et al., 2000; van der Knijff et al., 2008; Arnal et al., 2019), with the output from meteorological numerical weather prediction (NWP) systems. Whilst the operational EFAS system is a multi-model system with four sets of meteorological forcings, we focus only on the medium-range river discharge forecasts generated with meteorological forcings from the 51-member medium-range ensemble from the European Center for Medium-range Weather Forecasts (ECMWF) due to its large ensemble size. The meteorological forcings are interpolated to the EFAS grid. A single set of initial hydrological conditions is used for all ensemble members often leading to small ensemble spreads at short lead-times. The spread then increases as the different meteorological forcings propagate through the system. No data assimilation is performed in the generation of the initial hydrological conditions. Instead, the LISFLOOD hydrological model is forced with meteorological observations (and meteorological forecasts when observations are not available) to generate the initial conditions (Smith et al., 2016).

As an operational system, EFAS is constantly evolving. For the evaluation presented here we use EFAS version 4 (operational from 14 October 2020 to 20 September 2023) aggregated to daily timesteps with a maximum lead-time of 15 days. The ensembles have 51 members and predict the average river discharge for each timestep for each grid-box within the domain. The hindcasts have a spatial resolution of 5km $\times$ 5km with a ETRS89 Lambert Azimuthal Equal Area Coordinate Reference System. Hindcasts from the 00 UTC daily cycle are used resulting in a total of 365 hindcasts used in the evaluation.

5.6.2 Rhine-Meuse catchment

The Rhine-Meuse catchment has a drainage area of 195,300 km^2, a channel length of about 38,370 km in EFAS, and consists of 7812 grid-boxes. It is the 5th largest catchment in EFAS. The Rhine river originates in the Swiss Alps, flows through the Central Uplands and the North European Plain, before finally discharging into the North Sea. The Meuse river originates from the Langres Plateau in France, flows through the Ardennes Massif and the low-lying plains of the Netherlands, before merging with the Rhine and entering the North Sea. The catchment consists of rivers of different sizes, topologies, and levels of human influence, making it an ideal test catchment to see how the method deals with changes along the river network.

5.6.3 Observations

The Rhine-Meuse catchment has a dense river gauging station network. The main set of observations used in this study are daily river discharge observations from 89 stations across the Rhine-Meuse catchment for the time period from 21 December 2020 to 15 January 2022. The minimum value across the stations is < 1 $m^3 s^{-1}$ and the maximum value is 7663 $m^3 s^{-1}$. These observations were assimilated as part of the error-correction method to update the error ensemble and used in the evaluation of the corrected forecasts (Section 5.6.4 describes the cross-validation approach used). Whilst the error-correction method can adapt to missing observations, these 89 stations were selected as they have no missing data for the time period of interest allowing this analysis to focus on the spread of observational information to ungauged locations. The maximum distance between any grid-box and the closest of the 89 stations is 262 km which is set as our localisation length scale (cut-off distance is therefore 524 km; see Section 5.5.1). In addition to these stations, all available observations from across Europe were used to generate the initial error ensembles (total 505 stations). All river discharge observations were provided by local and national authorities and collated by the CEMS Hydrological Data Collection Centre (see `https://confluence.ecmwf.int/display/CEMS/EFAS+contributors`).

The construction of the non-localised observation error covariance matrix, $\mathbf{R}_k^{nl}$, is a key component of all data assimilation methods. The matrix describes the uncertainty associated with each observation (defined in Eq. (5.7)). This uncertainty arises due to instrument uncertainty, observation processing, observation operator error and scale mismatch between the observations and the model resolution (Janjić et al., 2018). The matrix also describes the correlation between errors of different observations (Stewart et al., 2013; Fowler et al., 2018b). In this study, we assume that the observation errors from different gauge stations are uncorrelated such that $\mathbf{R}_k^{nl}$ is a diagonal matrix with all off-diagonal elements set to 0. It is possible for the observation errors to be correlated (e.g., an agency uses the same measurement protocol for all gauges), however, due to limited information regarding the device specification, calibration, and maintenance for all gauges it is not possible to estimate this correlation at a large-scale (Coxon et al., 2015). As local factors, impacting individual gauges, typically dominate the observation errors (Ocio et al., 2017) this assumption is suitable for this proof-of-concept study. Observation errors are also assumed to be uncorrelated with the prior errors, which is a standard assumption in data assimilation (Janjić et al., 2018). Estimating the magnitude of the uncertainty in the observations requires information about the river and the gauge itself; information that is often not available at a large-scale (Westerberg and Karlsen, 2024; Coxon et al., 2015). Therefore, we assume that the observation errors are a percentage of the observation magnitude (Ocio et al., 2017; McMillan et al., 2010b). We estimate the standard deviation of the observation errors as 10% of the observation magnitude (Refsgaard et al., 2006; McMillan et al., 2018, 2012).

In the leave-one-out verification experiments (see Section 5.6.4) we use the observations from the non-assimilated station as validation data and assume they are the truth with no errors.

5.6.4 Experiments

We use three experimental schemes to investigate the effect that the error-correction scheme has on the ensemble hindcasts.

1. **Single station experiments:** Only observations from one of the 89 stations are assimilated when estimating the error-vector. All available observations are used in the generation of the initial error ensemble. These experiments allow the impact of an observation to be identified and allow the effects of localisation to be explored.

2. **All station experiments:** Observations from all stations are assimilated when estimating the error-vector and used in the generation of the initial error ensemble. These experiments allow the complete method to be assessed and for any spatiotemporal inconsistencies to be identified.

3. **Leave-one-out experiments:** Observations are withheld from one of the 89 stations and are not assimilated when estimating the error-vector nor used in the generation of the initial error ensemble. This cross-validation framework allows the skill of the adjusted hindcasts to be assessed at the locations of stations as if they were ungauged locations. To avoid confusion, we refer to withheld sites as 'proxy-ungauged' when comparing the corrected ensemble to observations. This distinction applies only in the evaluation context; from the perspective of error correction, these locations are treated as ungauged.

Each experimental scheme is applied to all hindcasts from 1 January 2021 to 31 December 2021. However, for brevity, for the single station and all station experiments we only discuss two hindcasts: 7 July 2021 and 8 October 2021. These dates represent high and normal flow conditions, respectively, allowing the ability of the method to be assessed for different circumstances.

5.6.5 Evaluation metrics

The following metrics are used to investigate the skill of the error-corrected hindcast ensemble mean and the reliability of the ensemble spread.

For the ensemble mean, the three components of the modified Kling-Gupta Efficiency: correlation, mean bias, and variability bias are used to assess different types of errors within

the ensemble mean (Kling et al., 2012; Gupta et al., 2009). Pearson's correlation coefficient measures the linear relationship between the simulated timeseries and the observations indicating timing errors (score range $[-1, 1]$). The mean bias given by the ratio between the mean of the simulated timeseries and mean of the observations indicates whether the flow is consistently over or under-estimated (score range $(-\infty, +\infty)$). The variability bias given by the ratio between the coefficient of variation of the simulation and the coefficient of variation of the observations indicates whether the variability in the flow is consistently over or under-estimated (score range $(-\infty, +\infty)$). All three components have a perfect score of 1. Additionally, to investigate whether the magnitude of the forecast mean error is reduced by the proposed method we use the Normalised Root Mean Square Error (N-RMSE Hodson, 2022; Jackson et al., 2019). The metric is normalised by dividing the RMSE by the mean of the observations for that station. Normalising the metric makes the scores at different stations comparable. The N-RMSE has a perfect score of 0.

To analyse the reliability of the spread of the ensemble forecast we use the rank histogram (Harrison et al., 1995; Anderson, 1996; Hamill and Colucci, 1997; Talagrand, 1999). To generate the histogram the rank of the observation relative to the sorted ensemble values is calculated for each hindcast. The frequencies with which the observation has a rank from 1 to M + 1 are plotted as a histogram. The shape of the histogram provides information about the reliability of the ensemble spread and bias of the ensemble (Hamill, 2001).

5.7 Results

In this section, we discuss the efficacy of the proposed error-correction method. In Section 5.7.1, we discuss how observation information is propagated along the river network and, in particular, we explore how the method reacts to different flow scenarios, both spatially and across different lead-times. In Section 5.7.2, we evaluate the skill of the resulting error-corrected ensembles in terms of their means and distributions.

5.7.1 How is observation information propagated along the river network?

5.7.1.1 Spatial propagation of the observation information

Here, we investigate how the observation information is propagated spatially from gauged locations to ungauged locations. We investigate the analysis increments of the mean — the difference between the ensemble mean before and after the update step (term 2 in Eq. (5.14)) — for single-station and all-station experiments (Fig. 5.2). Specifically, we focus on the single-station experiments for the Bonn station on the Rhine and the Uckange station on the Moselle. To investigate the impact of different flow scenarios, we study the hindcasts generated on 8 October 2021 (upper panels) and 7 July 2021 (lower panels), which represent normal and high flow scenarios, respectively.

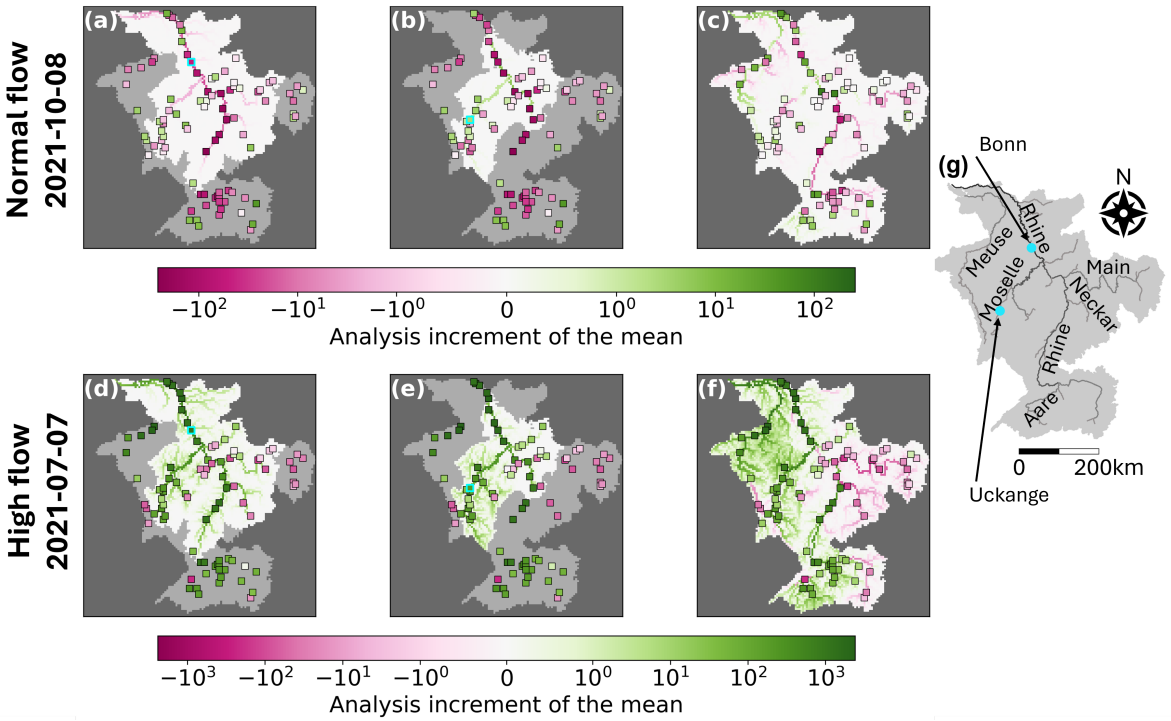

Figure 5.2: Analysis increments of the mean for a lead-time of 9 days for single station (a, b, d, and e) and all station (c and f) experiments for the hindcasts generated on 8 October 2021 (upper panels) and 7 July 2021 (lower panels). Assimilated stations for the single station experiments (cyan outline) are the Bonn station on the Rhine (a and d) and the Uckange station on the Moselle (b and e). The shaded region of the catchment is outside the localisation length of the assimilated station. Markers show the innovation at all stations. Catchment area: 195,300 km^2. Panel g shows the Rhine-Meuse catchment and highlights the rivers and stations discussed within this section.

In Fig. 5.2, the shaded regions show the parts of the catchment that are outside the localisation region for the assimilated observation. The number of grid-boxes within the localisation regions of the Bonn and Uckange stations differ because the distance is calculated along the river network and the channel length within each grid box is not constant (4662 grid boxes and 2451 grid boxes, respectively). Increasing (decreasing) the localisation length scale results in a more (less) gradual dampening of the analysis increments and more (fewer) grid-boxes being impacted by a single observation (not shown). The square markers indicate the innovation — the difference between the observation and the error-corrected ensemble mean prior to the update step (Fig. 5.2). Ideally, the analysis increment (background colour in Fig. 5.2) should reflect similar spatial behavior to the innovations within the localisation region. This would imply the ensemble is being adjusted towards the observations at each station.

For the October experiment, the innovation at Bonn is negative and results in negative analysis increments across the domain (Fig. 5.2a). For the Uckange experiment, the innovation is positive and the analysis increments are also all positive (Fig. 5.2b) indicating positive ensemble covariances. For both experiments, the analysis increments match the sign of the innovations at neighbouring stations (Figs. 5.2a and 5.2b), but at greater distances this is not the case. For example, the innovations along the Rhine in the Uckange experiment are negative whilst the analysis increments are positive.

The localisation implemented in this study allows the assimilated observations to influence the error ensemble both up- and downstream, although the influence is dampened at longer distances. We here discuss whether this choice of implementation is useful by studying the spatial patterns of the innovations. Focusing first on the Bonn experiment for October (Fig. 5.2a), we see that downstream (north) of the assimilated observations the innovations can be both positive and negative. Upstream of the assimilated observation the innovations are negative, matching the innovation at the Bonn station. The assimilation of the observation at Bonn is therefore primarily beneficial upstream, with some benefit also seen at specific locations downstream. For the Uckange experiment (Fig. 5.2b), the pattern is reversed with downstream innovations showing more consistency with the innovation at the

assimilated location. The inconsistent spatial patterns could be because, in the LETKF, we update the errors rather than the river discharge directly. The errors are dependent not only on the observed hydrological conditions but also the model structure and configuration. The spatial structure of the errors may therefore extend both up- and downstream. For example, if the drainage area within an upstream grid-box is overestimated due to the hydrological model spatial resolution, all grid-boxes downstream will be impacted by that overestimation. The benefit, in terms of consistency between the innovations and analysis increments, that is seen both up and downstream suggests that the localisation implementation is appropriate. However, we note that there may be additional factors other than distance, that could be included in the localisation to better modulate the observation influence (e.g., river confluences, regulation, or river size).

In the July experiments, we see that the innovations both up and downstream of the assimilated observations are positive, matching the innovations at the Bonn and Uckange stations, respectively. For the July experiment, the innovations are spatially homogeneous for greater distances along the river network (Figs. 5.2d and 5.2e). This indicates a greater spatial correlation length, likely due to the low-pressure system which covered large parts of the west of the catchment during this period (Mohr et al., 2022). The different correlation scales suggest that an adaptive localisation length scale may be beneficial.

The spatial heterogeneity for the October experiments suggests that assimilating a single observation cannot correct the entire domain. However, when all observation are assimilated the analysis increments vary across the domain, demonstrating the method's ability to adapt to the errors on different stretches of the river. In both all-station experiments (Figs. 5.2c and 5.2f), the analysis increments vary smoothly along the river network, which suggests the error-corrected ensemble will also change gradually. This is important because it ensures the hindcasts remain spatially consistent, with no abrupt transitions between adjacent grid boxes.

In general, for the July experiment, small rivers exhibit larger increments than in the October experiment. This indicates the assimilated observations have a greater impact across more of the domain. For October, the assimilation of an observation at Bonn results in the largest analysis increments near the observation location, with the increments diminishing to

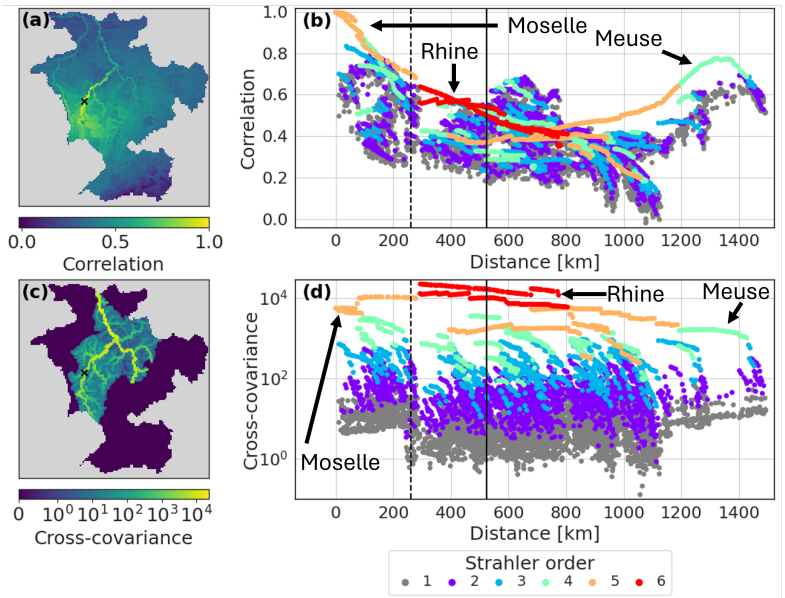

Figure 5.3: Ensemble correlations (upper panels) and cross-covariances (lower panels) between the error ensemble and the hindcast component of the augmented ensemble averaged across all all-station experiments. (a) Map of the correlation between the Uckange station and all other grid-boxes and (c) the same for the cross-covariances. (b) Scatter plot of the correlation between the Uckange station and all other grid-boxes and (d) the same for the cross-covariances. Grid-boxes on rivers discussed in the text are broadly indicated by the arrows. Dashed black line shows the localisation length scale and the solid black line shows the effective cut-off point beyond which the observation has no impact.

zero at distances greater than 524 km due to localisation (Fig. 5.2a). Interestingly, in the Uckange experiment, the largest increments occur not near the station, but along the Rhine near the confluence with the Moselle (Fig. 5.2b). In both experiments, the increments tend to be larger along bigger rivers, with smaller rivers showing smaller increments. This occurs due to large ensemble covariances between the location of the assimilated observation and locations along the bigger rivers (Fig. 5.3).

The spreading of observational information along the river network is dictated by the cross-covariances of the augmented ensemble prior to the update step. The magnitude of the cross-covariance between two locations depends on the correlation between the locations and the ensemble variance at both locations. The correlation between the location of the Uckange station and any grid-box is highest along the same river stretch (the Moselle) and decreases at longer distances from the station (Fig. 5.3a). Nearby grid boxes that are not on the same river stretch have lower correlations in general (Figs. 5.3a and 5.3b). Downstream

from the Uckange station the correlation is highest along the Moselle and downstream along the Rhine. On the other hand, the correlation upstream is more uniform across the grid-boxes (Fig. 5.3b). Whilst there are regions in the south of the catchment for which the correlation is small, in general there is a correlation of around 0.3 even with distant locations (Fig. 5.3b). This is likely spurious correlation and exemplifies the need for localisation. The correlations begin to rise again at longer distances due to grid-boxes that are geographically close to the station but the distance along the river network is large, such as the Meuse (Fig. 5.3b). Note that the similarity between the localisation length scale (dashed line) and the distance between the Uckange station and grid-boxes on the Rhine (change from a Strahler order of 5 to 6) is coincidental but does suggest that the method for defining the localisation length scale is capable of capturing the order of magnitude of the relevant spatial scales for the Rhine catchment.

Despite lower correlations, the magnitude of the cross-covariances are larger along the Rhine than for grid-boxes closer to the Uckange station on the Moselle (Fig. 5.3c). Whilst the correlation is dependent on distance, the magnitude of the cross-covariances is primarily dependent on the size of the river (note the horizontal bands of Strahler orders (a measure of stream size where larger orders indicate larger rivers Strahler, 1957) in Fig. 5.3d). Larger cross-covariances can lead to larger analysis increments as can be seen in Fig. 5.2b where the analysis increments along the Rhine are larger than those along parts of the Moselle. Localisation enforces a dependence on distance such that observations have less impact on large rivers very far from the station but this may not outweigh the larger cross-covariances.

5.7.1.2 Lead-time dependence of the analysis increments

Here, we investigate how the impact of assimilating observations changes over different lead-times. Figure 5.4 shows the trajectories of the three intermediate ensembles used in the LETKF for the 7 July hindcast for a single-station experiment where observations are assimilated at the Uckange station: the raw hindcast (left columns), the hindcast component of the augmented ensemble (middle columns), and the error component of the augmented ensemble (right columns). It should be noted that none of these ensembles are the final

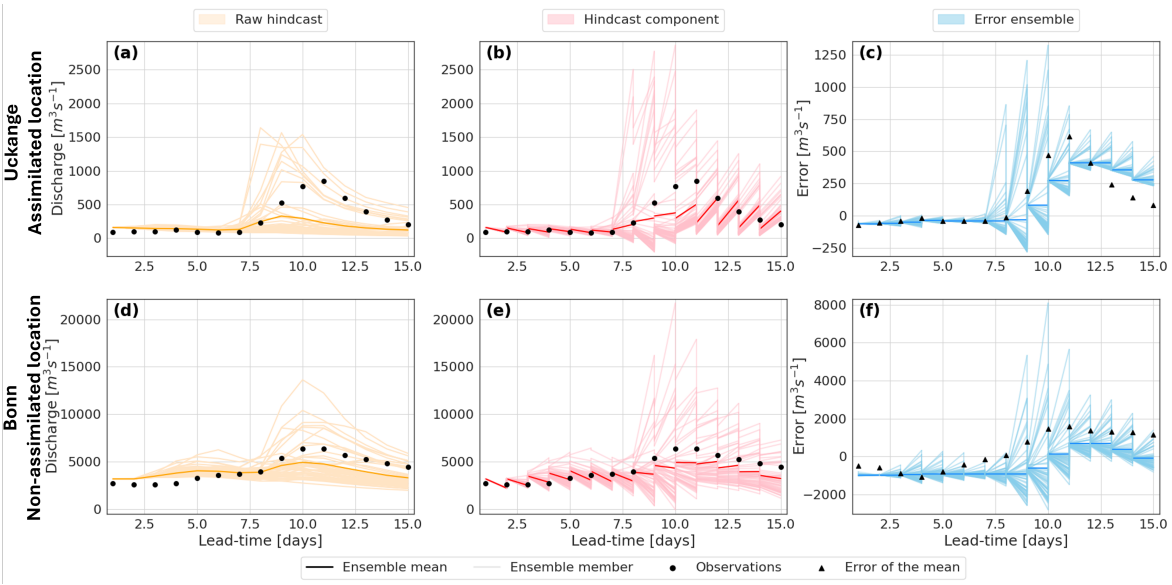

Figure 5.4: Ensemble trajectories for a single station experiment for the hindcast generated on 7 July 2021 for the location of the assimilated observation (Uckange station on the Moselle; panels a-c) and a location where an observations is not assimilated (Bonn station on the Rhine; panels d-f). The plots show the trajectory of all members and the ensemble mean of the raw hindcast ensemble (left panels), the hindcast component of the augmented ensemble (middle panels), and the error ensemble members (right panels; different y-axis scale). Markers show the river discharge observations (a, b, d, and e), and the error of the raw hindcast mean (c and f).

error-corrected ensemble but intermediate ensembles used in the LETKF. The lower panels show the trajectories at the Bonn station for which no observations are assimilated during this experiment. By plotting the raw hindcast trajectories and the observations we can visualise the errors to be estimated. We can see that for both stations the error of the hindcast mean is negative (observations are smaller than the hindcast mean) for lead-times up to 8 days, and positive at longer lead-times. Whilst this behaviour is similar for the Bonn station, the magnitude of the error is different by a factor of 10 at most lead-times.

The middle column of Fig. 5.4 shows the hindcast component of the augmented ensemble. To propagate this component between time steps without rerunning a hydrological model, we assume that the raw hindcast is a reasonable approximation of the analysis state (discussed in Section 5.4.1). As expected, this assumption results in a sub-optimal ensemble mean estimate. For example, at lead times beyond 10 days at Uckange, the update moves the ensemble mean further from the observations (Fig. 5.4b), and a similar effect is seen at Bonn (Fig. 5.4e). Also by using the precomputed ensemble, the assimilated observations do not update the ensemble

perturbations; although the perturbations do change between lead-times as the precomputed ensemble is lead-time dependent. This assumption ensures the analysis hindcast component is always physically plausible (e.g., the river discharge is always positive as this is a constraint within LISFLOOD), and provides a reasonable estimate of the uncertainty as the raw hindcast is generated using the output from an ensemble NWP. Additionally, at each timestep we aim to correct the raw hindcast, therefore this assumption provides consistency between the hindcast component and the error component of the augmented ensemble.

It is the error ensemble that is most important for our application (Figs. 5.4c and 5.4f). Despite the non-optimal formation of the analysis augmented ensemble, the error ensembles are updated beneficially, with the analysis error ensemble mean moving closer to the error of the raw hindcast mean (i.e., the difference between observations and the hindcast mean) at each lead-time for the assimilated location (Fig. 5.4c) and the non-assimilated location (Fig. 5.4f). At short lead-times, the updates to the error ensemble at the Bonn station do not appear to be beneficial (Fig. 5.4f). However, as this experiment only assimilates observations from one station this discussion should be considered a demonstration of how the method updates ungauged locations rather than an evaluation of the error-corrected ensemble (which is provided in Section 5.7.2). First we note that the updates at the assimilated location do not result in the error ensemble mean (dark blue line) matching the error of the mean (markers). This is expected and is due to the consideration of the observational uncertainty within the LETKF. This ensures spatial consistency across assimilated and non-assimilated locations, whilst combining the modelled and observed data to estimate the true state of the system across the domain.

The error ensemble is narrow after the update step and it is the covariance inflation that increases the spread between time steps. The spread of the hindcast is due to meteorological forcings, predominantly precipitation. Therefore, in general, the hindcast spread is larger for longer lead-times as the precipitation forecasts become more uncertain. Since the covariance inflation technique presented here results in the blending of the hindcast perturbation matrix with the error ensemble from the previous timestep, this behaviour in the hindcast spread is transferred to the error ensemble. As demonstrated in Figs. 5.4c and 5.4f, this can result

in the error ensemble spread being large for the rising limb of an event and smaller for the falling limb. This can result in the error not being updated sufficiently and the spread of the analysis state being too narrow, as seen after the peak in Fig. 5.4c and discussed later along with Fig. 5.5b.

5.7.2 How skillful are the error-corrected ensemble hindcasts?

In this section, we investigate whether the updated ensemble is more skillful than the raw hindcast ensemble. Using leave-one-out experiments we evaluate the ensemble mean and ensemble spread at proxy-ungauged locations. The hydrographs in Fig. 5.5 show the raw and error-corrected ensembles for three proxy-ungauged locations from the leave-one-out experiments. The hydrographs are used to illustrate the method's ability to correct the ensemble and some of the limitations.

5.7.2.1 Skill of the ensemble mean

To investigate the skill of the ensemble mean we calculate the correlation, mean bias, variability bias and the N-RMSE for each station and each lead-time. Figure 5.6 compares the skill of the ensemble means of the raw and the error-corrected ensembles focusing on the spatial dependency of the skill (a-d), and the lead-time dependency of the skill (e-h).

The error-corrected ensemble means show a stronger correlation with observations than the raw hindcast ensemble means, with an average increase from 0.82 to 0.92 (not shown). Figure 5.5a shows an example of how the error-corrected ensemble can better capture the dynamics of the river discharge resulting in an increased correlation. It can be seen that the resulting ensemble is temporally consistent (i.e., does not have improbable changes between time steps). The correlation is worsened compared to the raw hindcast ensemble at four stations (Fig. 5.6a). Focusing on the two most southern of these stations, we see that the correlation values of the raw ensembles at nearby stations are very different compared to the correlation at the two stations of interest (note the much lighter colours for nearby stations; Fig. 5.6a). The ensemble covariances are not capturing this change in regime correctly so

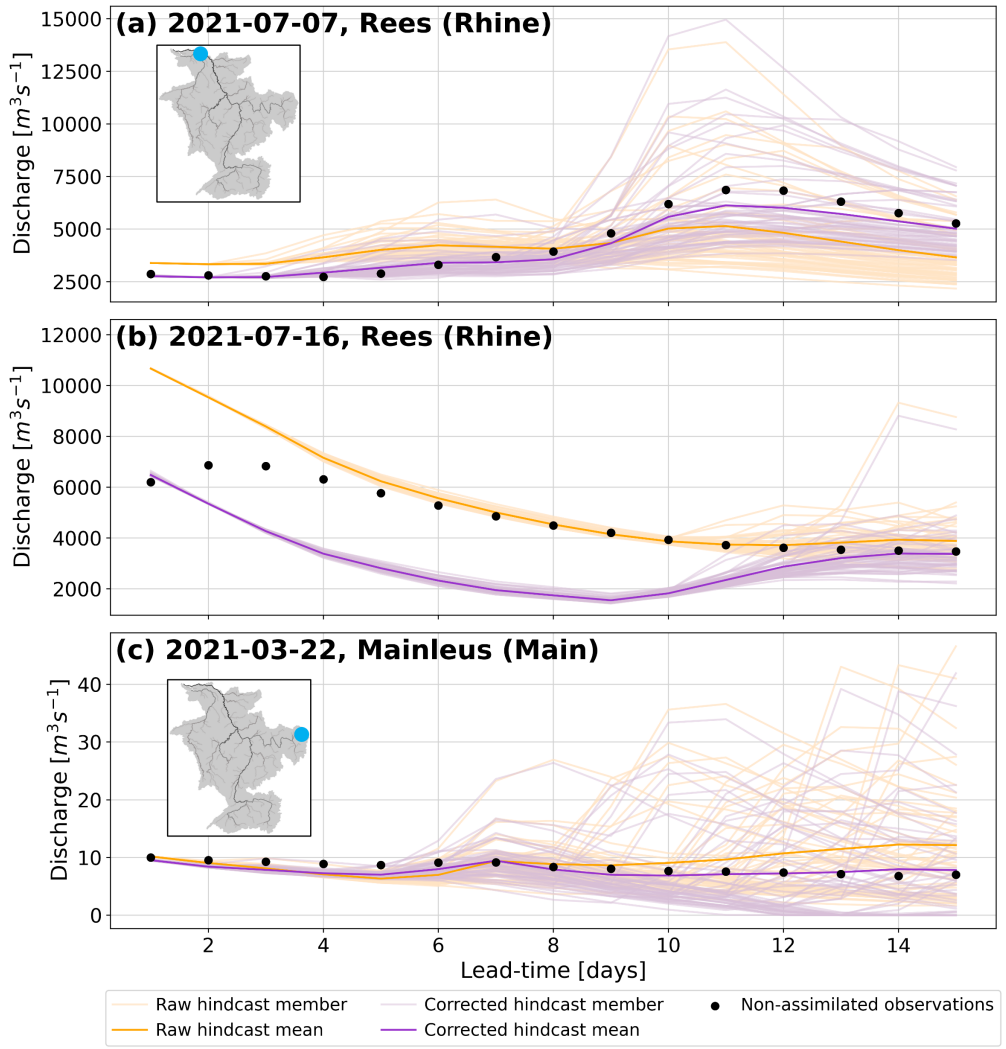

Figure 5.5: Raw and error-corrected hydrographs for proxy-ungauged locations in leave-one-out experiments at the Rees station on the Rhine (upstream area: 159,320 km^3) and the Mainleus station on the Main (upstream area: 1,164 km^3). Catchment illustrations indicate the location of the station (see Fig. 5.2 for river names and scale).

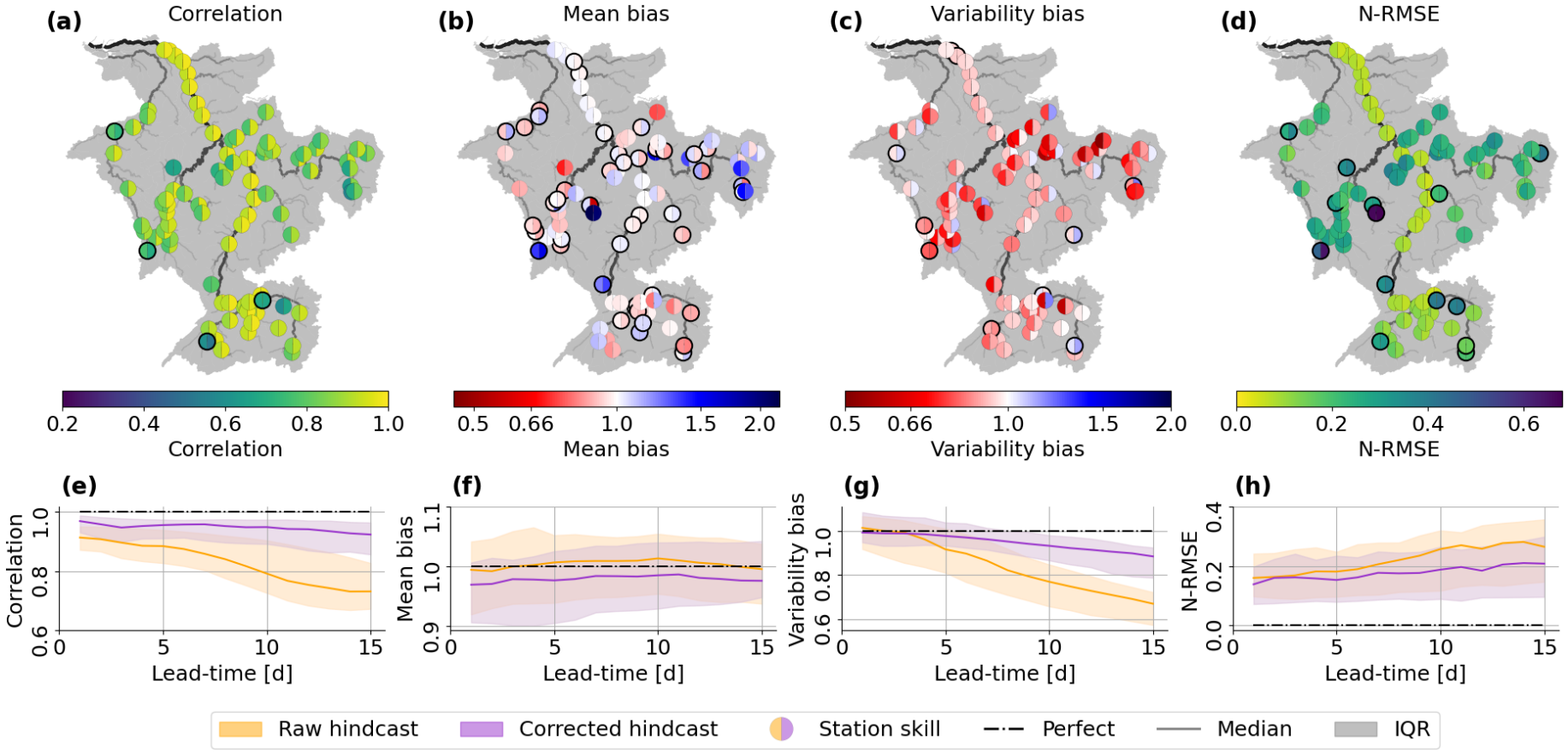

Figure 5.6: Skill of the ensemble means in terms of the correlation (a, e), mean bias (b, f), variability bias (c, g) and normalised root mean square error (N-RMSE; d, h). The catchment maps show the metric averaged across all lead-times at all 89 stations (a-d). The left (right) half of the markers show the skill for the raw (error-corrected) ensemble. Black outlines indicate stations for which the updated ensemble has worse skill than the raw hindcast ensemble (Correlation: 4/89, Mean bias: 42/89, Variability bias: 10/89, N-RMSE: 14/89). Line plots show the distribution of the metric pooled over all 89 stations for each lead-time for the raw (orange) and error-corrected (purple) ensembles (e-h). The solid line shows the median value of the metric and the shaded region shows the interquartile range (IQR) of the metric. A perfect score for the metric is shown by the dashed black line.

the observational information is not being advantageously spread between these rivers. The remaining two stations that have degraded correlation are the most upstream stations on their rivers. At these stations the updates made to the error-corrected ensemble are dependent on observations assimilated downstream. The assimilated observations are therefore providing information about a past state of the river upstream which could be the cause of the decreased correlation (a measure of timing errors) at these upstream stations. Whilst most upstream stations are improved by the error-correction method, stations which have much smaller upstream areas than their closest downstream station tend to be improved less than those that have a similar upstream area, particularly if the distance to the neighboring station is large.

Just over half of the stations (47) show improvement in the mean bias averaged across all lead times (Fig. 5.6b), but no clear spatial pattern emerges, as most rivers have a mix of improved and worsened stations. This spatial heterogeneity is also seen in the raw hindcast ensemble, with stations on the same river stretch often showing different biases. For example, stations on the Neckar and upstream of the Meuse show stations that are under- and overestimated, as well as stations with very little bias. The heterogeneity suggests local factors, which are not fully captured in the hydrological model, considerably influence flow bias. Stations showing the most improvement tend to have similar mean bias values to their neighboring stations in the raw hindcast ensemble, such as on the middle stretch of the Meuse, where four stations with similar biases show improvement (Fig. 5.6b). Spatial patterns of errors that are related to domain-wide model structure rather than local factors, such as weirs, are more likely to be portrayed by the ensemble covariances allowing observational information to be more helpfully spread along the river network.

The raw hindcast ensemble mean generally underestimates flow variability, with a variability bias below 1 (red in Fig. 5.6c). The error-corrected ensemble improves this, although there is an increase in the frequency of overestimation of the flow variability. Stations where the error-corrected ensemble overestimates the variability are often the most upstream station on their rivers (e.g., Plochingen station on the Neckar) or are closer to downstream neighbours (e.g., Chooz station on the Meuse). This suggests the hindcast covariances between downstream stations and upstream locations are too large, causing excessive adjustment at upstream

locations. Ten stations show worsened variability bias, including two stations downstream on the Rhine (Fig. 5.6h). For the two stations on the Rhine, the degradation is caused by the forecasts of the falling limb of a flood peak in July (Fig. 5.5b). Here, the hindcast uncertainty was very small at short lead-times, causing the analysis to ignore observations, leaving the error ensemble relatively unchanged, despite changes in the error behaviour after the peak (also shown in Fig. 5.4f).

Overall, the error-corrected ensemble reduces the N-RMSE but there are 14 stations where the skill is reduced. Typically, these stations are on the upstream reaches of their respective rivers (Fig. 5.6d; see discussion on correlation). Interestingly, the N-RMSE does not follow the same spatial pattern as the mean bias. This divergence indicates that the correction method is more effective at reducing large errors than at addressing systematic biases. One possible explanation is that the error vectors adjust too slowly to changes in forecast errors between time steps. This slow adjustment is particularly problematic when errors fluctuate around 0 $m^3 s^{-1}$, since alternating positive and negative deviations may not be corrected quickly enough and can accumulate into a worsening mean bias. When the error magnitude is large, the gradual adjustment is less detrimental because the sign of the error is usually captured correctly even if its magnitude is not. However, at upstream stations, where rivers are smaller and respond more quickly to rainfall, large errors often persist for shorter durations, making the slow adjustment of the error vectors more detrimental. This likely contributes to the increase in N-RMSE observed at these upstream stations. Further development of the method—for example, allowing the error vectors to evolve during the propagation step of the LETKF in addition to the update step—could enable faster adaptation to changing forecast errors.

The raw and error-corrected ensemble means both decrease in skill in terms of correlation, variability bias, and N-RMSE with increasing lead-times. The raw hindcast ensemble loses skill more quickly in particular for lead-times longer than 5 days (Figs. 5.6e-h). The uncertainty in the observations is not lead-time dependent. However, Fig. 5.3d shows that the ensemble covariances do change across lead-times, increasing for longer lead-times. The greater gain in skill for longer lead-times is likely due to larger covariances allowing the

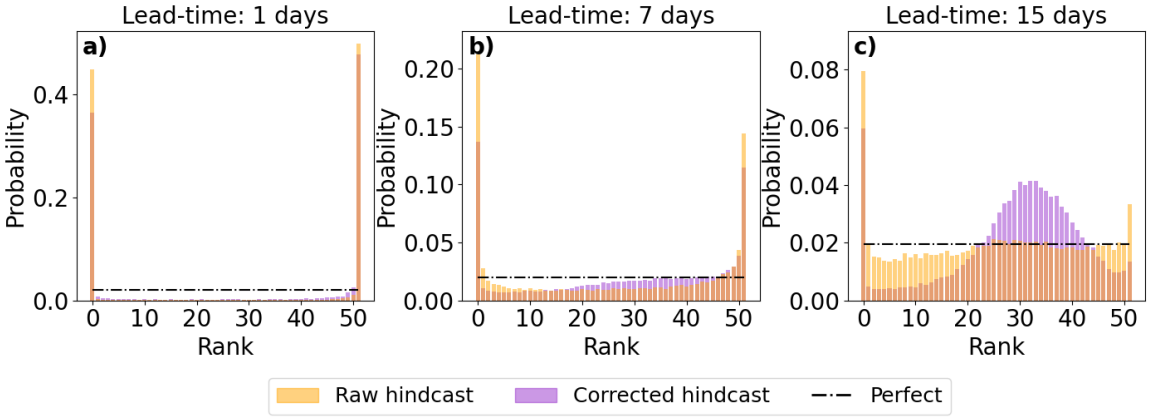

Figure 5.7: Reliability of the ensemble. Histograms show the rank of the ensemble pooled over all forecasts and stations for lead-times of 1 day (a), 7 days (b), and 15 days (c).

observations to have more influence (e.g., in Fig. 5.5b). However, the decrease in skill of the error-corrected ensemble means at longer lead-times suggests that the ensemble covariances are not as accurate at longer lead-times.

5.7.2.2 Skill of the ensemble distribution

The reliability of the ensemble distribution is assessed using rank histograms at different lead times (Fig. 5.7). At short lead times, the raw hindcast ensemble is underdispersed, due to the use of a single set of initial conditions (Fig. 5.7a). Although the error-corrected ensemble shows slight improvement, it remains overconfident with minimal correction to the spread. Both the raw and error-corrected ensembles generally appear unbiased, with observations falling both above and below the ensemble predictions at similar frequencies. However, some bias may be masked by the narrow ensemble spread as it is known that some stations are biased (Fig. 5.6b), likely contributing to the peaks at ranks 0 and 51 in the rank histograms.

As the lead-time increases, the spread of both ensembles becomes more reliable, and fewer observations fall outside the ensemble (Fig. 5.7b). However, even at a 15-day lead time, both ensembles show a tendency to overestimate observations, leading to a peak at rank 0, mostly due to a few stations consistently overestimating flow (Fig. 5.6b). Up to 7-day lead times, the rank histograms for both raw and error-corrected ensembles show similar shapes. Beyond 7 days, the raw hindcast ensemble's histogram flattens, suggesting a reliable ensemble, while the error-corrected ensemble shows a peak around ranks 25-35, suggesting

overdispersion (Fig. 5.7c). The left-skewness of the histograms is likely due to the inherent skewness in river discharge distributions. The LETKF update step seeks to minimise the difference between the ensemble mean and the true state of the system. The ensemble mean is often larger than the ensemble median leading to the observations falling in ranks above 25 if the adjustment method is successful at minimising the error of the mean (Figs. 5.5a and 5.5c).

As discussed in Section 5.4.1, the Kalman filter is not restricted to ensure positive discharge and there is therefore a need to adjust the error ensemble before correction of the hindcast. Enforcing non-negative discharge was necessary, for example, for the Mainleus station on the Main for the hindcast generated on the 22 March 2021 (Fig. 5.5c). Whilst the ensemble mean is error-corrected at most lead-times, several members indicate river discharge values of 0 $m^3 s^{-1}$. The river discharge is below 10 $m^3 s^{-1}$ but a zero flow is unlikely in reality. This suggests the ensemble spread is not sufficiently corrected even though the ensemble mean is improved as is also suggested by Fig. 5.7c.

5.8 Discussion

In general, the proposed data-assimilation-inspired method successfully spreads observational information along the river network, thereby improving the skill of the ensemble mean at proxy-ungauged locations (i.e, locations where observations were withheld for cross-validation). Locations downstream from assimilated observations are improved most although locations upstream are usually improved as well, even if they are far from neighbouring stations. This is likely due to two reasons: 1) constant biases in the river discharge estimates that are propagated downstream and hence can be accounted for when a downstream observation is assimilated, and 2) the daily aggregation of the river discharge extending the time period for which a downstream observation provides relevant information. If the error patterns of the ensemble mean at a location differ from those at nearby stations the method struggles to spread the observational information correctly. At shorter lead-times the reliability of the ensemble is slightly improved due to the decrease in the error of the ensemble

mean. However, at longer lead-times the ensemble spread is often too large leading to an under-confident forecast.

The ability of the method to correct the forecasts typically depends on the consistency of the error vectors between nearby locations. The localisation method implemented here depends only on the distance from the station along the river network. The method does successfully correct the forecast both up- and downstream; however, if the station is on a different river, or if there is a confluence between the station and the grid-box of interest, the errors are often not consistent for as long a distance along the river network. Therefore, it could be beneficial to investigate the impact of including information about the river stretch into the localisation length scale. Additionally, the errors were found to be more consistent when the catchment was impacted by large-scale weather systems. It may therefore be useful to incorporate information about the meteorological situation into the localisation function as well.

The covariance inflation method used here maintains consistency between the spread of the error ensemble and the spread of the hindcast (Section 5.5.2). This successfully stops the error ensemble from collapsing such that the observations are not ignored. However, in situations where the uncertainty of the hindcast ensemble is over- or under-estimated the covariance inflation does not correct the error ensemble covariances correctly. This can lead to the observations being ignored as for short lead-times in Fig. 5.5b. Additionally, if the hindcast perturbations do not provide an accurate estimate of the true error ensemble perturbations, this method may introduce errors which could be the cause for the slight degradation in skill of the ensemble mean with lead-time shown in Fig. 5.6e, 5.6g, and 5.6h. Correcting the spread of the hindcast before using it in the inflation of the error covariances could solve this issue. Covariance inflation techniques that use the innovation statistics could be used to first adjust the hindcast ensemble (e.g., Kotsuki et al., 2017). Alternatively, a lower threshold for the variance of the ensemble could be set - say 10% of the ensemble mean, similar to the observation error covariance matrix, or the root mean square-error of the initial conditions. However, caution is needed not to artificially inflate the covariances too much such that the analysis increments become too large.

As discussed in Section 5.7.2.2, the resulting ensemble must be adjusted in some cases to avoid negative discharge values (Section 5.4.1). This does in some cases lead to ensemble members close to $0\,m^3s^{-1}$ when a zero flow value is unlikely (e.g., Fig. 5.5c). This occurs due to the analysis increment being larger in magnitude than the value of some of the raw ensemble members. In general, this is due to the skewed distribution of discharge (Bogner et al., 2012). Future work could look into applying anamorphosis, or normalising transformations, to make the ensemble distribution more Gaussian-like (Nguyen et al., 2023; Bogner et al., 2012). This was not done in this proof-of-concept study for simplicity and to facilitate the interpretation of the errors. The results also showed that the covariances between grid boxes on larger rivers and the station locations tend to be large even when the correlation is small. This is due to larger rivers having larger variances which is partially due to their larger river discharge magnitudes. Localisation enforces a distance dependence on the covariance magnitudes. However, transforming the river discharge values to be comparable across the domain may help regulate the covariances based on river size. A transformation between river discharge and specific discharge (river discharge divided by upstream area) could be used to ensure that the ensemble covariances more accurately represent the true relationship between locations.

In this study, the initial estimate of the error ensemble mean is defined using the observations and the simulation forced with meteorological observations from the 10 days before the forecast. The average difference between the observations and simulations is calculated at gauged locations and interpolated to every grid-box using inverse distance weighted interpolation. The aim is to provide a physically plausible first guess of the errors which is then updated at each timestep. By taking the average over a 10-day period, we aim to capture the biases of the hydrological model but also to allow for seasonal/dynamic variation in this bias. However, the choice of 10 days has not been tuned, and may be more applicable to larger catchments with slowly changing errors than for smaller catchments (Matthews et al., 2022). Further research into the accuracy of the initial error ensemble, and how it influences the skill of the error-corrected ensemble, is needed. It should be noted that this component of the method is an implementation choice and can be adjusted depending on system configuration and data availability. The only requirement is that the initial error ensemble is physically

plausible as there is no warm-up period within this application (Kim et al., 2018).

We assume that the errors are sufficiently slowly changing such that a persistence model can be used to propagate the errors between timesteps. It should be noted that the LETKF updates the errors at each timestep so the analysis errors used to correct the hindcasts are not constant for all timesteps. However, the assumption that the errors are slowly changing is likely only true for larger rivers that respond more slowly. Future studies could investigate the use of a simple time-varying evolution model. The model would need to be simple enough that the calculations do not add too much computational time to the method. Additionally, the error values at every grid-box would need to be evolved; therefore, the evolution model should either rely only on the model output or must be spatially interpretable if using observations. For example, a model dependent on the hindcast river discharge magnitude could be used to evolve the errors between timesteps.

The leave-one-out approach used in this study allows the corrected ensembles to be assessed at proxy-ungauged locations. However, only one station is omitted at a time. Future work could use a block cross-validation strategy whereby multiple stations are omitted simultaneously (Roberts et al., 2017). This would allow the impact of the density of stations and their specific locations along the river network to be investigated more thoroughly (Rakovec et al., 2012). The impact of not having any observations along a river stretch could also be more thoroughly investigated. One benefit of this method is that the assimilated observations do not necessarily need to be traditional in-situ observations but could come from Earth Observation (EO; Durand et al., 2023), crowdsourced or community observations (Le Coz et al., 2016; Etter et al., 2020), or camera-based sensors (Vandaele et al., 2021). The key requirement is that an observation operator can be defined. Observation operators map the state of the system from state space to observations space. In our study the observation operator selects the grid box that represents the location of the station on the modelled river network. The mapping of the station locations from the physical river network to the modelled river network is not trivial and several studies have attempted to automate this step (Isikdogan et al., 2017; Li et al., 2018). If this mapping is incorrect then representation errors can be introduced (Janjić et al., 2018). For example, if a station on a bypass channel

is incorrectly located on the main channel, observations from the station will undoubtedly provide erroneous information in the update step.

The code developed for this study is designed to allow for research flexibility rather than operational efficiency. However, the error-adjustment of a single forecast took on average 8.5 minutes for the whole of the Rhine-Meuse catchment - a large catchment. This suggests that, with proper parallelisation, the method could be operationalized and applied to all gauged catchments in Europe. Before that, though, the method needs to be evaluated on additional catchments. The Rhine was selected because it is highly gauged, but this also means that the raw ensemble's skill is relatively high due to the effectiveness of the hydrological model calibration process. This could influence the method's performance in two ways: 1) the error ensemble may evolve more linearly than in less calibrated catchments, and 2) the hindcast ensemble's covariance may better represent the covariances between the estimated errors. The next step should be applying this method to a catchment with lower skill than the Rhine.

The method presented in this study spreads observation information along the river network but cannot yet be used as a post-processing method because observations from the hindcast period (the future) are assimilated. We envisage the method being developed further to make it applicable operationally as a hydrological forecast post-processing method. Nevertheless, it may still be useful in certain situations, such as post-event analysis. After a flood event an assessment is often performed estimating the severity of the event as well as potential causes and mitigating factors. However, in-situ river gauges only present a snapshot of the event at specific locations and are often damaged during flood events, resulting in missing or incorrect data. EO estimations of river discharge could fill in some of the gaps but this would depend on the satellite's orbit and its availability at the right time (Douben, 2006). Reanalysis is another option, but it requires additional hydrological model runs and may contain errors due to the structure of the hydrological model or errors in the meteorological observations. The method proposed here could offer a domain-wide estimate of observations without requiring additional model runs or a "warm-up" period typically needed in hydrological simulations to stabilize antecedent water storage within the catchment.

5.9 Conclusion

We present and evaluate a data-assimilation-inspired method for spreading observation information from gauged to ungauged locations in a post-processing environment. This method enables the error correction of ensemble simulations at all grid boxes. The method utilises state augmentation within an LETKF framework to estimate an ensemble of error vectors. The error vectors are then used to correct each hindcast ensemble member separately.

Overall, the method successfully reduces the errors of the ensemble mean at ungauged locations in leave-one-out experiments. The adjusted ensemble mean has a higher correlation with the observed river discharge and is more able to capture the variability of the river discharge at a point. Whilst the magnitude of the errors is reduced the ensemble spread is not adjusted sufficiently resulting in an under-confident ensemble spread at longer lead-times. The adjusted ensembles are spatially and temporally consistent with the river discharge predictions showing smooth evolution both between grid-boxes on the same river and between lead-times. The method is most limited at locations further upstream than the assimilated observations and for hindcasts with unrealistically small ensemble variance, which most often occurs at shorter lead times. These limitations could be reduced by further investigation into the localisation approach, for example having a different localisation length upstream and downstream from the assimilated observation, and the covariance inflation approach, which may involve applying a spread-correction to the hindcast ensemble as well as the error ensemble.

Our method of spreading observation information could be used to improve post-event analysis. However, as the computational requirements and processing time are both small the method could also be developed further to allow for its application to the post-processing of operational forecasts. The prediction of river discharge at ungauged locations is a crucial challenge for hydrological research and once successfully achieved will allow for better preparedness for floods.

5.10 Summary of Chapter 5

In this chapter, I develop a new method for propagating observation information from gauged to ungauged locations in a forecast post-processing environment, addressing *Objective 2*. The new method uses modified data assimilation techniques. State-augmentation is used within a Local Ensemble Transform Kalman Filter (LETKF) to error correct ensemble hindcasts at every grid-box. To avoid additional executions of the hydrological model I make the assumption that the raw ensemble hindcast is a good approximation of the analysis state of the river discharge. New methods of defining the localisation length, of covariance inflation, and of generating a spatially consistent initial estimate of the errors are defined. (Q2.1) In general, the observational information is successfully propagated along the river network correcting the ensemble mean at ungauged locations. (Q2.2) Almost all evaluated locations are improved irrespective of whether they are upstream or downstream of assimilated observations. However, when the errors at a proxy-ungauged location showed very different patterns to those at nearby stations the method struggled to correct the errors.

The method of propagating information along the river network using a LETKF is used in Chapter 6. In an operational forecasting system observations are not available during the forecast period (i.e., the future). Therefore, in Chapter 6 I combine the information propagation method developed in this chapter with the at-gauge post-processing method evaluated in Chapter 4. The at-gauge post-processed forecasts are treated as *proxy-observations* during the forecast period (*Objective 3*). The evaluation period and catchment are kept the same in Chapters 5 and 6 to allow direct comparison and to disentangle the strengths and weaknesses of the two components of the post-processing method for ungauged locations (the at-gauge post-processing and the information propagation). The results from this chapter that will be relevant for Chapter 6 are:

- The ensemble mean is generally improved across the domain in terms of correlation and variability.
- In terms of the mean bias the similarity of error patterns between neighbouring stations is a limiting factor. There are many rivers which show large variation in the mean bias

between stations (e.g., under and over-estimation at different stations on the same river stretch).

- The ensemble mean at the Domrémy-la-Pucelle station (the most upstream station on the Meuse) is not corrected in terms of any metric.
- The ensemble is over-confident at shorter lead-times (although less so that the raw forecast) and under-confident at longer lead-times.

Chapter 6

A new post-processing method for ensemble river discharge forecasts at ungauged locations

This chapter addresses the third objective of the thesis: *To explore the use of proxy-observations to facilitate the post-processing of ensemble river discharge forecasts at ungauged locations.*

A two-step approach to post-processing ensemble river discharge forecasts at ungauged locations is developed. The at-gauge post-processing method evaluated in Chapter 4 is used to generate proxy-observations. These proxy-observations are assimilated into the ensemble forecast using the LETKF-based error propagation method presented in Chapter 5. The new post-processing method is evaluated for the Rhine-Meuse catchment using a leave-one-out cross-verification strategy to evaluate the skill at ungauged locations.

In Section 6.7, I summarise the key results of the chapter and identify connections with the other chapters in this thesis.

Abstract

Accurate and timely river discharge forecasts provide crucial advanced notice to decision makers for disaster mitigation and water resource management. However, forecast errors can lead to suboptimal decision-making. Post-processing is commonly used to correct these errors, but it is typically limited to gauged locations where observations are available, leaving ungauged areas uncorrected. This study presents a novel post-processing approach for ensemble forecasts at ungauged locations. The method consists of two steps: (1) at-gauge estimation, where the Copernicus Emergency Management Service's European Flood Awareness System (EFAS) post-processing method is applied at gauged locations, and (2) knowledge transfer, which employs a Local Ensemble Transform Kalman Filter (LETKF) to propagate error corrections to ungauged locations using ensemble error covariances. The method is evaluated in the Rhine-Meuse catchment using a leave-one-out strategy to assess forecast skill. Results show improved forecasts at short lead times (up to three days) in both the ensemble mean and full distribution. The effectiveness of the method for post-processing at ungauged locations is constrained by the accuracy of the at-gauge corrections. Ungauged locations which benefit from post-processing typically occur along river stretches for which the at-gauge estimate is most accurate. However, the method's flexibility allows advances made for at-gauge post-processing method to be efficiently harnessed for post-processing of ungauged locations. A key challenge is the accurate quantification of uncertainty in the raw ensemble, which affects the propagation of error corrections along the river network and should be a focus of future research. Despite these challenges, the proposed approach demonstrates promising potential for improving river discharge forecasts at ungauged locations, ultimately providing more accurate information to support decision-making.

6.1 Introduction

Accurate river discharge forecasts are vital tools for flood mitigation and water resource management, helping decision-makers protect communities and allocate resources effectively (WMO, 2022). However, forecasts often contain errors which can undermine decision-

making potentially leading to inappropriate or untimely responses. To address forecast uncertainty, a range of techniques have been developed, including pre-processing of the meteorological forcings (Zhang et al., 2025b), hydrological model calibration (Althoff and Rodrigues, 2021), data assimilation (Camporese and Girotto, 2022), and post-processing of the river discharge forecast (Tyralis and Papacharalampous, 2023). Post-processing methods have gained significant interest for operational forecasting due to their ability to increase the skill of the forecasts whilst also being computationally efficient (Liu et al., 2022).

Post-processing methods typically define a statistical model relating the raw forecasts to observed river discharge, and use these models to correct future forecasts (Roulin and Vannitsem, 2015; Ye et al., 2014). The structure of these error models depends on the post-processing method of choice, but they always rely on river discharge observations—whether in near real-time or from historical records (Li et al., 2017). However, river gauges, the primary source of river discharge observations, are sparse and declining in number (Krabbenhoft et al., 2022; Lavers et al., 2019; Hannah et al., 2011). The scarcity of river gauges limits the spatial extent of post-processing, making methods for ungauged locations increasingly important.

Making use of the research conducted for at-gauge post-processing methods (Li et al., 2017), techniques for post-processing ungauged locations typically have two steps. First, an at-gauge post-processing method is used to estimate error-correction parameters or an improved estimate of the river discharge at gauged locations. Second, an information propagation technique is used to transfer the error-correction parameters or propagate the improved river discharge estimate to ungauged locations. Existing techniques for transferring information from gauged to ungauged locations include donor-receptor catchments (Hales et al., 2023), where correction parameters are defined for gauged locations and applied to hydrologically similar ungauged locations, and methods which spatially interpolate error-correction parameters from gauged to ungauged locations (Skøien et al., 2021). However, these approaches primarily transfer parameters rather than direct discharge estimates, so can only be used with post-processing methods that generate explicit error-correction parameters. An alternative approach involves using a river routing model to propagate error-corrected discharge

estimates downstream (Bennett et al., 2022). While effective, this method can reintroduce errors from the routing model and may require significant computational resources.

In this study, we use an approach based on the Local Ensemble Transform Kalman Filter (LETKF; Hunt et al., 2007) to propagate information from gauged to ungauged locations. This method has previously been shown to improve historic ensemble simulations at ungauged locations when observations are available at stations within the catchment during the forecast period (Matthews et al., 2025a, Chapter 5). However, the use of observations from the forecast period (i.e., the future) invalidates the ensemble as a true forecast. To address this limitation, we propose using at-gauge post-processed forecasts as proxies for observations during the forecast period. Near real-time observations are often used in at-gauge post-processing to condition the forecast on the recent state of the river. However, in contrast to the study presented in Chapter 5, these near real-time observations are from timesteps prior to the forecast generation time so the validity of the forecast is maintained. Therefore, using at-gauge post-processed forecasts as proxies for observations within the LETKF-based method allows us to maintain the forecast validity whilst also propagating the error information to ungauged locations. The research questions we aim to address are:

- Can error corrections derived from at-gauge post-processed forecasts be effectively propagated to ungauged locations using a data-assimilation-based method?
- What factors influence the method's ability to improve the forecast skill at ungauged locations?

This paper is outlined as follows. In Section 6.2 we provide an overview of the at-gauge post-processing method used in this study, and the information propagation method used to propagate the error correction from gauged to ungauged locations. In Section 6.3 we provide details about the experimental design including our chosen case-study (the Rhine-Meuse catchment; Section 6.3.1), the data used (ensemble forecasts, river discharge observations, and simulations forced with meteorological observations; Section 6.3.2), and the leave-one-out evaluation strategy used to assess the skill of the post-processed forecasts at ungauged locations (Section 6.3.3). In Section 6.4 we present the results of our evaluation beginning

with an assessment of the skill of the post-processed forecasts at ungauged locations in Section 6.4.1. In Section 6.4.2 we investigate the factors which impact the skill of the post-processed forecasts. We follow the results with a discussion on the advantages and limitations of the method, and potential future work. In Section 6.6 we conclude that the proposed method is capable of improving the ensemble forecasts at ungauged locations. However, further investigation and development is necessary before the method could be used operationally.

6.2 Post-processing methods

This section describes the method proposed in this study to post-process ensemble forecasts at ungauged locations. There are two steps: 1) an at-gauge post-processing method is applied at all available gauge locations (Section 6.2.1), and 2) using an LETKF-based method the error information from the at-gauge post-processed forecasts is propagated from gauged to ungauged locations (Section 6.2.2). Figure 6.1 outlines the full post-processing method for ungauged locations.

6.2.1 At-gauge post-processing method

The at-gauge post-processing method used in this study is the operational post-processing method of the CEMS European Flood Awareness System (EFAS; Matthews et al., 2025b). The post-processing method is described in detail in Chapter 4. Therefore, here we give an overview only before focusing on how the output is used in this study in Section 6.2.1.1.

The at-gauge post-processing method accounts for the hydrological uncertainty and meteorological uncertainty separately before combining them to form the final output. Hydrological uncertainty arises due to errors in the hydrological model and the initial hydrological conditions (Herrera et al., 2022). These errors are accounted for using a Model Conditional Processor (MCP; Todini, 2008). The form of the MCP used here estimates the river discharge for the 15 days of the forecast period by conditioning a climatological distribution on observed and simulated river discharge values from the 40 days prior to the forecast initialisation date. An offline calibration is required to define a joint distribution between historic simulated and

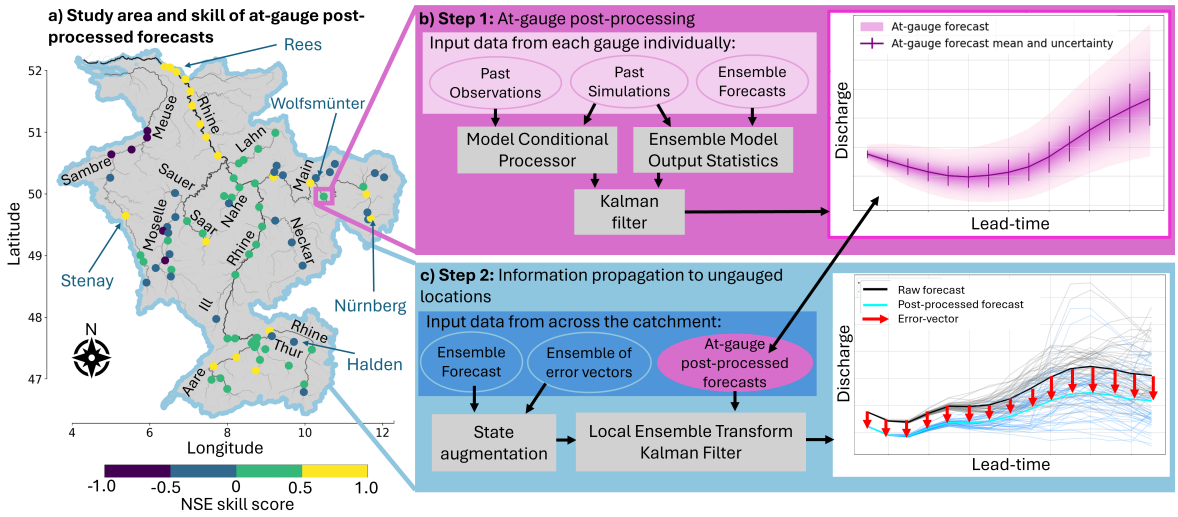

Figure 6.1: Diagram of the post-processing method for ungauged locations. a) The map shows the location of river gauges at which the at-gauge post-processing method is performed. The colour of the markers indicates the Nash Sutcliffe Efficiency (NSE) skill score for the at-gauge estimates (see Section 6.2.1.1) at a lead-time of 5 days. Positive (negative) NSE skill score values indicate the at-gauge post-processed forecasts are more (less) skillful than the raw forecasts. The five named stations (Rees, Wolfsmünter, Nürnberg, Halden, and Stenay) are discussed in Section 6.4.2. b) The at-gauge post-processing requires past observations and simulations forced with meteorological observations as well as the ensemble river discharge forecasts. The method consists of a Model Conditional Processor component used to quantify the hydrological uncertainty and an Ensemble Model Output Statistics component used to quantify the meteorological uncertainty. A Kalman filter is used to combine the output of both components. The river discharge output from the at-gauge post-processing method is used to define at-gauge estimates and their uncertainty. c) Augmented states are defined between the raw ensemble forecast and an ensemble of error vectors. To update the error vectors, the at-gauge post-processed forecasts are assimilated using a Local Ensemble Transform Kalman Filter (LETKF). The updated error vectors are added to the raw ensemble members to post-process the forecast at all grid-boxes (shown for one location).

observed variables which enables the MCP to be implemented.

Meteorological uncertainty arises due to errors in the meteorological forcings used to drive the hydrological models (Shu et al., 2023). Here, the meteorological forcings are the output from four numerical weather prediction systems (see Section 6.3.2.1) and their uncertainty is accounted for using an Ensemble Model Output Statistics method (EMOS; Gneiting et al., 2005). The EMOS method estimates spread correction parameters for the raw river discharge ensemble forecasts using data from the past 40 days. The spread correction parameters are used to adjust the raw ensemble. The output from the MCP component and the EMOS component are then combined using a Kalman filter (Kalman, 1960) to form the

final at-gauge post-processed forecast.

The MCP method, the EMOS method, and the Kalman filter assume the variables have a Gaussian distribution. Therefore, the normal quantile transform (NQT Bogner et al., 2012) is used to transform all data - river discharge observations, simulations forced with meteorological observations, and ensemble forecasts-to a standard Gaussian distribution, with mean 0 and standard deviation 1. The NQT requires the real, non-Gaussian Cumulative distribution Function (CDF) of the variable to be estimated before the transform can be applied. The estimation of distributions for the observed and simulated river discharge values is also performed during an offline calibration (Section 4.3.3).

The output of the at-gauge post-processing method is a 15-day probabilistic forecast for a specific gauged location. The forecast's probability distribution is defined via the 1st to the 99th percentiles for each lead-time.

6.2.1.1 At-gauge post-processed forecasts as proxy observations

In this study, we propose using at-gauge post-processed forecasts as proxy observations. This means observations from after the forecast initialisation time are not required, maintaining the validity of the forecasts. For each timestep of the forecast we require an estimate of the river discharge value at all river gauges, denoted $\mathbf{y}_k \in \mathbb{R}^{p_k}$, and the uncertainty associated with these estimates, denoted $\boldsymbol{\sigma}_k \in \mathbb{R}^{p_k}$ where p_k is the number of gauges where post-processing is performed. To define $\mathbf{y}_k$ and $\boldsymbol{\sigma}_k$ from the at-gauge post-processed forecasts we calculate the forecast means and standard deviations from the forecast percentiles. We then define $\mathbf{y}_k$ from the forecast means for lead-time k such that the i -th element corresponds to the i-th gauge, and similarly using the standard deviations for $\boldsymbol{\sigma}_k$.

It is assumed that the at-gauge estimates at different gauges are uncorrelated. The error-covariance matrix of the at-gauge estimates is therefore a diagonal matrix with all non-diagonal elements set to 0. The diagonal elements, which represent the uncertainty of the at-gauge estimates, are given by the variance, $\boldsymbol{\sigma}_k^2$. The at-gauge estimates and their error-covariance matrix are used within a LETKF in the information propagation step (Section 6.2.2). It is also assumed that the at-gauge post-processed forecasts are uncorrelated with the

prior augmented ensemble. This is likely not true as the at-gauge post-processed forecasts use the 51-member ensemble forecast that is also used to define the prior augmented ensemble. As the cross-correlation is unknown, this assumption simplifies the calculations considerably, but future work should account for the cross-correlation to avoid degrading the analysis (Waller et al., 2014).

6.2.2 Propagating error information from gauged to ungauged locations

In Chapter 5, a data-assimilation inspired method of propagating error information from gauged to ungauged locations using a Local Ensemble Transform Kalman Filter (LETKF; Hunt et al., 2007) was presented. Data assimilation is a technique used to combine multiple sources of information based on their respective uncertainties (Asch et al., 2016). Typically, in environmental studies, the sources of information are a prior forecast, generated using a model of the system of interest-here a hydrological model-and observations (Camporese and Girotto, 2022). Herein lies the main limitation of Chapter 5. The method relies on river discharge observations from the forecast period (i.e., the future), thereby preventing its application to post-process operational forecasts. We propose a solution by using real-time at-gauge post-processed forecasts (henceforth referred to as at-gauge estimates) in place of the observations. The method used to generate the at-gauge estimates is described in Section 6.2.1).

The core principle of the information propagation approach is to estimate an error vector, denoted $\mathbf{b}_k^{(i)}$, for each forecast ensemble member, $\mathbf{x}_k^{(i)}$, for each timestep k of the forecast period. The error vector represents the difference between the ensemble member and the true river discharge at each grid-box. The error vectors are then used to error-correct the forecast such that

$$\mathbf{z}_k^{(i)} = \mathbf{x}_k^{(i)} + \mathbf{b}_k^{(i)} \tag{6.1}$$

where $\mathbf{z}_k^{(i)}$ is the i-th ensemble member of the post-processed forecast at a timestep k.

To estimate the error vectors we use state augmentation, a common technique that allows the simultaneous estimation of the system state and parameters (e.g., Smith et al., 2013;

Martin et al., 2002). In the state augmentation step an error vector is appended to each of the ensemble members before the data assimilation method is applied, resulting in a river discharge component and an error vector component of an augmented ensemble. The data assimilation method used is the Local Ensemble Transform Kalman Filter (LETKF; Hunt et al., 2007) which consists of two steps: an update step and a propagation step.

In the update step the LETKF calculates a weighted average between the prior augmented ensemble and the at-gauge estimates, creating an updated ensemble called the analysis. First, the innovation vector, which contains the differences between the mean of the prior augmented ensemble and the at-gauge estimates, is calculated. The information provided by the innovation vector is then propagated to ungauged locations using the ensemble error-covariances calculated from the prior augmented ensemble (Lahoz and Schneider, 2014; Bannister, 2008a). The influence of each data source on the analysis is determined by their respective uncertainties. If the prior augmented ensemble spread is small—indicating high confidence in the prior augmented ensemble—the influence of the at-gauge estimates is small. Conversely, when uncertainty in the prior ensemble is high, the at-gauge estimates have a greater influence.

The propagation step evolves the analysis from timestep k to timestep $k + 1$. As we are applying the LETKF in a a post-processing environment, instead of propagating the river discharge component of the analysis between timesteps, we use the raw ensemble forecast values corresponding to timestep $k + 1$. Whilst mathematically the use of the raw forecast makes the LETKF sub-optimal, in-practice it removes the need for additional executions of the computationally expensive hydrological model. More information regarding the propagation of the forecasts is provided in Chapter 5. In addition, as we do not propagate the river discharge component of the analysis, this component does not need to be calculated. Therefore, we are able to only calculate the error vector component of the analysis reducing the required matrix computations and further reducing the computational expense of the post-processing method. The error vector component is assumed constant between timesteps. This is a common assumption when the true evolution of a variable is unknown (e.g., Martin et al., 2002; Pauwels et al., 2020; Rasmussen et al., 2016; Ridler et al., 2018). For the first timestep,

$k = 1$, an initial estimate of the error ensemble is generated using the same method as Chapter 5 and is based on recent observations and simulated values.

To improve the efficacy of the LETKF and to reduce the impact of spurious correlations caused by the small ensemble size, three additional modules are included in the method:

- **Localisation**: Localisation restricts the influence of the at-gauge estimates to geographically close grid-boxes for which the covariances are more likely to be correctly estimated (Hamill et al., 2001). Rather than using the Euclidean distance, the localisation distance is calculated along the river network which has been found to be beneficially in many hydrological studies (García-Pintado et al., 2015; El Gharamti et al., 2021b). The localisation length scale is set as the maximum distance between any grid-box and the closests river gauge. At-gauge estimates that are further than twice the localisation length scale have no influence within the LETKF. Defining the length scale in the way ensures all grid-boxes are updated, and reduces the potential for inconsistencies between grid-boxes (Section 5.5.1). The localisation is applied to the error-covariance matrix of the at-gauge estimates (Hunt et al., 2007).

- **Covariance inflation**: Covariance inflation is often used to ensure the uncertainty in the prior ensemble is not underestimated which can lead to filter divergence (Anderson, 2007). The covariance inflation method used in this study blends the analysis error-ensemble perturbations with an estimated future uncertainty based on knowledge of the system. The inflation method is inspired by the relaxation-prior-perturbations (Zhang et al., 2004). The estimated future uncertainty is derived from the raw ensemble covariance matrix. As we do not know the structure of the error-covariance matrix for the ensemble of error vectors, we assume that the raw forecast ensemble covariance matrix is a good approximation (e.g., as in Martin et al., 2002). The inflation parameter that determines the weighting between the analysis error ensemble perturbations and the estimated future uncertainty is calculated based on the magnitude of the changes in the raw forecast ensemble variance between timesteps (Section 5.5.2).

- **Non-negative discharge**: As river discharge is a positive-definite variable a non-

negativity constraint is applied at each lead-time after the error ensemble has been updated (Section 5.5.3).

At each timestep, after the error vectors have been updated by the LETKF they are used to post-process the raw forecast as described in Eq. (6.1). A leave-one-out framework is used to evaluate the skill of the post-processed forecasts at proxy-ungauged locations (Section 6.3.3).

6.3 Experiment design

We aim to evaluate the ability of the proposed post-processing method to correct ensemble forecasts at ungauged locations. This section describes the data and evaluation strategy employed to achieve this aim.

6.3.1 Case study

We investigate the efficacy of the post-processing method for the Rhine-Meuse catchment. The Rhine-Meuse catchment has a relatively dense river gauging station network with several gauges having long observation records (Krabbenhoft et al., 2022). This allows at-gauge post-processing, which requires historic observations (see Section 6.2.1), to be performed on most river stretches within the catchment (see locations of stations in Fig. 6.1a). Within the EFAS drainage network both the Rhine and the Meuse have the same outlet hence they are considered a single basin. The Rhine-Meuse catchment is the 5th largest catchment in EFAS with 7812 grid-boxes (resulting in a drainage area 195,300 km^2). The hydrological regime varies throughout the catchment due to differing river sizes, geology, and land and water usage (Dieperink, 1997). Starting in the Swiss Alps, the Rhine flows in a generally northwesterly direction through six countries (draining from nine countries) before reaching the North Sea (Uehlinger et al., 2009). Main tributaries of the Rhine include the Aare, Neckar, Main, and Moselle rivers. The Meuse rises in France and flows north joining with the Rhine at the Rhine–Meuse–Scheldt delta (Berger, 1992,).

6.3.2 Data

6.3.2.1 Ensemble river discharge forecasts

Ensemble forecasts consist of many possible future scenarios and aim to account for uncertainty in the forecast (Troin et al., 2021; Wu et al., 2020; Cloke and Pappenberger, 2009). This study uses ensemble forecasts from the Copernicus Emergency Management Service's European Flood Awareness System (EFAS) for 2021 (available from Barnard et al., 2020b). The EFAS forecasts are multi-model forecasts created by forcing the calibrated LISFLOOD hydrological model (De Roo et al., 2000; van der Knijff et al., 2008) with the output from four numerical weather prediction (NWP) systems, including two ensemble meteorological forecasts and two deterministic forecasts (Matthews et al., 2025b). The maximum lead-time differs between the NWPs with the medium-range ensemble of the European Center for Medium-range Weather Forecasts (ECMWF) having the longest forecast horizon of 15 days (Haiden et al., 2024). The ensemble river discharge forecasts are primarily initiated from a single set of initial conditions with the exception of the forecasts driven by the German Weather Service's deterministic NWP which have a separate set of initial conditions (Arnal et al., 2019). In general, the river discharge ensemble spread is narrow at short lead-times due to the restricted set of initial conditions, but increases as the different meteorological forcings propagate through the system (Smith et al., 2016).

The ensemble forecasts have three functions within this study: 1) they are post-processed at gauged locations using the post-processing method described in Section 6.2.1, 2) they are used to estimate the error-covariance matrix in the information propagation step of the post-processing method described in Section 6.2.2, and 3) they are the raw forecast that are corrected by the post-processing method and used as the benchmark for all skill scores. The full multi-model EFAS ensemble is used in the at-gauge post-processing step. However, only the ensemble members driven by the ECMWF ensemble are used as the benchmark forecast, and in the estimation of the error-covariance matrix due to its relatively large ensemble size (51 members). Using a single-model ensemble avoids issues caused by the different forecast horizons of the NWPs. In this study, we use the EFAS version 4 (operational from 14 October

2020 to 20 September 2023) ensemble forecasts from the 00 UTC cycle for each day of 2021 resulting in a total of 365 forecasts. The ensembles are aggregated to daily timesteps to reduce the computational demand within the study. The forecasts have a spatial resolution of 5km $\times$ 5km with a ETRS89 Lambert Azimuthal Equal Area Coordinate Reference System (Arnal et al., 2019). Each ensemble member predicts the average river discharge in each grid-box for each timestep.

6.3.2.2 Observations

The river discharge observations used here were provided by local authorities and collected by the CEMS Hydrological Data Collection Centre. More information is available at `https://european-flood.emergency.copernicus.eu/en/european-flood-awareness-system-efas`. Eighty-nine river gauges within the Rhine-Meuse catchment are used in the evaluation presented here. These stations were selected due to their complete timeseries for the evaluation period of this study (1 January 2021 to 15 January 2022). River discharge observations were used in four parts of this study:

1. Historic observations from the 89 stations of interest were used to calibrate error-models for use in the at-gauge post-processing method (see Section 6.2.1). The length of the calibration period varies between stations with a maximum period from 1 January 1990 to 20 December 2020 (median calibration period of 12 years).

2. Observations from the 40 days prior to a forecast initialisation time at the 89 stations of interest were used in the at-gauge post-processing method to condition the forecast on recent observations (see Section 6.2.1).

3. Observations available from any station for the 10 days prior to a forecast initialisation time (maximum of 505 stations) were used to estimate the initial error ensemble (see Section 6.2.2).

4. Observations from the evaluation period of the study were used to evaluate the skill of the forecasts at the 89 stations of interest (see Section 6.3.3.2)

The localisation length scale of the information propagation method is defined as outlined in Section 6.2.2 as the maximum distance between any grid-box and its closest river gauge. In this study the localisation length scale is set as 262 km.

6.3.2.3 Simulation forced with meteorological observations

CEMS generates a longterm simulation from 1990 to near real-time by driving the LISFLOOD hydrological model with meteorological observations (available from Mazzetti et al., 2020b). The simulation for the grid-box to which the river gauge is mapped is extracted from the full dataset and used for the offline calibration of the at-gauge error-models, within the at-gauge post-processing method, and to estimate the initial error ensemble (see points 1-3 of the use of observations in Section 6.3.2.2).

6.3.3 Evaluation strategy

6.3.3.1 Experiment framework

In this study we evaluate the forecast skill of a post-processing method intended for application at ungauged locations. A leave-one-out spatial cross-validation strategy is used to asses the skill of the forecasts at ungauged locations (Geroldinger et al., 2023). For each station we run an experiment where the at-gauge estimates (described in Section 6.2.1) for that station location are not assimilated imitating an ungauged location. The post-processed forecasts are then assessed at the proxy-ungauged location using river discharge observations as the 'truth'. Note we refer to the proxy-ungauged locations simply as ungauged locations throughout. Forecasts are created and evaluated for each day of 2021 resulting in 365 evaluated forecasts. The metrics used to evaluate the forecasts are described in Section 6.3.3.2.

We also conduct all station experiments where all the at-gauge estimates available are assimilated. These experiments are used to investigate the influence of the at-gauge estimates rather than to evaluate the skill of the post-processed forecasts.

6.3.3.2 Evaluation metrics

To assess the forecast skill, we use five metrics: the Continuous Ranked Probability Score (CRPS), the Nash-Sutcliffe Efficiency (NSE), the correlation, the bias ratio, and the variability ratio. Since both raw and post-processed forecasts are ensembles, we use the Continuous Ranked Probability Score (CRPS Hersbach, 2000) to assess the overall skill of the forecast distributions. The CRPS measures the difference between the forecast distribution and the observed values, providing an evaluation of probabilistic predictions. The CRPS has a perfect value of 0 and ranges from 0 to ∞. The Nash-Sutcliffe Efficiency (NSE Nash and Sutcliffe, 1970) is used to evaluate the skill of the ensemble means. As a normalised form of the root-mean-square error (RMSE), the NSE provides insight into the overall difference between forecasts and observations and is directly comparable across locations. The NSE penalizes large errors more heavily than small ones, making it particularly useful for identifying poor performing forecasts. The NSE has a perfect value of 1 when the observations and forecasts match exactly. Correlation, bias ratio, and variability ratio are components of the Kling-Gupta Efficiency (KGE) metric and each capture a different type of forecast error relevant to river discharge predictions (Gupta et al., 2009; Kling and Gupta, 2009). Specifically, correlation measures timing errors, the bias ratio quantifies systematic over- or underestimation of river discharge, and the variability ratio assesses the forecast's ability to capture discharge variability. Each component has a perfect value of 1 with correlation ranging from -1 to 1, and the bias and variability ratio ranging from $-\infty$ to ∞.

Each metric is computed separately for each station and lead-time by extracting the corresponding lead-time from each forecast. All 365 forecasts are used to calculate the metrics with the exception of Fig. 6.2b where we calculate the NSE and CRPS for each month separately. To assess the level of improvement provided by post-processing we use skill scores, SS, calculated as

$$SS_{metric} = \frac{\text{metric}_{post} - \text{metric}_{raw}}{\text{metric}_{perf} - \text{metric}_{raw}} \tag{6.2}$$

where metric_{post}, metric_{raw}, and metric_{perf} are the values of the metric of interest for the post-processed forecasts, the raw forecasts, and a perfect forecast, respectively (Harrigan et al., 2018b). Positive skill scores indicate post-processing has improved the skill of the forecast whereas negative skill scores indicate a worsening of the skill.

6.4 Results

In this section we present an evaluation of the post-processed forecasts at ungauged locations (Section 6.4.1) and an initial investigation into the factors that influence the skill of the post-processed forecasts (Section 6.4.2). For stations and river names mentioned in this section please see the map in Fig. 6.1a. For clarity, the at-gauge post-processed forecasts that are assimilated are referred to as the 'at-gauge estimates' (method is given in Section 6.2.1). The error-corrected forecasts derived using the LETKF-based method presented in this manuscript are referred to as the 'post-processed' forecasts (method is given in Section 6.2.2). The uncorrected forecasts are referred to as the 'raw' forecasts. Evaluation at ungauged locations is achieved using the cross-validation strategy outlined in Section 6.3.3.

6.4.1 Post-processed forecast skill

The skill of the post-processed forecasts at ungauged locations is assessed using spatial cross-validation and benchmarked against the raw forecasts. At a lead-time of 1 day, the skill of the ensemble mean (in terms of the NSE) and the ensemble distribution (in terms of the CRPS) are improved at over 75% of the ungauged locations (Fig. 6.2a). Both sets of skill scores show decreasing forecast skill with increasing lead-time. The post-processed forecast mean is improved compared to the raw forecast at the majority of stations up to a lead-time of 5 days. For the forecast distribution this level of improvement is seen until a lead-time of 6 days (not shown). The range of NSE skill scores is wide across all lead-times. At a lead-time of 3 days some stations show a 50% improvement of the ensemble mean (NSE skill score above 0.5) and some stations show a degradation of 100% (NSE skill score below -1). For lead-times greater than 5 days the range of NSE skill scores is reduced but this is due to a

decrease in the skill of the raw forecasts (i.e., naïve skill Pappenberger et al., 2015b). The range of CRPS skill scores at each lead-time is much smaller. For lead-times greater than 11 days, both skill scores have similar distributions as for an 11-day lead-time so are not shown for simplicity.

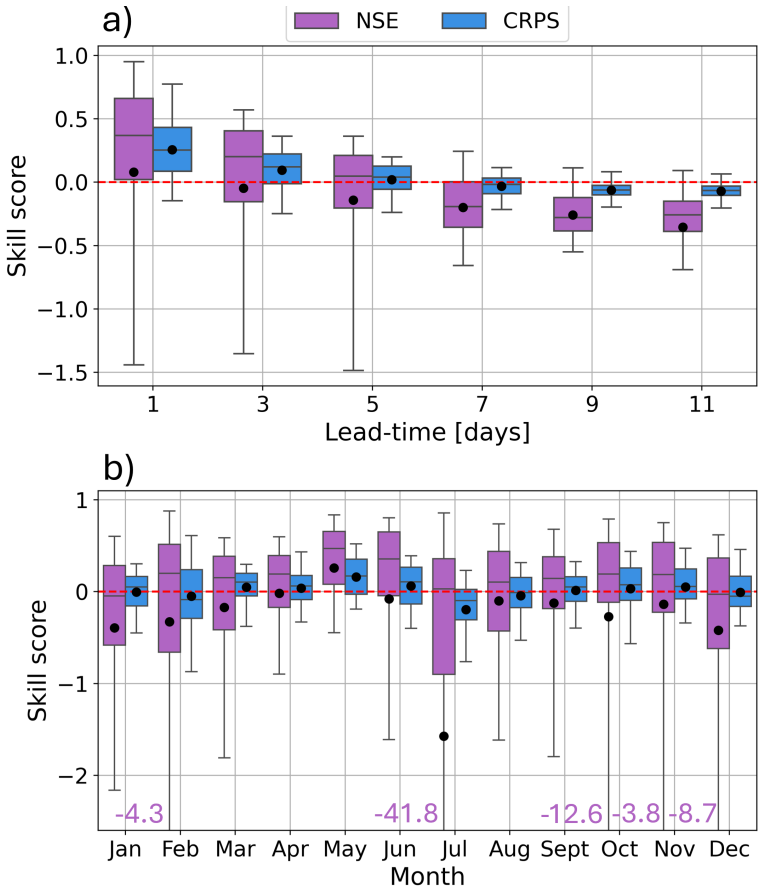

Figure 6.2: Skill scores of the post-processed ensemble mean (purple) and distribution (blue) at ungauged locations a) across lead-times and b) across months for a lead-time of 5 days. The red dashed lines show skill scores of 0. Positive skill scores indicate improvement and negative skill scores indicate degradation. Whiskers extend between the 5th and 95th percentiles. Black dots indicate the mean. Purple text states the extent of the whiskers where they extend beyond the axis limits.

We now investigate how the skill varies throughout the evaluation period focusing on a lead-time of 5 days (for which, overall, forecasts are improved at approximately half of stations; Fig. 6.2a). For most months both the NSE and CRPS values are improved by post-processing at more than 60% of stations with forecasts in May showing the most improvement. However, post-processing worsens the CRPS in general for February, July, August, and December. The NSE also decreases for these months with the exception of

February for which 61% of locations are still improved. Additionally, the range of NSE skill scores is large for these four months. These four months correspond to times when parts of the catchment experienced high flows. In February, the flooding primarily impacted the lower middle Rhine area, and in July the flooding primarily impacted the western tributaries of the Rhine (Mohr et al., 2022). Locations for which the post-processing worsens the forecast skill are typically those which experience the high flows (not shown) suggesting a limitation of the method for high flows.

To investigate the spatial variability of the forecast skill we again focus on a lead-time of 5 days allowing us to investigate the locations where the post-processing method is most beneficial (black outlines of the markers in Fig. 6.3). We compare the three components of the KGE (namely, the correlation, variability ratio, and bias ratio) for the ensemble means of the raw and post-processed forecasts at each location (Fig. 6.3). For the correlation, the spatial distribution of locations for which the post-processing is beneficial largely follows the spatial pattern of the skill of the adjusted at-gauge estimates (shown in Fig. 6.1a). Generally, the behaviour is regional with several consecutive stations being improved (or degraded) along a river. Notably, most locations along the Rhine are improved by post-processing, consistent with the locations of the at-gauge estimates with the highest skill. The correlation is also improved upstream on the Saar, the Nahe, and the Aare. In other places, such as the Main, the Necker, and the Moselle, the correlation is worsened by post-processing. These locations are consistent with locations of lower skill in the at-gauge estimates. For some of the degraded stations, in particular along the Main, the degradation is small (see similar colours of the left and right side of the markers). However, for some locations the degradation is larger. In general, the stations which show large degradation in the correlation tend to be the most upstream or only station on their river (e.g., the most upstream stations on the Meuse, the Sauer, the Ill, and the Thur).

The variability ratio shows an almost exact opposite spatial pattern of behaviour compared to the correlation. The majority of locations on the Moselle and Main are improved, as well as some locations on the Meuse, but few locations are improved along the Rhine (Fig. 6.3b). Whilst the specific locations that are improved are different, the variability ratio is improved

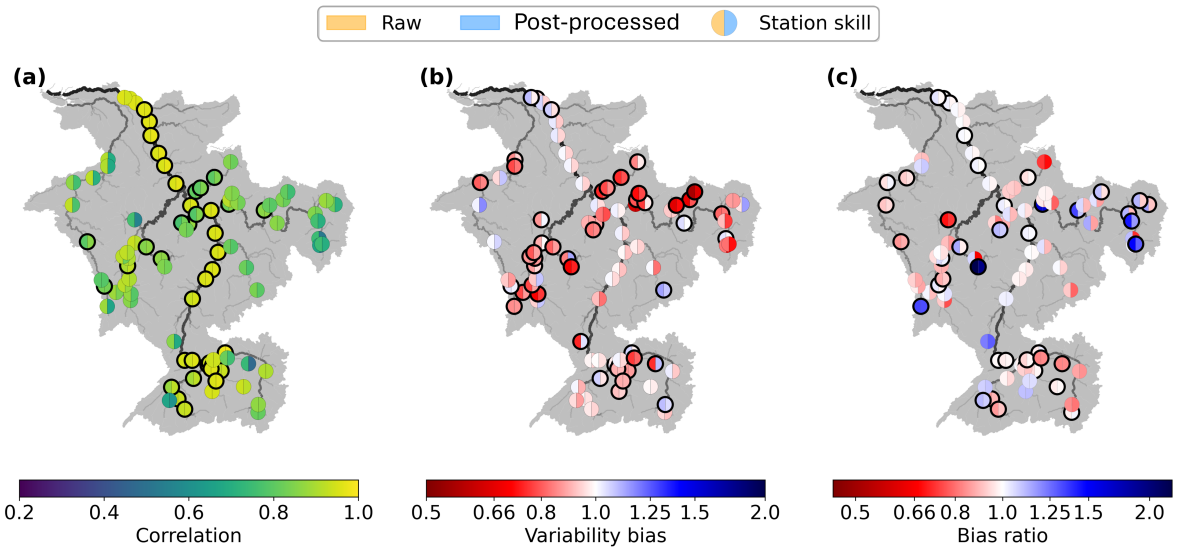

Figure 6.3: Components of the Kling Gupta Efficiency metric: a) correlation, b) variability ratio, and c) bias ratio, for a lead-time of 5 days for each station for the ensemble means of the raw forecasts (left half of markers) and the post-processed forecasts (right half of markers). A black outline to the marker indicates the forecasts are improved by post-processing. A perfect score is 1 for all three metrics.

for sequential stations creating regions of improvement and degradation. In general, locations with larger variability errors in the raw forecasts (indicated by darker colours on the left half of the marker) are corrected and locations with smaller variability errors, such as the Rhine, are worsened. The raw forecasts tend to underestimate the variability of the river discharge (red markers) and the post-processing is able to correct these errors. However, the worsening of the correlation along the Main and Moselle suggests that the timing of the peaks is incorrect even though the magnitude is better captured. It is notable that the Domrémy-la-Pucelle station, the most upstream station on the Meuse, shows an improvement in terms of the variability ratio, which it did not when observations were used as the at-gauge estimates in (Chapter 5) (see Fig. 6.7). The at-gauge estimate at the Domrémy-la-Pucelle station is also not skillful at a lead-time of 5-days (Fig. 6.1a). This suggests that the propagation of information from neighbouring stations is providing some useful information that is not captured in the error-model of the at-gauge post-processing method.

The bias ratio shows an entirely different spatial pattern to both the correlation and the variability ratio. The locations for which post-processing is beneficial are more sporadically located. For example, some locations along the Rhine are improved but the bias is increased at

others. The raw forecast of the Rhine is relatively unbiased particularly in the lower stretches of the river (shown by the near white of the left side of the marker). Therefore, small errors in a post-processed forecast can result in the forecast being degraded in terms of the bias. Many of the most upstream stations on the rivers (e.g., the Meuse, the Sauer, the Saar, the Aare, the Rhine, and the Main) are improved. These stations typically have larger biases in the raw forecasts and therefore the post-processing method is able to better estimate a consistent bias correction without being penalised for incorrectly estimating dynamic errors.

6.4.2 Factors impacting the forecast skill

In Section 6.4.1 we found that the improvement due to post-processing varied depending on the flow level and location as well as due to the metric of choice. In this section we will investigate some factors that influence the forecast skill gained by post-processing. First, using all station experiments (Fig. 6.4), we investigate the relationship between the analysis error ensemble mean which can be considered the *estimated error* of the raw forecast mean (background colour), the *true error* of the raw forecast mean (markers in left panels), and the difference between the raw forecast mean and the at-gauge estimates which are the *innovations* (markers in the right panels). Two flow scenarios are considered: 1) high flows on 15 July 2021 and 2) normal flows on 13 October 2021. We again look at the 5-day lead-time. Ideally, the estimated error would match the true error at the locations of markers which would result in a perfect forecast. Typically, the forecast is improved (shown by cyan marker outlines) when the estimated error has the same sign as the true error such as along the lower Rhine and Moselle in Fig. 6.4a. However, there are some cases where the estimated error over-corrects leading to a larger absolute error (e.g., along the Meuse in Fig. 6.4b).

Both the true and estimated errors are larger during the high flow event compared to the normal flow situation (note the different colour scales). This is expected as flood events are harder to predict and the EFAS forecasts are relatively unbiased in the Rhine catchment (see the 'Evaluation' layers at `https://european-flood.emergency.copernicus.eu/`). The estimated errors are also non-negligible for more of the river network with smaller rivers

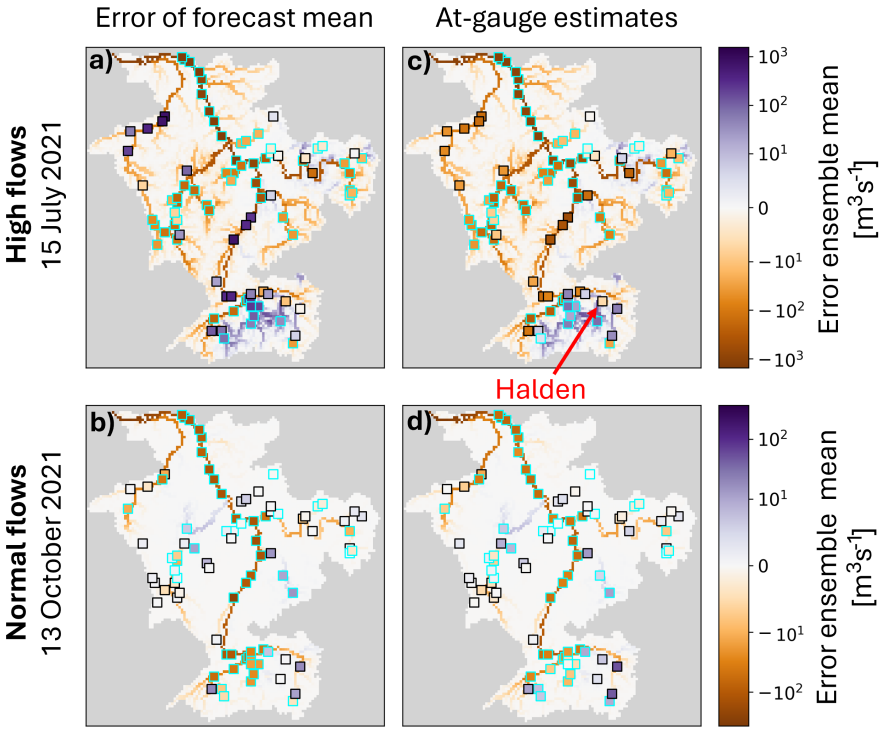

Figure 6.4: Error of the ensemble mean at a lead-time of 5 days for the forecasts generated on 10 July 2021 (a and c) and 8 October 2021 (b and d) for the all station experiments. The square markers show the difference between the raw forecast mean and the observations (a & b), and the raw forecast mean and the at-gauge estimates (c & d). A cyan marker outline indicates the forecast ensemble mean is improved by post-processing. The Halden station discussed in the text is labeled.

visible in Fig. 6.4a compared to Fig. 6.4b. Whilst we can not evaluate the forecasts at real ungauged locations, the true errors are also smaller in Fig. 6.4b especially for smaller rivers. This suggests that the estimated errors are capturing a real increase in the spatial extent of the errors.

There are regions where the sign of the error is incorrectly estimated. For example, in Fig. 6.4a locations on the lower Meuse are estimated to have negative errors (orange background) but the true errors are positive (purple markers). We can see the cause of this issue by looking at the innovations (Figs. 6.4c and 6.4d). The estimated errors almost exactly follow the spatial pattern of the innovations. This is beneficial when the innovations are similar to the true errors such as on the lower Rhine but is not beneficial when they differ such as along the middle Rhine (Fig. 6.4c). This change in behaviour along the Rhine for 15 July is due to the post-processing working better for the rising limb of a flood than for the falling

limb. An example of this behaviour is shown in Fig. 6.5a (rising) and Fig. 6.5b (falling). The post-processed forecast mean captures the flood event well 5-days before the peak (Fig. 6.5a). However, once the heavy rainfall has passed and the catchment response is captured in the initial conditions, the flood peak is routed along the river network with no spread in the ensemble forecast. The lack of ensemble spread means the prior ensemble is given more weight in the LETKF resulting in very little change in the estimated error, despite the true error of the forecast decreasing (Fig. 6.5b). This limitation was identified in (Chapter 5 and is discussed further in Section 6.5. Interestingly, in Fig. 6.4c, at the Halden station on the Thur river the innovation and the true error match in sign (negative) but the estimated error is positive. Neighboring innovations are positive which suggests the neighbouring at-gauge estimates are being given more weight in the LETKF than the at-gauge estimate at the Halden station itself. This could be because the uncertainty of the at-gauge estimate at Halden is incorrectly large and so is weighted lower in the LETKF, or because the distance based localisation has not fully removed the impact of spurious correlations between the nearby gauges and the Halden station. This leads to a positive bias in the post-processed forecast as shown in Fig. 6.5c.

Next, we use heatmaps to investigate the impact of three factors on the post-processed forecast skill (Fig. 6.6): 1) the skill of the at-gauge estimates, 2) the correlation between the ungauged location and the locations of the at-gauge estimate, and 3) the ratio of the average flow at both locations. In each heatmap, the rows represent the ungauged locations ranked by the NSE skill score of the post-processed forecasts at a lead-time of 5 days. The red line separates the locations where post-processing is beneficial and locations where it is not. The columns represent the distance rank of the at-gauge estimates from each ungauged location with the first column representing the closest station (which may be different for each ungauged location). If the at-gauge estimate is outside the localisation range of the ungauged location, the box is coloured grey as it has no impact on the error estimate. Note here we refer to the location of the at-gauge estimates simply as the gauges.

Typically, the post-processed forecasts have higher skill if nearby at-gauge estimates have a high skill. This is shown by the higher density of positive skill scores of the at-gauge

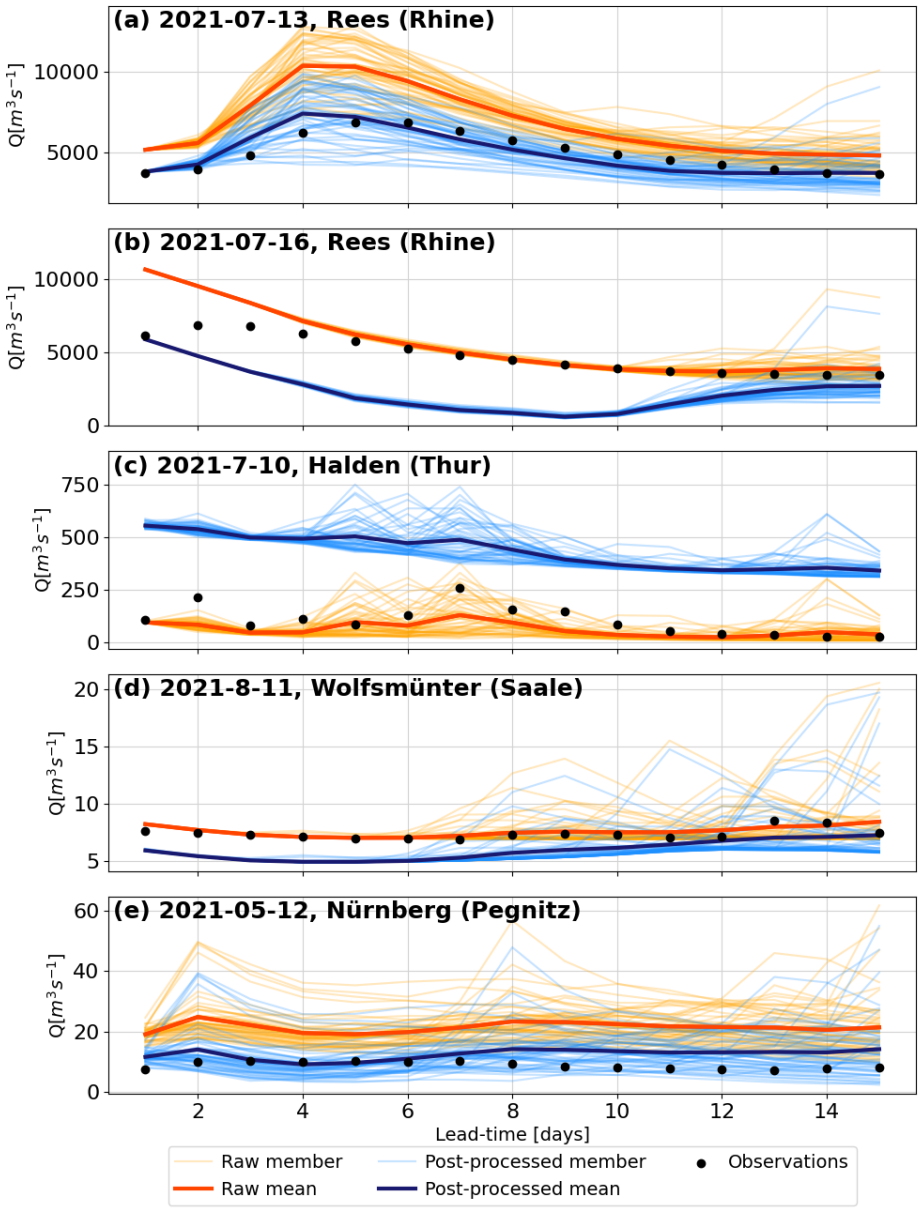

Figure 6.5: Hydrographs of the raw ensemble forecasts (orange) and the post-processed ensemble forecast (blue) at ungauged locations. Black markers show the observed values. Upstream drainage areas for the locations are a & b) Rees:159,683 km^2, c) Halden: 1,085 km^2, d) Wolfsmünter: 2,121 km^2, and e) Nürnberg: 1,198 km^2.

estimates (purple squares) in the top half of Fig. 6.6a compared to the lower half where negative skill scores (orange squares) are more common close to the ungauged locations. The at-gauge post-processing method used here is known to be less beneficial for rainfall-driven high flows (Chapter 4). Therefore, this dependency on the skill of the at-gauge estimates is likely the cause of the lower skill of the post-processed forecasts for months with flood events (Fig. 6.2b). Additionally, in general the post-processed forecasts have a higher skill when a larger number of gauges are within the localisation region. This is most noticeable for the lowest skilled ungauged locations (lowest rows) where the number of gauges decreases substantially. The exception to this is the Stenay station (indicated in Fig. 6.6a) which has a high skill in the leave-one-out experiment but only 9 gauges are within the localisation region and many have low skill.

The correlation between the ungauged location and the gauged locations is one factor determining how the information is propagated along the river network by the LETKF. Most of the ungauged locations assessed have gauge with which they have a high correlation (white/yellow squares in Fig. 6.6b). Whilst this is usually a nearby gauge it is not always the closest gauge. Typically higher correlations are found between locations on the same stretch of river but gauges on tributaries may be geographically closer (not shown). We note that the Stenay station is again an exception as all correlation values are comparatively low (less than 0.73) and drop further after the two closest gauges. The low correlations will give the at-gauge estimates less influence within the LETKF which may explain how the skill of the post-processed forecasts is high despite the skill of some of the nearby at-gauge estimates being low.

Finally, we look at the ratio between the mean observed flow throughout the evaluation period at gauged locations and the ungauged location. We take the log of the ratios such that negative values mean the flow at the ungauged location is larger than the flow at the gauges. In general, ungauged locations where post-processing is beneficial tend to have higher flows than at the nearby gauges shown by the largely orange upper half of Fig. 6.6c. Locations that are degraded by post-processing often have smaller flows than the surrounding gauges. For example, the Wolfsmünter station on the Saale has a lower flow than many of the nearby

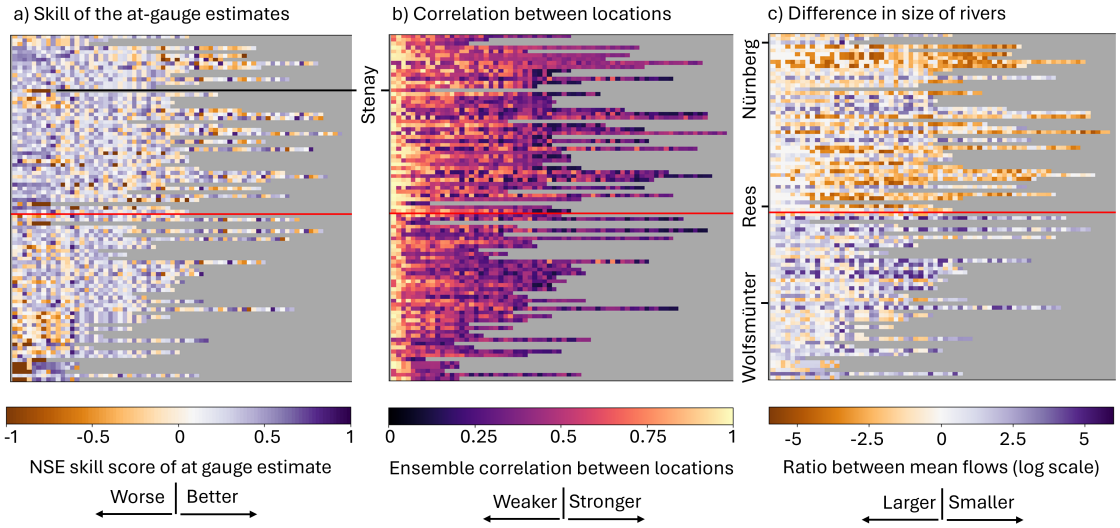

Figure 6.6: Heatmaps of the skill of post-processed forecasts at ungauged locations against the distance rank of the assimilated at-gauge estimates. Each row represents an ungauged location ordered by the NSE skill score at a lead-time of 5 days. Rows above the red horizontal line represent locations that are improved by post-processing and rows below represent degraded locations. Each column represents the ranking of the location of an assimilated at-gauge estimate in terms of distance from the ungauged location with the closest gauge on the left. The colour of the squares is determined by the a) NSE skill score of the at-gauge estimate, b) the correlation between the ungauged and gauged locations calculated from the raw forecast ensemble, and c) the ratio between the mean flow at the ungauged location and the gauged locations. Columns with grey squares are beyond the localisation range (524 km) of the ungauged location. Stations discussed in the text are labeled.

gauges, in particular, its nearest station. As such the estimate of the error is influenced by the larger rivers resulting in an incorrect error estimate and leading to degraded post-processed forecasts (Fig. 6.5d). On the other hand, the post-processed forecasts have a high skill at the Nürnberg station on the Pegnitz river. The drainage area of this station is only 1,195 km^2 and as such the average flow is much lower than for many of the other gauges in the Rhine-Meuse catchment. It was found that there is a fairly consistent bias at this location that the post-processing method is able to correct (Fig. 6.5e) leading to a high skill score. The at-gauge post-processing method has been shown to be more skillful for larger rivers (Chapter 4). As the skill of the at-gauge estimates influences the skill of the post-processed forecasts at ungauged locations (Fig. 6.6a), this is likely one reason why the post-processed forecasts have higher skill on larger rivers. Additionally, locations with smaller drainage areas are likely to show smaller variability due to rainfall resulting in smaller ensemble spreads. Therefore, smaller rivers are likely to be impacted more often by the limits of the uncertainty quantification in the prior ensemble. This is another reason for the Wolfsmünter station showing a poorer skill (e.g., early lead-times in Fig. 6.5d).

The skill of the at-gauge estimates influences the performance of the post-processed forecasts at ungauged locations (Fig. 6.3a). To further examine the spatial impact of the at-gauge estimates, Fig. 6.3 can be compared with the equivalent figure for the ensembles presented in Chapter 5 (see Fig. 6.7). In Chapter 5, observations rather than at-gauge post-processed forecasts (i.e., the at-gauge estimates) were assimilated. Comparing these two figures therefore isolates the impact of using at-gauge estimates in place of direct observations.

In terms of correlation (panels Fig. 6.3a and Fig. 6.7a), the use of at-gauge estimates has little influence on forecast skill along the Rhine and its upper tributaries. However, forecast skill on the Meuse, Moselle, and Main is reduced relative to the results from Chapter 5, as indicated by the absence of black-outlined stations along these rivers in Fig. 6.3a. The fact that higher skill is achieved when observations are assimilated suggests that the reduced performance in the post-processed forecasts of this chapter are due to inaccuracies in the at-gauge estimates.

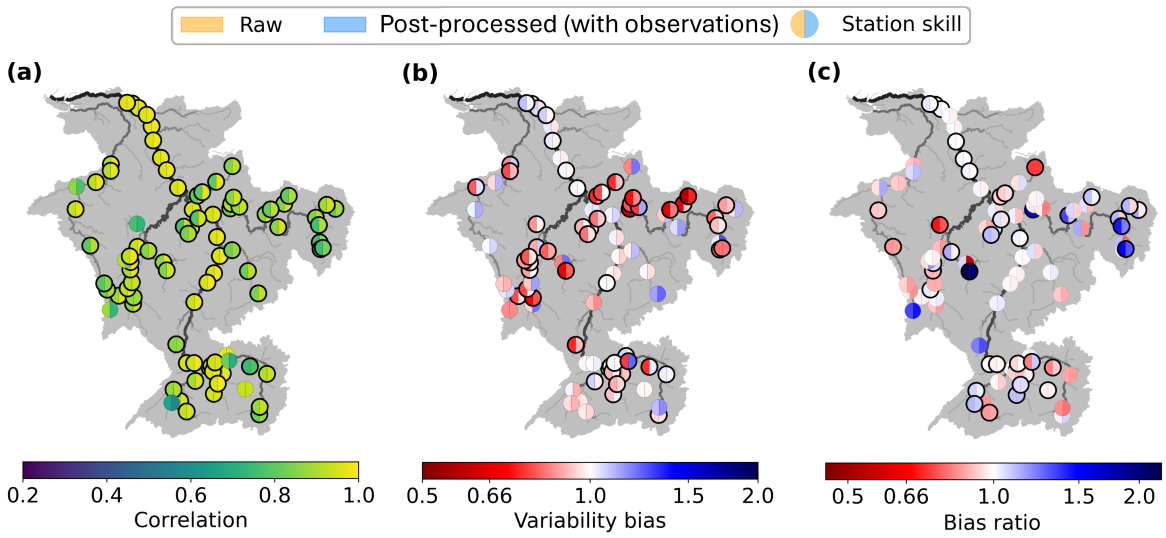

Figure 6.7: Same as Fig. 6.3 but for the error-corrected ensembles from (Chapter 5). In Chapter 5, observations are assimilated rather than the at-gauge post-processed forecasts that are used in this study. Each panel shows a components of the Kling Gupta Efficiency metric: a) correlation, b) variability ratio, and c) bias ratio, for a lead-time of 5 days for each station for the ensemble means of the raw forecasts (left half of markers) and the error-corrected forecasts (right half of markers). A black outline to the marker indicates the forecasts are improved by the error-correction. A perfect score is 1 for all three metrics. Please note this is different from Fig. 5.6 where the metrics are averaged across all lead-time.

For both variability bias and bias ratio (panels Fig. 6.3b,c and Fig. 6.7b,c), the two sets of forecasts exhibit broadly similar spatial patterns. Although fewer stations show improvement when at-gauge estimates are assimilated instead of observations, the similarity in the locations of degraded performance implies that the ensemble error covariances may not fully capture spatial dependencies between stations. Reducing the localisation length could help mitigate this issue by limiting the influence of distant stations with spuriously high correlations.

6.5 Discussion

The method proposed for post-processing ensemble forecasts at ungauged location combines two previously evaluated methods: 1) an at-gauge post-processing method - here the operational post-processing method of the CEMS' European Flood Awareness System (EFAS) - to estimate the river discharge during the forecast period at gauged locations, and 2) a data-assimilation inspired method to propagate the error correction from gauged to ungauged locations. The proposed post-processing method can correct the raw ensemble forecast at

most ungauged locations for lead-times up to five days, with some locations also benefitting at longer lead-times.

The improvement gained from post-processing the ensemble forecasts with this new method has been shown to depend on the skill of the at-gauge estimates. The at-gauge post-processing method used in this study has been shown to reduce hydrological model errors, but struggles with correcting errors due to the meteorological forcings (Matthews et al., 2022). As high flows in the Rhine-Meuse catchment are typically rainfall-driven (e.g., Mohr et al., 2022), this limitation in the at-gauge estimates leads to limited improvement at some ungauged locations during flood events. However, the correction of hydrological model biases are often correctly propagated to ungauged locations.

One advantage of this approach is that any at-gauge post-processing method can be used. Many information propagation techniques are restricted to transferring intermediate error-correction parameters between gauged and ungauged locations. However, the LETKF-based method uses the final output of the at-gauge post-processing method (i.e., the corrected river discharge). Therefore, any at-gauge post-processing method can be applied and the corrected river discharge estimates assimilated. This allows the at-gauge post-processing method to be developed independently to improve the skill at gauged locations, before combining it with the information propagation technique. Additionally, this flexibility allows this approach to be applied to any ensemble river discharge forecasting system, without the need to change the at-gauge post-processing method with which users may already be familiar.

A key challenge of the information propagation method is the accurate quantification of the uncertainty in the prior ensemble especially at shorter lead-times. The spread of the prior ensemble is largely determined by the raw ensemble forecast due to the state augmentation and covariance inflation technique used within the LETKF. This spread is driven by the meteorological forcings, in particular precipitation for the Rhine-Meuse catchment. When little or no precipitation is predicted, the ensemble spread remains narrow, giving greater weight to the prior ensemble and limiting the influence of the at-gauge estimates. There is therefore a need to account for additional sources of uncertainty, such as the initial conditions and the hydrological model. One potential solution is to estimate an error-covariance matrix repre-

senting the hydrological model uncertainties from historic simulations. This matrix could then be used within the covariance inflation module to ensure the inclusion of hydrological uncertainties within the prior-ensemble spread, similarly to some hybrid ensemble-variational data assimilation approaches (e.g., Kotsuki and Bishop, 2022).

Another component of the method that could be refined is the localisation. Localisation is used to reduce the impact of spurious correlations in the prior ensemble. Here, we define the localisation length scale as the maximum distance between any grid-box and its closest gauge. This is the minimum localisation length scale required to ensure every grid-box is updated which reduces the potential for spatially inconsistent forecasts - an important consideration for spatially distributed forecasts. However, in Fig. 6.5d, we see that the localisation does not always sufficiently limit the impact of spurious correlations which can lead to biases in the post-processed forecasts. River discharge downstream of a confluence is the convolution of the two merging rivers (Blöschl, 2022). However, upstream of the confluence the two rivers may be extremely different. Therefore, moderating the localisation further based on the location of confluences should be investigated.

We selected the Rhine-Meuse catchment as it is relatively highly gauged, allowing us to evaluate the method at many locations. However, we find that fewer gauges within the localisations region can lead to low skill in the post-processed forecasts. Therefore, further experiments are needed to understand the sensitivity of the method to the location and number of gauges. This will provide information on how the method will react if some at-gauge estimates are temporarily unavailable which could happen in an operational context. Additionally, the LISFLOOD hydrological model is well calibrated for the Rhine-Meuse catchment and typically has small hydrological errors making the raw forecasts a relatively high benchmark to beat. Further evaluations should be conducted on additional catchments to understand how the method responds to different sources of errors such as snowmelt and regulation.

6.6 Conclusion

In this study we evaluate a method for post-processing ensemble river discharge forecasts at ungauged locations. The ensemble forecasts are first post-processed at gauged locations. The at-gauge post-processed forecasts are then used within an LETKF-based method to estimate error vectors for each grid-box in the catchment. In this study, we use the operational at-gauge post-processing method of the Copernicus Emergency Management Service's European Flood Awareness System (EFAS; Matthews et al., 2025b), and post-process the EFAS ensemble forecasts at all locations across the Rhine-Meuse catchment.

We aim to answer two questions in this proof-of-concept study. First, can error corrections derived from at-gauge post-processed forecasts be effectively propagated to ungauged locations using a data-assimilation-based method? We show that the post-processing method has the potential to improve the forecast means and forecast distributions at ungauged locations for lead-times up to 5 days. However, the types of errors corrected vary spatially. The correlation of the forecast mean is improved along the Rhine. However, the bias and the ability of the forecast to capture the variability of the flow is improved for locations where the raw forecast is least skilled, which tend to be tributaries of the Rhine. The post-processing method showed less improvement for high flow events compared to normal flows situations.

Second, what factors influence the method's ability to improve the forecast skill at ungauged locations? The skill of the at-gauge estimation influences the skill of the post-processed forecasts at ungauged locations. Stretches of the river network for which the at-gauge estimates are skillful are typically improved by the post-processing method. On the other hand, when the at-gauge estimates perform poorly (e.g., for smaller rivers or during extreme floods) they can negatively impact the skill of the post-processed forecasts. The ensemble correlations (via the ensemble covariances) can help mitigate the influence of poorly skilled at-gauge estimates. However, when the spread of the raw forecast ensemble is incorrectly small, even skillful at-gauge estimates can be ignored within the LETKF. The number of at-gauge estimates influencing the estimation of the error vectors also impacts the skill of the post-processed forecasts with more at-gauge estimates being beneficial.

The results of this study demonstrate the potential of the ungauged post-processing method to improve river discharge forecasts. The flexibility of this approach, allowing any at-gauge post-processing method to be used, makes it widely applicable to different forecasting systems without requiring major changes to existing workflows. As advancements in at-gauge post-processing continue, these improvements can be integrated into the information propagation framework, improving forecast skill at both gauged and ungauged locations. With continued development, this approach offers a promising pathway towards more reliable ensemble river discharge forecasts at ungauged locations, ultimately supporting improved flood preparedness and water resource management.

6.7 Summary of Chapter 6

The new post-processing method presented in this chapter combines the methods of Chapters 4 and 5. At-gauge post-processed forecasts are used as proxy-observations within the LETKF-based error-propagation method (*Objective 3*). (Q3.1) The method shows potential as a post-processing method for ungauged locations particularly at short lead-times. The post-processed ensemble forecasts show improved skill at the majority of station up to a lead-time of 5 days, but the method is less skilled for high flows than for normal flows. (Q3.2) The skill of the post-processed forecasts is largely dependent on the accuracy of the proxy-observations. For example, the at-gauge post-processing method used to generate the proxy-observations was found to struggle for high-flows in catchments, such as the Rhine, where meteorological forcings are the main cause of the errors (Chapter 4; *Objective 1*). The location of the proxy-observations is also important with locations downstream of proxy-observations being improved more in general. Additionally, in Chapter 5, the data-assimilation-inspired information propagation method was found to struggle when the raw ensemble spread was incorrectly small (*Objective 2*). This behaviour is seen in this chapter (Chapter 6) as well. Forecasts for the falling limb of a flood event are most impacted by this limitation.

Whilst the new post-processing method has limitations that need to be addressed, it has shown potential to improve the skill of the forecasts at ungauged locations. To be considered

for operational implementation the usefulness of these forecasts must be considered. In Chapter 7, I will discuss a workshop that aims to identify developments to increase the usefulness of the EFAS at-gauge post-processed forecasts (*Objective 4*). At the end of Chapter 7, I will discuss how the outcomes of the workshop apply to the new post-processing method for ungauged locations (presented in this chapter) if it were to be made operational.

Chapter 7

User priorities for the development of the EFAS post-processed forecasts

This chapter addresses the fourth objective of the thesis: *To identify potential barriers for the use of the EFAS post-processed forecasts and determine future developmental priorities.*

A workshop was organised between developers and users of the post-processed flood forecast product of the European Flood Awareness System (EFAS) to determine priorities for future co-developments. The aim of the workshop was to ensure future developments are in line with user-requirements and the available resources. The workshop was held on the 27 September 2022 at the Joint Research Centre (JRC) Conference Centre, Italy and online (hybrid). The content of the chapter is reproduced from:

Matthews, G., H. L. Cloke, S. L. Dance, E. Hansford, C. Mazzetti, and C. Prudhomme, 2023: *Co-Design and Co-Production of Flood Forecast Products: Summary of a Hybrid Workshop*. Bull. Amer. Meteor. Soc., 104, E1058–E1066, `https://doi.org/10.1175/BAMS-D-23-0061.1`.

In Section 7.9, I summarise the key results of the chapter and discuss how they may apply to the post-processing method developed in Chapter 6.

7.1 Introduction

With flood damage increasing, the effective use of mitigation instruments is essential (, WMO). Flood Early Warning Systems are cost-effective tools to increase preparedness for floods and reduce the resulting damage (Pappenberger et al., 2015a; Verkade and Werner, 2011). The European Flood Awareness System (EFAS), part of the Copernicus Emergency Management Services (CEMS), is a pan-European flood forecasting system established to complement national systems by producing medium-range probabilistic flood forecasts particularly for large transnational rivers (Thielen et al., 2009; Demeritt et al., 2013; De Roo et al., 2011; Smith et al., 2016). EFAS is co-produced to ensure that the forecasts created provide decision-relevant information that can be incorporated into the procedures of the forecast users whilst fully utilizing the knowledge and resources of all involved (Lienert et al., 2022; Bierens et al., 2020). In practice, several mechanisms are employed to facilitate the required collaboration including webinars, online feedback forms, training sessions, user surveys, workshops, working groups, and the EFAS annual meetings.

The focus of the workshop discussed here was the co-design of the EFAS post-processed forecasts with the workshop forming part of a larger iterative co-production process (Fig. 7.1). The EFAS post-processed forecasts were initially introduced following a consultation with users which found that hydrographic forecast products would be beneficial. As with all forecasting systems, EFAS forecasts are subject to errors and uncertainties from several sources including the initial conditions, the meteorological forcings, and the hydrological model. Therefore, hydrographs created from the raw (or non-postprocessed) hydrological model output must be considered carefully. The post-processed forecast product aims to provide a bias-corrected hydrograph and to quantify the total predictive uncertainty (De Roo et al., 2011). The post-processed forecasts are available at river gauge locations for which historic and near real-time river discharge observations are available (Matthews et al., 2022). The number of locations has increased from three stations during initial testing (Bogner and Pappenberger, 2011) to 1608 stations as of 13 September 2022 across the EFAS domain.

During the 17th EFAS Annual Meeting in September 2022, a dedicated workshop was

organized for producers and users of the EFAS post-processed forecasts to discuss its future evolution. Here, we summaries the outcomes of the workshop and present the next steps in the co-developmental cycle. We also reflect on the broader workshop findings that should be applied to other EFAS forecast products and forecasting systems. Finally, we discuss the lessons learnt regarding stakeholder engagement in a "post-pandemic" world.

7.2 Workshop organization

The 17th EFAS Annual Meeting was held on the 27-28 September 2022. The organizers and host (the Joint Research Centre of the European Commission) opted for a hybrid format to encourage face-to-face discussions, as it was the first meeting since social and travel restrictions had eased, whilst still engaging with those unable to travel. The meeting was open to representatives from the EFAS partners (hydrometeorological authorities across Europe mandated to provide flood early warnings for their respective regions), and co-organized by all EFAS operational centers (organizations contracted to run EFAS). To complement the plenary sessions, four workshops were run focusing on specific topics voted for by attendees at registration, including a workshop on the EFAS post-processed forecasts.

The workshop on the EFAS post-processed forecasts was held on the first day of the meeting and lasted 1 hour and 15 minutes, with a total of 21 in-person attendees (approximately 27% of present) and 20 online attendees (approximately 40% of registered). The primary objectives of the workshop were to:

1. Present the operational EFAS post-processed forecasts and forecast products including a recent increase in temporal resolution,
2. Determine the current usage of the EFAS post-processed forecasts,
3. Identify limitations and barriers to the use of the EFAS post-processed forecasts, and
4. Determine priorities for future developments of the EFAS post-processed forecasts.

The usability of forecasts is dependent on the relevance of the forecast information content, the communication channel, the forecast visualization (also known as the forecast product),

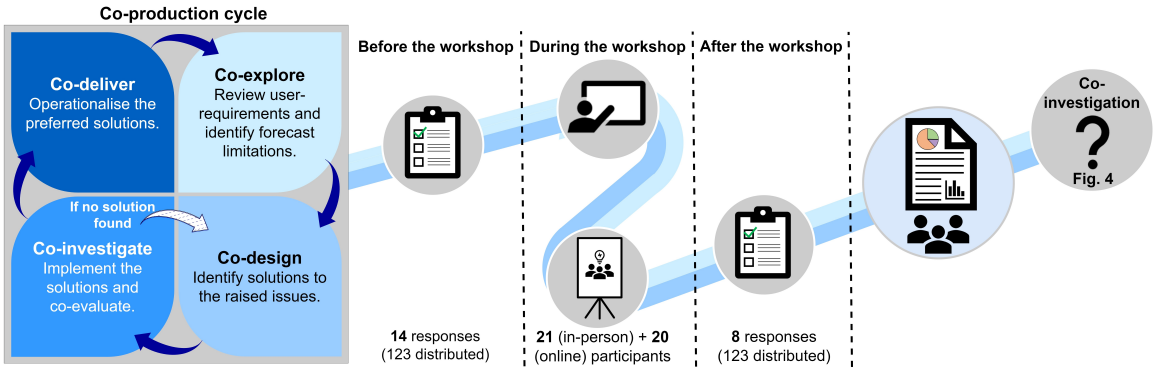

Figure 7.1: Co-production cycle for the EFAS post-processed forecasts and workshop agenda. The workshop was an opportunity to co-explore the limitation of the post-processed forecasts and co-design potential developments.

the quality of the forecast, and the expertise of the user (Vincent et al., 2020; WMO, 2022). Therefore, the workshop was organized around these five topics (Table 7.1). The format of the workshop and its position within the wider co-production process of the EFAS post-processed forecast is shown in Fig. 7.1. A pre-workshop questionnaire was distributed to all registered attendees one week before the workshop to comprehend the current usage and understanding of the post-processed forecasts. A post-workshop questionnaire was also distributed allowing participants to provide anonymous and individual feedback.

The workshop consisted of a presentation followed by a participatory group activity to encourage dialogue (Fig. 7.1). The presentation gave an overview of each of the five topics. Prompts were provided to guide group discussions (Table 7.1), but participants were encouraged to consider other ideas inspired by their own systems and experience using the EFAS post-processed forecasts in an operational setting. Mind-maps prepared ahead of the workshop were used to record the discussions (see Fig. 7.3). Online participants used Webex (`https://www.webex.com/`) and Miro (`https://miro.com/`) to participate in the workshop with breakout rooms used to allow for group discussions.

7.3 EFAS post-processed forecast

A key aim of the presentation was to ensure all participants had an understanding of the post-processed forecasts even if they did not regularly use the forecasts. The pre-

Topic	Discussion Prompts
Access	1. Do you find the post-processed forecast product easily accessible? 2. How could the post-processed forecast product or the data used to create the product be more easily accessed?
Method	1. Do you use post-processing in your system? What works and what does not work? 2. How about EFAS? What works and what does not work in EFAS post-processing? 3. Have you noticed any unusual cases of the EFAS post-processed forecast?
Product Design	1. Are the chosen thresholds useful? What other thresholds might be more useful? 2. Is the probability distribution presented in an appropriate and intuitive way? 3. Are the “probability of exceeding" graphs useful? How could they be improved?
Training and Documentation	1. Is the training provided sufficient? 2. What other training would you like to be provided? 3. Is the documentation provided sufficient?
Evaluation	1. Is the evaluation provided sufficient? 2. Would an evaluation product be useful? 3. What would be the most informative verification metric?

Table 7.1: Questions provided as prompts to facilitate the group discussions.

sentation also allowed the changes due to a recent increase in the temporal resolution to be presented (Mazzetti et al., 2020b). The full presentation is available at `https://www.efas.eu/en/news/17th-efas-annual-meeting` and details regarding each of the five topics are documented on the ‘CEMS-Flood Wiki’ (`https://confluence.ecmwf.int/display/CEMS/CEMS-Flood`). The EFAS post-processed forecasts use recent river discharge observations (provided by EFAS data providers which include several EFAS partners) to adjust the EFAS multi-model medium-range ensemble forecasts to account for errors and uncertainties in the ensemble forecasts. The method of post-processing has changed over time (Matthews et al., 2022; Bogner and Kalas, 2008; Bogner and Pappenberger, 2011; Raftery et al., 2005; Smith et al., 2016; De Roo et al., 2011) with the current method using a combination of the Model Conditional Processor (MCP; Todini, 2008; Coccia and Todini, 2011) and the Ensemble Model Output Statistics method (EMOS; Gneiting et al., 2005). The MCP requires an offline calibration procedure which is performed twice a year to incorporate the

most recent river discharge observations. The MCP quantifies the hydrological uncertainty and corrects biases due to the hydrological model. The EMOS method quantifies the uncertainty due to the meteorological forcings. The outputs from both methods are combined using the recursive Kalman Filter (Kalman, 1960).

The post-processed forecast is available to users via the EFAS Sensor Observation Service (SOS), a web-based API. The EFAS-SOS allows users to visualize each percentile of the post-processed forecast as well as to download the forecast data. Additionally, the post-processed forecast is transformed into a forecast product, called the 'Real-Time Hydrograph', available on the EFAS website (Fig. 7.2). The post-processed forecast product has been available since 2012 and in 2020 the product was merged with the medium-range reporting point layer so that all medium-range forecast information regarding a specific location was available in a single pop-out window.

The post-processed forecast product consists of three panels (pop-out window in Fig. 7.2): a hydrograph (left) where darker shaded percentiles are closer to the median, and two bar charts showing the probability of exceeding the mean flow (MQ, lower right) and the mean annual maximum (MHQ, upper right), respectively, at each lead-time. The flood thresholds are calculated from river discharge observations; a key distinction between the post-processed forecast product and other EFAS medium range forecast products which use thresholds calculated from simulated reference climatologies.

Results of the latest forecast evaluations were presented, showing that the post-processed forecasts could predict the exceedance probability of the flood thresholds more reliably (i.e., the forecast probabilities more accurately represented the exceedances of the thresholds) than the raw ensemble forecasts (Matthews et al., 2022), and that the update to 6-hourly timesteps increased the skill of the forecast probability distribution at shorter lead-times.

7.4 Current usage

All participants stated that, in the context of EFAS, post-processing is useful whilst 50% said it is essential. Most participants selected the key aims of post-processing to be the reduction

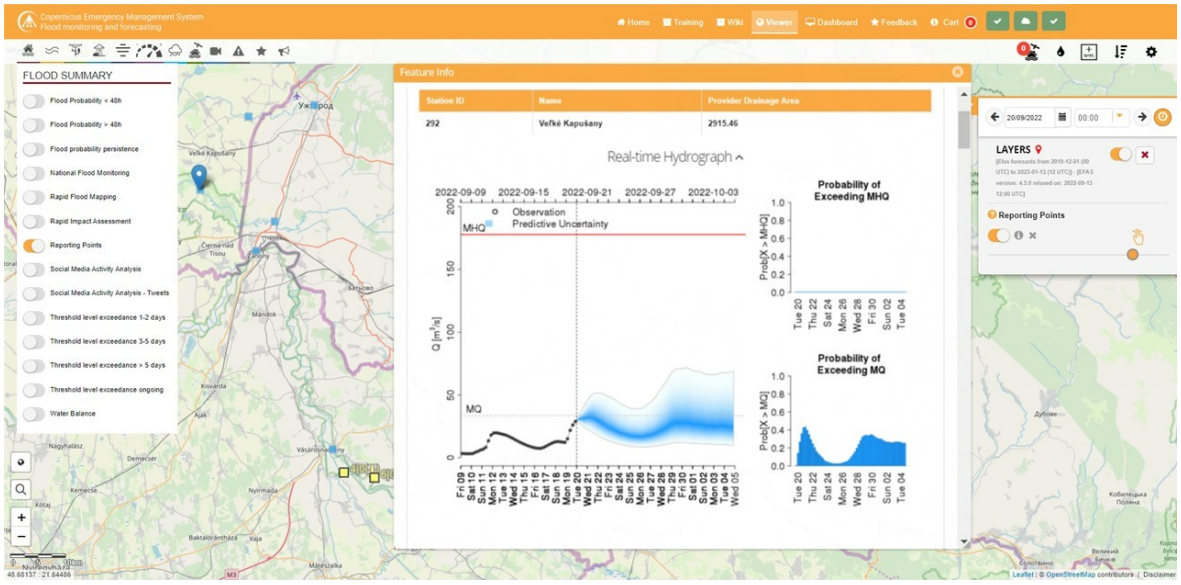

Figure 7.2: The so-called 'Real-time Hydrograph' forecast product accessed via the EFAS website. Forecast from the 20 September 2022 for the Veľké Kapušany station, Slovakia.

and quantification of forecast uncertainties. Over 70% of participants primarily use the EFAS post-processed forecasts to predict upcoming floods in the next 5 days whereas only 14% (2/14) use the forecasts to predict floods more than 10 days in the future. All participants indicated they use the EFAS post-processed forecasts in conjunction with other flood and meteorological forecast products.

7.5 Co-identifying limitations and future developments

Potential future development and limitations were discussed for each of the five topics. Here, we discuss the key responses from the questionnaires and the mind-map-based group discussions (Fig. 7.3).

Access

All participants indicated a preference for accessing the post-processed forecast via the EFAS website with most indicating a high confidence in accessing the product in this manner. Locating stations by name or identification number rather than by municipality was suggested as a more efficient search function. The EFAS SOS is useful for comparing EFAS forecasts with a user's own system. However, it can be slow and difficult to use, indicating the need for

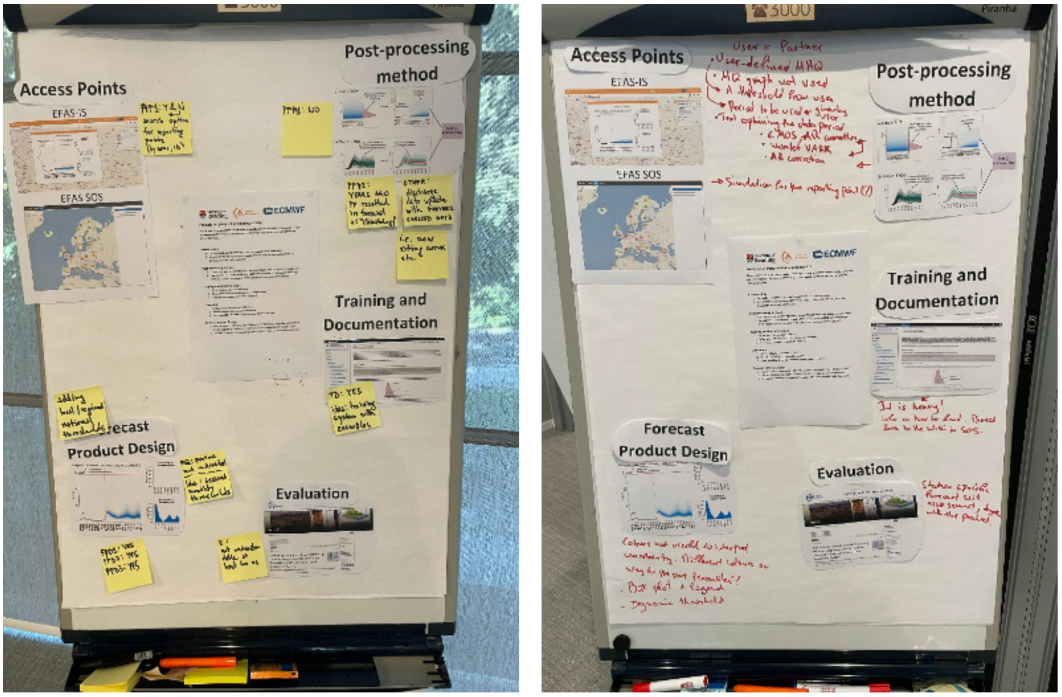

Figure 7.3: Mind-maps from two group discussions

improved efficiency of the EFAS SOS system and clearer guidelines on data access through this service.

Method

Improving the post-processing method was most frequently selected as a top priority for future developments, with both the MCP and the EMOS components of the method identified as possible limitations. For example, the MCP component requires a timeseries of recent (currently the previous 40 days) river discharge observations on which to condition the forecast probability. Incorporating local knowledge of the catchments may optimize the number of observations necessary for each station.

Water level is the main variable of interest for some EFAS partners. Therefore, extending the post-processing method to create forecasts of water level would be beneficial for some users. However, the suggestion was not unanimous so water level should not replace discharge and would need to be included as an additional product.

Product design

Two key changes to the forecast product were proposed during the workshop: the flood event thresholds, and the representation of the probability distribution. There was a consensus that locally-defined thresholds (provided by EFAS partners) would be more useful than the current flow-based thresholds. Alternative thresholds include frequency-based magnitudes (e.g., return periods), physically-based thresholds (e.g., defined using hydraulic models), or other locally-defined thresholds. Locally-defined thresholds would allow more consistency with local decision making and for direct comparison with users' own systems. The slowly changing color gradient in the forecast product (Fig. 7.2) makes it difficult to identify a specific percentile. Boxplots were suggested as a potential alternative to the current form of the hydrograph. Boxplots would also make the post-processed forecast product consistent with other EFAS forecast products allowing for direct comparison between the post-processed and raw forecasts

Training and documentation

In the post-workshop survey, most attendees said the workshop was useful, but that additional training was desired, with a webinar being the preferred format. Overall, participants were satisfied with the level of detail in the documentation provided on the EFAS Wiki pages but stated that it was not easy to find specific information on the post-processed forecasts. It was suggested that a direct link from the forecast product on the EFAS website would be beneficial. Additionally, some participants requested more practical guidance regarding the data requirements and station specific information regarding the calibration process (e.g., the calibration period).

Evaluation

Participants indicated a desire for the post-processed forecasts to be evaluated routinely with each calibration exercise performed (twice a year). There was a clear preference for the evaluation results to be station specific and available alongside the forecast products on the EFAS website as an 'evaluation product'. The importance of skill of different features of the forecasts varied between participants with some prioritizing the reliability of the forecast

exceeding a threshold and others prioritizing the peak magnitude and timing of the forecast median. This indicates that the evaluation product should be designed using a multi-metric approach.

7.6 Going forward

A set of potential future developments of the EFAS post-processed forecasts, to be researched in the co-investigation phase, has been defined from the workshop discussions (Fig. 7.4). During this process three key principles were found to overlap with many of the tasks:

1. **Locally relevant forecast products**: The information content of the forecast product should be relevant to the task of the users. For example, locally-defined flood thresholds allow users to connect the forecast to their local procedures.

2. **User-focused auxiliary information**: Auxiliary information should allow users to extract the information that is applicable to them, their usage of the forecast, and their region of interest. For example, forecast evaluations should be conducted at station level.

3. **Ease of access**: All forecast and auxiliary information should be easy and quick for the user to access. For example, providing links to the documentation alongside the forecast product means users can check the information without significantly detouring from their task.

These three principles can be applied to other EFAS forecast products and other early warning systems. In addition to user requirements, future developments are guided by available resources. Additionally, due to the number of locations being post-processed in EFAS each step must be automated once in operations.

EFAS partners have been invited to join a working group during the co-investigation phase, led by the CEMS Hydrological Forecast Computational Centre, to ensure that communication channels between producers and users remain open throughout the process. During the co-investigation phase feasibility studies will be conducted for each development in Fig. 7.4.

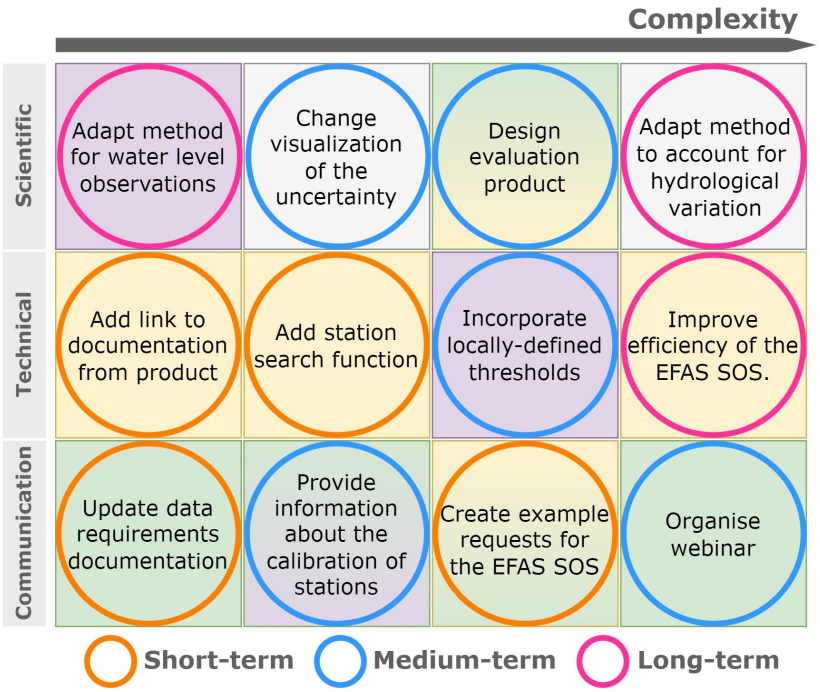

Figure 7.4: Potential future developments of the EFAS post-processed forecasts ordered by complexity. Square color indicates the principle of interest: locally relevant products (purple), user-focused auxiliary information (green), and ease of access (yellow). Grey indicates feedback that does not explicitly follow a guiding principle. Circle color indicates intended timeframe.

The study results will be discussed within the working group to identify any changing needs and to ensure the solutions are appropriate. Once a preferred solution has been selected by the working group, it will be presented to all stakeholders to provide the opportunity to raise concerns and offer alternative solutions before implementation in the operational system.

7.7 Reflections on a hybrid workshop

This was the first hybrid EFAS annual meeting following the pandemic. The aim was to allow in-person and online attendees to participate in the workshop equally. However, there were three key challenges in running the workshop discussed here. First, time constraints limited the workshop to 1 hour and 15 mins limiting opportunity for open discussion. Whilst extending the time within the annual meeting was not possible, a standalone workshop would allow more time but may not be attended by as many EFAS partners. Second, the dynamic differed between the in-person groups who discussed the topics, working through the prompts

together, and online groups who worked individually. When split into breakout rooms many online participants did not communicate with fellow participants. This could have been for several reasons including lack of technical equipment, being in a shared environment, multi-tasking or lack of engagement. Additionally, due to the explicit split between online and physical attendees there was no discussion between the two. Therefore, for the workshop to be more beneficial for both the organizers and the online participants, rather than replicating the group mind-mapping activity online, the activity could differ between online participants and in-person attendees. Finally, the number of responses to the questionnaires was low compared to participants. This is likely due to competing priorities outside of the workshop and as such, if time allows, questionnaires should be conducted during the workshop.

7.8 Concluding remarks

A workshop was held during the 17th EFAS Annual Meeting to co-produce future developmental priorities for the EFAS post-processed forecasts. Despite difficulties that emerged due to time constraints and the hybrid nature of the session, all workshop objectives were achieved. Whilst the specific outcomes from the workshop are applicable to the EFAS post-processed forecast product, the three derived principles are applicable to other EFAS forecast products and those of other forecasting systems, such as the Global Flood Awareness System (Alfieri et al., 2013).

7.9 Summary and discussion of Chapter 7

In this chapter, I present the outcomes from a workshop focused on identifying issues with the operational EFAS post-processed forecasts that may limit their use. Developmental priorities were then defined in order to overcome these barriers (*Objective 4*). As both developers and users of the forecasts participated in the workshop the challenges and solutions could be addressed dynamically. (Q4.1) The EFAS post-processed forecasts are primarily used to predict floods in the next 5 days, and usually alongside additional forecatsing products such

as weather forecasts, local flood forecasts, and the raw EFAS forecasts. (Q4.2) Discussions regarding the barriers of EFAS were centered around five topics: access, post-processing method, product design, training and documentation, forecast evaluation. Typically, these barriers were related to local relevancy of the products, useful auxiliary information, and ease of access. (Q4.3) These overarching principals where used to guide the identification of developmental priorities for the operational post-processed forecasts. These developments include changes to the post-processing method, updates to the forecast product (the visual representation of the forecast), and improvements to the data access system (Fig. 7.4).

* * *

In Chapter 6, I presented a post-processing method for ungauged locations which showed the potential of improving forecasts at ungauged locations. Now, I will discuss how easily each of the developments identified in this chapter (Chapter 7) could be applied to make the post-processing method for ungauged locations useful for users.

1) Forecast products should be locally relevant.

The first idea for how forecasts could be made more locally relevant was to include locally-defined thresholds in the forecast period. For at-gauge post-processed forecasts, this is possible, and was indeed implemented operationally as a direct outcome of this workshop in March 2024. For ungauged locations, this development will be more difficult. Whilst some users may have flood thresholds for the entire region of interest to them this is not the case for all users many of who only have thresholds at the locations of gauges. Even if thresholds are available everywhere mapping these thresholds to the EFAS river network in order to compare with the forecasts would not be trivial. In general, the flood threshold that should be used to define a flood event at an ungauged locations is a challenge.

The second idea to make the forecast products more locally relevant is to predict water level values. For many users of the EFAS forecasts water level is a key variable of interest as it dictates whether the river will exceed the riverbank. Currently, no water level forecasts are provided by EFAS. As the method is currently designed it would not be possible to create

water level forecasts at ungauged locations. However, if the at-gauge post-processed forecasts were for water level values, such that the proxy-observations are for water level instead of river discharge, then the method would be able to adapt to this change if an appropriate observation operator can be identified.

2) User-focused auxiliary information.

Main set of auxiliary information that users requested to be user-focused was the evaluation products. An evaluation product within EFAS is a product on the EFAS website that shows the skill of the forecast. The key benefit of the evaluation products is that they are available alongside the forecasts and can be displayed simultaneously. For ungauged locations, the evaluation products would face the same challenges as were faced in Chapters 5 and 6 in that the forecasts can't be evaluated at locations without observations (i.e., truly ungauged locations). It would be possible to take a similar approach as in these chapters of leaving-out a single station at a time; however, the evaluation would still only be available at gauged locations. Observations from alternative sources could be beneficial in this case. For example, EO data could be used to evaluate the skill of the forecasts at ungauged locations.

3) Ease of access.

The need for the forecasts to be easily accessible would not cause issues for the post-processed forecasts at ungauged locations in general. As the forecasts use the EFAS domain, any mechanism for disseminating the EFAS raw forecasts would be sufficient for the post-processed forecasts.

In this chapter, barriers to the use of the EFAS post-processed forecasts were identified, and users and developers co-designed solutions to overcome these barriers (*Objective 4*). Many of these implementations have now been implemented in the operational EFAS system. The post-processing method is able to adapt the these changes; however, continued interaction between users and developers would be critical to ensure the forecasts are useful.

Chapter 8

Discussion and future work

This PhD aims to conduct research to make river discharge forecasts more actionable with a focus on enhancing forecast skill and usability at both gauged and ungauged locations. Post-processing, a technique used at the end of the modelling chain to statistically correct the forecasts, is investigated to achieve this aim. Specifically, this thesis addresses four objectives which are introduced and motivated in Section 1.3. Each objective is discussed in detail in Sections 8.1-8.4 below. The research questions defined in Section 1.3 will be answered and the limitations of each study will be considered. In Section 8.5, I provide a synthesis of the thesis and highlight the main scientific and operational contributions. Finally, in Section 8.6, I outline the potential future work that could further improve the skill and value of post-processed forecasts.

8.1 Objective 1: To evaluate the skill of the operational EFAS at-gauge post-processing method.

This objective is addressed in Chapter 4 where I present an evaluation of post-processed forecasts created using the operational EFAS at-gauge post-processing method. The evaluation aims to answer two research questions which are answered in turn below

8.1.1 Summary of results

Q1.1: Does the operational post-processing method provide improved forecasts?

Yes, the study finds that post-processing has a positive overall effect on the forecasts. The forecast median is improved in terms of correlation, bias, and variability at the majority of the river gauging stations. The improvement is greatest at shorter lead-times (1-5 days) when the recent observations, incorporated into the forecast by the post-processing method, are most relevant. At longer lead-times (11-15 days), the post-processed forecasts tend to regress towards the mean flow, resulting in an overall underestimation of the river discharge. The mean annual maximum flood threshold was used to evaluate how well the post-processed forecast predicts floods. The post-processing method improved the reliability of the forecast uncertainty and increased the number of events detected at short lead-times compared to the raw ensemble forecasts. However, the ability of the post-processing method to correct the forecasts for high river discharge is limited compared to more normal river discharge values.

Q1.2: What affects the performance of the post-processing method?

Whilst forecasts were improved at most river gauging stations, the greatest improvements were observed at river gauge stations with larger upstream areas and slower hydrological response times (i.e., when the river discharge changes slowly). Similarly, rivers where hydrological errors, due to the initial conditions or the hydrological model, were the dominant source of uncertainty showed greater improvements than those where forecast uncertainty was primarily due to the meteorological forcings. These results are both due to the near real-time river discharge observations being more influential for rivers with high auto-correlation, because past conditions are more strongly predictive of future states. Errors due to regulation of the river were often corrected but primarily when the regulation was slowly changing. If the regulation differed between the forecast period and the days immediately prior to the forecast generation time then the regulation errors were not corrected effectively. The post-processing method also required a timeseries of historic observations and simulations in order to calibrate the post-processing method. For lead-times of 1 to 3 days, longer calibration timeseries led

to slightly better performance, but the forecasts were corrected even for river gauge stations with short timeseries (less than 5 years). Errors in the timeseries (e.g., due to changes in the rating curve) negatively impact the post-processed forecasts, making quality checking the observations a critical part of the process.

8.1.2 Limitations of the study

Limited availability of both forecasts and river discharge observations impacted this study in two ways. First, to be informative to users the evaluation needed to use the version of EFAS that was operational at the time. This was EFAS version 4.0, which was released 14 October 2020 (CEMS, 2025), less than half a year before this evaluation was conducted. Therefore, as so few operational forecasts were available, I used a common evaluation strategy of using reforecasts to allow for a longer evaluation period (e.g., Klehmet et al., 2024). Reforecasts are forecasts created for past dates but using the most recent operational forecasting system (Hamill et al., 2006). Whilst it is intended for the modelling chains of the forecasts and reforecasts to be identical, small differences occur due to computational constraints. The reforecasts used in Chapter 4 differ from the operational EFAS forecasts in the following ways: i) they include fewer ensemble members (11 from a single NWP system versus 73 from four NWP systems), and ii) their initial conditions are generated differently to operational forecasts (forced with meteorological observations rather than meteorological forecasts). Although these differences may affect the exact metric values, they are not expected to influence the overall conclusions, as the same reforecasts were used both as input to the post-processing method and as the benchmark (Pappenberger et al., 2015b). Additionally, the reforecasts were only available twice a week (Monday and Thursday) meaning the lead-times were evaluated against different observations timeseries (e.g., a 1-day lead-time was evaluated against Tuesdays and Fridays whereas a 2-day lead-time was evaluated against Wednesdays and Saturdays). This primarily impacted the assessment if a flood event occurred in the timeseries for some lead-times but not others. I aimed to minimise the impact of this limitation by grouping lead- times (1-5 days, 6-10 days, and 11-15 days) or considering

general trends across all river gauge stations where possible.

Second, the relatively short length of the evaluation period was determined by the availability of the reforecasts (up to the end of 2018), and the availability of river discharge observations which were needed for both the offline calibration and the evaluation. To ensure a fair assessment the calibration period (which requires a minimum of 2 years of observations) needed to end before the evaluation period began (temporal cross-validation; Robertson et al., 2016). An evaluation period from January 2017 to December 2018 was chosen to balance the length of the evaluation period, the availability of reforecasts, and the number of locations with sufficient observations. The two-year period captures all hydrological seasons and allows a good spatial distribution of river gauging stations across the EFAS domain. However, it may not be long enough to ensure the skill metrics converged, and additional years of evaluation could lead to changes in the metrics for individual locations (Shin and Jung, 2022). Additionally, the two-year period will not fully represent the climatological range of river discharge at each location - likely not including the rarer more extreme events (Brunner et al., 2021b). Whilst skill was shown to decrease with severity (Fig. 4.10), the extremeness of these values was calculated relative to the two-year period (to ensure large enough sample sizes) not the river discharge climatology. Therefore, these results may not be representative of the forecast skill for larger floods (Brunner et al., 2021a).

8.2 Objective 2: To investigate the applicability of data-assimilation in a post-processing environment to propagate observational information from gauged to ungauged locations.

This objective is addressed in Chapter 5 where I present a new data-assimilation-inspired method for propagating observational information from gauged to ungauged locations in a post-processing environment (i.e., with no additional use of the hydrological model). The

skill of the ensembles was evaluated at 89 proxy-ungauged locations using a leave-one-out verification strategy. The two research questions for this objective are discussed below.

8.2.1 Summary of results

Q2.1: Can data assimilation techniques be used in a forecast post-processing environment to propagate observation information to ungauged locations?

Yes, the new observation information propagation method uses state-augmentation within a Local Ensemble Transform Kalman Filter (LETKF; Hunt et al., 2007), and is applicable in a post-processing environment of ensemble river discharge simulations. An error vector is estimated for each ensemble member representing the difference between the ensemble member and the true river discharge at each grid-box. At each timestep, observations at gauged locations are assimilated, updating the error vectors via the ensemble covariances. To apply the data-assimilation technique in a post-processing environment, I assumed that the precomputed river discharge hindcast ensemble was a good estimate of the analysis ensemble, removing the need to propagate the hindcast forward in time with a computationally expensive hydrological model. To improve the efficacy of the method, and to facilitate its application to multiple catchments in a large-scale forecasting system, I define new techniques for determining the localisation length scale, inflating the covariance of the error vectors in line with the increase in the river discharge ensemble covariances, and defining initial error vectors which are spatially consistent across the domain allowing effective updates without the need for a warm-up period. The updated error vectors are added to the corresponding river discharge ensemble member creating an ensemble of corrected simulations at each lead-time.

Q2.2: Are the resulting ensemble predictions of river discharge more skillful than the raw ensemble?

The skill of the ensemble mean was improved compared to that of the raw (or uncorrected) ensemble. Both the variability ratio and the correlation of the ensemble mean were improved at almost all 89 evaluated locations. The ability of the method to correct the bias of the

ensemble mean depended on the consistency of the bias along the river network and whether the bias at neighbouring river gauge stations was similar. Locations both up and downstream of the assimilated observations were corrected if similar error behaviour was observed at both locations. The ensemble spread was found to be under confident at longer lead-times (11-15 days) with the spread of ensemble not correctly adjusting to the increased skill in the ensemble mean. This is in contrast to the raw ensemble which remains over confident (too narrow) at all lead-times. Overall, the corrected ensembles are more skillful that the raw ensembles but care would be needed in the interpretation of the uncertainty if used in decision making.

8.2.2 Limitations of the study

To avoid developments to the operational system impacting the assessment of the new observation propagation method, a one-year evaluation period with no major system upgrades was selected (CEMS, 2025). The evaluation period is short, therefore similar limitations to those discussed in Section 8.1.2 also apply here. Reforecasts could not be used in this study as the ensemble size impacts the efficacy of ensemble data assimilation techniques (Lin et al., 2025). To mitigate the impacts of the relatively short evaluation period I identify and analyse two specific forecasts corresponding to two flow scenarios: 7 July 2021 (high flow) and 8 October 2021 (normal flow). One of these scenarios (July 2021) corresponds to an extreme flood in the Rhine estimated to be above the 500-year return period threshold, and potentially much higher along different tributaries (Mohr et al., 2022; Ludwig et al., 2023). By identifying different scenarios I was able to focus on how the method responded under different conditions. However, the evaluation results may still be biased due to the short evaluation period, and a longer period may provide additional insight.

The evaluation was performed on the Rhine-Meuse catchment as it has a relatively high number of river gauge stations allowing assessment at many locations. Additionally, the Rhine-Meuse is one of the largest catchments in EFAS, allowing the computational efficiency of the method to be judged. However, the results may not be representative of

the method's behaviour in other hydroclimatological regions (Du et al., 2023). Given the positive results, the method should be assessed for other catchments particularly those with different physical characteristics to the Rhine Meuse (e.g., snowmelt-driven catchments, arid catchments, catchments with fewer gauges, and catchments where the hydrological model error is dominant).

Finally, in the leave-one-out experiments used to evaluate the performance of the corrected ensembles at ungauged locations, I only consider the removal of one river gauge station. This reflects a typical situation where an individual station fails temporarily. However, multiple stations can become unavailable simultaneously—for instance, during widespread flood damage or if the provider experiences technical difficulties. To better understand the system's sensitivity to missing data, a more comprehensive data denial experiment, where groups of river gauge stations are withheld, should be conducted (Roberts et al., 2017). This experimental design would assess the observation propagation method under different river observation network configurations.

8.3 Objective 3: To explore the use of proxy-observations to facilitate the post-processing of ensemble river discharge forecasts at ungauged locations.

This objective is addressed in Chapter 6, where I investigate a two-step post-processing method for ungauged locations. First, the post-processing method evaluated in Chapter 4 is used to estimate proxy-observations during the forecast period at gauged locations. Then, the proxy-observations are assimilated using the LETKF-based method presented in Chapter 5 to propagate the error correction to ungauged locations. The two research questions for this objective are discussed below.

8.3.1 Summary of results

Q3.1: Can error corrections derived from at-gauge post-processed forecasts be effectively propagated to ungauged locations using a data-assimilation-based method?

Yes, at a lead-time of 1 day the post-processed forecasts at ungauged locations showed improved skill at over 75% of the assessed locations. The skill of the post-processed forecasts decreased with increasing lead-time. Improvement over the raw forecasts is shown up to a lead-time of five days at most locations. The correlation of the forecast mean with the river discharge observations is improved along the Rhine, the largest river in the test catchment. On the other hand, the variability ratio and the mean bias ratio are improved most at locations where the raw forecast does not perform well, which are typically on smaller tributaries of the Rhine. The post-processed forecasts showed less skill for flood events than for normal flow conditions, in particular for forecasts when the flood was receding (i.e., the falling limb).

Q3.2: What factors influence the method's ability to improve the forecast skill at ungauged locations?

The accuracy of proxy-observations directly affects the performance of post-processed forecasts at ungauged locations. River stretches where proxy-observations are skillful typically see improvements from the post-processing method. On the other hand, when these estimates are less accurate—such as on smaller rivers or during extreme flood events—they can degrade the quality of the resulting forecasts. Ensemble covariances can help to mitigate the impact of incorrect nearby proxy-observations by reducing their influence within the LETKF. A benefit of the data-assimilation-inspired propagation step is that proxy-observations from multiple locations can inform the error correction allowing it to compensate for limitations at any one river gauge station. Another influential factor is the number of proxy-observations within the localisation region of an ungauged location, with a greater number generally leading to better performance. Typically, if the majority of the assimilated proxy-observations are from smaller rivers than the ungauged location, the performance is improved. On the other hand, if the assimilated observations are from larger rivers, the post-processed forecasts tends to be

degraded by a lead-time of 5 days.

8.3.2 Limitations of the study

To facilitate comparison between Chapters 5, where observations are assimilated, and Chapter 6, where proxy-observations are assimilated, I conducted the evaluation over the same time period and for the same catchment. While the limitations discussed in Section 8.2.2 also apply here, in this section I focus on a limitation specific to the analysis of Chapter 6. In this study I investigate the impact that different factors have on the skill of the post-processed forecasts. In particular, I investigate the impact of the accuracy of the proxy-observations, the number of river gauges within the localisation region, the ensemble correlation between gauged and ungauged locations, and the ratio between mean flows at the two locations. Whilst the impact of each of these factors is different, they may be correlated. For example, the at-gauge post-processing method used to generate the proxy-observations is known to have a higher skill on larger rivers (Chapter 4), and larger rivers may have more gauges. Further research is needed to fully understand the impact each factor has on the post-processed forecast.

8.4 Objective 4: To identify potential barriers for the use of the EFAS post-processed forecasts and determine future developmental priorities.

This objective is addressed in Chapter 7 where I present the outcomes from a workshop between producers and users of the EFAS post-processed forecasts. The workshop was focused on answering three research questions discussed below.

8.4.1 Summary of results

Q4.1: What is the current usage of the EFAS post-processed forecasts?

The most common use of EFAS post-processed forecasts (which have a forecast period of 15

days) was to predict upcoming floods in the next 5 days with only 2 participants indicating that they use the forecasts for longer lead-times of 10 to 15 days. Users primarily used the EFAS post-processed forecasts in conjuction with other forecast products such as the EFAS raw forecasts, meteorological forecasts or their own river discharge forecasts.

Q4.2: What are the barriers to the use of the EFAS post-processed forecasts?

A key barrier was lack of comparability between forecasts due to the different flood thresholds used and/or different variables being forecast (river discharge or water level). Users also found it difficult to find the relevant documentation and to identify auxillary information that is relevant to their regions (e.g., skill or length of the calibration period of a river gauge station). Whilst accessing the forecasts via the EFAS website was preferred by many users, those who wanted to access the forecast data found that the data dissemmination system was too slow.

Q4.3: What developments should be prioritised to better satisfy user requirements?

During the workshop developmental priorities were also discussed. A road map of developments was created including short, medium, and long term developments. These developments included smaller changes, such as improving the sign posting to the documentation for each forecast product and adding additional axillary data for the post-processed forecasts, and larger changes such as adding local flood thresholds, changing the forecast product visualisation, and adapting the method to use water level observations.

8.4.2 Limitations of the study

As of May 2025, there are over 100 EFAS partner organizations, each with multiple analysts who rely on the forecasts, but only 41 participants attended the workshop. Whilst this is a relatively large number of attendees for a co-production workshops (Varsha and Srinidhi, 2025; production Network, 2024), it means that many forecast users were not part of the discussion—even if each participant were assumed to fully represent their organization.

Logistical constraints, such as pandemic-related restrictions, parallel workshops, potential language barriers (the workshop was conducted in English), and competing priorities, limited participation. Efforts were made to improve accessibility, including a hybrid format, but ensuring broader partner engagement is an ongoing challenge (Landström et al., 2024). Moreover, communication should extend beyond the workshop itself. To address this, an additional event was held the following year to engage more partners and select prototypes for operational implementation.

The EFAS post-processed forecasts are just one component of a much larger system, as described in Chapter 3. This complexity makes it difficult to gather input from both forecast developers and users for every system component. In this study, these challenges were reflected in the workshop's duration and the need to run multiple parallel sessions focusing on different forecast products. While continuous dialogue between producers and users is the goal, the time and resources necessary to do this make it impractical to conduct regular discussions for every component (Lemos et al., 2018). The constant invitations to join workshops, webinars, and working groups also risks leading to fatigue among producers and users who are asked to participate (Lemos et al., 2018). As a result, most interactions take place at the EFAS Annual Meetings. However, attendance varies each year, leading to an inconsistent group of participants which makes it harder to build partnerships (Norström et al., 2020).

The workshop included two questionnaires and a mind-mapping activity, designed to give all participants an opportunity to contribute (Lune and Berg, 2017). However, these methods have constraints. Questionnaires capture participants' perspectives at a single point in time. To address this, two were used—one to collect initial thoughts and another after discussions. Still, spreading questionnaires throughout the year might allow more time for idea development (Iwaniec, 2019). A general feedback form is always available, but it does not specifically focus on EFAS post-processed forecasts. Additionally, the workshop setting differed from the real-world environment in which users interact with the forecasts (Curdt-Christiansen, 2019). As a result, only major challenges may have been identified, while more frequent, smaller barriers to effective forecast use may have been overlooked.

8.5 Thesis synthesis

This thesis explores how post-processing methods can improve river discharge forecasts within a large-scale operational flood forecasting system. Specifically, two post-processing approaches are investigated: (1) the existing at-gauge method used operationally in the European Flood Awareness System (EFAS), and (2) a new method for post-processing forecasts at ungauged locations, developed as part of this thesis. Figure 8.1 illustrates how information propagates through this thesis.

The at-gauge post-processing method was assessed in terms of both forecast skill (Chapter 4) and value (Chapter 7). The evaluation presented in Chapter 4 represents the first formal evaluation of the operational EFAS post-processing method. Using data from 522 river gauge stations across the EFAS domain, the analysis identified conditions under which post-processing improved forecast accuracy and when its impact was more limited. Large-scale evaluation studies such as this have become more feasible in recent years due to improved availability of hydrological datasets (Nearing et al., 2024). However, most previous studies have focused primarily on catchment characteristics, without examining the influence of error type or data availability (e.g., Du et al., 2023). This thesis extends current practice by considering these additional factors. The results, which were shared with EFAS users, highlighted the benefits of post-processing in improving both forecast accuracy and reliability.

Additionally, user engagement through a co-production framework helped identify development priorities for the operational at-gauge post-processing system. These priorities focused on increasing local relevance, improving access to forecasts and documentation, and providing supplementary information to aid interpretation. Many of the priorities have since been implemented operationally, following a second collaborative session in 2023 (CEMS, 2025). These developments include the inclusion of local thresholds, changes to visualisations of the probability distribution, updates to documentation, and providing calibration-related metadata. Chapter 7 highlights how research embedded within an operational context can lead to tangible and timely improvements.

The second main contribution of this thesis is the development of a new post-processing

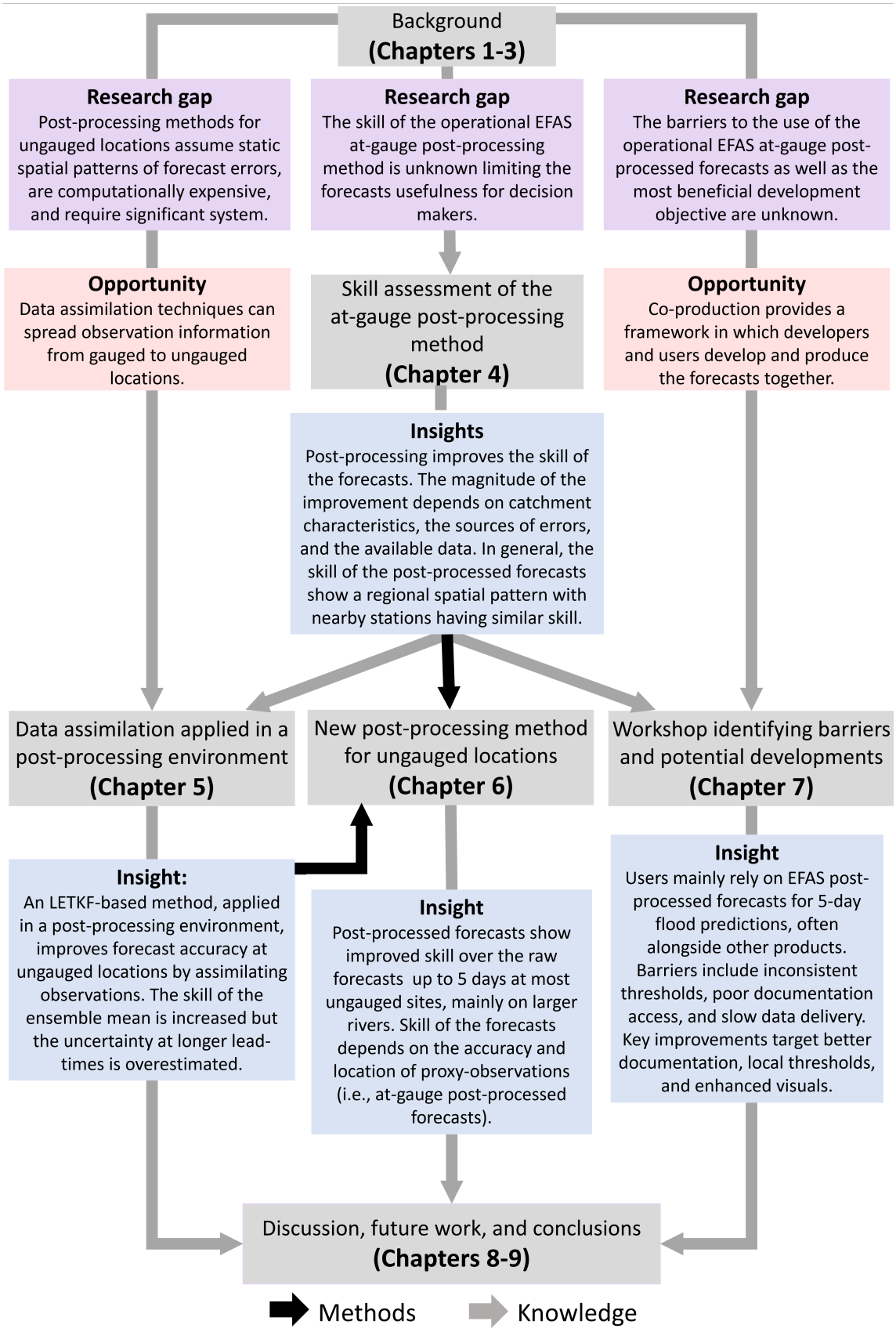

Figure 8.1: Diagram of information flow through this thesis: Chapters (grey), research gaps (purple), opportunities (red), insights (blue).

method for ungauged locations. This approach was implemented in two steps. First, a new information propagation method, inspired by data assimilation techniques, was developed and tested (Chapter 5). Using state augmentation within a Local Ensemble Transform Kalman Filter (LETKF; Hunt et al., 2007) framework, the method propagates observational information from gauged to ungauged locations in a post-processing environment, without relying on a hydrological model. Although based on established data assimilation principles, the LETKF has not, to the author's knowledge, been adapted for application in a medium-range (2–15 day) post-processing environment. The approach is computationally efficient and improves forecast accuracy relative to the raw ensemble simulations. This method could support post-event analysis (e.g., Mohr et al., 2022), offering domain-wide river discharge estimates at every timestep—without the computational cost of running a full reanalysis system.

Chapter 6 presents the full post-processing method for ungauged locations, which combines the operational at-gauge approach with the LETKF-based propagation technique. This combined method was evaluated in the Rhine–Meuse catchment and demonstrated forecast improvements up to a 5-day lead time. Although not directly comparable due to different case studies and evaluation strategies, this performance exceeds that reported for previous methods, which typically show improvements up to 2 days (e.g., Skøien et al., 2021), or are not evaluated in a forecasting context at all (Bennett et al., 2022; Farmer et al., 2018; Hales et al., 2023; Choi and Kim, 2025). Moreover, many EFAS users use the post-processed forecasts to predict floods in the next 1 to 5 days, as identified during the stakeholder workshop (Chapter 7). Making post-processed forecasts available across both gauged and ungauged locations ensures wider coverage and supports more effective flood preparedness.

While this work is centered on the EFAS system, the methods and conclusions are broadly applicable. The generalisability of the methods is discussed further in Section 8.6. The close collaboration with EFAS has enabled the research to directly impact the operational system but has also introduced challenges, particularly around data availability. Frequent system updates mean that it is common to have only a short period (1-2 years) of consistent forecasts, which complicates evaluations and the development of new methods (Hossain et al., 2025).

In summary, this thesis demonstrates how post-processing methods can improve forecast accuracy and usability. It contributes a new post-processing method for ungauged locations, the first comprehensive evaluation of the EFAS post-processed forecasts and highlights the value of stakeholder-driven development. This thesis balances scientific research with operational constraints and provides insights into how post-processing can lead to more reliable, user-focused flood forecasting services.

8.6 Future research directions

In this section I discuss some potential avenues for future work. I split the suggestions into four sections. The first section is related to the currently operational EFAS at-gauge post-processing method, the second section is related to the processes by which developments are identified, and the latter two sections are related to the post-processing method for ungauged locations.

8.6.1 Developing the current operational EFAS post-processing method

Based on the results of the evaluation presented in Chapter 4 and the outcomes of the workshop presented in Chapter 7, the following improvements for the at-gauge post-processing method were identified:

- **Ensemble Model Output Statistics (EMOS) component:** Lead-time dependence and a bias component should be included in the EMOS component to better improve errors due to the meteorological forcings (Skøien et al., 2021).
- **Model Conditional Processor (MCP) component:** The Model Conditional Processor (MCP) component of the method should be temporally localised to ensure the recent observations are only influencing lead-times for which the correlation is physically plausible (Hamill et al., 2001). This will aid the prediction of high river discharge values, particularly in areas with low hydrological variability.

- **Water level observations:** Adaptations to the method to allow the generation of water level forecasts should be investigated. As many EFAS partners use water level thresholds in their decision making, this development could make the forecasts more useful to a wider group of users.

8.6.2 Strengthening the process of identifying future developments

Potential improvements of the operational at-gauge post-processing method were identified using two approaches in this thesis: a skill assessment (Chapter 4), and a co-production workshop (Chapter 7). These processes should be repeated regularly as part of an iterative development cycle. In particular, in this thesis I assessed the post-processed forecasts of EFAS version 4, the operational version at the time of the studies. However, EFAS version 5 is now operational and these assessments should be repeated for this new version, which includes an increase in the spatial resolution of the raw forecasts (Chapter 3). To improve the ability to identify limitations these approaches should be adapted in the following ways:

- **Iterative co-development:** More workshops between developers and users of the post-processed forecasts should be organised ensuring sufficient time for discussion. Serious-games could be used within these workshops to replicate real-world scenarios (Crochemore et al., 2021; Neumann et al., 2018). Additionally, smaller focus groups could be conducted for each interested EFAS partner organisations. These focus groups would allow a more detailed discussion on their specific application, the opinions of all stakeholders to be better collected, and any language barrier to be navigated (Cairns et al., 2017).

- **Event-based evaluations:** Continuous evaluations (i.e., for every forecast within a long evaluation period) aim to capture the average skill of the forecasts (Brunner et al., 2021b). However, many forecasting systems perform differently for flood events (Brunner et al., 2021a). Evaluations focused on specific events could help better identify issues that only impact the forecast during high flows. These event-based evaluations could be conducted in collaboration with users helping to further refine the

priorities (Hyytinen et al., 2019).

8.6.3 Improving the post-processing method for ungauged locations

In Chapter 5 and Chapter 6, two components of the new post-processing method for ungauged locations were highlighted for further investigation:

- **Localisation:** The localisation applied in this thesis is symmetric for distances up- and downstream. However, this was shown to occasionally cause biases in the post-processed forecasts at locations close to confluences between large and small rivers (Chapter 6). Moderating the localisation by river size and direction along the river network should therefore be investigated to address this challenge. Additionally, covariances were found to be (correctly) larger for greater distances during precipitation events, an adaptive localisation length scale depending on weather systems should also be investigated.

- **Covariance inflation:** The holistic covariance inflation technique used in Chapters 5 and 6 is dependent on the raw ensemble perturbation matrix. However, the raw ensemble spread is not always reliable due to the single set of initial conditions. One way to address this limitation could be to perform a spread-correction on the raw ensemble before it is used in the covariance inflation (Desroziers et al., 2005; Anderson, 2007). Alternatively, a climatological hydrological model error-covariance matrix could be defined, and blended with the ensemble covariance matrix to account for more sources of uncertainty (Kotsuki and Bishop, 2022). These developments may aid the method's ability to correct the spread of the ensemble and improve the forecasts for the falling limb of flood events (Chapter 6).

8.6.4 Evaluating the transferability of the post-processing method for ungauged locations

One advantage of the post-processing method for ungauged locations is its flexibility. The localisation length scale, the covariance inflation, and the generation of initial error estimates are defined such that they can be applied to different systems and domains. Additionally, the estimation of the proxy-observations can be performed using any at-gauge estimation technique. Therefore, the transferability of the method should be assessed in the following ways:

- **Different catchments:** The method should be applied on more catchments across all hydroclimatological regions in Europe (Du et al., 2023). Selected catchments should have different error characteristics (e.g., meteorologically driven errors, or those due to the hydrological model) and physical characteristics (e.g., small/large catchments, snowmelt-driven/rainfall-driven, regulated/unregulated, etc.). This would allow the technical transferability to be tested and the impact on the skill of the forecast to be evaluated.

- **Different proxy-observations:** Alternative at-gauge post-processing methods should be used to estimate the proxy-observations leveraging the scientific research that has been conducted in recent years (Li et al., 2017). Different at-gauge post-processing methods could be used at each river gauge station depending which method provides the most accurate estimates. Alternatively, point forecasts from alternative systems (e.g., a data-driven point forecast; Nearing et al., 2024) could be used as proxy-observations. This approach could allow local forecasts from EFAS partner organisations to be combined with the EFAS forecasts, merging the two sources of information.

- **Different forecasting system:** The method can theoreticaly be applied to any ensemble river discharge forecasts. To test this the method should be used within different systems. The Global Flood Awareness System (GloFAS; Matthews et al., 2025c) could be a useful test case as it is similar in design to EFAS. By using the same

proxy-observations as in Chapter 6, this experiment could allow the impact of different ensemble covariances to be investigated further.

Chapter 9

Conclusions

The primary aim of this thesis is to improve the actionability of ensemble river discharge forecasts at gauged and ungauged locations using post-processing. Post-processing techniques typically rely on river discharge observations to estimate the relationship between the forecast and the true river discharge. Therefore, post-processing is much more challenging for locations where river discharge observations are not available. Nevertheless, decisions regarding water resources and flood management must still be made, and forecast inaccuracies can introduce additional complexity, particularly in time-sensitive situations. To address this challenge, this thesis has four key objectives, which were addressed through a forecast skill assessment, two offline prototyping experiments, and a workshop.

The new post-processing method for ungauged locations follows a two-step approach: (1) estimating proxy observations at gauged locations and (2) propagating error-corrections along the river network. In this thesis the proxy observations are generated using the operational at-gauge post-processing method used in the European Flood Awareness System (EFAS). Therefore, the skill of this at-gauge post-processing method was assessed in the first robust evaluation of the EFAS post-processed forecasts (*Objective 1*). Constraints on the number of forecasts available for the evaluation necessitated the use of reforecasts. Even with the use of reforecasts, the evaluation period was limited to 2 years. Therefore, despite post-processing providing added skill for floods exceeding the mean annual maximum, the full river discharge distribution may not have been represented and the added skill for larger floods may be more

limited.

Next, a new data-assimilation-inspired technique for propagating observational information from gauged to ungauged locations within a post-processing framework was developed (*Objective 2*). Since both components demonstrated satisfactory skill, they were combined to create a new post-processing method for ungauged locations (*Objective 3*). This method was evaluated at proxy-ungauged locations in the Rhine-Meuse catchment, where it was found to be effective. However, its overall skill was largely dependent on the accuracy of the proxy observations, further motivating the development of the operational at-gauge post-processing method. To fully understand the limitations of the method it should be evaluated across other hydrologically diverse catchments.

A workshop was organized to bring together forecast producers and users to identify key priorities for the development of EFAS's operational post-processed forecasts (*Objective 4*). While these developments were initially focused on the existing post-processing approach for gauged locations, their implications for the newly developed method were also considered. Considering the operational restrictions and user requirements throughout the thesis increases the potential uptake of the method by developers of systems like EFAS.

Further research has been identified to build on the progress made in this thesis, with priority given to improving the at-gauge post-processing method and the assessing the new post-processing method for ungauged locations for different hydroclimatological regions and for different forecasting systems. Overall, this research advances post-processing techniques, contributing to more skillful and actionable river discharge forecasts to better support decision-makers responsible for protecting lives, infrastructure, and the economy.

Bibliography

Abaza, M., F. Anctil, V. Fortin, and R. Turcotte, 2014: Sequential streamflow assimilation for short-term hydrological ensemble forecasting. *Journal of Hydrology*, **519**, 2692–2706, https://doi.org/10.1016/j.jhydrol.2014.08.038.

Abramowitz, M., and I. A. Stegun, 1972: *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications, Inc., New York.

Adams, T. E., C. Gangodagamage, and T. C. Pagano, 2024: *Flood forecasting: A global perspective*. Elsevier.

Alfieri, L., P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, and F. Pappenberger, 2013: GloFAS – global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, **17**, 1161–1175, https://doi.org/10.5194/hess-17-1161-2013.

Alfieri, L., F. Pappenberger, F. Wetterhall, T. Haiden, D. Richardson, and P. Salamon, 2014a: Evaluation of ensemble streamflow predictions in Europe. *Journal of Hydrology*, **517**, 913–922, https://doi.org/10.1016/j.jhydrol.2014.06.035.

Alfieri, L., P. Salamon, A. Bianchi, J. Neal, P. Bates, and L. Feyen, 2014b: Advances in pan-European flood hazard mapping. *Hydrological Processes*, **28**, 4067–4077, https://doi.org/10.1002/hyp.9947.

Alfieri, L., P. Salamon, F. Pappenberger, F. Wetterhall, and J. Thielen, 2012: Operational early warning systems for water-related hazards in Europe. *Environmental Science & Policy*, **21**, 35–49, https://doi.org/10.1016/j.envsci.2012.01.008.

Alfieri, L., and J. Thielen, 2015: A European precipitation index for extreme rain-storm and flash flood early warning. *Meteorological Applications*, **22**, 3–13, https://doi.org/10.1002/met.1328.

Alizadeh, B., R. A. Limon, D.-J. Seo, H. Lee, and J. Brown, 2020: Multiscale postprocessor for ensemble streamflow prediction for short to long ranges. *Journal of Hydrometeorology*, **21**, 265–285, https://doi.org/10.1175/JHM-D-19-0164.1.

Alley, R. B., K. A. Emanuel, and F. Zhang, 2019: Advances in weather prediction. *Science*, **363**, 342–344, https://doi.org/10.1126/science.aav7274.

Althoff, D., and L. N. Rodrigues, 2021: Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment. *Journal of Hydrology*, **600**, 126 674, https://doi.org/10.1016/j.jhydrol.2021.126674.

Alvisi, S., and M. Franchini, 2017: Assessment of predictive uncertainty within the framework of water demand forecasting using the Model Conditional Processor (MCP). *Urban Water Journal*, **14**, 1–10, https://doi.org/10.1016/j.proeng.2014.11.522.

Aminyavari, S., and B. Saghafian, 2019: Probabilistic streamflow forecast based on spatial post-processing of TIGGE precipitation forecasts. *Stochastic Environmental Research and Risk Assessment*, **33**, 1939–1950, https://doi.org/10.1007/s00477-019-01737-4.

Anderson, E. P., and Coauthors, 2019: Understanding rivers and their social relations: A critical step to advance environmental water management. *Wiley Interdisciplinary Reviews: Water*, **6**, e1381, https://doi.org/10.1002/wat2.1381.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of climate*, **9**, 1518–1530, https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Anderson, J. L., 2007: An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A: Dynamic meteorology and oceanography*, **59**, 210–224, https://doi.org/10.1111/j.1600-0870.2006.00216.x.

Anderson, J. L., 2012: Localization and sampling error correction in ensemble Kalman filter data assimilation. *Monthly Weather Review*, **140**, 2359–2371, https://doi.org/10.1175/MWR-D-11-00013.1.

Andrews, L., and T. E. Grantham, 2024: Strategic stream gauging network design for sustainable water management. *Nature Sustainability*, **7**, 714–723, https://doi.org/10.1038/s41893-024-01357-z.

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? *Hydrology and Earth System Sciences*, **22**, 2057–2072, https://doi.org/10.5194/hess-22-2057-2018.

Arnal, L., and Coauthors, 2019: EFAS upgrade for the extended model domain–technical documentation. Tech. rep., JRC Technical Reports, EUR 29323 EN, Publications Office of the European Union. https://doi.org/10.2760/806324.

Arnott, J. C., R. J. Neuenfeldt, and M. C. Lemos, 2020: Co-producing science for sustainability: can funding change knowledge use? *Global Environmental Change*, **60**, 101 979, https://doi.org/10.1016/j.gloenvcha.2019.101979.

Asch, M., M. Bocquet, and M. Nodet, 2016: *Data assimilation: methods, algorithms, and applications*. SIAM.

Ashby, K. R., R. C. Hales, J. Nelson, D. P. Ames, and G. P. Williams, 2021: Hydroviewer: A web application to localize global hydrologic forecasts. *Open Water Journal*, **7**, 9.

Bank, W., 2010: Costs and Benefits of Early Warning Systems. Economic and Sector Work (ESW) Paper 69358, World Bank. URL https://www.preventionweb.net/english/hyogo/gar/2011/en/bgdocs/Rogers_&_Tsirkunov_2011.pdff, accessed: 2025-05-10.

Bannister, R. N., 2008a: A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **134**, 1951–1970, https://doi.org/10.1002/qj.339.

Bannister, R. N., 2008b: A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **134**, 1971–1996, https://doi.org/10.1002/qj.339.

Bannister, R. N., 2017: A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **143**, 607–633, https://doi.org/10.1002/qj.2982.

Bao, H.-J., L.-N. Zhao, Y. He, Z.-J. Li, F. Wetterhall, H. Cloke, F. Pappenberger, and D. Manful, 2011: Coupling ensemble weather predictions based on TIGGE database with Grid-Xinanjiang model for flood forecast. *Advances in Geosciences*, **29**, 61–67, https://doi.org/10.5194/adgeo-29-61-2011.

Barnard, C., and Coauthors, 2020a: Reforecasts of river discharge and related data by the European Flood Awareness System version 4.0. [Dataset]. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessed: 2021-03-04, https://doi.org/10.24381/cds.c83f560f.

Barnard, C., and Coauthors, 2020b: River discharge and related forecasted data from the European Flood Awareness System, v4.0. European Commission, Joint Research Center (JRC), [Dataset], https://doi.org/10.24381/cds.9f696a7a.

Bartholmes, and Todini, 2005: Coupling meteorological and hydrological models for flood forecasting. *Hydrology and Earth System Sciences*, **9**, 333–346, https://doi.org/10.5194/hess-9-333-2005.

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. *Hydrology and Earth System Sciences*, **13**, 141–153, https://doi.org/10.5194/hess-13-141-2009.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Baugh, C., P. de Rosnay, H. Lawrence, T. Jurlina, M. Drusch, E. Zsoter, and C. Prudhomme, 2020: The impact of SMOS soil moisture data assimilation within the Operational Global Flood Awareness System (GloFAS). *Remote Sensing*, **12**, 1490, https://doi.org/10.3390/rs12091490.

Bell, M. J., M. J. Martin, and N. K. Nichols, 2004: Assimilation of data into an ocean model with systematic errors near the equator. *Quarterly Journal of the Royal Meteorological Society*, **130**, 873–893, https://doi.org/10.1256/qj.02.109.

Bennett, A., A. Stein, Y. Cheng, B. Nijssen, and M. McGuire, 2022: A process-conditioned and spatially consistent method for reducing systematic biases in modeled streamflow. *Journal of Hydrometeorology*, **23**, 769–783, https://doi.org/10.1175/JHM-D-21-0174.1.

Berenguer, M., D. Sempere-Torres, and G. G. Pegram, 2011: SBMcast–An ensemble nowcasting technique to assess the uncertainty in rainfall forecasts by Lagrangian extrapolation. *Journal of Hydrology*, **404**, 226–240, https://doi.org/10.1016/j.jhydrol.2011.04.033.

Berger, H., 1992,: Flow Forecasting for the River Meuse. Ph.D. thesis, Civil Engineering and Geosciences (CEG) (TU Delft).

Berghuijs, W. R., S. Harrigan, P. Molnar, L. J. Slater, and J. W. Kirchner, 2019: The relative importance of different flood-generating mechanisms across Europe. *Water Resources Research*, **55**, 4582–4593, https://doi.org/10.1029/2019WR024841.

Beven, K., 2012a: *Rainfall-Runoff Modelling: The Primer*, chap. Evolution of rainfall–runoff models: Survival of the fittest, 25–50. John Wiley & Sons, Ltd.: Chichester, UK, https://doi.org/10.1002/9781119951001.ch2.

Beven, K., 2012b: *Rainfall-Runoff Modelling: The Primer*, chap. Down to basics: Runoff processes and the modelling process, 1–23. John Wiley & Sons, Ltd.: Chichester, UK, https://doi.org/10.1002/9781119951001.

Beven, K., and P. Smith, 2015: Concepts of information content and likelihood in parameter calibration for hydrological simulation models. *Journal of Hydrologic Engineering*, **20**, A4014 010, https://doi.org/10.1061/(ASCE)HE.1943-5584.000099.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

Bierens, S., K. Boersma, and M. J. van den Homberg, 2020: The legitimacy, accountability, and ownership of an impact-based forecasting model in disaster governance. *Politics and Governance*, **8**, 445–455, https://doi.org/10.17645/pag.v8i4.3161.

Bischiniotis, K., B. van den Hurk, E. C. de Perez, T. Veldkamp, G. G. Nobre, and J. Aerts, 2019: Assessing time, cost and quality trade-offs in forecast-based action for floods. *International Journal of Disaster Risk Reduction*, **40**, 101 252, https://doi.org/10.1016/j.ijdrr.2019.101252.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Monthly weather review*, **129**, 420–436, https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.

Blöschl, G., 2022: Flood generation: process patterns from the raindrop to the ocean. *Hydrology and Earth System Sciences*, **26**, 2469–2480, https://doi.org/10.5194/hess-26-2469-2022.

Blöschl, G., and Coauthors, 2015: Increasing river floods: fiction or reality? *Wiley Interdisciplinary Reviews: Water*, **2**, 329–344, https://doi.org/10.1002/wat2.1079.

Blöschl, G., and Coauthors, 2019: Twenty-three unsolved problems in hydrology (UPH)–a community perspective. *Hydrological sciences journal*, **64**, 1141–1158, https://doi.org/10.1080/02626667.2019.1620507.

Boelee, L., D. M. Lumbroso, P. G. Samuels, and H. L. Cloke, 2019: Estimation of uncertainty in flood forecasts—A comparison of methods. *Journal of Flood Risk Management*, **12**, e12 516, https://doi.org/10.1111/jfr3.12516.

Bogner, K., and M. Kalas, 2008: Error-correction methods and evaluation of an ensemble based hydrological forecasting system for the Upper Danube catchment. *Atmospheric Science Letters*, **9**, 95–102, https://doi.org/10.1002/asl.180.

Bogner, K., K. Liechti, and M. Zappa, 2017: Combining quantile forecasts and predictive distributions of streamflows. *Hydrology and Earth System Sciences*, **21**, 5493–5502, https://doi.org/10.5194/hess-21-5493-2017.

Bogner, K., and F. Pappenberger, 2011: Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. *Water Resources Research*, **47**, https://doi.org/10.1029/2010WR009137.

Bogner, K., F. Pappenberger, and H. Cloke, 2012: The normal quantile transformation and its application in a flood forecasting system. *Hydrology and Earth System Sciences*, **16**, 1085–1094, https://doi.org/10.5194/hess-16-1085-2012.

Boucher, M.-A., L. Perreault, F. Anctil, and A.-C. Favre, 2015: Exploratory analysis of statistical post-processing methods for hydrological ensemble forecasts. *Hydrological Processes*, **29**, 1141–1155, https://doi.org/10.1002/hyp.10234.

Bourdin, D. R., S. W. Fleming, and R. B. Stull, 2012: Streamflow modelling: a primer on applications, approaches and challenges. *Atmosphere-Ocean*, **50**, 507–536, https://doi.org/10.1080/07055900.2012.734276.

Bourgin, F., M.-H. Ramos, G. Thirel, and V. Andreassian, 2014: Investigating the interactions between data assimilation and post-processing in hydrological ensemble forecasting. *Journal of Hydrology*, **519**, 2775–2784, https://doi.org/10.1016/j.jhydrol.2014.07.054.

Bremer, S., and S. Meisch, 2017: Co-production in climate change research: reviewing different perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, **8**, e482, https://doi.org/10.1002/wcc.482.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Monthly weather review*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brown, J., M.-H. Ramos, and N. Voisin, 2013: Intercomparison of streamflow postprocessing techniques: first results of a HEPEX community experiment. *EGU General Assembly Conference Abstracts*, EGU2013–8221, URL https://meetingorganizer.copernicus.org/EGU2013/EGU2013-8221.pdf.

Brown, J. D., M. He, S. Regonda, L. Wu, H. Lee, and D.-J. Seo, 2014: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. *Journal of Hydrology*, **519**, 2847–2868, https://doi.org/10.1016/j.jhydrol.2014.05.028.

Brown, J. D., and D.-J. Seo, 2010: A nonparametric postprocessor for bias correction of hydrometeorological and hydrologic ensemble forecasts. *Journal of Hydrometeorology*, **11**, 642–665, https://doi.org/10.1175/2009JHM1188.1.

Brown, J. D., and D.-J. Seo, 2013: Evaluation of a nonparametric post-processor for bias correction and uncertainty estimation of hydrologic predictions. *Hydrological Processes*, **27**, 83–105, https://doi.org/10.1002/hyp.9263.

Brunner, M. I., L. A. Melsen, A. W. Wood, O. Rakovec, N. Mizukami, W. J. Knoben, and M. P. Clark, 2021a: Flood spatial coherence, triggers, and performance in hydrological simulations: large-sample evaluation of four streamflow-calibrated models. *Hydrology and Earth System Sciences*, **25**, 105–119, https://doi.org/10.5194/hess-25-105-2021.

Brunner, M. I., L. Slater, L. M. Tallaksen, and M. Clark, 2021b: Challenges in modeling and predicting floods and droughts: A review. *Wiley Interdisciplinary Reviews: Water*, **8**, e1520, https://doi.org/10.1002/wat2.1520.

Bubeck, P., A. Otto, and J. Weichselgartner, 2017: Societal impacts of flood hazards. *Oxford research encyclopedia of natural Hazard science*, Oxford University Press, https://doi.org/10.1093/acrefore/9780199389407.013.281.

Buizza, R., F. Pappenberger, P. Salamon, J. Thielen, and A. de Roo, 2009: EPS/EFAS probabilistic flood prediction for Northern Italy: the case of 30 April 2009. *ECMWF Newsletter*, 10–16, https://doi.org/10.21957/yv858u3v4q.

Burek, P., J. D. Pozo, V. Thiemig, P. Salamon, and A. De Roo, 2011: The European Flood Alert System (EFAS). *KW- Korrespondenz Wasserwirtschaft*, **4**, 193–199, https://doi.org/10.3243/kwe2011.04.001.

Burek, P., J. Van Der Knijff, and A. De Roo, 2013: LISFLOOD - Distributed Water Balance and Flood Simulation Model - Revised User Manual 2013. Tech. Rep. EUR 26162, Publications Office of the European Union, Luxembourg. URL https://op.europa.eu/en/publication-detail/-/publication/e3b3c713-c832-4614-8527-f3ab720192f8/language-en, jRC78917.

Burgers, G., P. J. Van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Monthly weather review*, **126**, 1719–1724, https://doi.org/10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2.

Cairns, M. R., C. L. Workman, and I. Tandon, 2017: Gender mainstreaming and water development projects: analyzing unexpected enviro-social impacts in Bolivia, India, and Lesotho. *Gender, Place & Culture*, **24**, 325–342, https://doi.org/10.1080/0966369X.2017.1314945.

Camporese, M., and M. Girotto, 2022: Recent advances and opportunities in data assimilation for physics-based hydrological modeling. *Frontiers in Water*, **4**, 948 832.

Cantone, C., H. I. Grape, S. El Habash, and I. G. Pechlivanidis, 2023: A co-generation success story: improving drinking water management through hydro-climate services. *Climate Services*, **31**, 100 399, https://doi.org/10.1016/j.cliser.2023.100399.

Carrassi, A., M. Bocquet, L. Bertino, and G. Evensen, 2018: Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, **9**, e535, https://doi.org/10.1002/wcc.535.

Carter, S., A. Steynor, K. Vincent, E. Visman, and K. Waagsaether, 2019: Co-production of African weather and climate services. *Manual. Cape Town: Future Climate for Africa and Weather and Climate Information Services for Africa*.

CEMS, 2025: EFAS Versioning System. European Centre for Medium-Range Weather Forecasts (ECMWF), URL https://confluence.ecmwf.int/display/CEMS/EFAS+versioning+system, accessed: 2025-05-17, https://confluence.ecmwf.int/display/CEMS/EFAS+versioning+system.

Ceppi, A., N. A. Chaves González, S. Davolio, and G. Ravazzani, 2023: Can meteorological model forecasts initialize hydrological simulations rather than observed data in ungauged basins? *Meteorological Applications*, **30**, e2165, https://doi.org/10.1002/met.2165.

Chambers, J. M., and Coauthors, 2021: Six modes of co-production for sustainability. *Nature sustainability*, **4**, 983–996, https://doi.org/10.1038/s41893-021-00755-x.

Chen, K., S. Guo, J. Wang, P. Qin, S. He, S. Sun, and M. R. Naeini, 2019: Evaluation of GloFAS-Seasonal Forecasts for cascade reservoir impoundment operation in the upper Yangtze River. *Water*, **11**, 2539, https://doi.org/10.3390/w11122539.

Choi, J., and S. Kim, 2025: Data-driven model as a post-process for daily streamflow prediction in ungauged basins. *Heliyon*, https://doi.org/10.1016/j.heliyon.2025.e42512.

Choulga, M., F. Moschini, C. Mazzetti, S. Grimaldi, J. Disperati, H. Beck, P. Salamon, and C. Prudhomme, 2023: LISFLOOD static and parameter maps for Europe. European Commission, Joint Research Centre (JRC), [Dataset], https://doi.org/http://data.europa.eu/89h/f572c443-7466-4adf-87aa-c0847a169f23.

Choulga, M., F. Moschini, C. Mazzetti, S. Grimaldi, J. Disperati, H. Beck, P. Salamon, and C. Prudhomme, 2024: Surface fields for global environmental modelling. *Hydrology and Earth System Sciences*, **28**, 2991–3036, https://doi.org/10.5194/hess-28-2991-2024.

Clark, M. P., D. E. Rupp, R. A. Woods, X. Zheng, R. P. Ibbitt, A. G. Slater, J. Schmidt, and M. J. Uddstrom, 2008: Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in water resources*, **31**, 1309–1324, https://doi.org/10.1016/j.advwatres.2008.06.005.

Cloke, H., and F. Pappenberger, 2009: Ensemble flood forecasting: A review. *Journal of hydrology*, **375**, 613–626, https://doi.org/10.1016/j.jhydrol.2009.06.005.

Cloke, H. L., and F. Pappenberger, 2008: Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, **15**, 181–197, https://doi.org/10.1002/met.58.

Cloke, H. L., J. Thielen, F. Pappenberger, S. Nobert, G. Balint, C. Edlund, A. Koistinen, and C. de Saint-Aubin, 2009: Progress in the implementation of Hydrological Ensemble Prediction Systems (HEPS) in Europe for operational flood forecasting. *ECMWF Newsletter*, 20–24, https://doi.org/10.21957/bn6mx5nxfq.

Coccia, G., 2011: Analysis and developments of uncertainty processors for real time flood forecasting. Ph.D. thesis, Alma Mater Studiorum University of Bologna, https://doi.org/10.6092/unibo/amsdottorato/3423.

Coccia, G., and E. Todini, 2011: Recent developments in predictive uncertainty assessment based on the model conditional processor approach. *Hydrology and Earth System Sciences*, **15**, 3253–3274, https://doi.org/10.5194/hess-15-3253-2011.

Commission, E., 2002: Communication from the commission to the European Parliament and the Council: The European Community response to the flooding in Austria, Germany and several applicant countries - A solidarity-based initiative. Publications Office of the European Union, URL https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=CELEX:52004DC0472, cOM(2002) 481, Brussels, 28.8.2002.

Commission, E., 2005: An European flood forecasting system. Publications Office of the European Union, URL https://cordis.europa.eu/project/id/EVG1-CT-1999-00011.

Commission, E., 2010: Communication from the commission to the European Parliament and the Council: Towards a stronger European disaster response: the role of civil protection and humanitarian assistance. Publications Office of the European Union, URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52010DC0600, cOM(2010) 600, Brussels, 26.10.2010.

Commission, E., 2013: Decision No 1313/2013/EU of the European Parliament and of the Council of 17 December 2013 on a Union Civil Protection Mechanism (1). Publications Office of the European Union, URL http://data.europa.eu/eli/dec/2013/1313/oj, 924–947 pp.

Commission, E., 2014: Regulation (EU) No 377/2014 of the European Parliament and of the Council of 3 April 2014 establishing the Copernicus Programme and repealing Regulation (EU) No 911/2010 (1). Publications Office of the European Union, URL https://eur-lex.europa.eu/eli/reg/2014/377/oj, 44–66 pp.

Conacher, A., 2002: A role for geomorphology in integrated catchment management. *Australian Geographical Studies*, **40**, 179–195, https://doi.org/10.1111/1467-8470.00173.

Cooper, E. S., S. L. Dance, J. García-Pintado, N. K. Nichols, and P. J. Smith, 2019: Observation operators for assimilation of satellite observations in fluvial inundation forecasting. *Hydrology and Earth System Sciences*, **23**, 2541–2559, https://doi.org/10.5194/hess-23-2541-2019.

Coughlan de Perez, E., B. van den Hurk, M. Van Aalst, B. Jongman, T. Klose, and P. Suarez, 2015: Forecast-based financing: an approach for catalyzing humanitarian action based on extreme weather and climate forecasts. *Natural Hazards and Earth System Sciences*, **15**, 895–904, https://doi.org/10.5194/nhess-15-895-2015.

Coxon, G., J. Freer, I. K. Westerberg, T. Wagener, R. Woods, and P. Smith, 2015: A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water resources research*, **51**, 5531–5546, https://doi.org/10.1002/2014WR016532.

Crochemore, L., C. Cantone, I. G. Pechlivanidis, and C. S. Photiadou, 2021: How does seasonal forecast performance influence decision-making? Insights from a serious game. *Bulletin of the American Meteorological Society*, **102**, E1682–E1699, https://doi.org/10.1175/BAMS-D-20-0169.1.

Crochemore, L., M.-H. Ramos, and F. Pappenberger, 2016: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, **20**, 3601–3618, https://doi.org/10.5194/hess-20-3601-2016.

Curdt-Christiansen, X. L., 2019: Observations and Field Notes. *The Routledge Handbook of Research Methods in Applied Linguistics*, J. McKinley, and H. Rose, Eds., Routledge, 336–347, https://doi.org/10.4324/9780367824471-29, URL https://doi.org/10.4324/9780367824471-29.

Dance, S. L., and Coauthors, 2019: Improvements in forecasting intense rainfall: Results from the FRANC (Forecasting Rainfall exploiting new data Assimilation techniques and Novel observations of Convection) project. *Atmosphere*, **10**, 125, https://doi.org/10.3390/atmos10030125.

Daniels, E., S. Bharwani, Å. G. Swartling, G. Vulturius, and K. Brandon, 2020: Refocusing the climate services lens: Introducing a framework for co-designing "transdisciplinary knowledge integration processes" to build climate resilience. *Climate Services*, **19**, 100 181.

Darbandsari, P., and P. Coulibaly, 2021: HUP-BMA: An Integration of Hydrologic Uncertainty Processor and Bayesian Model Averaging for Streamflow Forecasting. *Water Resources Research*, **57**, e2020WR029 433, https://doi.org/10.1029/2020WR029433.

DARC, 2025: DARE Data Assimilation Demonstrations. Data Assimilation Research Centre GitHub, accessed: 2025-05-13, https://github.com/darc-reading/dare-da-demos.

Dasgupta, A., R. Hostache, R. Ramsankaran, G. J.-P. Schumann, S. Grimaldi, V. R. Pauwels, and J. P. Walker, 2021: On the impacts of observation location, timing, and frequency on flood extent assimilation performance. *Water Resources Research*, **57**, e2020WR028 238, https://doi.org/10.1029/2020WR028238.

Dasgupta, A., and Coauthors, 2022: Connecting hydrological modelling and forecasting from global to local scales: Perspectives from an international joint virtual workshop. *Journal of Flood Risk Management*, e12880, https://doi.org/10.1111/jfr3.12880.

Davie, T., 2019: *Fundamentals of hydrology*. Routledge.

De Roo, A., M. Odijk, G. Schmuck, E. Koster, and A. Lucieer, 2001: Assessing the effects of land use changes on floods in the Meuse and Oder catchment. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, **26**, 593–599, https://doi.org/10.1016/S1464-1909(01)00054-5.

De Roo, A., C. Wesseling, and W. Van Deursen, 2000: Physically based river basin modelling within a GIS: the LISFLOOD model. *Hydrological Processes*, **14**, 1981–1992, https://doi.org/10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F.

De Roo, A., and Coauthors, 2003: Development of a European flood forecasting system. *International Journal of River Basin Management*, **1**, 49–59, https://doi.org/10.1080/15715124.2003.9635192.

De Roo, A. P. J., and Coauthors, 2011: Quality control, validation and user feedback of the European Flood Alert System (EFAS). *International Journal of Digital Earth*, **4**, 77–90, https://doi.org/10.1080/17538947.2010.510302.

de Zea Bermudez, P., and S. Kotz, 2010: Parameter estimation of the generalized Pareto distribution—Part I. *Journal of Statistical Planning and Inference*, **140**, 1353–1373, https://doi.org/10.1016/j.jspi.2008.11.019.

DeChant, C. M., 2014: Quantifying the impacts of initial condition and model uncertainty on hydrological forecasts. Ph.D. thesis, Portland State University.

Dee, D. P., 2005: Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **131**, 3323–3343, https://doi.org/10.1256/qj.05.137.

Demargne, J., and Coauthors, 2014: The science of NOAA's operational hydrologic ensemble forecast service. *Bulletin of the American Meteorological Society*, **95**, 79–98, https://doi.org/10.1175/BAMS-D-12-00081.1.

Demeritt, D., H. Cloke, F. Pappenberger, J. Thielen, J. Bartholmes, and M. Ramos, 2007: Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environmental Hazards*, **7**, 115–127, https://doi.org/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. *Hydrological Processes*, **27**, 147–157, https://doi.org/10.1002/hyp.9419.

Desroziers, G., L. Berre, B. Chapnik, and P. Poli, 2005: Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **131**, 3385–3396, https://doi.org/10.1256/qj.05.108.

Dey, D., and C. Rao, 2006: Handbook of Statistics, vol. 25. Amsterdam: North-Holland, https://doi.org/10.1016/S0169-7161(05)25037-X.

Di Baldassarre, G., L. Brandimarte, and K. Beven, 2016: The seventh facet of uncertainty: wrong assumptions, unknowns and surprises in the dynamics of human–water systems. *Hydrological Sciences Journal*, **61**, 1748–1758, https://doi.org/10.1080/02626667.2015.1091460.

Di Baldassarre, G., and A. Montanari, 2009: Uncertainty in river discharge observations: a quantitative analysis. *Hydrology and Earth System Sciences*, **13**, 913–921, https://doi.org/10.5194/hess-13-913-2009.

Dieperink, C., 1997: International regime development: lessons from the Rhine catchment area. *TDRI Quarterly Review*, **12**, 27–35.

Dottori, F., M. Kalas, P. Salamon, A. Bianchi, L. Alfieri, and L. Feyen, 2017: An operational procedure for rapid flood risk assessment in Europe. *Natural Hazards and Earth System Sciences*, **17**, 1111–1126, https://doi.org/10.5194/nhess-17-1111-2017.

Dottori, F., P. Salamon, A. Bianchi, L. Alfieri, F. A. Hirpa, and L. Feyen, 2016: Development and evaluation of a framework for global flood hazard mapping. *Advances in water resources*, **94**, 87–102, https://doi.org/10.1016/j.advwatres.2016.05.002.

Douben, K.-J., 2006: Characteristics of river floods and flooding: a global overview, 1985–2003. *Irrigation and Drainage: The journal of the International Commission on Irrigation and Drainage*, **55**, S9–S21, https://doi.org/10.1002/ird.239.

Du, Y., I. Clemenzi, and I. G. Pechlivanidis, 2023: Hydrological regimes explain the seasonal predictability of streamflow extremes. *Environmental Research Letters*, **18**, 094 060, https://doi.org/10.1088/1748-9326/acf678.

Du, Y., and I. G. Pechlivanidis, 2025: Hybrid approaches enhance hydrological model usability for local streamflow prediction. *Communications Earth & Environment*, **6**, 334, https://doi.org/10.1038/s43247-025-02324-y.

Duc, L., K. Saito, and D. Hotta, 2020: Analysis and design of covariance inflation methods using inflation functions. Part 1: Theoretical framework. *Quarterly Journal of the Royal Meteorological Society*, **146**, 3638–3660, https://doi.org/10.1002/qj.3864.

Durand, M., and Coauthors, 2023: A framework for estimating global river discharge from the Surface Water and Ocean Topography satellite mission. *Water Resources Research*, **59**, e2021WR031 614, https://doi.org/10.1029/2021WR031614.

ECMWF, 2023: IFS Documentation CY48R1 (Vol. 1). ECMWF, URL https://www.ecmwf.int/en/publications/ifs-documentation.

Edwards, C. A., A. M. Moore, I. Hoteit, and B. D. Cornuelle, 2015: Regional ocean data assimilation. *Annual review of marine science*, **7**, 21–42, https://doi.org/10.1146/annurev-marine-010814-015821.

EFAS, 2020: Meteorological forecasts. ECMWF, accessed: 2021-04-30, https://www.efas.eu/en/meteorological-forecasts.

Ehrendorfer, M., 2007: A review of issues in ensemble-based Kalman filtering. *Meteorologische Zeitschrift (Berlin)*, **16**, https://doi.org/10.1127/0941-2948/2007/0256.

El Gharamti, M., J. L. McCreight, S. J. Noh, T. J. Hoar, A. RafieeiNasab, and B. K. Johnson, 2021a: Ensemble streamflow data assimilation using wrf-hydro and dart: novel localization and inflation techniques applied to hurricane florence flooding. *Hydrology and Earth System Sciences*, **25**, 5315–5336, https://doi.org/10.5194/hess-25-5315-2021.

El Gharamti, M., J. L. McCreight, S. J. Noh, T. J. Hoar, A. RafieeiNasab, and B. K. Johnson, 2021b: Ensemble streamflow data assimilation using wrf-hydro and dart: novel localization and inflation techniques applied to hurricane florence flooding. *Hydrology and Earth System Sciences*, **25**, 5315–5336, https://doi.org/10.5194/hess-25-5315-2021.

Emerton, R. E., and Coauthors, 2016: Continental and global scale flood forecasting systems. *Wiley Interdisciplinary Reviews: Water*, **3**, 391–418, https://doi.org/10.1002/wat2.1137.

Emery, C. M., A. Paris, S. Biancamaria, A. Boone, S. Calmant, P.-A. Garambois, and J. Santos da Silva, 2018: Large-scale hydrological model river storage and discharge correction using a satellite altimetry-based discharge product. *Hydrology and Earth System Sciences*, **22**, 2135–2162, https://doi.org/10.5194/hess-22-2135-2018.

Emery, C. M., and Coauthors, 2020: Underlying fundamentals of Kalman filtering for river network modeling. *Journal of Hydrometeorology*, **21**, 453–474, https://doi.org/10.1175/JHM-D-19-0084.1.

Engeland, K., and I. Steinsland, 2014: Probabilistic postprocessing models for flow forecasts for a system of catchments and several lead times. *Water Resources Research*, **50**, 182–197, https://doi.org/10.1002/2012WR012757.

Etter, S., B. Strobl, I. van Meerveld, and J. Seibert, 2020: Quality and timing of crowd-based water level class observations. *Hydrological Processes*, **34**, 4365–4378, https://doi.org/10.1002/hyp.13864.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, **99**, 10 143–10 162, https://doi.org/10.1029/94JC00572.

Evensen, G., 2009: Spurious correlations, localization, and inflation. *Data Assimilation: The Ensemble Kalman Filter*, Springer, 237–253, https://doi.org/10.1007/978-3-642-03711-5_15.

Evensen, G., F. C. Vossepoel, and P. J. Van Leeuwen, 2022: *Data assimilation fundamentals: A unified formulation of the state and parameter estimation problem*. Springer Nature, 120-121 pp., https://doi.org/10.1007/978-3-030-96709-3.

Fairbairn, D., 2009: Comparison of the ensemble transform Kalman filter with the ensemble transform Kalman smoother. M.S. thesis, Department of Mathematics, School of Mathematics, Meteorology and Physics, University of Reading.

Fakhruddin, S., 2014: Applications of medium range probabilistic flood forecast for societal benefits: Lessons learnt from Bangladesh. *Reducing disaster: Early warning systems for climate change*, 167–183, https://doi.org/10.1007/978-94-017-8598-3_9.

Färber, C., H. Plessow, S. Mischel, F. Kratzert, N. Addor, G. Shalev, and U. Looser, 2024: GRDC-Caravan: extending Caravan with data from the Global Runoff Data Centre. *Earth System Science Data Discussions*, **2024**, 1–17, https://doi.org/10.5194/essd-2024-427.

Farmer, W. H., T. M. Over, and J. E. Kiang, 2018: Bias correction of simulated historical daily streamflow at ungauged locations by using independently estimated flow duration curves. *Hydrology and Earth System Sciences*, **22**, 5741–5758, https://doi.org/10.5194/hess-22-5741-2018.

Fearnhead, P., and H. R. Künsch, 2018: Particle filters and data assimilation. *Annual Review of Statistics and Its Application*, **5**, 421–449, https://doi.org/10.1146/annurev-statistics-031017-100232.

Ferro, C. A., D. S. Richardson, and A. P. Weigel, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, **15**, 19–24, https://doi.org/10.1002/met.45.

Field, C. B., V. Barros, T. F. Stocker, and Q. Dahe, 2012: Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change. Tech. rep., IPCC. URL https://www.ipcc.ch/report/managing-the-risks-of-extreme-events-and-disasters-to-advance-climate-change-adaptation/.

Flack, D. L., and Coauthors, 2019: Recommendations for improving integration in national end-to-end flood forecasting systems: An overview of the FFIR (Flooding From Intense Rainfall) programme. *Water*, **11**, 725, https://doi.org/10.3390/w11040725.

Fowler, A., S. Dance, and J. Waller, 2018a: On the interaction of observation and prior error correlations in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **144**, 48–62, https://doi.org/10.1002/qj.3183.

Fowler, A. M., S. L. Dance, and J. A. Waller, 2018b: On the interaction of observation and prior error correlations in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **144**, 48–62, https://doi.org/https://doi.org/10.1002/qj.3183, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3183.

Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review*, **138**, 190–202, https://doi.org/10.1175/2009MWR3046.1.

Frame, J. M., F. Kratzert, A. Raney, M. Rahman, F. R. Salas, and G. S. Nearing, 2021: Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics. *JAWRA Journal of the American Water Resources Association*, **57**, 885–905, https://doi.org/10.1111/1752-1688.12964.

Fujisaki-Manome, A., D. Gill, T. Guo, E. J. Anderson, and M. C. Lemos, 2022: Knowledge co-production in a research-to-operation (R2O) process for development of a Great Lakes ice forecast: Reflection from a stakeholder engagement workshop. *Authorea Preprints*, https://doi.org/10.1002/essoar.10501135.1.

Furrer, R., and T. Bengtsson, 2007: Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, **98**, 227–255, https://doi.org/10.1016/j.jmva.2006.08.003.

García-Pintado, J., D. C. Mason, S. L. Dance, H. L. Cloke, J. C. Neal, J. Freer, and P. D. Bates, 2015: Satellite-supported flood forecasting in river networks: A real case study. *Journal of Hydrology*, **523**, 706–724.

García-Pintado, J., J. C. Neal, D. C. Mason, S. L. Dance, and P. D. Bates, 2013: Scheduling satellite-based SAR acquisition for sequential assimilation of water level observations into flood modelling. *Journal of Hydrology*, **495**, 252–266, https://doi.org/10.1016/j.jhydrol.2013.03.050.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, **125**, 723–757, https://doi.org/10.1002/qj.49712555417.

Gebhardt, O., and C. Kuhlicke, 2024: Co-Evaluation: How to Measure Achievements in Complex Co-Production Projects? ANYWHERE's Contribution to Enhance Emergency Management of Weather and Climate Events. *Responding to Extreme Weather Events*, 163–180, https://doi.org/0.1002/9781119741374.ch8.

Georgakakos, K. P., D.-J. Seo, H. Gupta, J. Schaake, and M. B. Butts, 2004: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *Journal of hydrology*, **298**, 222–241, https://doi.org/10.1016/j.jhydrol.2004.03.037.

Geroldinger, A., L. Lusa, M. Nold, and G. Heinze, 2023: Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures—a simulation study. *Diagnostic and Prognostic Research*, **7**, 9, https://doi.org/10.1186/s41512-023-00146-0.

Gharamti, M., and I. Hoteit, 2014: Complex step-based low-rank extended Kalman filtering for state-parameter estimation in subsurface transport models. *Journal of Hydrology*, **509**, 588–600, https://doi.org/10.1016/j.jhydrol.2013.12.004.

Giuliani, M., L. Crochemore, I. Pechlivanidis, and A. Castelletti, 2020: From skill to value: isolating the influence of end user behavior on seasonal forecast assessment. *Hydrology and Earth System Sciences*, **24**, 5891–5902, https://doi.org/10.5194/hess-24-5891-2020.

Gneiting, T., 2011: Making and evaluating point forecasts. *Journal of the American Statistical Association*, **106**, 746–762, https://doi.org/10.1198/jasa.2011.r10138.

Gneiting, T., 2014: Calibration of medium-range weather forecasts. *ECMWF Technical Memoranda*, https://doi.org/10.21957/8xna7glta.

Gneiting, T., A. Raftery, A. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Gold, S., E. White, W. Roeder, M. McAleenan, C. S. Kabban, and D. Ahner, 2020: Probabilistic contingency tables: an improvement to verify probability forecasts. *Weather and Forecasting*, **35**, 609–621, https://doi.org/10.1175/WAF-D-19-0116.1.

Golding, B., 2022: *Towards the "perfect" weather warning: Bridging disciplinary gaps through partnership and communication*. Springer Nature, https://doi.org/10.1007/978-3-030-98989-7.

Golub, G. H., and C. F. Van Loan, 2013: *Matrix computations*. JHU press.

Gouweleeuw, B., J. Thielen, G. Franchello, A. De Roo, and R. Buizza, 2005: Flood forecasting using medium-range probabilistic weather prediction. *Hydrology and Earth System Sciences*, **9**, 365–380, https://doi.org/10.5194/hess-9-365-2005.

Grainger, S., C. Murphy, and S. M. Vicente-Serrano, 2021: Barriers and opportunities for actionable knowledge production in drought risk management: Embracing the frontiers of co-production. *Frontiers in Environmental Science*, **9**, 602 128, https://doi.org/10.3389/fenvs.2021.602128.

Grewal, M. S., and A. P. Andrews, 2014: *Kalman filtering: Theory and Practice with MATLAB*. John Wiley & Sons, https://doi.org/10.1002/9780470377819.

Gupta, A., and R. S. Govindaraju, 2023: Uncertainty quantification in watershed hydrology: Which method to use? *Journal of Hydrology*, **616**, 128 749, https://doi.org/10.1016/j.jhydrol.2022.128749.

Gupta, H., H. Kling, K. Yilmaz, and G. Martinez, 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, **377**, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003.

Haiden, T., M. Janousek, F. Vitart, Z. Ben Bouallegue, L. Ferranti, F. Prates, and D. Richardson, 2021: Evaluation of ECMWF forecasts, including the 2020 upgrade. Tech. rep., European Centre for Medium Range Weather Forecasts. URL https://www.ecmwf.int/sites/default/files/elibrary/2021/19879-evaluation-ecmwf-forecasts-including-2020-upgrade.pdf.

Haiden, T., M. Janousek, F. Vitart, M. Tanguy, F. Prates, and M. Chevalier, 2024: Evaluation of ECMWF forecasts. Tech. Rep. 918, ECMWF. https://doi.org/10.21957/52f2f31351.

Haiden, T., and Coauthors, 2014: ECMWF forecast performance during the June 2013 flood in Central Europe. Tech. rep., European Centre for Medium-Range Weather Forecasts Reading, MA. https://doi.org/10.21957/4p9ebc2r3.

Hales, R. C., G. P. Williams, E. J. Nelson, R. B. Sowby, D. P. Ames, and J. L. S. Lozano, 2023: Bias correcting discharge simulations from the GEOGloWS global hydrologic model. *Journal of Hydrology*, **626**, 130 279, https://doi.org/10.1016/j.jhydrol.2023.130279.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Hamill, T. M., 2002: Ensemble-based data assimilation: a review. *University of Colorado and NOAA-CIRES Climate Diagnostics Center Boulder*.

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, **125**, 1312–1327, https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bulletin of the American Meteorological Society*, **87**, 33–46, https://doi.org/10.1175/BAMS-87-1-33.

Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, **129**, 2776–2790, https://doi.org/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2.

Hamilton, A., and R. Moore, 2012: Quantifying uncertainty in streamflow records. *Canadian Water Resources Journal/Revue canadienne des ressources hydriques*, **37**, 3–21, https://doi.org/10.4296/cwrj3701865.

Hannah, D., S. Demuth, H. van Lanen, U. Looser, C. Prudhomme, R. Rees, K. Stahl, and L. Tallaksen, 2011: Large-scale river flow archives: importance, current status and future needs. *Hydrological Processes*, **25**, 1191–1200, https://doi.org/10.1002/hyp.7794.

Harmel, R. D., and P. K. Smith, 2007: Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. *Journal of Hydrology*, **337**, 326–336, https://doi.org/10.1016/j.jhydrol.2007.01.043.

Harrigan, S., J. Hannaford, K. Muchan, and T. J. Marsh, 2018a: Designation and trend analysis of the updated UK Benchmark Network of river flow stations: The UKBN2 dataset. *Hydrology Research*, **49**, 552–567, https://doi.org/10.2166/nh.2017.058.

Harrigan, S., C. Prudhomme, S. Parry, K. Smith, and M. Tanguy, 2018b: Benchmarking ensemble streamflow prediction skill in the UK. *Hydrology and Earth System Sciences*, **22**, 2023–2039, https://doi.org/10.5194/hess-22-2023-2018.

Harrigan, S., E. Zoster, H. Cloke, P. Salamon, and C. Prudhomme, 2020: Daily ensemble river discharge reforecasts and real-time forecasts from the operational Global Flood Awareness System. *Hydrology and Earth System Sciences Discussions*, 1–22, https://doi.org/10.5194/hess-27-1-2023.

Harrison, M., D. Richardson, K. Robertson, and A. Woodcock, 1995: Medium-range ensembles using both the ECMWF T63 and unified models—An initial report. *UK Meteorological Office Tech. Rep*, **153**, 25.

Hasan, F., P. Medley, J. Drake, and G. Chen, 2024: Advancing hydrology through machine learning: insights, challenges, and future directions using the CAMELS, caravan, GRDC, CHIRPS, PERSIANN, NLDAS, GLDAS, and GRACE datasets. *Water*, **16**, 1904, https://doi.org/10.3390/w16131904.

Hashino, T., A. Bradley, and S. Schwartz, 2007: Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrology and Earth System Sciences*, **11**, 939–950, https://doi.org/10.5194/hess-11-939-2007.

Hayes, F., 2021: *Flood Modeling, Prediction and Mitigation*. Callisto Reference, New York, NY, accessed: 2025-05-10.

He, Y., and Coauthors, 2010: Ensemble forecasting using TIGGE for the July–September 2008 floods in the Upper Huai catchment: a case study. *Atmospheric Science Letters*, **11**, 132–138, https://doi.org/10.1002/asl.270.

Hemri, S., 2018: Applications of postprocessing for hydrological forecasts. *Statistical Postprocessing of Ensemble Forecasts*, 219–240, https://doi.org/10.1016/B978-0-12-812372-0.00008-X.

Hemri, S., D. Lisniak, and B. Klein, 2015: Multivariate postprocessing techniques for probabilistic hydrological forecasting. *Water Resources Research*, **51**, 7436–7451, https://doi.org/10.1002/2014WR016473.

Herrera, P. A., M. A. Marazuela, and T. Hofmann, 2022: Parameter estimation and uncertainty analysis in hydrological modeling. *Wiley Interdisciplinary Reviews: Water*, **9**, e1569, https://doi.org/10.1002/wat2.1569.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Herschy, R., 1993: The velocity-area method. *Flow Measurement and Instrumentation*, **4**, 7–10, https://doi.org/10.1016/0955-5986(93)90004-3.

Hirons, L., and Coauthors, 2021: Using co-production to improve the appropriate use of sub-seasonal forecasts in Africa. *Climate Services*, **23**, 100 246, https://doi.org/10.1016/j.cliser.2021.100246.

Hirpa, F. A., P. Salamon, H. E. Beck, V. Lorini, L. Alfieri, E. Zsoter, and S. J. Dadson, 2018: Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data. *Journal of Hydrology*, **566**, 595–606, https://doi.org/10.1016/j.jhydrol.2018.09.052.

Hodson, T. O., 2022: Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, 5481–5487, https://doi.org/10.5194/gmd-15-5481-2022.

Hofmann, H., H. Wickham, and K. Kafadar, 2017: Value plots: Boxplots for large data. *Journal of Computational and Graphical Statistics*, **26**, 469–477, https://doi.org/10.1080/10618600.2017.1305277.

Horner, I., J. Le Coz, B. Renard, F. Branger, and M. Lagouy, 2022: Streamflow uncertainty due to the limited sensitivity of controls at hydrometric stations. *Hydrological Processes*, **36**, e14 497, https://doi.org/10.1002/hyp.14497.

Horton, P., B. Schaefli, and M. Kauzlaric, 2022: Why do we have so many different hydrological models? A review based on the case of Switzerland. *Wiley Interdisciplinary Reviews: Water*, **9**, e1574, https://doi.org/10.1002/wat2.1574.

Hossain, S., H. L. Cloke, A. Ficchì, H. Gupta, L. Speight, A. Hassan, and E. M. Stephens, 2025: A decision-led evaluation approach for flood forecasting system developments: An application to the Global Flood Awareness System in Bangladesh. *Journal of Flood Risk Management*, **18**, e12 959, https://doi.org/10.1111/jfr3.12959.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Monthly weather review*, **126**, 796–811, https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

Howarth, C., C. Howarth, and S. Morse-Jones, 2019: The importance of communication, collaboration and co-production. *Resilience to Climate Change: Communication, Collaboration and Co-production*, 65–86, https://doi.org/10.1007/978-3-319-94691-7_4.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, **230**, 112–126, https://doi.org/10.1016/j.physd.2006.11.008.

Hunt, K. M., G. R. Matthews, F. Pappenberger, and C. Prudhomme, 2022: Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. *Hydrology and Earth System Sciences*, **26**, 5449–5472, https://doi.org/10.5194/hess-26-5449-2022.

Hyytinen, K., E. Saari, and M. Elg, 2019: Human-centered co-evaluation method as a means for sustainable service innovations. *Human-centered digitalization and services*, Springer, 57–75, https://doi.org/10.1007/978-981-13-7725-9_4.

ICPDR, 2021: Danube Flood Risk Management Plan (DFRMP) Update 2021. International Commission for the Protection of the Danube River, URL https://icpdr.org/main/publications/danube-flood-risk-management-plan-dfrmp-update-2021.

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc, 1997: Unified notation for data assimilation: Operational, sequential and variational (gtspecial issueltdata assimilation in meteology and oceanography: Theory and practice). *Journal of the Meteorological Society of Japan. Ser. II*, **75**, 181–189.

Isikdogan, F., A. Bovik, and P. Passalacqua, 2017: RivaMap: An automated river analysis and mapping engine. *Remote Sensing of Environment*, **202**, 88–97, https://doi.org/10.1016/j.rse.2017.03.044.

Islam, S. A., A. F. Anwar, and M. Bari, 2023: A simple method of bias correction for GCM derived streamflow at catchment scale. *Hydrological Sciences Journal*, **68**, 1409–1425, https://doi.org/10.1080/02626667.2023.2218036.

Iwaniec, J., 2019: Questionnaires: Implications for Effective Implementation. *The Routledge Handbook of Research Methods in Applied Linguistics*, J. McKinley, and H. Rose, Eds., Routledge, 324–335, https://doi.org/10.4324/9780367824471-28, URL https://doi.org/10.4324/9780367824471-28.

Jackson, E. K., W. Roberts, B. Nelsen, G. P. Williams, E. J. Nelson, and D. P. Ames, 2019: Introductory overview: Error metrics for hydrologic modelling - A review of common practices and an open source library to facilitate use and adoption. *Environmental Modelling & Software*, **119**, 32–48, https://doi.org/10.1016/j.envsoft.2019.05.001.

Jadidoleslam, N., R. Mantilla, and W. F. Krajewski, 2021: Data assimilation of satellite-based soil moisture into a distributed hydrological model for streamflow predictions. *Hydrology*, **8**, 52, https://doi.org/10.3390/hydrology8010052.

Jagannathan, K., A. D. Jones, and I. Ray, 2021: The making of a metric: Co-producing decision-relevant climate science. *Bulletin of the American Meteorological Society*, **102**, E1579–E1590, https://doi.org/10.1175/BAMS-D-19-0296.1.

Jang, P. A., and D. S. Matteson, 2023: Spatial correlation in weather forecast accuracy: a functional time series approach. *Computational Statistics*, **38**, 1215–1229, https://doi.org/10.1007/s00180-023-01338-4.

Janjić, T., and Coauthors, 2018: On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **144**, 1257–1278, https://doi.org/10.1002/qj.3130.

Johnson, C., B. J. Hoskins, and N. K. Nichols, 2005: A singular vector perspective of 4D-Var: Filtering and interpolation. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **131**, 1–19, https://doi.org/10.1256/qj.03.231.

Johnson, S., and Coauthors, 2019: SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, **12**, 1087–1117, https://doi.org/10.5194/GMD-12-1087-2019.

Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons, https://doi.org/10.1002/9781119960003.

Jones, A. S., T. L. Jones, and J. S. Horsburgh, 2022: Toward automating post processing of aquatic sensor data. *Environmental Modelling & Software*, **151**, 105 364, https://doi.org/10.1016/j.envsoft.2022.105364.

Jonkman, S., A. Curran, and L. M. Bouwer, 2024: Floods have become less deadly: an analysis of global flood fatalities 1975–2022. *Natural Hazards*, **120**, 6327–6342, https://doi.org/10.1007/s11069-024-06444-0.

Jordan, A., F. Krüger, and S. Lerch, 2019: Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software*, **90**, 1–37, https://doi.org/10.18637/jss.v090.i12.

Juston, J., A. Kauffeldt, B. Quesada Montano, J. Seibert, K. Beven, and I. Westerberg, 2013: Smiling in the rain: Seven reasons to be positive about uncertainty in hydrological modelling. *Hydrological Processes*, **27**, 1117–1122, https://doi.org/10.1002/hyp.9625.

Kalas, M., F. Dottori, C. Baught, B. Krzeminski, C. Vitolo, P. Salamon, and C. Prudhomme, 2019: Flood Hazard Impact Forecasts in the Global Flood Awareness System (GloFAS). *Geophysical Research Abstracts*, Vol. 21, URL https://meetingorganizer.copernicus.org/EGU2019/EGU2019-17518-2.pdf.

Kalas, M., M. Ramos, J. Thielen, and G. Babiakova, 2008: Evaluation of the medium-range European flood forecasts for the March–April 2006 flood in the Morava River. *Journal of Hydrology and Hydromechanics*, **56**, 116–132.

Kalman, R., 1960: A new approach to linear filtering and prediction problems. *Trans. ASME, D*, **82**, 35–44, https://doi.org/10.1115/1.3662552.

Kan, G., X. He, J. Li, L. Ding, Y. Hong, H. Zhang, K. Liang, and M. Zhang, 2019: Computer aided numerical methods for hydrological model calibration: An overview and recent development. *Archives of Computational Methods in Engineering*, **26**, 35–59, https://doi.org/10.1007/s11831-017-9224-5.

Kang, T.-H., Y.-O. Kim, and I.-P. Hong, 2010: Comparison of pre-and post-processors for ensemble streamflow prediction. *Atmospheric Science Letters*, **11**, 153–159, https://doi.org/10.1002/asl.276.

Kats, C., 2024: Co-production in mitigating the urban peak discharge. M.S. thesis, School of Geosciences, Faculty of Geosciences, URL https://studenttheses.uu.nl/handle/20.500.12932/47035.

Katz, R. W., and A. H. Murphy, 1997: *Economic value of weather and climate forecasts*. Cambridge University Press, https://doi.org/10.1093/oxfordhb/9780195398649.013.0021.

Kauffeldt, A., S. Halldin, A. Rodhe, C.-Y. Xu, and I. K. Westerberg, 2013: Disinformative data in large-scale hydrological modelling. *Hydrology and Earth System Sciences*, **17**, 2845–2857, https://doi.org/10.5194/hess-17-2845-2013.

Käyhkö, J., M. Hildén, I. Hyttinen, and K. Korhonen-Kurki, 2025: The emerging institutionalisation of knowledge co-production in sustainability research. *Ambio*, 1–14, https://doi.org/10.1007/s13280-025-02161-5.

Khaki, M., I. Hoteit, M. Kuhn, J. Awange, E. Forootan, A. I. Van Dijk, M. Schumacher, and C. Pattiaratchi, 2017: Assessing sequential data assimilation techniques for integrating GRACE data into a hydrological model. *Advances in Water Resources*, **107**, 301–316, https://doi.org/10.1016/j.advwatres.2017.07.001.

Khaniya, M., Y. Tachikawa, Y. Ichikawa, and K. Yorozu, 2022: Impact of assimilating dam outflow measurements to update distributed hydrological model states: Localization for improving ensemble Kalman filter performance. *Journal of Hydrology*, **608**, 127 651, https://doi.org/10.1016/j.jhydrol.2022.127651.

Kharin, V. V., and F. W. Zwiers, 2003: On the ROC score of probability forecasts. *Journal of Climate*, **16**, 4145–4150, https://doi.org/10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2.

Khodarahmi, M., and V. Maihami, 2023: A review on Kalman filter models. *Archives of Computational Methods in Engineering*, **30**, 727–747, https://doi.org/10.1007/s11831-022-09815-7.

Kiang, J. E., and Coauthors, 2018: A comparison of methods for streamflow uncertainty estimation. *Water Resources Research*, **54**, 7149–7176, https://doi.org/10.1029/2018WR022708.

Kim, K. B., H.-H. Kwon, and D. Han, 2018: Exploration of warm-up period in conceptual hydrological modelling. *Journal of Hydrology*, **556**, 194–210.

Kim, K. B., H.-H. Kwon, and D. Han, 2022: Intercomparison of joint bias correction methods for precipitation and flow from a hydrological perspective. *Journal of Hydrology*, **604**, 127 261, https://doi.org/10.1016/j.jhydrol.2021.127261.

Kiptum, A., and Coauthors, 2025: Advancing operational flood forecasting, early warning and risk management with new emerging science: Gaps, opportunities and barriers in Kenya. *Journal of flood risk management*, **18**, e12 884, https://doi.org/10.1111/jfr3.12884.

Klehmet, K., P. Berg, D. Bozhinova, L. Crochemore, Y. Du, I. Pechlivanidis, C. Photiadou, and W. Yang, 2024: Robustness of hydrometeorological extremes in surrogated seasonal forecasts. *International Journal of Climatology*, **44**, 1725–1738, https://doi.org/10.1002/joc.8407.

Kleiber, C., and S. Kotz, 2003: *Statistical size distributions in economics and actuarial sciences*, Vol. 470. John Wiley & Sons, https://doi.org/10.1002/0471457175.

Klein, B., I. Pechlivanidis, L. Arnal, L. Crochemore, D. Meissner, and B. Frielingsdorf, 2020: Does the application of multiple hydrological models improve seasonal streamflow forecasting skill? *EGU General Assembly Conference Abstracts*, 20187, https://doi.org/10.5194/egusphere-egu2020-20187.

Kling, H., M. Fuchs, and M. Paulin, 2012: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, **424–425**, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011.

Kling, H., and H. Gupta, 2009: On the development of regionalization relationships for lumped watershed models: The impact of ignoring sub-basin scale variability. *Journal of Hydrology*, **373**, 337–351, https://doi.org/10.1016/j.jhydrol.2009.04.031.

Kliskey, A., and Coauthors, 2023: Building trust, building futures: Knowledge co-production as relationship, design, and process in transdisciplinary science. *Frontiers in Environmental Science*, **11**, 1007 105, https://doi.org/10.3389/fenvs.2023.1007105.

Knoben, W., Y. Fan, I. Garousi-Nejad, J. Masterman, H. McMillan, J. Read, K. van Werkhoven, and M. Clark, 2025: Towards a synthesis of perceptual models of dominant hydrologic processes across North America. *European Geophysical Union 2025*, https://doi.org/10.5194/egusphere-egu25-7228.

Kolling Neto, A., and Coauthors, 2023: Advancing medium-range streamflow forecasting for large hydropower reservoirs in brazil by means of continental-scale hydrological modeling. *Water*, **15**, 1693, https://doi.org/10.3390/w15091693.

Konapala, G., S.-C. Kao, S. L. Painter, and D. Lu, 2020: Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. *Environmental Research Letters*, **15**, 104 022, https://doi.org/10.1088/1748-9326/aba927.

Kotsuki, S., and C. H. Bishop, 2022: Implementing hybrid background error covariance into the LETKF with attenuation-based localization: Experiments with a simplified AGCM. *Monthly Weather Review*, **150**, 283–302, https://doi.org/10.1175/MWR-D-21-0174.1.

Kotsuki, S., Y. Ota, and T. Miyoshi, 2017: Adaptive covariance relaxation methods for ensemble data assimilation: Experiments in the real atmosphere. *Quarterly Journal of the Royal Meteorological Society*, **143**, 2001–2015, https://doi.org/10.1002/qj.3060.

Krabbenhoft, C. A., and Coauthors, 2022: Assessing placement bias of the global river gauge network. *Nature Sustainability*, **5**, 586–592, https://doi.org/10.1038/s41893-022-00873-0.

Kratzert, F., D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, 2018: Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, **22**, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018.

Kratzert, F., and Coauthors, 2023: Caravan-A global community dataset for large-sample hydrology. *Scientific Data*, **10**, 61, https://doi.org/10.1038/s41597-023-01975-w.

Krengel, F., and Coauthors, 2018: Challenges for transboundary river management in Eastern Europe–Three case studies. *Erde*, **149**, 157–172, https://doi.org/10.12854/ERDE-2018-389.

Krzysztofowicz, R., 1999: Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research*, **35**, 2739–2750, https://doi.org/10.1029/1999WR900099.

Krzysztofowicz, R., 2001: The case for probabilistic forecasting in hydrology. *Journal of hydrology*, **249**, 2–9, https://doi.org/10.1016/S0022-1694(01)00420-6.

Krzysztofowicz, R., and H. D. Herr, 2001: Hydrologic uncertainty processor for probabilistic river stage forecasting: precipitation-dependent model. *Journal of hydrology*, **249**, 46–68, https://doi.org/10.1016/S0022-1694(01)00412-7.

Krzysztofowicz, R., and K. S. Kelly, 2000: Hydrologic uncertainty processor for probabilistic river stage forecasting. *Water resources research*, **36**, 3265–3277, https://doi.org/10.1029/2000WR900108.

Krzysztofowicz, R., and C. J. Maranzano, 2004: Hydrologic uncertainty processor for probabilistic stage transition forecasting. *Journal of Hydrology*, **293**, 57–73, https://doi.org/10.1016/j.jhydrol.2004.01.003.

Lahoz, W. A., and P. Schneider, 2014: Data assimilation: making sense of Earth Observation. *Frontiers in Environmental Science*, **2**, 16, https://doi.org/10.3389/fenvs.2014.00016.

Landström, C., E. Sarmiento, and S. J. Whatmore, 2024: Stakeholder engagement does not guarantee impact: A co-productionist perspective on model-based drought research. *Social Studies of Science*, **54**, 210–230, https://doi.org/10.1177/03063127231199220.

Lang, S., and Coauthors, 2024: AIFS–ECMWF's data-driven forecasting system. *arXiv preprint arXiv:2406.01465*, https://doi.org/10.48550/arXiv.2406.01465.

Lavers, D. A., S. Harrigan, E. Andersson, D. S. Richardson, C. Prudhomme, and F. Pappenberger, 2019: A vision for improving global flood forecasting. *Environmental Research Letters*, **14**, 121 002, https://doi.org/10.1088/1748-9326/ab52b2.

Lavers, D. A., S. Harrigan, and C. Prudhomme, 2021: Precipitation biases in the ECMWF integrated forecasting system. *Journal of Hydrometeorology*, **22**, 1187–1198, https://doi.org/10.1175/JHM-D-20-0308.1.

Le Coz, J., and Coauthors, 2016: Crowdsourced data for flood hydrology: Feedback from recent citizen science projects in Argentina, France and New Zealand. *Journal of Hydrology*, **541**, 766–777, https://doi.org/10.1016/j.jhydrol.2016.07.036.

Le Cozannet, G., M. Kervyn, S. Russo, C. Ifejika Speranza, P. Ferrier, M. Foumelis, T. Lopez, and H. Modaressi, 2020: Space-based earth observations for disaster risk management. *Surveys in geophysics*, **41**, 1209–1235, https://doi.org/10.1007/s10712-020-09586-5.

Lee, D.-G., and K.-H. Ahn, 2024: Improving medium-range streamflow forecasts over South Korea with a dual-encoder transformer model. *Journal of Environmental Management*, **368**, 122 114, https://doi.org/10.1016/j.jenvman.2024.122114.

Lemos, M. C., C. J. Kirchhoff, and V. Ramprasad, 2012: Narrowing the climate information usability gap. *Nature climate change*, **2**, 789–794, https://doi.org/10.1038/nclimate1614.

Lemos, M. C., and R. B. Rood, 2010: Climate projections and their impact on policy and practice. *Wiley interdisciplinary reviews: climate change*, **1**, 670–682, https://doi.org/10.1002/wcc.71.

Lemos, M. C., and Coauthors, 2018: To co-produce or not to co-produce. *Nature sustainability*, **1**, 722–724, https://doi.org/10.1038/s41893-018-0191-0.

Leutbecher, M., 2019: Ensemble size: How suboptimal is less than infinity? *Quarterly Journal of the Royal Meteorological Society*, **145**, 107–128, https://doi.org/10.1002/qj.3387.

Levine, S., E. Wilkinson, L. Weingärtner, and P. Mall, 2020: Anticipatory action for livelihood protection. Tech. rep., Overseas Development Institute. URL https://www.anticipation-hub.org/Documents/Policy_Papers/Anticipatory_action_for_livelihood_protection_a_collective_endeavour.pdf.

Li, B., M. Rodell, B. F. Zaitchik, R. H. Reichle, R. D. Koster, and T. M. van Dam, 2012: Assimilation of GRACE terrestrial water storage into a land surface model: Evaluation and potential value for drought monitoring in western and central Europe. *Journal of Hydrology*, **446**, 103–115, https://doi.org/10.1016/j.jhydrol.2012.04.035.

Li, D., K. M. Andreadis, S. A. Margulis, and D. P. Lettenmaier, 2020: A data assimilation framework for generating space-time continuous daily SWOT river discharge data products. *Water Resources Research*, **56**, e2019WR026 999, https://doi.org/10.1029/2019WR026999.

Li, J., T. Li, S. Liu, and H. Shi, 2018: An efficient method for mapping high-resolution global river discharge based on the algorithms of drainage network extraction. *Water*, **10**, 533, https://doi.org/10.3390/w10040533.

Li, M., Q. Wang, J. C. Bennett, and D. E. Robertson, 2016: Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting. *Hydrology and Earth System Sciences*, **20**, 3561–3579, https://doi.org/10.5194/hess-20-3561-2016.

Li, W., Q. Duan, C. Miao, A. Ye, W. Gong, and Z. Di, 2017: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews: Water*, **4**, e1246, https://doi.org/10.1002/wat2.1246.

Li, Y., Z. Cong, and D. Yang, 2023: Remotely sensed soil moisture assimilation in the distributed hydrological model based on the error subspace transform Kalman filter. *Remote Sensing*, **15**, 1852, https://doi.org/10.3390/rs15071852.

Lienert, J., J. C. M. Andersson, D. Hofmann, F. Silva Pinto, and M. Kuller, 2022: The role of multi-criteria decision analysis in a transdisciplinary process: co-developing a flood forecasting system in western Africa. *Hydrology and Earth System Sciences*, **26**, 2899–2922, https://doi.org/10.5194/hess-26-2899-2022.

Lin, J., T. Dai, L. Sheng, W. Zhang, S. Hai, and Y. Kong, 2025: Sensitivity studies of a four-dimensional local ensemble transform Kalman filter coupled with WRF-Chem version 3.9. 1 for improving particulate matter simulation accuracy. *Geoscientific Model Development*, **18**, 2231–2248, https://doi.org/10.5194/gmd-18-2231-2025.

Linsenmeier, M., and J. G. Shrader, 2023: Global inequalities in weather forecasts. *SocArXiv 7e2jf, Center for Open Science*, https://doi.org/10.31219/osf.io/7e2jf.

Liu, S., J. Wang, H. Wang, and Y. Wu, 2022: Post-processing of hydrological model simulations using the convolutional neural network and support vector regression. *Hydrology Research*, **53**, 605–621, https://doi.org/10.2166/nh.2022.004.

Liu, Y., and H. V. Gupta, 2007: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water resources research*, **43**, https://doi.org/10.1029/2006WR005756.

Liu, Y., and Coauthors, 2012: Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. *Hydrology and Earth System Sciences*, **16**, 3863–3887, https://doi.org/10.5194/hess-16-3863-2012.

Liu, Z., Y. Wang, Z. Xu, and Q. Duan, 2017: Conceptual Hydrological Models. *Handbook of Hydrometeorological Ensemble Forecasting*, Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. L. Cloke, and J. C. Schaake, Eds., Springer Berlin Heidelberg, Berlin, Heidelberg, 1–23, https://doi.org/10.1007/978-3-642-40457-3_22-1, URL https://doi.org/10.1007/978-3-642-40457-3_22-1.

Livings, D., 2005: Aspects of the ensemble Kalman filter. M.S. thesis, Department of Mathematics and Statistics, Reading University, URL https://www.reading.ac.uk/maths-and-stats/-/media/project/uor-main/schools-departments/maths/documents/livingspdf.pdf.

Livings, D. M., S. L. Dance, and N. K. Nichols, 2008: Unbiased ensemble square root filters. *Physica D: Nonlinear Phenomena*, **237**, 1021–1028, https://doi.org/10.1016/j.physd.2008.01.005.

Lorenz, E. N., 1963: Deterministic Nonperiodic Flow. *Journal of Atmospheric Sciences*, **20**, 130 – 141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Lozano, J. S., and Coauthors, 2025: Historical simulation performance evaluation and monthly flow duration curve quantile-mapping (MFDC-QM) of the GEOGLOWS ECMWF streamflow hydrologic model. *Environmental Modelling & Software*, **183**, 106 235, https://doi.org/10.1016/j.envsoft.2024.106235.

Lu, G. Y., and D. W. Wong, 2008: An adaptive inverse-distance weighting spatial interpolation technique. *Computers & Geosciences*, **34**, 1044–1055, https://doi.org/10.1016/j.cageo.2007.07.010.

Lucas, M., B. Renard, J. Le Coz, M. Lang, A. Bard, and G. Pierrefeu, 2023: Are historical stage records useful to decrease the uncertainty of flood frequency analysis? A 200-year long case study. *Journal of Hydrology*, **624**, 129 840, https://doi.org/10.1016/j.jhydrol.2023.129840.

Ludwig, P., and Coauthors, 2023: A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe–Part 2: Historical context and relation to climate change. *Natural Hazards and Earth System Sciences*, **23**, 1287–1311, https://doi.org/10.5194/nhess-23-1287-2023.

Lune, H., and B. L. Berg, 2017: *Qualitative Research Methods for the Social Sciences*. 9th ed., Pearson, Boston, MA.

Maass, C., 2021: Implementation of IFS Cycle 47r2. ECMWF, accessed: 2021-10-07, https://confluence.ecmwf.int/display/COPSRV/Implementation of IFS cycle 47r2.

MacDonald, A., C. J. Scarrott, D. Lee, B. Darlow, M. Reale, and G. Russell, 2011: A flexible extreme value mixture model. *Computational Statistics & Data Analysis*, **55**, 2137–2157, https://doi.org/10.1016/j.csda.2011.01.005.

Madu, C. N., C.-H. Kuei, I. E. Madu, B. C. Ozumba, V. E. Nnadi, U. L. Odinkonigbo, and I. C. Ezeasor, 2018: Introduction to disaster risk reduction and management. *Handbook of disaster risk reduction & management*, World Scientific, 1–29, https://doi.org/10.1142/9789813207950_0001.

Magni, M., E. H. Sutanudjaja, Y. Shen, and D. Karssenberg, 2023: Global streamflow modelling using process-informed machine learning. *Journal of Hydroinformatics*, **25**, 1648–1666, https://doi.org/10.2166/hydro.2023.217.

Mansanarez, V., B. Renard, J. L. Coz, M. Lang, and M. Darienzo, 2019: Shift happens! Adjusting stage-discharge rating curves to morphological changes at known times. *Water Resources Research*, **55**, 2876–2899, https://doi.org/10.1029/2018WR023389.

Martin, M., M. Bell, and N. K. Nichols, 2002: Estimation of systematic error in an equatorial ocean model using data assimilation. *International Journal for Numerical Methods in Fluids*, **40**, 435–444, https://doi.org/10.1002/fld.298.

Martin, M. J., 2001: Data assimilation in ocean circulation models with systematic errors. Ph.D. thesis, University of Reading.

Mason, D., J. Garcia Pintado, H. L. Cloke, S. Dance, and J. Munoz-Sabater, 2020: Assimilating high resolution remotely sensed soil moisture into a distributed hydrologic model to improve runoff prediction. *ECMWF Technical Memorandum*, https://doi.org/10.21957/5isuz4a91.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting*, **14**, 713–725, https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2.

Matthews, G., C. Barnard, H. Cloke, S. Dance, T. Jurlina, C. Mazzetti, and C. Prudhomme, 2022: Evaluating the impact of post-processing medium-range ensemble streamflow forecasts from the European Flood Awareness

System. *Hydrology and Earth System Sciences*, **26**, 2939–2968, https://doi.org/10.5194/HESS-26-2939-2022.

Matthews, G., H. L. Cloke, S. L. Dance, E. Hansford, C. Mazzetti, and C. Prudhomme, 2023: Co-design and co-production of flood forecast products: Summary of a hybrid workshop. *Bulletin of the American Meteorological Society*, **104**, E1058–E1066, https://doi.org/10.1175/BAMS-D-23-0061.1.

Matthews, G., H. L. Cloke, S. L. Dance, and C. Prudhomme, 2025a: Error-correction across gauged and ungauged locations: A data assimilation-inspired approach to post-processing river discharge forecasts. *EGUsphere*, **2025**, 1–35, https://doi.org/10.5194/hess-2024-3989.

Matthews, G., and Coauthors, 2025b: Chapter 14 - On the operational implementation of the European Flood Awareness System (EFAS). *Flood Forecasting (Second Edition)*, T. E. Adams, C. Gangodagamage, and T. C. Pagano, Eds., second edition ed., Academic Press, 251–298, https://doi.org/10.1016/B978-0-443-14009-9.00005-5, URL https://www.sciencedirect.com/science/article/pii/B9780443140099000055.

Matthews, G., and Coauthors, 2025c: Chapter 15 - On the operational implementation of the Global Flood Awareness System (GloFAS). *Flood Forecasting (Second Edition)*, T. E. Adams, C. Gangodagamage, and T. C. Pagano, Eds., second edition ed., Academic Press, 299–350, https://doi.org/10.1016/B978-0-443-14009-9.00014-6, URL https://www.sciencedirect.com/science/article/pii/B9780443140099000146.

Maurer, T., 2005: The Global Terrestrial Network for River Discharge (GTN-R): Near Real-Time Data Acquisition and Dissemination Tool for Online River Discharge and Water Level Information. *Geo-information for Disaster Management*, Springer, 1297–1314, https://doi.org/10.1007/3-540-27468-5_90.

Mazzetti, C., D. Decremer, C. Barnard, M. Blick, C. Carton de Wiart, W. F, and C. Prudhomme, 2020a: River discharge and related historical data from the European Flood Awareness System v4.0. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessed: 2021-03-04, https://doi.org/10.24381/cds.e3458969.

Mazzetti, C., D. Decremer, C. Barnard, M. Blick, C. Carton de Wiart, F. Wetterhall, and C. Prudhomme, 2020b: River discharge and related historical data from the European Flood Awareness System, v4.0 [dataset]. Joint Research Center (JRC), accessed on 10-10-2024, https://doi.org/10.24381/cds.e3458969.

Mazzetti, C., D. Decremer, and C. Prudhomme, 2021a: Challenges of the European Flood Awareness System (EFAS) hydrological calibration. ECMWF, presented at Joint Virtual Workshop on "Connecting global to local hydrological modelling and forecasting: scientific advances and challenges" [Online].

Mazzetti, C., D. Decremer, and C. Prudhomme, 2021b: Major upgrade of the European Flood Awareness System. *ECMWF Newsletter*.

Mazzetti, C., and S. Harrigan, 2020: What's new in EFAS 4.0? Model improvements, 6-hourly calibration, new evaluation layers & reporting points. Copernicus Emergency Management Service, URL https://www.efas.eu/sites/default/files/AM/AM2020/EFAS_AM_2020_2_What%20is%20new%20in%20EFAS4.pdf, presented at EFAS Annual Meeting [Online].

Mazzoleni, M., M. Verlaan, L. Alfonso, M. Monego, D. Norbiato, M. Ferri, and D. P. Solomatine, 2017: Can assimilation of crowdsourced data in hydrological modelling improve flood prediction? *Hydrology and Earth System Sciences*, **21**, 839–861, https://doi.org/10.5194/hess-21-839-2017.

McMillan, H., J. Freer, F. Pappenberger, T. Krueger, and M. Clark, 2010a: Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrological Processes: An International Journal*, **24**, 1270–1284, https://doi.org/10.1002/hyp.7587.

McMillan, H., J. Freer, F. Pappenberger, T. Krueger, and M. Clark, 2010b: Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrological Processes: An International Journal*, **24**, 1270–1284, https://doi.org/10.1002/hyp.7587.

McMillan, H., T. Krueger, and J. Freer, 2012: Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrological Processes*, **26**, 4078–4111, https://doi.org/10.1002/hyp.9384.

McMillan, H. K., G. Coxon, A. E. Sikorska-Senoner, and I. K. Westerberg, 2022: Impacts of observational uncertainty on analysis and modelling of hydrological processes: Preface. *Hydrological Processes*, e14481, https://doi.org/10.1002/hyp.14481.

McMillan, H. K., I. K. Westerberg, and T. Krueger, 2018: Hydrological data uncertainty and its implications. *Wiley Interdisciplinary Reviews: Water*, **5**, e1319, https://doi.org/10.1002/wat2.1319.

McNie, E. C., 2007: Reconciling the supply of scientific information with user demands: an analysis of the problem and review of the literature. *Environmental science & policy*, **10**, 17–38, https://doi.org/10.1016/j.envsci.2006.10.004.

Meadow, A. M., D. B. Ferguson, Z. Guido, A. Horangic, G. Owen, and T. Wall, 2015: Moving toward the deliberate coproduction of climate science knowledge. *Weather, Climate, and Society*, **7**, 179–191, https://doi.org/10.1175/WCAS-D-14-00050.1.

Micheletty, P., D. Perrot, G. Day, and K. Rittger, 2022: Assimilation of ground and satellite snow observations in a distributed hydrologic model for water supply forecasting. *JAWRA Journal of the American Water Resources Association*, **58**, 1030–1048, https://doi.org/10.1111/1752-1688.12975.

Miyoshi, T., S. Yamane, and T. Enomoto, 2007: Localizing the error covariance by physical distances within a local ensemble transform Kalman filter (LETKF). *Sola*, **3**, 89–92, https://doi.org/10.2151/sola.2007-023.

Mockler, E. M., F. E. O'Loughlin, and M. Bruen, 2016: Understanding hydrological flow paths in conceptual catchment models using uncertainty and sensitivity analysis. *Computers & Geosciences*, **90**, 66–77, https://doi.org/10.1016/j.cageo.2015.08.015.

Moges, E., Y. Demissie, L. Larsen, and F. Yassin, 2021: Sources of hydrological model uncertainties and advances in their analysis. *Water*, **13**, 28, https://doi.org/10.3390/w13010028.

Mohr, S., and Coauthors, 2022: A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe. Part 1: Event description and analysis. *Natural Hazards and Earth System Sciences Discussions*, **2022**, 1–44, https://doi.org/10.5194/nhess-23-525-2023.

Monhart, S., M. Zappa, C. Spirig, C. Schär, and K. Bogner, 2019: Subseasonal hydrometeorological ensemble predictions in small-and medium-sized mountainous catchments: benefits of the NWP approach. *Hydrology and Earth System Sciences*, **23**, 493–513, https://doi.org/10.5194/hess-23-493-2019.

Montanari, A., 2007: What do we mean by 'uncertainty'? The need for a consistent wording about uncertainty assessment in hydrology. *Hydrological Processes: An International Journal*, **21**, 841–845, https://doi.org/10.1002/hyp.6623.

Montani, A., D. Cesari, C. Marsigli, and T. Paccagnella, 2016: Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges. *Tellus A: Dynamic Meteorology and Oceanography*, **63**, 605–624, https://doi.org/10.1111/j.1600-0870.2010.00499.x.

Montani, A., and Coauthors, 2003: Operational limited-area ensemble forecasts based on the Lokal Modell. *ECMWF Newsletter*, **98**, 2–7.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Muste, M., and T. Hoitink, 2017: *Measuring Flood Discharge*. Oxford University Press, https://doi.org/10.1093/acrefore/9780199389407.013.121, URL https://oxfordre.com/naturalhazardscience/view/10.1093/acrefore/9780199389407.001.0001/acrefore-9780199389407-e-121.

Musuuza, J. L., L. Crochemore, and I. G. Pechlivanidis, 2023: Evaluation of earth observations and in situ data assimilation for seasonal hydrological forecasting. *Water Resources Research*, **59**, e2022WR033 655, https://doi.org/10.1029/2022WR033655.

Musuuza, J. L., D. Gustafsson, R. Pimentel, L. Crochemore, and I. Pechlivanidis, 2020: Impact of satellite and in situ data assimilation on hydrological predictions. *Remote Sensing*, **12**, 811, https://doi.org/10.3390/rs12050811.

Nash, J., and J. Sutcliffe, 1970: River flow forecasting through conceptual models part I: A discussion of principles. *Journal of Hydrology*, **10**, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6.

Nearing, G., and Coauthors, 2024: Global prediction of extreme floods in ungauged watersheds. *Nature*, **627**, 559–563, https://doi.org/10.1038/s41586-024-07145-1.

Nearing, G. S., D. Klotz, A. K. Sampson, F. Kratzert, M. Gauch, J. M. Frame, G. Shalev, and S. Nevo, 2021: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks. *Hydrology and earth system sciences discussions*, **2021**, 1–25, https://doi.org/10.5194/hess-26-5493-2022.

Nearing, G. S., Y. Tian, H. V. Gupta, M. P. Clark, K. W. Harrison, and S. V. Weijs, 2016: A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal*, **61**, 1666–1678, https://doi.org/10.1080/02626667.2016.1183009.

Nelson, B., 2009: Data sharing: Empty archives. *Nature*, **461**, 160–163, https://doi.org/10.1038/461160a.

Nerger, L., T. Janjić, J. Schröter, and W. Hiller, 2012: A unification of ensemble square root Kalman filters. *Monthly Weather Review*, **140**, 2335–2345, https://doi.org/10.1175/MWR-D-11-00102.1.

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

Nguyen, T. H., S. Ricci, C. Fatras, A. Piacentini, A. Delmotte, E. Lavergne, and P. Kettig, 2022: Improvement of flood extent representation with remote sensing data and data assimilation. *IEEE Transactions on Geoscience and Remote Sensing*, **60**, 1–22, https://doi.org/10.1109/TGRS.2022.3147429.

Nguyen, T. H., S. Ricci, A. Piacentini, E. Simon, R. R. Suquet, and S. P. Luque, 2023: Gaussian anamorphosis for ensemble kalman filter analysis of SAR-derived wet surface ratio observations. *IEEE Transactions on Geoscience and Remote Sensing*, **62**, 1–21, https://doi.org/10.48550/arXiv.2304.01058.

Nichols, N. K., 2003: Data Assimilation: Aims and Basic Concepts. *Data Assimilation for the Earth System*, R. Swinbank, V. Shutyaev, and W. A. Lahoz, Eds., Springer Netherlands, Dordrecht, 9–20.

Nichols, N. K., 2010: Mathematical concepts of data assimilation. *Data assimilation: making sense of observations*, 13–39, https://doi.org/10.1007/978-3-540-74703-1_2.

Norström, A. V., and Coauthors, 2020: Principles for knowledge co-production in sustainability research. *Nature sustainability*, **3**, 182–190, https://doi.org/10.1038/s41893-019-0448-2.

Ocio, D., N. Le Vine, I. Westerberg, F. Pappenberger, and W. Buytaert, 2017: The role of rating curve uncertainty in real-time flood forecasting. *Water Resources Research*, **53**, 4197–4213, https://doi.org/10.1002/2016WR020225.

Ott, E., and Coauthors, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A: Dynamic Meteorology and Oceanography*, **56**, 415–428, https://doi.org/10.3402/tellusa.v56i5.14462.

Oubanas, H., I. Gejadze, P.-O. Malaterre, M. Durand, R. Wei, R. P. Frasson, and A. Domeneghetti, 2018: Discharge estimation in ungauged basins through variational data assimilation: The potential of the SWOT mission. *Water Resources Research*, **54**, 2405–2423, https://doi.org/10.1002/2017WR021735.

Pagano, T. C., D. L. Shrestha, Q. Wang, D. Robertson, and P. Hapuarachchi, 2013: Ensemble dressing for hydrological applications. *Hydrological Processes*, **27**, 106–116, https://doi.org/10.1002/hyp.9313.

Page, R., and L. Dilling, 2019: The critical role of communities of practice and peer learning in scaling hydroclimatic information adoption. *Weather, Climate, and Society*, **11**, 851–862, https://doi.org/10.1175/WCAS-D-18-0130.1.

Panchanthan, A., A. H. Ahrari, K. Ghag, S. M. T. Mustafa, A. T. Haghighi, B. Kløve, and M. Oussalah, 2024: An overview of approaches for reducing uncertainties in hydrological forecasting: Progress and challenges. *Earth-Science Reviews*, 104956, https://doi.org/10.1016/j.earscirev.2024.104956.

Pappenberger, F., J. Bartholmes, J. Thielen, H. Cloke, R. Buizza, and A. de Roo, 2008: New dimensions in early flood warning across the globe using grand-ensemble weather predictions. *Geophysical Research Letters*, **35**, https://doi.org/10.1029/2008GL033837.

Pappenberger, F., K. Beven, N. Hunter, P. Bates, B. Gouweleeuw, J. Thielen, and A. de Roo, 2005: Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). *Hydrology and Earth System Sciences*, **9**, 381–393, https://doi.org/10.5194/HESS-9-381-2005.

Pappenberger, F., and K. J. Beven, 2006: Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water resources research*, **42**, https://doi.org/10.1029/2005WR004820.

Pappenberger, F., H. L. Cloke, D. J. Parker, F. Wetterhall, D. S. Richardson, and J. Thielen, 2015a: The monetary benefit of early flood warnings in Europe. *Environmental Science & Policy*, **51**, 278–291, https://doi.org/10.1016/j.envsci.2015.04.016.

Pappenberger, F., E. Dutra, F. Wetterhall, and H. Cloke, 2012: Deriving global flood hazard maps of fluvial floods through a physical model cascade. *Hydrology and Earth System Sciences*, **16**, 4143–4156, https://doi.org/10.5194/hess-16-4143-2012.

Pappenberger, F., M.-H. Ramos, H. L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salamon, 2015b: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology*, **522**, 697–713, https://doi.org/10.1016/j.jhydrol.2015.01.024.

Pappenberger, F., J. Thielen, and M. del Medico, 2011: The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European Flood Alert System. *Hydrological Processes*, **25**, 1091–1113, https://doi.org/10.1002/hyp.7772.

Pappenberger, F., and Coauthors, 2019: Hydrological Ensemble Prediction Systems Around the Globe. *Handbook of Hydrometeorological Ensemble Forecasting*, Springer Berlin/Heidelberg, 1187–1221, https://doi.org/10.1007/978-3-642-40457-3_47-1.

Park, S., M. Berenguer, and D. Sempere-Torres, 2019: Long-term analysis of gauge-adjusted radar rainfall accumulations at European scale. *Journal of Hydrology*, **573**, 768–777, https://doi.org/10.1016/j.jhydrol.2019.03.093.

Park, S., M. Berenguer, D. Sempere-Torres, and A. von Lerber, 2024: The EUMETNET OPERA Radar Network – European-Wide Precipitation Composites Supporting Rainfall-Induced Flash Flood Emergency Management. *Responding to Extreme Weather Events*, D. Sempere-Torres, A. Karakostas, C. Rossi, and P. Quevauviller, Eds., 1st ed., John Wiley & Sons Ltd.

Parker, D., S. Tunstall, and T. Wilson, 2005: Socio-economic benefits of flood forecasting and warning. *International conference on innovation advances and implementation of flood forecasting technology*, Vol. 17, URL https://filelist.tudelft.nl/TBM/Over%20faculteit/Afdelingen/Values%2C%20Technology%20and%20Innovation/People/Full%20Professors/Pieter%20van%20Gelder/Citations/citatie_Parker_Dennis.pdf.

Parkes, B., F. Wetterhall, F. Pappenberger, Y. He, B. Malamud, and H. Cloke, 2013: Assessment of a 1-hour gridded precipitation dataset to drive a hydrological model: a case study of the summer 2007 floods in the Upper Severn, UK. *Hydrology Research*, **44**, 89–105, https://doi.org/10.2166/nh.2011.025.

Pathiraja, S., L. Marshall, A. Sharma, and H. Moradkhani, 2016: Detecting non-stationary hydrologic model parameters in a paired catchment system using data assimilation. *Advances in water resources*, **94**, 103–119, https://doi.org/10.1016/j.advwatres.2016.04.021.

Pauwels, V. R., H.-J. Hendricks Franssen, and G. J. De Lannoy, 2020: Evaluation of State and Bias Estimates for Assimilation of SMOS Retrievals Into Conceptual Rainfall-Runoff Models. *Frontiers in Water*, **2**, 4, https://doi.org/10.3389/frwa.2020.00004.

Pechlivanidis, I. G., and Coauthors, 2025: Enhancing research-to-operations in hydrological forecasting: innovations across scales and horizons. *Bulletin of the American Meteorological Society*, **106**, E894–E919.

Pedinotti, V., A. Boone, S. Ricci, S. Biancamaria, and N. Mognard, 2014: Assimilation of satellite data to optimize large-scale hydrological model parameters: a case study for the SWOT mission. *Hydrology and Earth System Sciences*, **18**, 4485–4507, https://doi.org/10.5194/hess-18-4485-2014.

Peel, M. C., and T. A. McMahon, 2020: Historical development of rainfall-runoff modeling. *Wiley Interdisciplinary Reviews: Water*, **7**, e1471, https://doi.org/10.1002/wat2.1471.

Perera, D., O. Seidou, J. Agnihotri, M. Ramsy, V. Smakhtin, P. Coulibaly, and H. Mehmood, 2019: Flood Early Warning Systems: A Review of Benefits, Challenges and Prospects. Tech. Rep. 8, United Nations University Institute for Water, Environment and Health (UNU-INWEH), Hamilton, Canada, 30 pp. URL https://collections.unu.edu/view/UNU:8272, accessed: 2025-05-10.

Petrie, R. E., and S. L. Dance, 2010: Ensemble-based data assimilation and the localisation problem. *Weather*, **65**, 65–69, https://doi.org/10.1002/wea.505.

Piazzi, G., G. Thirel, C. Perrin, and O. Delaigue, 2021: Sequential data assimilation for streamflow forecasting: Assessing the sensitivity to uncertainties and updated variables of a conceptual hydrological model at basin scale. *Water Resources Research*, **57**.

Pretorius, L., A. Taylor, K. Iipinge, B. Mwalukanga, H. Mucavele, R. Mamombe, S. Zenda, and A. McClure, 2019: An Embedded Researcher approach to integrate climate information into decision making in southern African cities: lessons from FRACTAL. Tech. rep., FRACTAL Working Paper, 8, Future Climate for Africa. URL https://www.fractal.org.za/wp-content/uploads/2019/07/Pretorius-L-et-al-Embedded-Researcher-approach.pdf.

production Network, S. C., 2024: Methods and techniques, URL https://www.coproductionscotland.org.uk/guide/coproduction-in-practice/methods-and-techniques, accessed: 2025-05-17.

Pugliese, A., and Coauthors, 2018: A geostatistical data-assimilation technique for enhancing macro-scale rainfall–runoff simulations. *Hydrology and Earth System Sciences*, **22**, 4633–4648, https://doi.org/10.5194/hess-22-4633-2018.

Qu, B., X. Zhang, F. Pappenberger, T. Zhang, and Y. Fang, 2017: Multi-model grand ensemble hydrologic forecasting in the Fu river basin using Bayesian model averaging. *Water*, **9**, 74, https://doi.org/10.3390/w9020074.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Monthly weather review*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Rai, S. P., A. T. Wolf, N. Sharma, and H. Tiwari, 2017: Hydropolitics in transboundary water conflict and cooperation. *River system analysis and management*, 353–368, https://doi.org/10.1007/978-981-10-1472-7_19.

Rakovec, O., A. Weerts, P. Hazenberg, P. Torfs, and R. Uijlenhoet, 2012: State updating of a distributed hydrological model with Ensemble Kalman Filtering: effects of updating frequency and observation network density on forecast accuracy. *Hydrology and Earth System Sciences*, **16**, 3435–3449.

Ramos, M., J. Thielen Del Pozo, and A. de Roo, 2009: Prévision Hydrologique d'Ensemble et Alerte avec le Système Européen d'Alerte aux Crues (EFAS): Cas des Crues du Bassin du Danube en Août 2005. *Applications Des Modèles Numériques En Ingénierie 1*, **7**, 69–86.

Ramos, M.-H., J. Bartholmes, and J. Thielen-del Pozo, 2007: Development of decision support products based on ensemble forecasts in the European flood alert system. *Atmospheric Science Letters*, **8**, 113–119, https://doi.org/10.1002/asl.161.

Ramos, M. H., S. J. Van Andel, and F. Pappenberger, 2013: Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, **17**, 2219–2232, https://doi.org/10.5194/hess-17-2219-2013.

Rasmussen, J., H. Madsen, K. H. Jensen, and J. C. Refsgaard, 2016: Data assimilation in integrated hydrological modelling in the presence of observation bias. *Hydrology and earth system sciences*, **20**, 2103–2118, https://doi.org/10.5194/hess-20-2103-2016.

Raynaud, D., J. Thielen, P. Salamon, P. Burek, S. Anquetin, and L. Alfieri, 2015: A dynamic runoff co-efficient to improve flash flood early warning in Europe: evaluation on the 2013 central European floods in Germany. *Meteorological Applications*, **22**, 410–418, https://doi.org/10.1002/met.1469.

Raynaud, L., and F. Bouttier, 2017: The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, **143**, 3037–3047, https://doi.org/10.1002/qj.3159.

Refsgaard, J., P. Van der Keur, B. Nilsson, D.-I. Müller-Wohlfeil, and J. Brown, 2006: Uncertainties in river basin data at various support scales–Example from Odense Pilot River Basin. *Hydrology and Earth System Sciences Discussions*, **3**, 1943–1985, https://doi.org/10.5194/hessd-3-1943-2006.

Reggiani, P., M. Renner, A. Weerts, and P. Van Gelder, 2009: Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system. *Water Resources Research*, **45**, https://doi.org/10.1029/2007WR006758.

Reggiani, P., A. Talbi, and E. Todini, 2022: Towards informed water resources planning and management. *Hydrology*, **9**, 136, https://doi.org/10.3390/hydrology9080136.

Reichle, R. H., 2008: Data assimilation methods in the Earth sciences. *Advances in water resources*, **31**, 1411–1418, https://doi.org/10.1016/j.advwatres.2008.01.001.

Reinert, D., and Coauthors, 2021: DWD Database Reference for the Global and Regional ICON and ICON-EPS Forecasting System Version 2.1.7. Deutscher Wetterdienst, URL https://www.dwd.de/DWD/forschung/nwv/fepub/icon_database_main.pdf.

Remesan, R., and J. Mathew, 2015: *Data Based Rainfall-Runoff Modelling*, 151–182. Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-09235-5_6, URL https://doi.org/10.1007/978-3-319-09235-5_6.

Revel, M., D. Ikeshima, D. Yamazaki, and S. Kanae, 2019: A physically based empirical localization method for assimilating synthetic SWOT observations of a continental-scale river: A case study in the Congo Basin. *Water*, **11**, 829, https://doi.org/10.3390/w11040829.

Revel, M., D. Ikeshima, D. Yamazaki, and S. Kanae, 2021: A framework for estimating global-scale river discharge by assimilating satellite altimetry. *Water Resources Research*, **57**, e2020WR027 876, https://doi.org/10.1029/2020WR027876.

Revel, M., X. Zhou, D. Yamazaki, and S. Kanae, 2023: Assimilation of transformed water surface elevation to improve river discharge estimation in a continental-scale river. *Hydrology and Earth System Sciences*, **27**, 647–671, https://doi.org/10.5194/hess-27-647-2023.

Rhea, S., N. Gubbins, A. G. DelVecchia, M. R. Ross, and E. S. Bernhardt, 2023: User-focused evaluation of National Ecological Observatory Network streamflow estimates. *Scientific Data*, **10**, 89, https://doi.org/10.1038/s41597-023-01983-w.

Ridler, M.-E., D. Zhang, H. Madsen, J. Kidmose, J. C. Refsgaard, and K. H. Jensen, 2018: Bias-aware data assimilation in integrated hydrological modelling. *Hydrology Research*, **49**, 989–1004, https://doi.org/10.2166/nh.2017.117.

Robbins, J., E. Bee, A. Sneddon, S. Brown, E. Stephens, and I. Amuron, 2022: Gaining user insights into the research-to-operational elements of Impact-based Forecasting (IbF) from within the SHEAR programme: summary of findings. Tech. rep., NERC. URL https://nora.nerc.ac.uk/id/eprint/532837.

Roberts, D. R., and Coauthors, 2017: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, **40**, 913–929, https://doi.org/10.1111/ecog.02881.

Robertson, D., J. Bennett, and A. Schepen, 2016: How good is my forecasting method? Some thoughts on forecast evaluation using cross-validation based on Australian experiences. HEPEX Blog, URL https://hepex.org.au/how-good-is-my-forecasting-method-some-thoughts-on-forecast-evaluation-using-cross-validation-based-on-australian-experiences.

Robinson, M., and R. C. Ward, 2017: *Hydrology: principles and processes*. Iwa Publishing, https://doi.org/10.2166/9781780407296.

Rogers, J. S., M. M. Maneta, S. R. Sain, L. E. Madaus, and J. P. Hacker, 2025: The role of climate and population change in global flood exposure and vulnerability. *Nature Communications*, **16**, 1287, https://doi.org/10.1038/s41467-025-56654-8.

Romero-Cuellar, J., C. J. Gastulo-Tapia, M. R. Hernández-López, C. Prieto Sierra, and F. Francés, 2022: Towards an extension of the model conditional processor: predictive uncertainty quantification of monthly streamflow via Gaussian mixture models and clusters. *Water*, **14**, 1261, https://doi.org/10.3390/w14081261.

Roulin, E., and S. Vannitsem, 2015: Post-processing of medium-range probabilistic hydrological forecasting: impact of forcing, initial conditions and model errors. *Hydrological Processes*, **29**, 1434–1449, https://doi.org/10.1002/hyp.10259.

Roundy, J., Q. Duan, and J. Schaake, 2019: Hydrological predictability, scales, and uncertainty issues. *Handbook of Hydrometeorological Ensemble Forecasting*, 3–31, https://doi.org/10.1007/978-3-642-39925-1_8.

Rouzies, E., C. Lauvernet, and A. Vidard, 2024: Comparison of different ensemble assimilation methods in a modular hydrological model dedicated to water quality management. *Hydrology and Earth System Sciences Discussions*, **2024**, 1–33, https://doi.org/10.5194/hess-2024-219.

Saltikoff, E., and Coauthors, 2019: OPERA the Radar Project. *Atmosphere*, **10**, 320, https://doi.org/10.3390/ATMOS10060320.

Samaniego, L., R. Kumar, and S. Attinger, 2010: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, **46**, https://doi.org/10.1029/2008WR007327.

Sanchez Lozano, J., G. Romero Bustamante, R. C. Hales, E. J. Nelson, G. P. Williams, D. P. Ames, and N. L. Jones, 2021: A streamflow bias correction and performance evaluation web application for GEOGloWS ECMWF streamflow services. *Hydrology*, **8**, 71, https://doi.org/10.3390/hydrology8020071.

Sawada, Y., T. Nakaegawa, and T. Miyoshi, 2018: Hydrometeorology as an inversion problem: Can river discharge observations improve the atmosphere by ensemble data assimilation? *Journal of Geophysical Research: Atmospheres*, **123**, 848–860, https://doi.org/10.1002/2017JD027531.

Schaake, J. C., T. M. Hamill, R. Buizza, and M. Clark, 2007: HEPEX: the hydrological ensemble prediction experiment. *Bulletin of the American Meteorological Society*, **88**, 1541–1548, https://doi.org/10.5194/hessd-3-3321-2006.

Schaeybroeck, B. V., and S. Vannitsem, 2011: Post-processing through linear regression. *Nonlinear Processes in Geophysics*, **18**, 147–160, https://doi.org/10.5194/npg-18-147-2011.

Scheffler, G., A. Carrassi, J. Ruiz, and M. Pulido, 2022: Dynamical effects of inflation in ensemble-based data assimilation under the presence of model error. *Quarterly Journal of the Royal Meteorological Society*, **148**, 2368–2383, https://doi.org/10.1002/qj.4307.

Schumacher, M., J. Kusche, and P. Döll, 2016: A systematic impact assessment of GRACE error correlation on data assimilation in hydrological models. *Journal of Geodesy*, **90**, 537–559, https://doi.org/10.1007/s00190-016-0892-y.

Seo, D.-J., H. Herr, and J. Schaake, 2006: A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrology and Earth System Sciences Discussions*, **3**, 1987–2035, https://doi.org/10.5194/hessd-3-1987-2006.

Shamseldin, A. Y., and K. M. O'Connor, 2001: A non-linear neural network technique for updating of river flow forecasts. *Hydrology and Earth System Sciences*, **5**, 577–598, https://doi.org/10.5194/hess-5-577-2001.

Sharma, S., G. Raj Ghimire, and R. Siddique, 2023: Machine learning for postprocessing ensemble streamflow forecasts. *Journal of Hydroinformatics*, **25**, 126–139, https://doi.org/10.2166/hydro.2022.114.

Sharma, S., R. Siddique, S. Reed, P. Ahnert, P. Mendoza, and A. Mejia, 2018: Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system. *Hydrology and Earth System Sciences*, **22**, 1831–1849, https://doi.org/10.5194/hess-22-1831-2018.

Shen, C., X. Chen, and E. Laloy, 2021: Broadening the use of machine learning in hydrology. Frontiers Media SA, 681023 pp., https://doi.org/10.3389/frwa.2021.681023.

Sherman, J., and W. J. Morrison, 1950: Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, **21**, 124–127, https://doi.org/10.1214/aoms/1177729893.

Shin, M.-J., and Y. Jung, 2022: Using a global sensitivity analysis to estimate the appropriate length of calibration period in the presence of high hydrological model uncertainty. *Journal of Hydrology*, **607**, 127 546, https://doi.org/10.1016/j.jhydrol.2022.127546.

Shrestha, D. L., T. Pagano, Q. Wang, and D. Robertson, 2011: Application of Ensemble Dressing for Hydrological Applications. *Geophysical Research Abstracts*, Vol. 13, URL https://meetingorganizer.copernicus.org/EGU2011/EGU2011-5397.pdf.

Shu, Z., and Coauthors, 2023: Evaluation of the impact of multi-source uncertainties on meteorological and hydrological ensemble forecasting. *Engineering*, **24**, 212–228, https://doi.org/10.1016/j.eng.2022.06.007.

Silverman, B. W., 1984: Spline smoothing: the equivalent variable kernel method. *The annals of Statistics*, 898–916, https://doi.org/10.1214/aos/1176346710.

Singletary, L., and K. Sterle, 2020: Supporting local adaptation through the co-production of climate information: an evaluation of collaborative research processes and outcomes. *Climate Services*, **20**, 100 201, https://doi.org/10.1016/j.cliser.2020.100201.

Siqueira, V. A., A. Weerts, B. Klein, F. M. Fan, R. C. D. de Paiva, and W. Collischonn, 2021: Postprocessing continental-scale, medium-range ensemble streamflow forecasts in South America using ensemble model output statistics and ensemble copula coupling. *Journal of Hydrology*, **600**, 126 520, https://doi.org/10.1016/j.jhydrol.2021.126520.

Skøien, J. O., K. Bogner, P. Salamon, P. Smith, and F. Pappenberger, 2016: Regionalization of post-processed ensemble runoff forecasts. *Proceedings of the International Association of Hydrological Sciences*, **373**, 109–114, https://doi.org/10.5194/piahs-373-109-2016.

Skøien, J. O., K. Bogner, P. Salamon, and F. Wetterhall, 2021: On the Implementation of Postprocessing of Runoff Forecast Ensembles. *Journal of Hydrometeorology*, **22**, 2731–2749, https://doi.org/10.1175/JHM-D-21-0008.1.

Skøien, J. O., R. Merz, and G. Blöschl, 2006: Top-kriging-geostatistics on stream networks. *Hydrology and Earth System Sciences*, **10**, 277–287, https://doi.org/10.5194/hess-10-277-2006.

Skoulikaris, C., P. Venetsanou, G. Lazoglou, C. Anagnostopoulou, and K. Voudouris, 2022: Spatio-temporal interpolation and bias correction ordering analysis for hydrological simulations: An assessment on a Mountainous River Basin. *Water*, **14**, 660, https://doi.org/10.3390/w14040660.

Skute, A., D. Gruberts, J. Soms, and J. Paidere, 2008: Ecological and hydrological functions of the biggest natural floodplain in Latvia. *Ecohydrology & Hydrobiology*, **8**, 291–306, https://doi.org/10.2478/v10104-009-0023-y.

Slater, L., and Coauthors, 2024: Challenges and opportunities of ML and explainable AI in large-sample hydrology. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, https://doi.org/10.31223/X5069W.

Smith, P., G. Thornhill, S. Dance, A. Lawless, D. Mason, and N. Nichols, 2013: Data assimilation for state and parameter estimation: application to morphodynamic modelling. *Quarterly Journal of the Royal Meteorological Society*, **139**, 314–327, https://doi.org/10.1002/qj.1944.

Smith, P., and Coauthors, 2016: Chapter 11 - On the Operational Implementation of the European Flood Awareness System (EFAS). *Flood Forecasting*, T. E. Adams, and T. C. Pagano, Eds., Academic Press, Boston, 313–348, https://doi.org/10.1016/B978-0-12-801884-2.00011-6.

Smith, P. J., S. L. Dance, M. J. Baines, N. K. Nichols, and T. R. Scott, 2009: Variational data assimilation for parameter estimation: application to a simple morphodynamic model. *Ocean Dynamics*, **59**, 697–708, https://doi.org/10.1007/s10236-009-0205-6.

Speight, L., and Coauthors, 2025: Recommendations to improve the interpretation of global flood forecasts to support international humanitarian operations for tropical cyclones. *Journal of Flood Risk Management*, **18**, e12 952, https://doi.org/10.1111/jfr3.12952.

Stephens, E., and H. Cloke, 2014: Improving flood forecasts for better flood preparedness in the UK (and beyond). *The Geographical Journal*, **180**, 310–316, https://doi.org/10.1111/geoj.12103.

Stern, H., and N. E. Davidson, 2015: Trends in the skill of weather prediction at lead times of 1–14 days. *Quarterly Journal of the Royal Meteorological Society*, **141**, 2726–2736, https://doi.org/10.1002/qj.2559.

Stewart, L. M., S. L. Dance, and N. K. Nichols, 2013: Data assimilation with correlated observation errors: experiments with a 1-D shallow water model. *Tellus A: Dynamic Meteorology and Oceanography*, **65**, 19 546, https://doi.org/10.3402/tellusa.v65i0.19546.

Strahler, A. N., 1957: Quantitative analysis of watershed geomorphology. *Eos, Transactions American Geophysical Union*, **38**, 913–920, https://doi.org/10.1029/TR038i006p00913.

Sun, L., I. Nistor, and O. Seidou, 2015: Streamflow data assimilation in SWAT model using Extended Kalman Filter. *Journal of Hydrology*, **531**, 671–684, https://doi.org/10.1016/j.jhydrol.2015.10.060.

Sun, L., O. Seidou, I. Nistor, and K. Liu, 2016: Review of the Kalman-type hydrological data assimilation. *Hydrological Sciences Journal*, **61**, 2348–2366, https://doi.org/10.1080/02626667.2015.1127376.

Tabeart, J. M., S. L. Dance, A. S. Lawless, N. K. Nichols, and J. A. Waller, 2020: Improving the condition number of estimated covariance matrices. *Tellus A: Dynamic Meteorology and Oceanography*, **72**, 1–19, https://doi.org/10.1080/16000870.2019.1696646.

Takeshi, A., 1985: *Advanced econometrics*, 125–127. Harvard university press.

Talagrand, O., 1999: Evaluation of probabilistic prediction systems. *Workshop Proceedings" Workshop on Predictability", 20-22 October 1997, ECMWF, Reading, UK*, URL https://www.ecmwf.int/en/elibrary/76596-evaluation-probabilistic-prediction-systems.

Tanguy, M., M. Eastman, A. Chevuturi, E. Magee, E. Cooper, R. H. Johnson, K. Facer-Childs, and J. Hannaford, 2025: Optimising ensemble streamflow predictions with bias correction and data assimilation techniques. *Hydrology and Earth System Sciences*, **29**, 1587–1614.

Team, R. C., 2019: *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing, URL https://www.R-project.org/.

Thiboult, A., F. Anctil, and M. Ramos, 2017: How does the quantification of uncertainties affect the quality and value of flood early warning systems? *Journal of Hydrology*, **551**, 365–373, https://doi.org/10.1016/j.jhydrol.2017.05.014.

Thielen, J., J. Bartholmes, M. H. Ramos, and A. de Roo, 2009: The European flood alert system - Part 1: Concept and development. *Hydrology and Earth System Sciences*, **13**, 125–140, https://doi.org/10.5194/HESS-13-125-2009.

Thielen-del Pozo, J., V. Thiemig, F. Pappenberger, B. Revilla-Romero, P. Salamon, T. De Groeve, and F. Hirpa, 2015: The Benefit of Continental Flood Early Warning Systems to Reduce the Impact of Flood Disasters: An Assessment for Europe and an Outlook for Africa. EUR - Scientific and Technical Research Reports JRC97266, European Commission, Joint Research Centre. https://doi.org/10.2788/46941, URL https://op.europa.eu/en/publication-detail/-/publication/8fde9714-c00f-11e5-9e54-01aa75ed71a1/language-en, accessed: 2025-05-10.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble square root filters. *Monthly weather review*, **131**, 1485–1490, https://doi.org/10.1175/1520-0493(2003)131<1485:ESRF>2.0.CO;2.

Tiwari, A. D., P. Mukhopadhyay, and V. Mishra, 2022: Influence of bias correction of meteorological and streamflow forecast on hydrological prediction in India. *Journal of Hydrometeorology*, **23**, 1171–1192, https://doi.org/10.1175/JHM-D-20-0235.1.

Todini, E., 2008: A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, **6**, 123–137, https://doi.org/10.1080/15715124.2008.9635342.

Todini, E., 2009: Predictive uncertainty assessment in real time flood forecasting. *Uncertainties in environmental modelling and consequences for policy making*, Springer, 205–228, https://doi.org/10.1007/978-90-481-2636-1_9.

Todini, E., 2013: From HUP to MCP: Analogies and extended performances. *Journal of hydrology*, **477**, 33–42, https://doi.org/10.1016/j.jhydrol.2012.10.037.

Todini, E., 2017: Flood forecasting and decision making in the new millennium. Where are we? *Water Resources Management*, **31**, 3111–3129, https://doi.org/10.1007/s11269-017-1693-7.

Todini, E., G. Coccia, and E. Ortiz, 2015: On the proper use of ensembles for predictive uncertainty assessment. *EGU General Assembly Conference Abstracts*, 10365, URL https://meetingorganizer.copernicus.org/EGU2015/EGU2015-10365.pdf.

Tomkins, K. M., 2014: Uncertainty in streamflow rating curves: methods, controls and consequences. *Hydrological processes*, **28**, 464–481, https://doi.org/10.1002/hyp.9567.

Troin, M., R. Arsenault, A. Wood, F. Brissette, and J.-L. Martel, 2021: Generating ensemble streamflow forecasts: A review of methods and approaches over the past 40 years. *Water Resources Research*, **57**, e2020WR028 392, https://doi.org/10.1029/2020WR028392.

Turnipseed, D., and V. Sauer, 2010: Discharge Measurements at Gaging Stations. Tech. rep., U.S. Geological Survey, 87 pp. URL http://pubs.usgs.gov/tm/tm3-a8/.

Tyralis, H., and G. Papacharalampous, 2023: Hydrological post-processing for predicting extreme quantiles. *Journal of Hydrology*, **617**, 129 082, https://doi.org/10.1016/j.jhydrol.2023.129082.

Uehlinger, U., K. M. Wantzen, R. Leuven, and H. Arndt, 2009: *The Rhine river basin*. Bibliothek der Universität Konstanz, https://doi.org/10.1016/B978-0-12-369449-2.00006-0.

UNDRR, 2015: Sendai Framework for Disaster Risk Reduction 2015-2030. United Nations Office for Disaster Risk Reduction, Geneva, Switzerland, URL https://www.undrr.org/publication/sendai-framework-disaster-risk-reduction-2015-2030, accessed: 2025-03-14.

UNDRR, 2024: Global Status of Multi-Hazard Early Warning Systems 2024. Tech. rep., https://www.undrr.org/reports/global-status-MHEWS-2024, United Nations Office for Disaster Risk Reduction (UNDRR) and World Meteorological Organization (WMO). Accessed: 2025-05-10.

UNDRR, C. ., 2020: The Human Cost of Disasters: An Overview of the Last 20 Years (2000-2019). Tech. rep., Centre for Research on the Epidemiology of Disasters (CRED) and United Nations Office for Disaster Risk Reduction (UNDRR). URL https://www.preventionweb.net/files/74124_humancostofdisasters20002019reportu.pdf.

Valdez, E. S., F. Anctil, and M.-H. Ramos, 2022: Choosing between post-processing precipitation forecasts or chaining several uncertainty quantification tools in hydrological forecasting systems. *Hydrology and Earth System Sciences*, **26**, 197–220, https://doi.org/10.5194/hess-26-197-2022.

Valler, V., J. Franke, and S. Brönnimann, 2019: Impact of different estimations of the background-error covariance matrix on climate reconstructions based on data assimilation. *Climate of the Past*, **15**, 1427–1441, https://doi.org/10.5194/cp-15-1427-2019.

van Andel, S. J., A. Weerts, J. Schaake, and K. Bogner, 2013: Post-processing hydrological ensemble predictions intercomparison experiment. *Hydrological Processes*, **27**, 158–161, https://doi.org/10.1002/hyp.9595.

van der Graaf, P., M. Cheetham, S. Redgate, C. Humble, and A. Adamson, 2021: Co-production in local government: process, codification and capacity building of new knowledge in collective reflection spaces. Workshops findings from a UK mixed methods study. *Health Research Policy and Systems*, **19**, 1–13, https://doi.org/10.1186/s12961-021-00677-2.

van der Knijff, J. M., J. Younis, and A. P. J. de Roo, 2008: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, **24**, 189–212, https://doi.org/10.1080/13658810802549154.

Van Houtven, G., 2024: Economic Value of Flood Forecasts and Early Warning Systems: A Review. *Natural Hazards Review*, **25**, 03124 002, https://doi.org/10.1061/NHREFO.NHENG-2094.

Vandaele, R., S. L. Dance, and V. Ojha, 2021: Deep learning for automated river-level monitoring through river-camera images: an approach based on water segmentation and transfer learning. *Hydrology and Earth System Sciences*, **25**, 4435–4453, https://doi.org/10.5194/hess-25-4435-2021.

Vannitsem, S., and Coauthors, 2021: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World. *Bulletin of the American Meteorological Society*, **102**, E681 – E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Varsha, N., and A. Srinidhi, 2025: Bridging Knowledge and Action: The Community Hydrology Approach. WELL LABS, URL https://welllabs.org/bridging-knowledge-and-action-the-community-hydrology-approach/, accessed: 2025-05-17.

Velazquez, J. A., F. Anctil, M.-H. Ramos, and C. Perrin, 2011: Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. *Advances in Geosciences*, **29**, 33–42, https://doi.org/10.5194/adgeo-29-33-2011.

Venables, W. N., and B. D. Ripley, 2002: *Modern Applied Statistics with S*. 4th ed., Springer, New York, URL http://www.stats.ox.ac.uk/pub/MASS4.

Verkade, J., J. Brown, F. Davids, P. Reggiani, and A. Weerts, 2017: Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine. *Journal of Hydrology*, **555**, 257–277, https://doi.org/10.1016/j.jhydrol.2017.10.024.

Verkade, J., J. Brown, P. Reggiani, and A. Weerts, 2013: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, **501**, 73–91, https://doi.org/10.1016/j.jhydrol.2013.07.039.

Verkade, J. S., and M. G. F. Werner, 2011: Estimating the benefits of single value and probability forecasting for flood warning. *Hydrology and Earth System Sciences*, **15**, 3751–3765, https://doi.org/10.5194/hess-15-3751-2011.

Vincent, K., M. Daly, C. Scannell, and B. Leathes, 2018: What can climate services learn from theory and practice of co-production? *Climate Services*, **12**, 48–58.

Vincent, K., A. Steynor, A. McClure, E. Visman, K. L. Waagsaether, S. Carter, and N. Mittal, 2021: Co-production: learning from contexts. *Climate Risk in Africa: Adaptation and Resilience*, 37–56, https://doi.org/10.1007/978-3-030-61160-6_3.

Vincent, K., and Coauthors, 2020: Re-balancing climate services to inform climate-resilient planning–A conceptual framework and illustrations from sub-Saharan Africa. *Climate Risk Management*, **29**, 100 242, https://doi.org/10.1016/j.crm.2020.100242.

Viney, N. R., and B. C. Bates, 2004: It never rains on Sunday: The prevalence and implications of untagged multi-day rainfall accumulations in the Australian high quality data set. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **24**, 1171–1192, https://doi.org/10.1002/joc.1053.

Visman, E., and Coauthors, 2022: Institutionalising co-production of weather and climate services: learning from the African SWIFT and ForPAc projects. *White Rose Research Online*.

Vitart, F., and Coauthors, 2019: Extended-range prediction. ECMWF Technical Memoranda 854, ECMWF. https://doi.org/10.21957/PDIVP3T9M, URL https://doi.org/10.21957/PDIVP3T9M.

Vyas, J. K., M. Perumal, and T. Moramarco, 2024: Non-contact discharge estimation at a river site by using only the maximum surface flow velocity. *Journal of Hydrology*, **638**, 131 505, https://doi.org/10.1016/j.jhydrol.2024.131505.

Waller, J. A., S. L. Dance, A. S. Lawless, and N. K. Nichols, 2014: Estimating correlated observation error statistics using an ensemble transform Kalman filter. *Tellus A: Dynamic Meteorology and Oceanography*, **66**, 23 294, https://doi.org/10.3402/tellusa.v66.23294, https://doi.org/10.3402/tellusa.v66.23294.

Wan, E. A., and R. Van Der Merwe, 2000: The unscented Kalman filter for nonlinear estimation. *Proceedings of the IEEE 2000 adaptive systems for signal processing, communications, and control symposium (Cat. No. 00EX373)*, IEEE, 153–158, https://doi.org/10.1109/ASSPCC.2000.882463.

Wang, B., Z. Sun, X. Jiang, J. Zeng, and R. Liu, 2023: Kalman filter and its application in data assimilation. *Atmosphere*, **14**, 1319, https://doi.org/10.3390/atmos14081319.

Wang, F., J. Polcher, P. Peylin, and V. Bastrikov, 2018: Assimilation of river discharge in a land surface model to improve estimates of the continental water cycles. *Hydrology and Earth System Sciences*, **22**, 3863–3882, https://doi.org/10.5194/hess-22-3863-2018.

Wang, H., and G. He, 2022: Rivers: Linking nature, life, and civilization. *River*, **1**, 25–36, https://doi.org/doi.org/10.1002/rvr2.7.

Wazed, M. A., S. K. Saha, A. Rahman, and S. Barmon, 2021: Forecast-based Financing in Bangladesh: Scopes and Challenges. Tech. rep., SUFAL Consortium. URL https://www.anticipation-hub.org/Documents/Reports/Forecast_based_financing_in_Bangladesh_-_scoping_and_challenges.pdf, accessed: 2025-05-10.

Weerts, A., H. Winsemius, and J. Verkade, 2011: Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales). *Hydrology and Earth System Sciences*, **15**, 255–265, https://doi.org/10.5194/hess-15-255-2011.

Westerberg, I., V. Mansanarez, S. Lyon, and N. Lam, 2020: Comparison of rating-curve uncertainty estimation using hydraulic modelling and power-law methods. *EGU General Assembly Conference Abstracts*, 6779, https://doi.org/10.5194/egusphere-egu2020-6779.

Westerberg, I. K., and R. H. Karlsen, 2024: Sharing perceptual models of uncertainty: On the use of soft information about discharge data. *Hydrological Processes*, **38**, e15 145, https://doi.org/10.1002/hyp.15145.

Wetterhall, F., and F. di Giuseppe, 2018: The benefit of seamless forecasts for hydrological predictions over Europe. *Hydrology and Earth System Sciences*, **22**, 3409–3420, https://doi.org/10.5194/HESS-22-3409-2018.

Wetterhall, F., and Coauthors, 2013: HESS Opinions "Forecaster priorities for improving probabilistic flood forecasts". *Hydrology and Earth System Sciences*, **17**, 4389–4399, https://doi.org/10.5194/hess-17-4389-2013.

Wishner, R. P., J. A. Tabaczynski, and M. Athans, 1969: A comparison of three non-linear filters. *Automatica*, **5**, 487–496, https://doi.org/10.1016/0005-1098(69)90110-1.

Węglarczyk, S., 2018: Kernel Density Estimation and Its Application. *ITM Web of Conferences*, EDP Sciences, Vol. 23, 00037, https://doi.org/10.1051/itmconf/20182300037.

WMO, 2011: Manual on Flood Forecasting and Warning. Tech. rep., World Meteorological Organization, Geneva. URL https://library.wmo.int/idurl/4/35881.

WMO, 2022: *Assessment Guidelines for End-to-End Flood Forecasting and Early Warning Systems*. WMO-No. 1286, 36 pp., URL https://library.wmo.int/doc_num.php?explnum_id=11379.

WMO, 2023: Manual on the Global Telecommunication System. Tech. rep., World Meteorological Organization, Geneva. URL https://library.wmo.int/viewer/35800/download?file=386_2023-edition_en.pdf&type=pdf&navigator=1, annex III to the WMO Technical Regulations.

WMO, 2024a: Hydromet Gap Report 2024. Tech. rep., World Meteorological Organization and Alliance for Hydromet Development. URL https://library.wmo.int/idurl/4/68926, accessed: 2025-01-10.

WMO, 2024b: State of Global Water Resources Report 2023. Tech. Rep. WMO-No. 1362, World Meteorological Organization, Geneva, 81 pp. URL https://library.wmo.int/idurl/4/69033, accessed: 2025-05-10.

(WMO), W. M. O., 2021: *WMO Atlas of Mortality and Economic Losses from Weather, Climate, and Water Extremes (1970–2019)*. WMO-No. 1267, 90 pp., URL https://library.wmo.int/doc_num.php?explnum_id=10989.

Woldemeskel, F., D. McInerney, J. Lerat, M. Thyer, D. Kavetski, D. Shin, N. Tuteja, and G. Kuczera, 2018: Evaluating post-processing approaches for monthly and seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, **22**, 6257–6278, https://doi.org/10.5194/hess-22-6257-2018.

Won, J., J. Seo, J. Lee, J. Choi, Y. Park, O. Lee, and S. Kim, 2023: Streamflow predictions in ungauged basins using recurrent neural network and decision tree-based algorithm: application to the southern region of the Korean peninsula. *Water*, **15**, 2485, https://doi.org/10.3390/w15132485.

Wood, A. W., and J. C. Schaake, 2008: Correcting errors in streamflow forecast ensemble mean and spread. *Journal of Hydrometeorology*, **9**, 132–148, https://doi.org/10.1175/2007JHM862.1.

Wu, R., L. Yang, C. Chen, S. Ahmad, S. M. Dascalu, and F. C. Harris Jr, 2018: Modeling error learning based post-processor framework for hydrologic models accuracy improvement. *Geoscientific Model Development Discussions*, **1**, https://doi.org/10.5194/gmd-12-4115-2019.

Wu, W., R. Emerton, Q. Duan, A. W. Wood, F. Wetterhall, and D. E. Robertson, 2020: Ensemble flood forecasting: Current status and future opportunities. *Wiley Interdisciplinary Reviews: Water*, **7**, e1432, https://doi.org/10.1002/wat2.1432.

Xiao, M., and Coauthors, 2025: LSTM-based Post-Processing Improves Streamflow Prediction in the Sierra Nevada. *Authorea Preprints*, https://doi.org/10.22541/essoar.174534330.09680378/v1.

Xie, X., and D. Zhang, 2010: Data assimilation for distributed hydrological catchment modeling via ensemble Kalman filter. *Advances in Water Resources*, **33**, 678–690, https://doi.org/10.1016/j.advwatres.2010.03.012.

Xu, J., F. Anctil, and M.-A. Boucher, 2019: Hydrological post-processing of streamflow forecasts issued from multimodel ensemble prediction systems. *Journal of Hydrology*, **578**, 124 002, https://doi.org/10.1016/j.jhydrol.2019.124002.

Xu, J., F. Anctil, and M.-A. Boucher, 2020: Exploring hydrologic post-processing of ensemble stream flow forecasts based on Affine kernel dressing and Nondominated sorting genetic algorithm II. *Hydrology and Earth System Sciences Discussions*, **2020**, 1–34, https://doi.org/10.5194/hess-26-1001-2022.

Ye, A., Q. Duan, X. Yuan, E. F. Wood, and J. Schaake, 2014: Hydrologic post-processing of MOPEX streamflow simulations. *Journal of hydrology*, **508**, 147–156, https://doi.org/10.1016/j.jhydrol.2013.10.055.

Younis, J., M.-H. Ramos, and J. Thielen, 2008: EFAS forecasts for the March–April 2006 flood in the Czech part of the Elbe River Basin—a case study. *Atmospheric Science Letters*, **9**, 88–94, https://doi.org/10.1002/ASL.179.

Zalachori, I., M.-H. Ramos, R. Garçon, T. Mathevet, and J. Gailhard, 2012: Statistical processing of forecasts for hydrological ensemble prediction: A comparative study of different bias correction strategies. *Advances in Science and Research*, **8**, 135–141, https://doi.org/10.5194/asr-8-135-2012.

Zamo, M., and P. Naveau, 2018: Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, **50**, 209–234, https://doi.org/10.1007/s11004-017-9709-7.

Zängl, G., D. Reinert, P. Rípodas, and M. Baldauf, 2015: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, **141**, 563–579, https://doi.org/10.1002/qj.2378.

Zappa, M., S. Jaun, U. Germann, A. Walser, and F. Fundel, 2011: Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmospheric Research*, **100**, 246–262, https://doi.org/10.1016/j.atmosres.2010.12.005.

Zappa, M., and Coauthors, 2010: Propagation of uncertainty from observing systems and NWP into hydrological models: COST-731 Working Group 2. *Atmospheric Science Letters*, **11**, 83–91, https://doi.org/10.1002/asl.248.

Zeitoun, M., M. Goulden, and D. Tickner, 2013: Current and future challenges facing transboundary river basin management. *Wiley Interdisciplinary Reviews: Climate Change*, **4**, 331–349, https://doi.org/10.1002/WCC.228.

Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of Initial Estimate and Observation Availability on Convective-Scale Data Assimilation with an Ensemble Kalman Filter. *Monthly Weather Review*, **132**, 1238 – 1253, https://doi.org/10.1175/1520-0493(2004)132<1238:IOIEAO>2.0.CO;2.

Zhang, J., W. Li, and Q. Duan, 2025a: Quantifying the contributions of hydrological pre-processor, post-processor, and data assimilator to ensemble streamflow prediction skill. *Journal of Hydrology*, **651**, 132 611, https://doi.org/10.1016/j.jhydrol.2024.132611.

Zhang, Y., A. Ye, J. Li, P. Nguyen, B. Analui, K. Hsu, and S. Sorooshian, 2025b: Improve streamflow simulations by combining machine learning pre-processing and post-processing. *Journal of Hydrology*, 132904, https://doi.org/10.1016/j.jhydrol.2025.132904.

Zhao, L., Q. Duan, J. Schaake, A. Ye, and J. Xia, 2011: A hydrologic post-processor for ensemble streamflow predictions. *Advances in Geosciences*, **29**, 51–59, https://doi.org/10.5194/adgeo-29-51-2011.

Zhong, Y., S. Guo, F. Xiong, D. Liu, H. Ba, and X. Wu, 2020: Probabilistic forecasting based on ensemble forecasts and EMOS method for TGR inflow. *Frontiers of Earth Science*, **14**, 188–200, https://doi.org/10.1007/s11707-019-0773-9.

Zijl, F., J. Sumihar, and M. Verlaan, 2015: Application of data assimilation for improved operational water level forecasting on the northwest European shelf and North Sea. *Ocean Dynamics*, **65**, 1699–1716, https://doi.org/10.1007/s10236-015-0898-7.

Appendix A

Supplementary material for Chapter 5

A.1 Kalman gain matrix decomposition

The Kalman gain matrix has the following form for timestep k:

$$\mathbf{K}_k = \mathbf{P}_k^f \mathbf{H}_k^\top \left(\mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k^\top + \mathbf{R}_k \right)^{-1} \tag{A.1}$$

where $\mathbf{P}_k^f$ is the prior ensemble covariance matrix, $\mathbf{R}_k$ is the error-covariance matrix of the observations, and $\mathbf{H}_k$ is the observation operator (Livings, 2005; Hunt et al., 2007; Kalman, 1960). Substituting the definitions of the perturbation matrix and the observation operator for the augmented state given in Eqs. (5.9) and (5.11) gives:

$$\mathbf{K}_k = \mathbf{W}_k^f \mathbf{W}_k^{fT} \mathbf{H}_k^\top \left(\mathbf{H}_k \mathbf{W}_k^f \mathbf{W}_k^{fT} \mathbf{H}_k^\top + \mathbf{R}_k \right)^{-1}.$$

This can then be decomposed into the hindcast and error components as

$$\mathbf{K}_k = \begin{pmatrix} \mathbf{K}_{\mathbf{x}_k} \\ \mathbf{K}_{\mathbf{b}_k} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_k^f + \mathbf{B}_k^f \\ \mathbf{B}_k^f \end{pmatrix} (\mathbf{H}_k \mathbf{X}_k^f + \mathbf{H}_k \mathbf{B}_k^f)^T ((\mathbf{H}_k \mathbf{X}_k^f + \mathbf{H}_k \mathbf{B}_k^f)(\mathbf{H}_k \mathbf{X}_k^f + \mathbf{H}_k \mathbf{B}_k^f)^T + \mathbf{R}_k)^{-1} \tag{A.2}$$

The analysis of the ensemble mean of the augmented states is therefore given by

$$\overline{\mathbf{w}}_k^a = \begin{pmatrix} \overline{\mathbf{x}}_k^f + \overline{\mathbf{b}}_k^f \\ \overline{\mathbf{b}}_k^f \end{pmatrix} + \begin{pmatrix} \mathbf{K}_{\mathbf{x}_k} \\ \mathbf{K}_{\mathbf{b}_k} \end{pmatrix} (\mathbf{y}_k - \overline{\mathbf{y}}_k^{\mathbf{x}}), \tag{A.3}$$

where $\overline{\mathbf{x}}_k^f$ and $\overline{\mathbf{b}}_k^f$ are the ensembles mean of the raw hindcast and the prior error ensemble at timestep k, $\mathbf{y}_k$ is the observation vector, and $\overline{\mathbf{y}}_k^{\mathbf{x}}$ is the model-observation ensemble mean.

A.2 Gaspari-Cohn function

The Gaspari-Cohn function is correlation function commonly used in data assimilation to define the localisation weights (Gaspari and Cohn, 1999). It has the following form:

$$\rho(r) = \begin{cases} 1 - \frac{5}{3}r^2 + \frac{5}{8}r^3 + \frac{1}{2}r^4 - \frac{1}{4}r^5, & 0 \le r \le 1 \\ -\frac{2}{3}r^{-1} + 4 - 5r + \frac{5}{3}r^2 + \frac{5}{8}r^3 - \frac{1}{2}r^4 + \frac{1}{12}r^5, & 1 < r \le 2 \\ 0, & r > 2 \end{cases} \tag{A.4}$$

where $r = d/c$ where d is the physical distance between two points, and c is the localisation length scale. The function has a value of 1 when $r = 0$ and a value of 0 when $r > 2$.