Evaluating and improving flood inundation forecasts using satellite data

---

HELEN HOOKER

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

Doctor of Philosophy

PhD Atmosphere, Oceans and Climate

Department of Meteorology

School of Mathematical, Physical and Computational Sciences

January 2024

University of Reading

Declaration

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

Helen Hooker

Publications

Chapters 3, 4 and 5 of this thesis are reproduced from the following publications respectively:

- Hooker, H., Dance, S. L., Mason, D. C., Bevington, J., & Shelton, K. (2022). Spatial scale evaluation of forecast flood inundation maps. Journal of Hydrology, 128170. https://doi.org/https://doi.org/10.1016/j.jhydrol.2022.128170
- Hooker, H., Dance, S. L., Mason, D. C., Bevington, J., & Shelton, K. (2023). Assessing the spatial spread–skill of ensemble flood maps with remote-sensing observations. Natural Hazards and Earth System Sciences, 23(8), 2769–2785. https://doi.org/10.5194/nhess-23-2769-2023
- Hooker, H., Dance, S. L., Mason, D. C., Bevington, J., & Shelton, K. (2023). A multi-system comparison of forecast flooding extent using a scale-selective approach. Hydrology Research, 54(10), 1115–1133. https://doi.org/10.2166/nh.2023.025

All work undertaken in these publications was carried out by Helen Hooker with coauthors providing guidance and review.

Abstract

Flood inundation forecast maps provide an essential tool to disaster management teams for planning and preparation ahead of a flood event in order to mitigate the impacts of flooding. The maps can be used to inform forecast-based financing schemes to release funds ahead of a predicted flood event. Evaluating the accuracy of forecast flood maps is essential for model development and improving future flood predictions. The goal of this thesis is to develop spatial verification methods for deterministic and ensemble flood map forecasts and to improve forecasts using satellite data. Binary verification measures typically provide a domain-averaged score of forecast skill. The skill score is dependent on the magnitude of the flood and the spatial scale of the flood map. In this thesis, a new scale-selective approach is presented to evaluate both deterministic and ensemble forecast flood maps against remotely observed flood extents. The flood-edge location accuracy proves to be more sensitive to variations in forecast skill and spatial scale compared to the accuracy of the entire flood extent. Both the ensemble spatial-skill and spread-skill relationship vary with location and can be linked to the physical characteristics of the flooding event. We find that a scale-selective verification approach can quantify the skill of three systems operating at different spatial scales, so that the benefits and limitations of each system can be evaluated. A new data assimilation framework is presented to update the flood map selection from a static library of flood maps using satellite data, taking account of observation uncertainties. Results show that the flood map selection could be

triggered in four out of five sub-catchments tested. The resultant analysis flood map has the potential to be used to trigger a secondary finance scheme during a flood event and avoid missed financing opportunities for humanitarian action. Overall, sensitive spatial verification methods that are location specific and can evaluate ensemble performance will aid future model development for flood inundation prediction.

Dedication

For Richard, Ebony, Kian and Maple

Acknowledgements

I would like to thank my previous work colleagues for their support and encouragement to apply for and pursue a PhD. Thank you to my lead supervisor Sarah Dance for believing I could do it, trusting in my ideas and correcting my grammar (no more greengrocer's apostrophe)! Thank you to David Mason for his wisdom and support. My work always felt meaningful because of the direct links to my CASE sponsor the JBA Trust and it was great to work with John Bevington and Kay Shelton throughout, thank you for hosting me at your Wallingford office too.

It was always reassuring to be part of something bigger, thank you to SCENARIO DTP and Wendy Neale along with NCEO and Uzma Saeed. I'm grateful for the fun training opportunities both provided. It was great to be part of Cohort 7 PhD students in Meteorology, and thank you to my office buddies Charlie and Natalie for listening to me ramble on. Thank you to DARC for the challenging seminars and to my fellow PhD DARCies Gwyn, Devon, Harriet, Ieuan and Laura. Thank you to Hugo Dalton for sharing his art. I'd also like to thank Hannah Cloke and Liz Stephens for the Water group seminars, journal club meetings and watery activities.

Finally, I'd like to thank my Mum and Dad for teaching me to keep on learning and to Rich, Ebony and Kian for encouraging me to get on with it and keep on going. Thank you to Maple (my English Springer) for her company throughout.

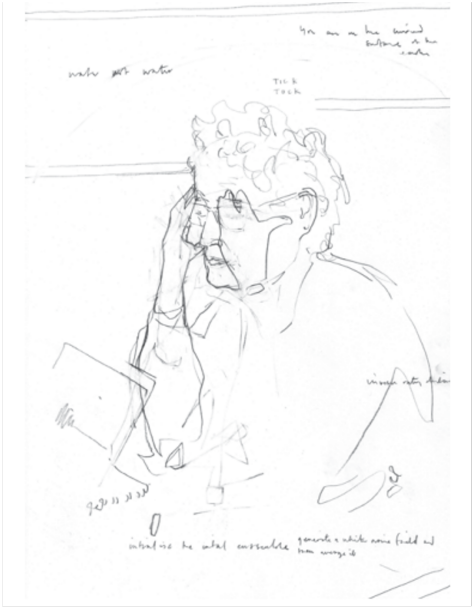
Assimilated Watercolours: Pop up art exhibitions in Care Homes

During my PhD I took part in a sci-art collaboration. The project created a series of exhibitions that explore current research at the University of Reading's Department of Meteorology through historical and contemporary works of art. Scientists at Reading aim to develop new ways to make flood forecasting more accurate by assimilating visual and mathematical data from sources including satellite imagery, aerial photographs and existing flooding maps.

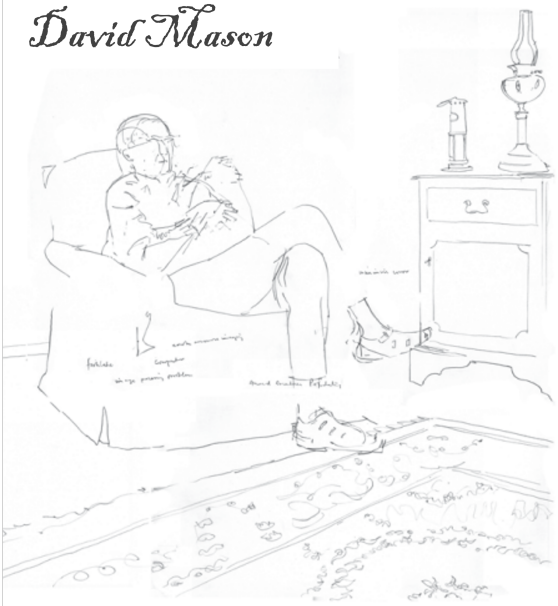
The artist Hugo Dalton has made a series of watercolour paintings of the research locations and the scientists at work. These have been merged with existing reproductions of historic artworks from the Royal Collection at nearby Windsor, which also show views of the surrounding area. The idea is to show how different kinds of information are used to create a better understanding of flooding (Dance et al., 2022). The objective is to offer elderly people in care homes a connection to current science research in their local area. The project provides them with a link to places they may no longer be able to physically access.

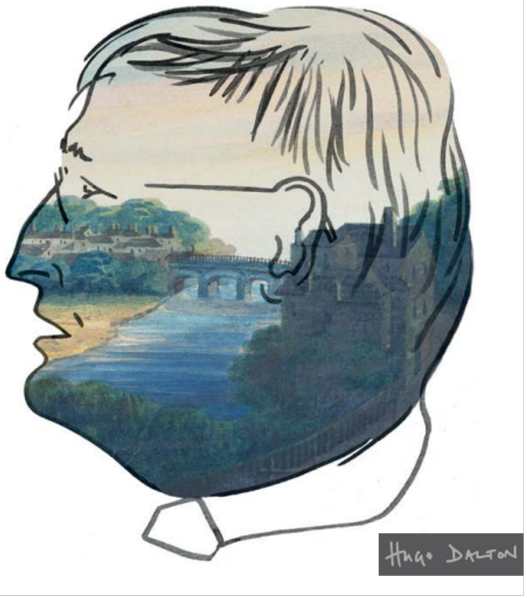
A selection of the artwork is included in this thesis as chapter illustrations with permission granted by Hugo Dalton.

Sarah Dance



David Mason





Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Globally, an estimated 1.8 billion people (23% of the world's population) live in areas that are directly exposed to a 100-year return period flood (Rentschler et al., 2022). The vast majority (89%) of these people live in low- and middle-income countries where infrastructure systems, including flood protection and drainage, and early warning systems, tend to be less developed. Satellite data shows that the number of people exposed to flooding has increased by 20 to 24% globally from 2000 to 2018. Increasing exposure to flooding is predicted to continue with climate change (Tellman et al., 2021). Early warning systems can significantly improve the outcomes following disasters, reducing deaths and damage and enabling faster recovery (UNDRR, 2022a). The 2023 UN Global Assessment Report (GAR) on Disaster Risk Reduction (UNDRR, 2023) states that the benefits of early warning systems triple in vulnerable contexts. Despite these benefits, the GAR2022 report shows that just 5.8% ($5.5 billion USD) of official development assistance contributes to disaster prevention and preparedness compared to 90.1% ($119.8 billion USD) for emergency response. Yet it has been demonstrated (for Europe) that financing for mitigation purposes such as flood forecasting systems can lead to overall cost savings (Pappenberger

et al., 2015).

Global-scale flood forecasting systems can support the early warnings for all initiative (UNDRR, 2022a) by providing flood forecasts for large rivers around the world. Advances in flood forecasting systems link together meteorological and hydrological forecasts to hydrodynamic models, simulating flood-wave propagation (Emerton et al., 2016; Wu et al., 2020; Apel et al., 2022). A simulation library forecasting system saves computation time by storing static flood extent and depth maps at various return periods. Depending on the forecast river discharge, the maps are looked-up per sub-catchment and mosaicked together. The resulting flood maps can be used to inform disaster risk reduction schemes such as forecast-based financing (FbF). FbF works by quantifying risks in advance of crises or disasters, prepositioning funds, and agreeing in advance how funds will be released based on forecasts, ahead of an event (OCHA, 2020). This results in global-scale models being used to inform local-scale action. Hoch and Trigg (2019) outline a Global Flood Model Validation Framework, which includes a recommendation to routinely validate flood extent. Quantitative performance evaluation forms an important part of fitness-for-purpose assessment and continual system improvement. Currently, there is limited quantitative validation of operational flood forecasting systems producing flood maps.

The accuracy of forecasts of flood extent can be verified by comparing with observations of flooding from drones or satellite-based sensors. Typically, binary performance measures are calculated and provide an average measure of skill across a region (Stephens et al., 2014). In this thesis we address several limitations of binary performance measures by applying a new scale-selective approach to flood map verification. The approach is developed further to evaluate how well an ensemble flood map forecast represents the uncertainties involved. We apply scale-selective verification methods to a multi-system comparison where each of the systems' forecasts are presented at different spatial scales.

Some limitations to the flood forecasting system are addressed by developing a data assimilation framework using satellite data to improve the flood map analysis.

1.2 Thesis aims

The aims of this thesis are to address the following research questions:

1. **What are the skilful spatial scales in flood inundation forecasts made using a simulation library approach?**
   How can we determine the skilful spatial scales of forecast flood maps by comparing against satellite-derived observations of flooding? Does the skilful spatial scale vary with location and how can this be visualised? How does validation of the flood edge location alone compare to validation of the entire flood extent? How can the skilful spatial scale results be used in operational practice?

2. **How skilfully does an ensemble of forecast flood maps represent the spatial uncertainty within the flood forecast?**
   How can we summarise the spatial predictability information in ensemble flood map forecasts? How can we evaluate the spatial spread-skill of an ensemble flood map forecast? How does the spatial spread-skill vary with location and how can this be presented?

3. **How useful are scale-selective evaluation approaches when applied to multiple flood forecasting systems?**
   How can we evaluate the performance of flood forecasting systems predicting flood inundation extent at different spatial scales? What can we learn about the flood forecasting system performance and how does each compare?

4. **Does a data assimilation framework improve the analysis of flood inundation from a simulation library system?**
   Can we incorporate probabilistic information from remotely observed flood inundation into a data assimilation framework to improve the flood map selection within a simulation library flood forecasting system? How does the analysis flood map compare to independent validation data?

1.3 Principal new results

The outcomes of this thesis provide the following answers to the research questions:

1. A skilful spatial scale for forecast flood maps can be found by calculating the Fraction Skill Score, a validation metric, found by comparing a deterministic forecast flood map against a satellite SAR-derived observation of flooding across a range of neighbourhood sizes. A target skill score can be calculated and this depends on the magnitude of the observed flood. The skilful scale determined for the flood edge is more sensitive to changes in spatial accuracy and spatial scale compared to the skilful scale found by evaluating the entire flood extent. Categorical scale maps developed show that the skilful scale varies with location across a domain.

2. We present a new scale-selective approach to assess the spatial predictability and spread-skill of an ensemble flood map forecast that accounts for the individual spatial prediction of flood extent held within each ensemble member flood map. The method determines, at specific locations within the domain, whether the ensemble forecast is over-, under- or well-spread. The spatial spread-skill relationship can be mapped onto a Spatial spread-skill map.

3. We investigate a new application of scale-selective verification by evaluating the performance of three flood forecasting systems. Two simulation library systems, Flood

Foresight (30 m) and GloFAS Rapid Flood Mapping (1000 m) and one hydrodynamically modelled system, the Bangladesh FFWC Super Model (300 m), all made predictions of flood extent at different spatial scales (grid lengths, shown in brackets) for the Jamuna River flood, Bangladesh, July 2020. Our results show that the simulation library system accuracy critically depends on the discharge return period threshold set to trigger a flood map selection and the number of hydrological model ensemble members that must exceed it.

4. A data assimilation (DA) framework is developed to integrate probabilistic flood extent maps from satellite-based SAR sensors into the simulation library flood map selection process. The method is tested on the severe flood event in Pakistan, 2022, where several sub-catchments resulted in a non-trigger of the forecast-based financing system deployed here, despite significant flooding evident from earth observation data. The DA successfully triggered flood maps in 4 out of 5 sub-catchments tested and we found that evaluating sub-catchments at the flood edge gave the best results.

1.4 Thesis outline

This thesis is structured as follows:

- Chapter 2 introduces relevant background information for the thesis including simulation library flood inundation forecasting systems, observing flooding from satellites, flood extent verification and data assimilation approaches in flood inundation forecasting.

- Chapter 3 addresses the first research question in Section 1.2. A new approach to forecast flood map spatial verification against satellite-derived observations of flooding is presented. A scale-selective verification method is applied to evaluate the performance of a simulation library flood forecasting system at predicting flood ex-

tent on the Rivers Wye and Lugg following storm Dennis in February 2020. The scale-selective approach addresses multiple limitations of conventional binary performance measures and we find several applications of the evaluation approach that benefit operational flood forecasting practice. Chapter 3 is reproduced from Hooker et al. (2022).

- Chapter 4 addresses the second research question. We present a new approach to evaluate and visualise the spatial spread-skill of an ensemble flood map forecast. The method can be used to assess how well the probabilistic flood maps represent the uncertainty present within the forecast-chain and how this may change with updates to the forecasting system such as including additional observations. Chapter 4 is reproduced from Hooker et al. (2023a).

- Chapter 5 addresses the third research question. Through application of scale-selective evaluation methods we can directly compare forecast flood maps from three flood forecasting systems, each predicting flood extent at different spatial scales (grid lengths). This quantitative spatial validation means that the benefits and limitations of the forecast systems can be evaluated. Chapter 5 is reproduced from Hooker et al. (2023b).

- Chapter 6 addresses the final research question. We present a new data assimilation framework to incorporate satellite-derived probabilistic flood extent information into the simulation library flood map selection process. This overcomes limitations of the simulation library system and has the potential to improve forecast-based financing schemes.

- Chapter 7 summarises the main findings from the thesis and makes recommendations for future work.

Chapter 2

Background

In this chapter we introduce some of the key topics used in this thesis. In Section 2.1 we introduce simulation library flood inundation forecasting systems and their application in disaster risk reduction. Observing flooding from satellite data is explained in Section 2.2 and using these observations to verify flood forecasts is discussed in Section 2.3. Data assimilation approaches used in flood inundation forecasting are introduced in Section 2.4.

2.1 Fluvial simulation library flood inundation forecasting systems

The current state-of-the-art in operational flood inundation forecasting at national or transnational scales uses a simulation library system (Revilla-Romero et al., 2017). Flood extent and depth maps are precomputed using a hydrodynamic model at a range of return period thresholds. Together, these flood maps form a simulation library for a particular country or river basin.

The system links together a chain of models that begins with a numerical weather prediction (NWP) model providing meteorological inputs such as forecast precipitation to

a hydrological model and subsequent hydrodynamic model. The NWP model, combined with recent observations also provides initial conditions for the hydrological model. The forecast river discharge from the hydrological model, along with the return period thresholds set (from historical observations or reanalysis data-sets of river discharge (Grimaldi, 2022)) determines which flood map from the simulation library is selected and presented (for an example diagram see Figure 4.3).

Pre-computing the flood maps reduces the model run-time and means that the flood forecasting system can operate in near real-time. Additionally, the system can handle multiple inputs from ensemble NWP forecasts (Cloke & Pappenberger, 2009; Emerton et al., 2016). An ensemble of meteorological inputs means that some of the uncertainty in the NWP forecast can be accounted for in the hydrological model and ultimately in the forecast flood maps. The forecast flood maps can be presented probabilistically as an ensemble flood map forecast indicating the probability of flooding at a specific location within a catchment. An ensemble NWP forecast lengthens the forecast lead-time where the forecast is deemed skilful compared to a deterministic forecast (Emerton et al., 2016). Information on flood inundation uncertainty is particularly useful for disaster management teams operating before a flood event occurs and can be directly linked to flood impacts such as maps of vulnerable infrastructure. Probabilistic flood maps can also be used to inform disaster risk reduction schemes such as Forecast-based Financing (FbF). FbF schemes work by quantifying risks in advance of disasters, prepositioning funds, and agreeing in advance how funds will be released based on forecasts, ahead of an event (OCHA, 2020). FbF allows time for local action utilising the insurance funds for flood mitigation purposes.

In Chapters 3 to 6 we use forecast flood maps from JBA's Flood Foresight system. Flood Foresight is a simulation library flood inundation forecasting system deployed in several international countries for FbF applications.

2.2 Observing flooding from Space

Flooding events are usually observed using in situ ground-based gauging stations recording river discharge or water level. The recorded flood discharge does not linearly correlate with the observed inundation extent and is highly uncertain due to instrument error and rating curve extrapolation uncertainties (Beven, 2016). Globally, there is limited coverage of catchments with maintained gauging stations and the data is not always openly available. Satellite-derived observations of flood extent have the potential to bring additional spatial information into flood inundation forecasts compared to in situ point gauging stations. Satellite-based Synthetic Aperture Radar (SAR) sensors are well known for their flood detection capability (Grimaldi et al., 2016). Unobstructed flood waters appear dark on SAR images due to the low backscatter return from the relatively smooth water surface. SAR sensors also have an advantage over optical instruments as they can scan at night and are not impacted by cloud and weather, usually associated with a flooding situation. Optical instruments rely on solar energy and cannot penetrate cloud, making them less useful during a flooding situation. Recent studies have investigated the flood detection benefits from combining both optical and SAR imagery (Konapala et al., 2021; Tavus et al., 2020).

Due to improvements in spatial resolution and more frequent revisit times, SAR data have been used successfully to calibrate and validate hydrodynamic and hydraulic forecast models (Schumann et al., 2009; Grimaldi et al., 2016). Dasgupta et al. (2018) detail some of the challenges along with approaches to solutions of flood detection using SAR. Examples of these challenges include: roughening of the water surface by heavy rain and strong wind, emergent or partially submerged vegetation and flood detection in urban areas. Accurate flood detection in urban areas particularly due to surface water flooding has become increasingly important (Speight et al., 2021) and recent techniques have led

to improved flood detection (Mason et al., 2018, 2021a, 2021b).

The Copernicus Emergency Management Service (CEMS) (Copernicus Programme, 2021) offers freely available, open access Sentinel-1 SAR data. Currently (due to the malfunction of Sentinel-1B in December, 2021) one satellite is in orbit, at 10 m ground resolution and a six day revisit time (for the mid-latitudes). Sentinel-1C is due to launch in April 2024 to replace Sentinel-1B. Nevertheless, Sentinel-1 data offers good coverage of a potential flood event. For a major flood event CEMS can be triggered to offer additional rapid flood mapping. Since late 2021, SAR-derived flood maps are produced for every Sentinel-1 image detecting flooding around the world by the Global Flood Monitoring (GFM) service (EU Science Hub, 2021; GFM, 2021; Hostache et al., 2021). Within eight hours of the Sentinel-1 image acquisition, three flood detection algorithms are combined to give the flood class (flooded or unflooded), the likelihood of a flood class representing the uncertainty estimation per grid cell along with an exclusion mask where flooding cannot be reliably detected from SAR. In Chapters 3, 4 and 5 we make use of SAR-derived flood maps from Sentinel-1 satellite data for deterministic and ensemble forecast flood map verification. In Chapter 6 we assimilate the GFM flood likelihood data to improve the flood extent analysis from a simulation library flood forecasting system.

2.3 Flood extent verification

Verification is an essential part of model understanding and improvement, but has only received limited attention over the past decade (Schumann, 2019). Forecast flood maps can be verified against satellite-derived observations of flood extent. Validation of forecast flood maps against remotely observed flood extent is typically carried out by labelling each grid cell using a contingency table with categories: correctly predicted flooded, under-prediction (miss), over-prediction (false alarm) and correctly predicted unflooded. Follow-

ing this categorisation, a variety of conventional binary performance measures such as the Critical Success Index (CSI) can be calculated (see Chapter 3). Within a domain of interest, a flood covering a significant area of the domain will be easier to accurately predict (by chance) compared to a flood of a smaller extent. Thus, flood magnitude can create a bias in the skill scores. It has been suggested by Stephens et al. (2014); Pappenberger et al. (2007) that it is less important to validate all flooded cells, when only cells that are close to the flood margin are difficult to predict. The flood edge location is an important consideration for flood risk mitigation and response activities.

Flood maps at different spatial scales (grid lengths) will also impact conventional skill scores. A high resolution, fine scale forecast flood map will show greater detail of the flood extent and the flood-edge location compared to a low resolution, coarse scale flood map. At a high resolution the discrepancy between the forecast and observed flood maps may be closer in terms of distance, however a small mismatch will lead to a double penalty impact on forecast verification. The model is penalised twice for the over-prediction (false alarm) and the under-prediction (miss) (Stein & Stoop, 2019). When high resolution forecasts are verified against observations at grid level, the predictability can appear to worsen and the high resolution forecast would need to perform better than the low resolution forecast to achieve the same verification score. Hence, it is not meaningful to compare verification scores across different spatial scales. Conventional binary performance measures (reviewed in Chapter 3) give a single, domain averaged, skill score. The averaged score is less sensitive to changes in accuracy and does not indicate where location specific improvements could be made. The Fraction Skill Score (FSS; see Chapter 3 for details) uses a neighbourhood approach to overcome the double penalty impact problem in convective precipitation verification (Roberts & Lean, 2008). In Chapter 3 we apply the FSS scale-selective verification approach to evaluate the whole flood, and the flood edge of forecast flood maps and in Chapter 5 we use the same approach to compare three flood forecast-

ing systems, each producing forecast flood maps at different spatial scales. In Chapter 6 we use scale-selective verification methods to compare the analysis flood map, following DA, against independent observations of flood extent from Sentinel-2 optical satellite data.

Ensemble flood maps require an extra dimension of verification, a measure of *spread* as well as *skill*. A perfect ensemble should encompass forecast uncertainties such that the ensemble spread is correlated to the RMSE of the forecast (Hopson, 2014). The verification of ensemble forecasts usually involves comparing the RMSE of the ensemble mean against an observed quantity to assess the skill of the forecast with the ensemble standard deviation used as a measure of spread. To evaluate the accuracy of an ensemble forecast, a number of verification measures have been proposed. Anderson et al. (2019) developed a joint verification framework for end-to-end assessment of the England and Wales Flood Forecasting Centre (FFC) ensemble flood forecasting system. Anderson et al. (2019) describe verification metrics such as the continuous rank probability score (CRPS), rank histograms, Brier Skill Score (BSS) and the relative operative characteristics (ROC) diagrams that are commonly applied to assess the main ensemble attributes desirable in both precipitation and streamflow ensemble forecasts (e.g., Renner et al., 2009). These metrics refer to flooding events as part of a time series evaluated against a reference benchmark, such as climatology, to produce an average skill score. In contrast, in Chapter 4 we consider ensemble *spatial* verification at a single time point. The spatial spread-skill of the ensemble forecast is determined by evaluating the full ensemble against remote observations of flooding. For a flood map ensemble to be considered spatially well-spread, the spread or variation between ensemble members should equal the spatial predictability, or skill of the ensemble members (Dey et al., 2014).

2.4 Introduction to data assimilation

Data assimilation (DA) typically finds an optimal *state* (such as river water level) and/or model parameter values (such as floodplain roughness coefficients) of a dynamical system, taking account of the previous forecast, the observations available, and both of their associated uncertainties. The updated state, the *analysis*, and/or parameter values are used to initiate the next forecast in a feedback loop or cycle (Figure 2.1).

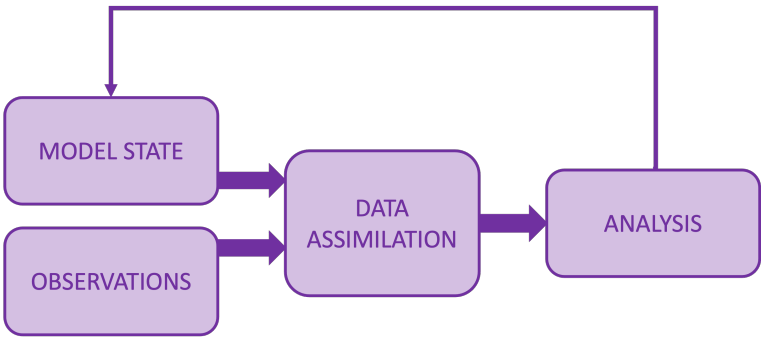


Figure 2.1: Data assimilation cycle

Bayesian estimation forms the basis of most data assimilation techniques and this assumes that errors in the forecast and observations can be represented by a Gaussian distribution. The uncertainties associated with the *prior* or *background*, usually the previous forecast, are represented by the covariance of the prior error distribution. Similarly, the uncertainties in the observations depend on the covariance of the observation *likelihood*. The aim of DA is to find the optimal state that maximises the *posterior* probability and thus minimises the variance of the posterior error distribution (Bouttier & Courtier, 2002). The optimal state is found by minimising a cost function (derived from the probability distributions), or in other words finding the state or parameter variables where the gradient of the cost function is equal to zero.

Variational DA methods use numerical minimisation methods such as iterative gradient descent to minimise the cost function. Three-dimensional variational DA (3D-Var) methods assimilate observations at a fixed point in time (Lorenc et al., 2000), whereas 4D-Var methods assimilate observations over a window of time (Bannister, 2017). Filtering methods such as particle filtering (PF) and ensemble Kalman filter (EnKF) methods sample (using Monte Carlo methods) and weight the prior distribution according to the observation likelihood (van Leeuwen, 2009; Evensen, 1994).

Previously, SAR data have been used in several different ways to improve hydraulic models and flood prediction through data assimilation (DA). A review of approaches used to assimilate satellite-derived data into hydraulic models (from 2007 until 2015) can be found in Table 7 of Grimaldi et al. (2016) and Table 1 of Revilla-Romero et al. (2016). In Chapter 6 we review more recent DA approaches used to improve flood inundation forecasts. A new DA framework is presented to improve the flood inundation analysis from a simulation library forecasting system. The analysis is evaluated using scale selective verification methods described in Chapter 3.

2.5 Chapter summary

In this chapter we have introduced and discussed a number of topics relevant to the thesis. In Section 2.1 we described a simulation library flood forecasting system and its application to disaster risk reduction. Forecast data from a simulation library system is used in all of the case studies in this thesis. Satellite-derived observations of flooding outlined in Section 2.2 are used to compare against flood forecasts. The spatial verification methods introduced in Section 2.3 are calculated alongside new scale-selective methods in Chapters 3, 4, and 5. Data assimilation methods are introduced as background for Chapter 6 in Section 2.4. In the next chapter we develop a scale-selective verification approach to

evaluate forecast flood extent maps.

Chapter 3

Spatial scale evaluation of forecast flood inundation maps

In this chapter we address the first research question outlined in Chapter 1; What are the skilful spatial scales in flood inundation forecasts made using a simulation library approach? In particular we wish to find out:

- How can we determine the skilful spatial scales of forecast flood maps by comparing against satellite-derived observations of flooding?
- Does the skilful spatial scale vary with location and how can this be visualised?
- How does validation of the flood edge location alone compare to validation of the entire flood extent?
- How can the skilful spatial scale results be used in operational practice?

The remainder of this chapter (except for the chapter summary, Section 3.8), has been published and is reproduced from (Hooker et al., 2022).

3.1 Abstract

Flood inundation forecast maps provide an essential tool to disaster management teams for planning and preparation ahead of a flood event in order to mitigate the impacts of flooding on the community. Evaluating the accuracy of forecast flood maps is essential for model development and improving future flood predictions. Conventional, quantitative binary verification measures typically provide a domain-averaged score, at grid level, of forecast skill. This score is dependent on the magnitude of the flood and the spatial scale of the flood map. Binary scores have limited physical meaning and do not indicate location-specific variations in forecast skill that enable targeted model improvements to be made. A new, scale-selective approach is presented here to evaluate forecast flood inundation maps against remotely observed flood extents. A neighbourhood approach based on the Fraction Skill Score is applied to assess the spatial scale at which the forecast becomes skilful at capturing the observed flood. This skilful scale varies with location and when combined with a contingency map creates a novel categorical scale map, a valuable visual tool for model evaluation and development. The impact of model improvements on forecast flood map accuracy skill scores are often masked by large areas of correctly

predicted flooded/unflooded cells. To address this, the accuracy of the flood-edge location is evaluated. The flood-edge location accuracy proves to be more sensitive to variations in forecast skill and spatial scale compared to the accuracy of the entire flood extent. Additionally, the resulting skilful scale of the flood-edge provides a physically meaningful verification measure of the forecast flood-edge discrepancy. The methods are illustrated by application to a case study flood event (with an estimated return period of 120 to 550 years) of the River Wye and River Lugg (UK) in February 2020.

Representation errors are introduced where remote sensing observations capture flood extent at different spatial resolutions in comparison with the model. The sensitivity of the verified skilful scale to the resolution of the observations is investigated. Re-scaling and interpolating observations leads to a small reduction in skill score compared with the observation flood map derived at the model resolution. The domain-averaged skilful scale remains the same with slight location-specific variations in skilful scale evident on the categorical scale map. Overall, our novel emphasis on scale, rather than domain-average score, means that comparisons can be made across different flooding scenarios and forecast systems and between forecasts at different spatial scales.

Highlights

- A novel spatial scale-selective approach to evaluate forecast flood maps against Synthetic Aperture Radar data.
- Validation of the flood edge gives a physically meaningful measure of prediction accuracy.
- Conventional contingency flood maps are improved by including a location-specific skilful spatial scale.

3.2 Introduction

Timely predictions of flood extent and depth from flood forecasting systems provide essential information to flood risk managers that enable anticipatory action prior to the occurrence of a potential flooding event. Evaluating the accuracy of flood extent forecasts against observations forms an essential part of model development (Schumann, 2019). Forecast flood inundation footprints are typically validated against remote sensing images using binary performance measures (Stephens et al., 2014) calculated at grid level.

In order to produce a forecast flood map, hydrodynamic or hydraulic flood models in two-dimensions simulate the flow of water using a local digital terrain model (DTM). The spatial resolution of DTMs has increased over recent years and is important for accurate flood mapping. For example, in the UK, the Environment Agency National LIDAR Programme offers open source 1 m surface elevation data for the whole of England (Environment Agency, 2021). Additional surface detail to 0.3 m spatial resolution from unmanned aerial vehicle UAV-LIDAR data acquired in urban areas is now possible (Trepekli et al., 2021). This means forecast flood maps could be presented at this very high resolution. It is questionable how meaningful it is to present highly detailed flood maps as a deterministic forecast (Savage et al., 2016), particularly at longer lead times where the skill of the flood forecasting system becomes increasingly dependent on the accuracy of the meteorological forecast (ECMWF, 2022). Speight et al. (2021) note for surface water flooding that more detail is included in local scale flood maps than can be justified by the predictability of the forecast. A high resolution, fine scale forecast flood map will show greater detail of the flood extent and the flood-edge location compared to a low resolution, coarse scale flood map. At a high resolution the discrepancy between the forecast and observed flood maps may be closer in terms of distance, however a small mismatch will lead to a double penalty impact on forecast verification. The model is pe-

nalised twice for the over-prediction (false alarm) and the under-prediction (miss) (Stein & Stoop, 2019). When high resolution forecasts are verified against observations at grid level, the predictability can appear to worsen and the high resolution forecast would need to perform better than the low resolution forecast to achieve the same verification score. It is not meaningful to compare verification scores across different spatial scales. Spatial verification methods for flood inundation mapping have only received limited attention over the past decade (Schumann, 2019).

Verification approaches that account for uncertainties in observations and small discrepancies in gridded data using a fuzzy set approach (Hagen, 2003) have previously been applied to flood mapping (Pappenberger et al., 2007; Dasgupta et al., 2018). However, the fuzzy set method does not incorporate variations in spatial scale (Cloke & Pappenberger, 2008). In atmospheric sciences, verification approaches that account for changes in spatial scale are well established. These approaches include the Fraction Skill Score (FSS), which applies a neighbourhood approach to assess a useful/skilful scale (Roberts & Lean, 2008) of a precipitation forecast. Dey et al. (2014); Dey, Roberts, et al. (2016) developed the FSS approach to produce location-specific agreement scales between the forecast and observed fields to understand the spatial predictability of an ensemble forecast. Other spatial scale approaches include the wavelet method of scale decomposition, where the forecast and observed fields are decomposed into maps at different scales by wavelet transformation and subsequently verified (Briggs & Levine, 1997; Casati & Wilson, 2007). Cloke and Pappenberger (2008) note that this method is extremely sensitive to offsetting of maps.

In general, the performance of forecast flood maps are evaluated for the entire flood extent, regardless of flood magnitude, adding bias to binary performance measures (Stephens et al., 2014). Stephens et al. (2014) question whether it is important to validate all flooded cells, when only cells that are close to the flood margin are difficult to predict.

Pappenberger et al. (2007) evaluated model performance only on cells that were subject to change between differing model runs to address the issue of large areas of correctly predicted flooded/unflooded cells masking variations in forecast skill scores.

Satellite based Synthetic Aperture Radar (SAR) sensors are well known for their flood detection capability. Unobstructed flood waters appear dark on SAR images due to the low backscatter return from the relatively smooth water surface. SAR sensors also have an advantage over optical instruments as they can scan at night and are not impacted by cloud and weather, usually associated with a flooding situation. Due to improvements in spatial resolution and more frequent revisit times, SAR data has been used successfully to calibrate and validate hydrodynamic and hydraulic forecast models (Schumann et al., 2009; Grimaldi et al., 2016). Further model improvements have been shown through the assimilation of SAR data (e.g., García-Pintado et al., 2015; Hostache et al., 2018; Cooper et al., 2019; Di Mauro et al., 2020; Dasgupta et al., 2018, 2021a, 2021b). Recent techniques have improved the flood detection in urban areas using medium and high resolution SAR (Mason et al., 2018, 2021a, 2021b). The Copernicus Emergency Management Service (CEMS) (Copernicus Programme, 2021) offers freely available, open access Sentinel-1 SAR data. Currently (due to the malfunction of Sentinel-1B in December, 2021) one satellite is in orbit, at 10 m ground resolution and a six day revisit time (for the mid-latitudes). Nevertheless, Sentinel-1 data offers good coverage of a potential flood event. For a major flood event CEMS can be triggered to offer additional rapid flood mapping. From 2022, the new Global Flood Monitoring (GFM) product (GFM, 2021; Hostache et al., 2021) of the Copernicus Emergency Management Service (CEMS) (Copernicus Programme, 2021) produces Sentinel-1 SAR-derived flood inundation maps using three flood detection algorithms providing uncertainty and population affected estimates within 8 hours of the image acquisition.

Representation errors arise where observation spatial scales are different from the model spatial scale (Janjić et al., 2018). The spatial resolution of SAR imagery suitable for flood detection varies across satellite constellations both historically and presently and continues to improve. Very high resolution (less than 3 m) imaging capabilities are increasingly available including TerraSAR-X, ALOS-2/PALSAR-2, and the COSMO-SkyMed, RADARSAT-2, and ICEYE constellations (Mason et al., 2021a). It is common practice to re-scale SAR-derived flood maps to match the model grid size for validation or assimilation with model data.

The objective of this paper is to present a scale-selective approach to evaluate flood inundation forecast maps and to develop a physically meaningful measure of flood-edge location accuracy that can be automated and easily applied in practice. The method has been developed with operational forecast verification in mind, but it is applicable to all flood inundation maps. A new approach is described and applied here to evaluate the spatial scale at which the forecast becomes useful/skilful at capturing the remotely observed flood extent and specifically the flood-edge location. The spatial skill of a forecast flood map varies with location. We aim to improve the conventional contingency map by incorporating the skilful scale to create a new *categorical scale map*. Also, we address how representation errors arising from observation spatial scale variations and interpolation have an impact on model evaluation.

In the rest of this paper we explore the features of a novel scale-selective evaluation approach illustrated through application to a case study. In Section 3.3 we describe the case study, a recent flooding event in the UK following Storm Dennis, February 2020, along with catchment descriptions for three chosen domains. The flood inundation forecasting system developed by JBA Consulting, Flood Foresight, (Revilla-Romero et al., 2017) is used to produce forecast flood maps for the event and is detailed in Section 3.4.1. Section

3.4.2 explains two methods that are used to derive remotely observed flood maps from SAR imagery. Our new approach to the spatial evaluation of flood maps is detailed in Section 3.5 along with descriptions of other binary performance measures. The novel categorical scale map is applied to the case studies in Section 3.6, and the evaluation results are discussed. We conclude in Section 3.7 and discuss the wider applications of a spatial scale approach to flood map skill evaluation.

3.3 Flood event

This extreme flooding event is chosen here as a case study to demonstrate the features of a spatial scale approach to forecast flood map evaluation. During February 2020, three named Storms, Ciara, Dennis and Jorge, arrived in quick succession delivered by a powerful and ideally positioned jet-stream that enabled rapid cyclogenesis (Davies et al., 2021). Each storm rapidly intensified and deepened bringing damaging winds and exceptionally heavy rainfall across the UK (Met Office, 2020). This led to the River Wye reaching its highest ever recorded water level at the Old Bridge in Hereford (riverlevels.uk, 2020). The annual exceedance probability (AEP) for the recorded peak flow of the Lugg and Wye rivers was 0.2 - 0.8 % (return period 120-550 years) and 0.6 - 2.0 % (160-550 years) respectively (Sefton et al., 2021).

3.3.1 February 2020

February 2020 was the UK's wettest February on record and the fifth wettest month ever recorded. The UK average rainfall total exceeded the 1981 – 2010 average by 237% (Kendon, 2020). Locally, in northwest England and north Wales the rainfall exceedance was three to four times the typical monthly average rainfall. During this period around 4000 to 5000 properties were flooded in the UK, with significant river water levels recorded in Wales, west and northwest England (Sefton et al., 2021). With six days between Ciara

and Dennis, groundwater and river levels were high and soils saturated. The Environment Agency issued a record number of over 600 flood alerts and warnings for England (JBA, 2021).

3.3.2 Catchment location and description

Three domains, each differing in hydrological characteristics, have been selected for forecast flood map evaluation during the storm Dennis flooding event. Two domains (A and B) have been chosen from the Wye catchment (Fig. 3.1), a 28.4 km length centred upon Ross-on-Wye (A) and the Wye at Hereford (B), a 5.8 km section. A third domain (C) includes 4 km of the River Lugg.

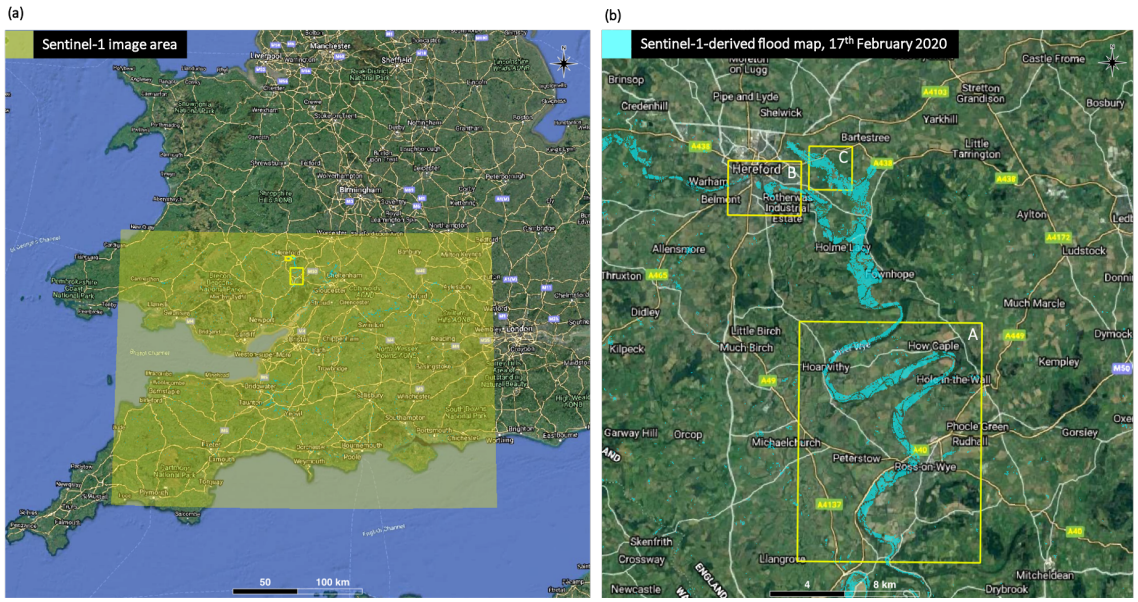


Figure 3.1: Location of Sentinel-1 image acquisition over southeast UK (a) and flood map evaluation domains (b). Domain A: 28.4 km length of the River Wye centred at Ross-on-Wye, domain size 9.8 x 12.8 km. Domain B: 5.8 km of the River Wye at Hereford, domain size 3.0 x 4.0 km. Domain C: 4 km of the River Lugg at Lugwardine, domain size 2.3 x 2.3 km. Base map from Google Maps.

3.3.2.1 The River Wye (domains A and B)

The River Wye flows for approximately 215 km from Plynlimon at 750 meters above ordnance datum (mAOD) in the Cambrian Mountains, mid Wales. It initially travels southeastwards into England where it meanders southwards to ultimately join the Severn Estuary. The upper catchment land cover is predominantly grassland with some forest cover with highly impermeable bedrock and superficial deposits of sand and gravel in the Hereford area (National River Flow Archive, 2021). The upstream catchment area of Hereford is 1896 km^2. At Hereford, the only city situated on the Wye, the river is embanked on the north side by a deep flood wall with further embankments on the opposite side. Hereford is characterised by the Old Bridge, a 15^{th} century stone bridge that creates a damming effect during high river flows. As the Wye flows south of Hereford, the topography flattens and the floodplain widens, with large river meanders and a distinctive U-shaped valley.

3.3.2.2 River Lugg at Lugwardine (domain C)

The River Lugg has an upstream catchment area of 886 km^2 and a maximum altitude of 660 mAOD and flows across the grasslands and agricultural fields of the Herefordshire plain. It has similar bedrock to the Wye catchment and a higher proportion of more permeable superficial fluvial deposits of sand and gravel. This is particularly evident in the Lugwardine region where the topography is relatively flat with little to impede the flow of floodwaters across the plain. The Lugg flows into the River Wye, 2 km south of domain C.

3.3.2.3 Event hydrology

The observed catchment rainfall (which also includes a downstream section of the River Wye) shows that 50 mm fell on the 15^{th}, 10 mm on the 16^{th} and 1 mm on the 17^{th} February

2020 (UK Water Resources Portal, 2022). There were further heavy showers forecast for the 16/17th and whilst these have not been captured by the rain gauges on the 17th, they cannot be ruled out as contributing to surface water flooding in Hereford. The nearest hourly rainfall-rate observation is a citizen science observation from the Met Office WOW database (Met Office, 2022) for a site at Sutton St Nicholas near the River Lugg and this shows the highest rainfall rate of 5.8 mm/hr at 0300 on the 16th and a total accumulation of 12.5 mm on the 16th and 0.3 mm on the 17th.

Daily maximum river levels recorded at Ross-on-Wye, the Old Bridge, Hereford and Lugwardine for January to March 2020 are plotted in Figure 3.2 (riverlevels.uk, 2020). The impact of the three storms on the River Wye is indicated by a very sharp rise in water levels from the 8th to the 10th February following storm Ciara. Further heavy showers maintained high water levels before storm Dennis brought an exceptional rise in water levels, peaking on the morning of the 17th February with record levels recorded at Hereford (6.11 m at 9.30 am UCT) and Ross-on-Wye (4.77 m at 5.45 am UTC). Unfortunately there are two days of missing data at Ross-on-Wye following the flood event. By analysing the trend between the Hereford and Ross-on-Wye river levels, the peak level at Ross-on-Wye was likely higher and later than recorded. The response of the Wye at Hereford is faster than at Ross-on-Wye, most likely due to the upstream location of Hereford and a more constrained embankment with the city center located either side of the river. In comparison to the fast, rapid response of the Wye, the River Lugg displays a distinctively dampened response. Whilst the Lugg initially responded quickly to the heavy rainfall, once bankfull was reached and overtopping occurred the water levels remained consistently high, with floodwaters extending across the relatively flat flood plain.

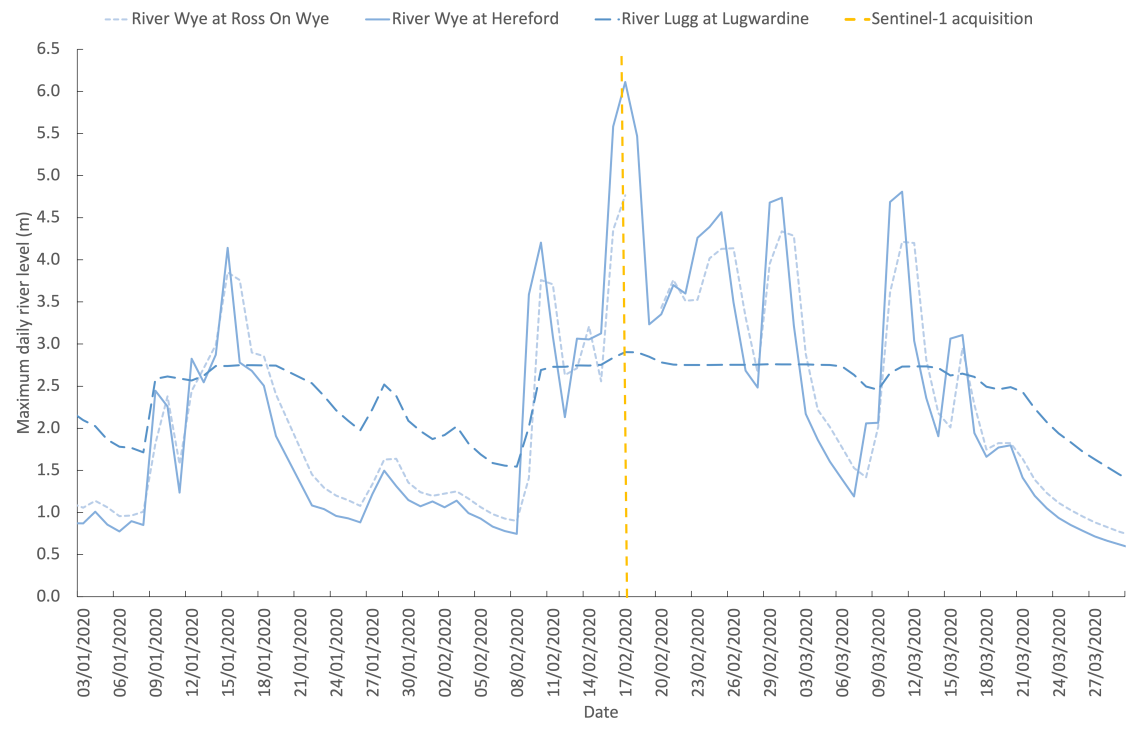


Figure 3.2: Daily maximum river levels (m) at Ross-on-Wye, Hereford and Lugwardine. The dashed yellow line indicates Sentinel-1 SAR acquisition date.

3.4 Data

In this section we describe the model and observation data that we will use to illustrate our novel scale selective verification approach.

3.4.1 Flood Foresight

Flood Foresight (Fig. 3.3), developed and run routinely by JBA Consulting, is a fluvial flood inundation mapping system that can be implemented in any catchment around the globe. Flood Foresight utilises a simulation library approach to generate maps of real time and forecast flood inundation and water depth. The simulation library approach saves valuable computing time and allows the application of Flood Foresight in near continuous real-time at national and international scales. A library of flood maps is pre-computed

using JFlow®, a 2D hydrodynamic model (Bradbrook, 2006). Note that in this study the flood maps are undefended i.e. temporary flood defences are not included. JFlow uses a raster-based approach with a detailed underlying DTM and a simplified form of the full 2D hydrodynamic equations that capture the main controls of the flood routing for shallow, topographically driven flow. Five flood maps at 5 m resolution are created for 20, 75, 100, 200 and 1000 year return period flood events (corresponding to annual exceedance probabilities (AEPs) of 5%, 1.3%, 1%, 0.5% and 0.1% respectively). These are interpolated to derive five intermediate maps between each adjacent pair of the JFlow maps, equally spaced in return period creating a total library of thirty flood maps. Flood Foresight takes inputs of rainfall from numerical weather prediction (NWP) models, river gauge data (both historical and real-time) and forecast streamflow and uses these to select the most appropriate flood map for the location and forecast time period. The UK and Ireland configurations of the Flood Forecasting Module use deterministic streamflow forecast data from the Swedish Meteorological and Hydrological Institute (SMHI) European HYdrological Predictions for the Environment (E-HYPE). The meteorological input data for the E-HYPE model is the European Centre for Medium-range Weather Forecasts (ECMWF) Atmospheric Model high resolution (HRES) numerical weather prediction (NWP) model on a 0.1° x 0.1° grid with forecasts issued daily out to 10 days lead time. Forecast flood maps for the UK are produced on a 25 m grid length out to 10 days ahead (see Mason et al. (2021b) Section 2.1 for additional details).

3.4.2 SAR-derived flood maps

Two methods are applied to derive a flood map from SAR backscatter values captured close to the flood peak. The second method was included as it provides derivation of flood maps at different spatial resolutions. A Sentinel-1 (S1B) image was acquired in interferometric wide swath mode (swath width 250 km) just prior to the flood peak at 0622 on the 17th February. A pre-flood image (September 2019) from the same satellite sensor and

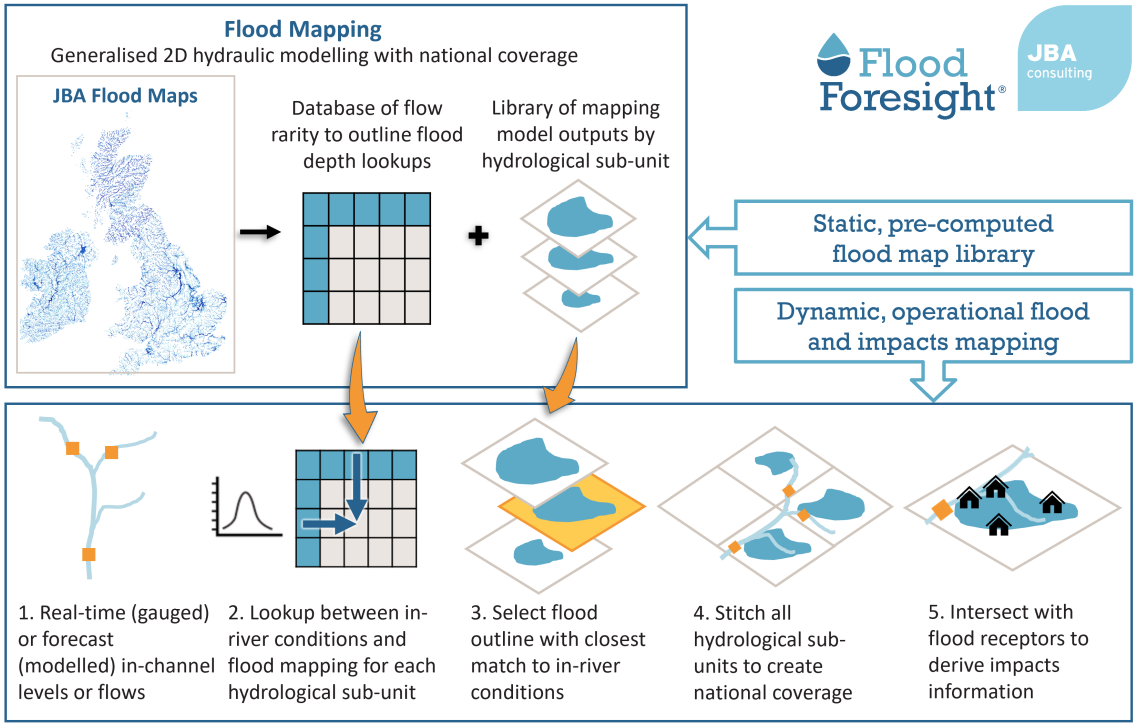


Figure 3.3: Flood Foresight flood map simulation library selection process. Source JBA Consulting.

track was used to derive the flood map in both methods.

In the first method, the ESA Grid Processing on Demand (GPOD) HASARD service (`http://gpod.eo.esa.int/`) has been utilised. The automated flood mapping algorithm (Chini et al., 2017) uses a statistical, hierarchical split-based approach to distinguish the two classes (flood and background) using a pre-flood and flood image. Level-1 GRD product SAR images (VV) are preprocessed, which involves; precise orbit correction, radiometric calibration, thermal noise removal, speckle reduction, terrain correction, and reprojection to the WGS84 coordinate system. The HASARD mapping algorithm removes permanent water bodies, including the river water. Flooded areas beneath vegetation, bridges and near to buildings are not detected using this method. The HASARD flood map at 20 m spatial scale is used to evaluate the performance of Flood Foresight for each

of the three domains out to 10 days lead time.

In the second method, the same Sentinel-1 SAR image (in this case using both VV and VH) was processed using Google Earth Engine (GEE) to derive flood maps at a range of spatial resolutions (5 m to 25 m). GEE holds a catalogue of level-1 preprocessed Sentinel-1 SAR images (Google Earth Engine Catalog, 2021). A smoothing filter is applied to reduce speckle and a pre and post flood image are used to train a Classification And Regression Tree (CART) classifier (Breiman et al., 1984; Google Earth Engine CART, 2021). The classifier is applied to the whole image to produce a flood map at a specified scale. GEE uses an image pyramid approach to scale, or pixel resolution, analysis. This means variations in the scale selected are determined from the scale of the input image (Google Earth Engine Scale, 2021). The variation of the flood extent detected at a range of spatial resolutions and the impact of re-scaling and interpolation errors on performance measures are investigated.

Flood Foresight forecast flood maps include the river channel and exclude surface features such as vegetation and buildings. To smooth the HASARD and GEE flood maps and allow a fairer comparison we apply a morphological closing operation (without impacting the location of the flood extent) to flood fill vegetation and buildings.

3.5 Flood map evaluation methods

The following subsections detail a new spatial scale-selective approach to forecast flood map evaluation. The Fraction Skill Score (FSS) developed by Roberts and Lean (2008) for validation of convective precipitation forecasts in atmospheric science uses a neighbourhood approach to determine the scale at which the forecast becomes skilful. Dey, Roberts, et al. (2016) developed this approach to determine an agreement scale between an ensem-

ble forecast and observations at each grid cell to add location-specific information. Here we extend the technique to apply it to the new application of flood inundation mapping, and further develop a novel categorical scale map that combines an agreement scale map with a conventional contingency map.

3.5.1 Spatial scale-selective approach

Initially, the observed flood extent derived from SAR data is re-scaled to match the forecast flood map grid size using spline interpolation and both are converted into binary fields. A threshold approach is determined for the situation. For a flood map verification of spatial skill, the simplest example applied here is to assign each grid cell as flooded (1) or unflooded (0) for the whole domain. Alternative future threshold approaches for flood inundation maps could include applying thresholds to water depth percentiles. The location of the flood-edge cells can be extracted from the observed and modelled binary flood maps.

Given a domain of interest, we number all of the grid cells according to their spatial coordinates (i, j), $i = 1 \dots N_x$ and $j = 1 \dots N_y$ where N_x is the number of columns in the domain and N_y is the number of rows. For each grid cell a square of length n forms an $n \times n$ neighbourhood surrounding the grid cell. The fraction of $1s$ in the square neighbourhood is calculated for each grid cell. This creates two fields of fractions over the domain for both the forecast M_{nij} and observed O_{nij} data. The fraction fields are compared against one another to calculate the mean squared error (MSE) for the neighbourhood

$$MSE_n = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [O_{nij} - M_{nij}]^2. \tag{3.1}$$

Based on the fractions calculated for the model and observed fields a worst possible MSE

is calculated

$$MSE_{n(ref)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [O_{nij}^2 + M_{nij}^2]. \tag{3.2}$$

The FSS is given by

$$FSS_n = 1 - \frac{MSE_n}{MSE_{n(ref)}}. \tag{3.3}$$

Figure 3.4 illustrates an example of the FSS application at grid level ($n = 1$) and at the next neighbourhood size $n = 3$. In this simple example, there is no agreement between the model and observation at grid level but at $n = 3$, the skill score improves to 0.92.

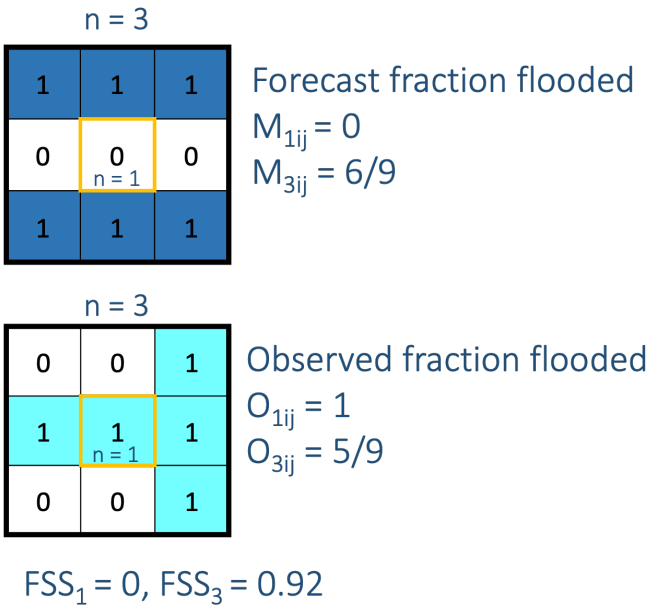


Figure 3.4: FSS (see subsection 3.5.1 for calculation details) example applied to a binary flooded (1) / unflooded (0) field at grid scale (yellow box, n = 1) and a 3 x 3 neighbourhood (black box, n = 3). The observed SAR-derived forecast is in turquoise and the forecast is shown in blue.

In general, the FSS is calculated for each length of neighbourhood n. For a given neighbourhood size an FSS of 1 is said to have perfect skill and 0 means no skill. The FSS will increase as n increases up to an asymptote (see Fig. 3 from Roberts and Lean (2008)). If there is no model bias across the whole domain of interest (observed and forecast flooded areas are the same) then the asymptotic fraction skill score (AFSS) at $n = 2N - 1$, where

N is the number of grid cells along the longest side of the domain, will equal 1. Plotting FSS against spatial scale can indicate a range of scales where the model is deemed to be the most useful. This usefulness is a trade-off between being too smooth (larger n) or too fine, where the forecast skill is lost and the computation time lengthy. The gradient of the FSS curve versus neighbourhood size is another indicator of forecast skill with respect to spatial scale. A steeper gradient indicates more rapidly improving skill over smaller grid sizes compared with a flatter curve, indicating a much wider neighbourhood is required to reach the same skill score. A target FSS score (FSS_T) is defined as

$$FSS_T \geq 0.5 + \frac{f_o}{2}, \tag{3.4}$$

where f_0 is the fraction of flood observed across the whole domain of interest and can be thought of as being equidistant between the skill of a random forecast and perfect skill. FSS_T will vary depending on the magnitude of the observed flood, relative to the domain area. This allows the comparison of the FSS_T scale across different domain sizes and floods of different magnitudes.

When the FSS is plotted against spatial scale (neighbourhood size), we can identify a spatial scale when the FSS first equals or exceeds FSS_T (Fig. 3.6 shows an example of this plot). The spatial scale (neighbourhood size) reached at FSS_T can tell us the displacement distance (D_T) between the observed and forecast flood, or more meaningfully the flood-edge locations. As the flood-edge represents a very small fraction of the domain, the scale at FSS_T will tend to $2D_T$, meaning the displacement distance is half of this scale (see Figure 4 in Roberts and Lean (2008)).

It has been shown by Skok and Roberts (2016) that care must be taken when calculating the FSS near to the domain boundary since increasingly larger neighbourhood sizes would

extend further beyond the boundary edge. Skok and Roberts (2016) concluded that as long as the domain was sufficiently large, relative to the spatial errors, then the boundary effect could be considered to be insignificant. For flood mapping verification purposes the domain area should be selected to include the area of interest (e.g. the floodplain) with the neighbourhoods considered extending beyond the domain at the boundary. This assumes that the observations available allow this. If this is not that case then another boundary method could be applied, such as cropping at the domain edge.

3.5.2 Location dependent agreement scales

The FSS gives an overall domain-averaged measure of forecast performance and an average minimum scale at which the forecast is deemed skilful. Dey, Roberts, et al. (2016) describe a method for calculating an agreement scale at each grid cell located at coordinate position (i, j). A brief summary of the method is presented here. Two fields are considered f_{1ij} and f_{2ij}. In this application these are the forecast and observed fields. In alternative applications the method could be applied to measure similarity between members of an ensemble. The fields in this instance are not required to be thresholded and can be applied to flood depths. The aim is to find a minimum neighbourhood size (or scale) for every grid point such that there is an agreement between f_{1ij} and f_{2ij}. This is known as the agreement scale S_{ij}. The relationship between the agreement scale and the neighbourhood size described in Section 3.5.1 is given by $S_{ij} = (n-1)/2$.

Firstly, all grid points are compared by calculating the relative MSE D_{ij}^S at the grid scale, $S = 0$ $(n = 1)$,

$$D_{ij}^S = \frac{(f_{1ij}^S - f_{2ij}^S)^2}{(f_{1ij}^S)^2 + (f_{2ij}^S)^2}. \tag{3.5}$$

If $f_{1ij} = 0$ and $f_{2ij} = 0$ (both dry) then $D_{ij}^S = 0$ (correct at grid level). Note that D_{ij}^S varies from zero to 1. The fields are considered to be in agreement at the scale being

tested if:

$$D_{ij}^{S} \leq D_{crit,ij}^{S_{ij}} \quad \text{where} \quad D_{crit,ij}^{S} = \alpha + (1-\alpha)\frac{S}{S_{lim}} \tag{3.6}$$

and S_{lim} is a predetermined, fixed maximum scale. The parameter value α is chosen to indicate the acceptable bias at grid level such that $0 \leq \alpha \leq 1$. Here we set $\alpha = 0$ (no background bias). If $D_{ij}^{S} \geq D_{crit,ij}^{S}$ then the next neighbourhood size up is considered ($S = 1$, a 3 by 3 square). The process continues with increasingly larger neighbourhoods until the agreement scale, or S_{lim} is reached for every cell in the domain of interest. The agreement scale at each grid cell is then mapped onto the domain of interest.

3.5.3 Categorical scale map

Currently, the agreement scale map proposed by Dey, Roberts, et al. (2016) provides a location-specific scale of agreement between the forecast and observed flood map. However, it does not show whether the model is over- or under-predicting the flood extent. In our work, we develop the agreement scale map further by combining with a contingency map for the forecast to create a new *categorical scale map*. This highlights the agreement scale for areas of over- or under-prediction. In a contingency map, each cell in the forecast and observed flood map are compared and classified using a contingency table (Table 3.1). The categories are re-classified numerically in the array for automated updating of the agreement scale map. Over-predicted cells (B) are set to -1, under-predicted cells (C) are set to +1, correctly predicted flooded cells (A) are assigned NaN and correctly predicted unflooded cells are set to 0. The array element-wise product of the agreement scale map and the numerical contingency map produces the new categorical scale map.

Table 3.1: Contingency table

	Forecast flooded	Forecast unflooded
Observed flooded	A (correct wet)	C (under-prediction/miss)
Observed unflooded	B (over-prediction/false alarm)	D (correct dry)

3.5.4 Binary performance measures

It has been suggested by Cloke and Pappenberger (2008) that a range of performance measures should be applied so that a forecast can be assessed as rigorously as possible. A selection of commonly applied binary performance measures, each focusing on a different aspect of performance have been included here for comparison with the Fraction Skill Score results. Following the application of a contingency table (Table 3.1) to the forecast flood map, a number of binary performance measures can be calculated (Table 3.2). Table 3.2 describes the range of performance value, the ideal score and a description of which aspects of the forecast flood map performance each binary measure assesses.

Table 3.2: Binary performance measures and formula based on contingency Table 3.1.

Performance measure	Formula	Description [range min, range max, perfect score]
Bias	$\frac{A+B}{A+C}$	[0, ∞, 1] 1 implies forecast and observed flooded areas are equal >1 indicates over-prediction, <1 indicates under-prediction
Critical Success Index/Threat score $F^{<2>}$ (CSI)	$\frac{A}{A+B+C}$	[0, 1, 1] Fraction correct of observed and forecast flooded cells
$F^{<1>}$ Proportion correct	$\frac{A+D}{A+B+C+D}$	[0, 1, 1] Proportion correct (wet and dry) of total domain area
$F^{<3>}$	$\frac{A-C}{A+B+C}$	[-1, 1, 1] Score reduced by over-prediction
$F^{<4>}$	$\frac{A-B}{A+B+C}$	[-1, 1, 1] Score reduced by under-prediction
False Alarm Rate (FAR)	$\frac{B}{B+D}$	[0, 1, 0] Proportion of over-prediction of dry areas
Hit Rate (HR)	$\frac{A}{A+C}$	[0, 1, 1] Fraction correct of observed flooded area
Pierce Skill Score (PSS)	$HR - FAR$	[-1, 1, 1] Incorporates both under and over-prediction

3.6 Results

We illustrate and discuss our new method applied to the flood event in subsection 3.6.1 and 3.6.2. The scale-selective approach is applied to an extreme flooding event in the UK to determine a useful/skilful spatial scale for both the entire flood extent and the flood-edge location for three domains out to 10-days lead time. An example forecast flood map for 0-day lead time compared with the SAR-derived flood map is presented as a contingency map in Figure 3.5. The zoomed in perspective shows the double penalty impact described in Section 3.2. The discrepancy at the flood-edge depends on the spatial scale of the forecast flood maps along with the model performance. Next, in subsection 3.6.3 location-specific agreement scales are presented on categorical scale maps. The final subsection 3.6.4 addresses the question of the impact of representation error caused by variations in SAR-derived flood map spatial resolution on the evaluation results.

3.6.1 Spatial scale variability of forecast flood extent and flood-edge location

An evaluation of the spatial skill of the Flood Foresight forecast flood maps against the SAR-derived flood map for the flood peak on the 17$^{\text{th}}$ February 2020 has been calculated for each domain (Fig. 3.1) for both the entire flood extent and the flood-edge location. The Fraction Skill Score (FSS) is applied to increasing neighbourhood sizes (n) to determine the spatial scale at which the forecast becomes skilful at capturing the observed flood. Figure 3.6 shows FSS against n for one example, the River Lugg (domain C) for the entire flood (a) and the flood-edge (b). Each line represents a different model run date from the 10/02/2020 (7-day lead time) to the 17/02/2020 (0-day lead time). With the exception of the 7-day lead time, all forecasts for the whole flood (Fig. 3.6a) exceed the FSS_T at grid level ($n = 1$) with gradually improving skill as n increases. In contrast to this, the FSS applied to the flood-edge (Fig. 3.6b) shows all forecasts below FSS_T at grid level

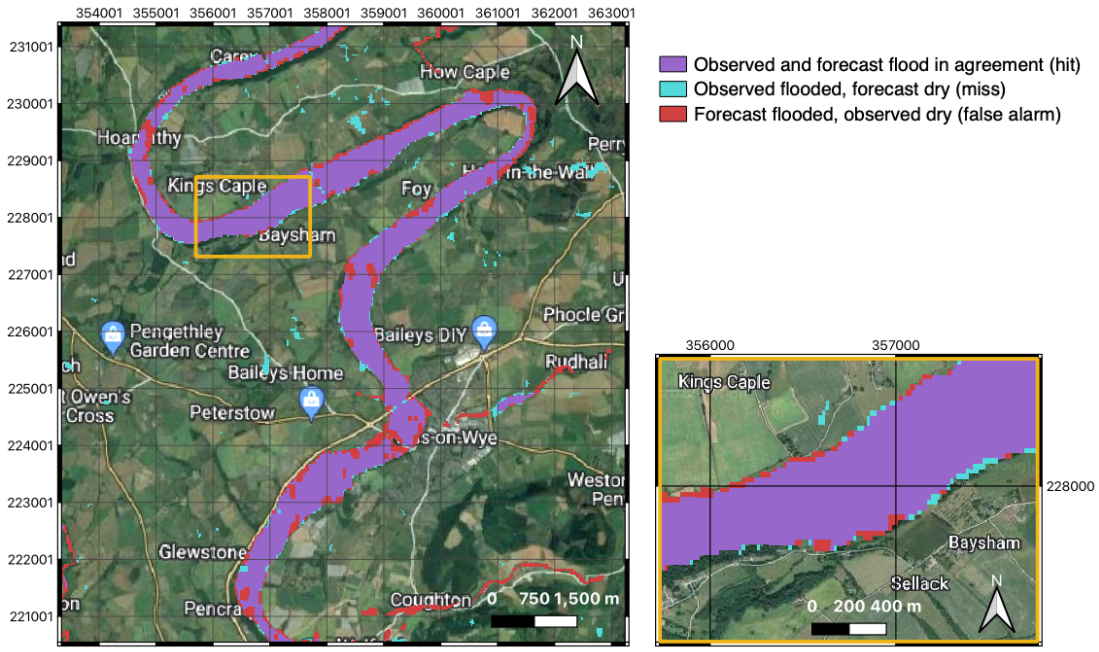


Figure 3.5: Left panel: contingency map of a 0-day lead time forecast verses the HASARD SAR-derived flood map for the Wye valley indicates the model is predicting the flood extent accurately, including the position of the flood-edge. Right panel: Zoom of yellow box on the left panel. On closer inspection, at grid level, the flood-edge in many places is over- or under-predicted by around one grid length. Base map from Google Maps.

and $n = 3$ with the skill increasing more rapidly compared with the whole flood to reach FSS_T at $n = 5$ for all run dates within a 5-day lead time (except for 16/02/2020, which is just below FSS_T). This indicates that the flood-edge is forecast to be around 62.5 m from the observed flood-edge, on average, for a 5-day lead time. The difference between the gradients of the plots indicate the flood-edge is more sensitive to changes in spatial scale compared with evaluation of the whole flooded area. The whole flood verification here indicates a strong model performance. However, verifying the whole flood alone could mask the flood-edge location performance, which in this case has a coarser scale at FSS_T. Similar trends in FSS with neighbourhood size and comparisons between the entire flood

and the flood-edge verification scales are found for all domains. The rate of FSS increase, or FSS gradient with n, tells us how quickly the forecast skill improves with increasing scale. A more spatially accurate forecast of the flood-edge will demonstrate a steeper gradient, reaching FSS_{T} at a smaller neighbourhood size.

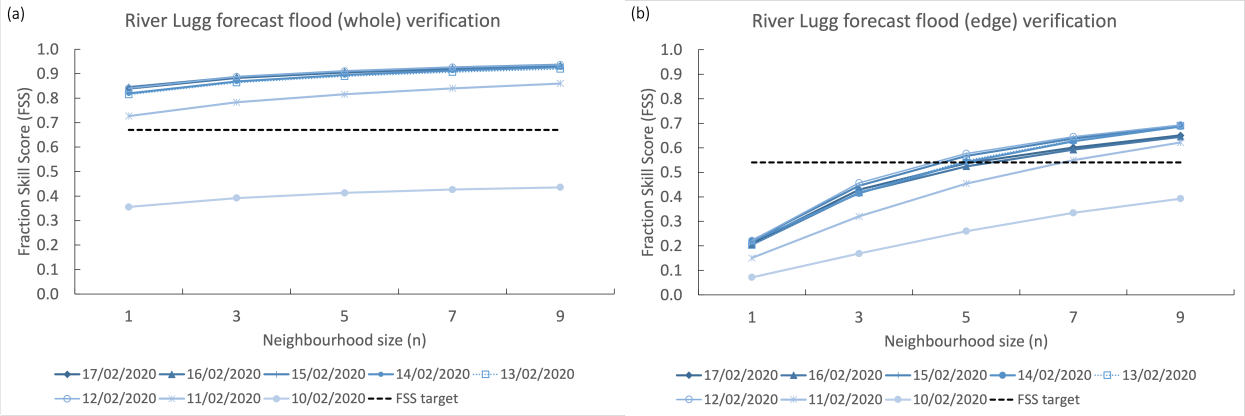


Figure 3.6: FSS calculated for the River Lugg at Lugwardine for (a) entire flood extent and (b) the flood-edge for increasing neighbourhood sizes for daily forecast lead times up to 7 days.

3.6.2 Comparison of spatial scales at differing lead times and domain location

The performance measures for each domain for daily lead times out to 10 days are presented in Figure 3.7. The FSS at $n = 1$, 3, and 5 are shown along with Critical Success Index (CSI), Hit Rate (HR), Pierce Skill Score (PSS) and the Bias (see Table 3.2 for definitions). The Bias score is an indicator of over- or under-prediction of the flood extent and is plotted on a separate axis to account for the larger range. For lead times within 5-days of the flood peak, $FSS > 0.8$ for the entire flooded area at grid level for the River Wye (domain A) indicates a strong model performance (Fig. 3.7a). There is a dip in the FSS on the 16/02/2020 where the forecast over-predicts the flood extent. This is also reflected in the CSI score. In contrast to this the HR and PSS increase, despite the over-prediction, as more observed flood cells are correctly predicted wet. We note that

the PSS (HR - FAR) does account for over-prediction, however the FAR is the fraction of the dry area incorrectly predicted wet, which is very small relative to the HR (0.03 versus 0.90). Validation of the River Wye flood-edge (Fig. 3.7b) is more sensitive to changes in neighbourhood size compared with the whole flood validation. Here the flood-edge is very well forecast in terms of spatial location and exceeds FSS_T at $n = 3$ (on average, 37.5 m displacement) for a 5-day lead time (except for 1-day lead time where FSS_T is exceeded at $n = 5$). As shown previously in Subsection 3.6.1, the forecast of the River Lugg flood-edge is skilful at $n = 5$ (Fig. 3.7f) (on average, 62.5 m displacement) for a 5-day lead time. Differences in the hydrological characteristics might explain differences in model performance. The Wye valley flood plain is well defined with distinctive valley sides and this event proved to be valley filling in contrast to the Lugg flood plain which is relatively flat and extensive. This could explain the increased skill shown for the prediction of the Wye flood-edge. The average observed flood top width for the Lugg (domain C) is 740 m and for the Wye (domain A) 430 m. This gives a flood-edge displacement as a fraction of the flood top width of 7.4% for the Lugg and 7.8% for the Wye.

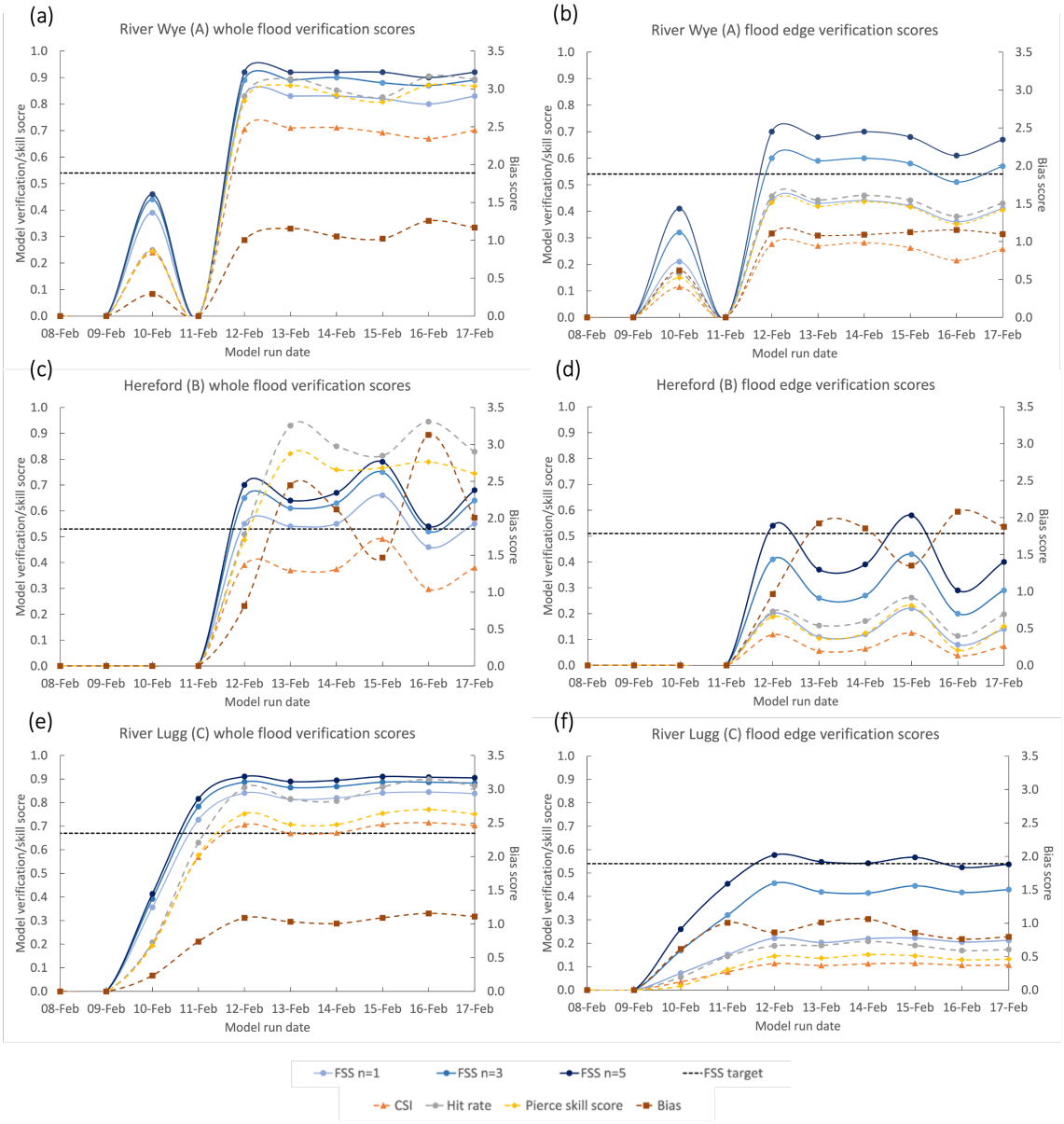


Figure 3.7: Conventional binary performance measures (dashed lines) and FSS (solid lines) at n = 1, 3, and 5 for each domain for both the whole flooded area and the flood-edge for daily lead times out to 10 days for the River Wye (domain A, (a) and (b), Hereford (domain B, (c) and (d)) and the River Lugg (domain C, (e) and (f). Plots on the left show the verification scores applied to the entire flood extent and plots on the right show the flood-edge scores.

The results for all three domains show that for this case study the forecasting system has limited skill beyond a five-day lead time. The forecast accuracy of the meteorological driving data diminishes with increasing lead time (ECMWF, 2022). Extratropical cyclones (ETCs) are the dominant meteorological driver of major winter flooding in the UK. This is particularly true when an Atmospheric River is associated with an ETC and when ETCs arrive in clusters (as was the case here) bringing multiple spells of heavy precipitation (Lavers et al., 2011; Griffith et al., 2020). The typical formation time of ETCs is 3-5 days, occasionally up to 10 days (Ulbrich et al., 2009) which limits the predictability of the meteorological system, particularly when the jet stream is very strong (as was the case here). The atmospheric (and precipitation) predictability will vary depending on the situation, for example a slow moving ETC close to the UK would potentially have a longer lead time of useful prediction. Conversely, flooding in the summer associated with convection would likely have a shorter skilful lead time. The scale selective approach presented here can be used to determine a meaningful scale to present flood inundation maps. This scale will vary with forecast lead time and will depend on the predictability of the meteorological situation.

There is more variation in skilful scale with lead time evident for the Wye at Hereford (domain B) in Figure 3.7c and d compared with domain A and C. To achieve the same FSS for the whole flood as domain A and C up to a 5-day lead time, the neighbourhood size would need to exceed $n = 5$. The model is over-predicting the flood extent, in particular on the 16/02/2020 (1-day) lead time. This overprediction at 1-day lead time is evident for all domains as can be seen in the Bias scores but the impact of this is most noticeable at Hereford. Hereford has more complex topography compared to the other domains, particularly along the river bank with bridges, buildings, permanent and temporary flood defences deployed during the event affecting the flow of the flood wave through the city. The maps used in the simulation library of Flood Foresight are produced using a

bare-earth DTM. Despite this, the model performs well, exceeding FSS_T at $n = 5$ at the 5-day and 2-day lead times for the flood-edge forecast.

Overall, the FSS indicates a similar trend in performance across all results as the commonly applied CSI. The value of FSS_T is determined by the magnitude of the observed flood, which means the skilful scale determined at FSS_T can be meaningfully compared across the domains. The skilful scale of the forecast flood-edge location gives an average discrepancy distance. A physically meaningful evaluation measure provides additional information compared to a conventional verification score.

3.6.3 Categorical scale maps

Location dependent categorical scale maps (Subsection 3.5.3) have been calculated for all run dates for both the entire flooded area and the flood-edge. Figure 3.8 shows categorical scale maps for the whole flood for three different lead times for each domain, longer lead times are on the left. The run dates vary with domain to present the most informative maps such that variation in forecast skill can be seen across the different lead times. The colours on the map indicate grid cell specific agreement scales (Subsection 3.5.2) between the forecast flood map and the SAR-derived flood map. Grey/white regions indicate correctly predicted flooded/unflooded cells, red shows the forecast flood extent is under-predicted (miss) and blue indicates over-prediction (false alarm). Increasingly darker shades of red/blue show that larger scales were needed for the agreement criteria to be met. The darkest blue at $S = 10$ indicates a total mismatch between forecast and observed flooding. The addition of the agreement scale information in comparison to a conventional contingency map (for an example, see Fig. 3.5) quickly highlights regions of total mismatch through the darkest shading, with areas that are slightly misaligned in lighter shades. The agreement scale indicated gives a physical measure of distance at specific locations between the forecast and the observed flood map (where $S < S_{lim}$).

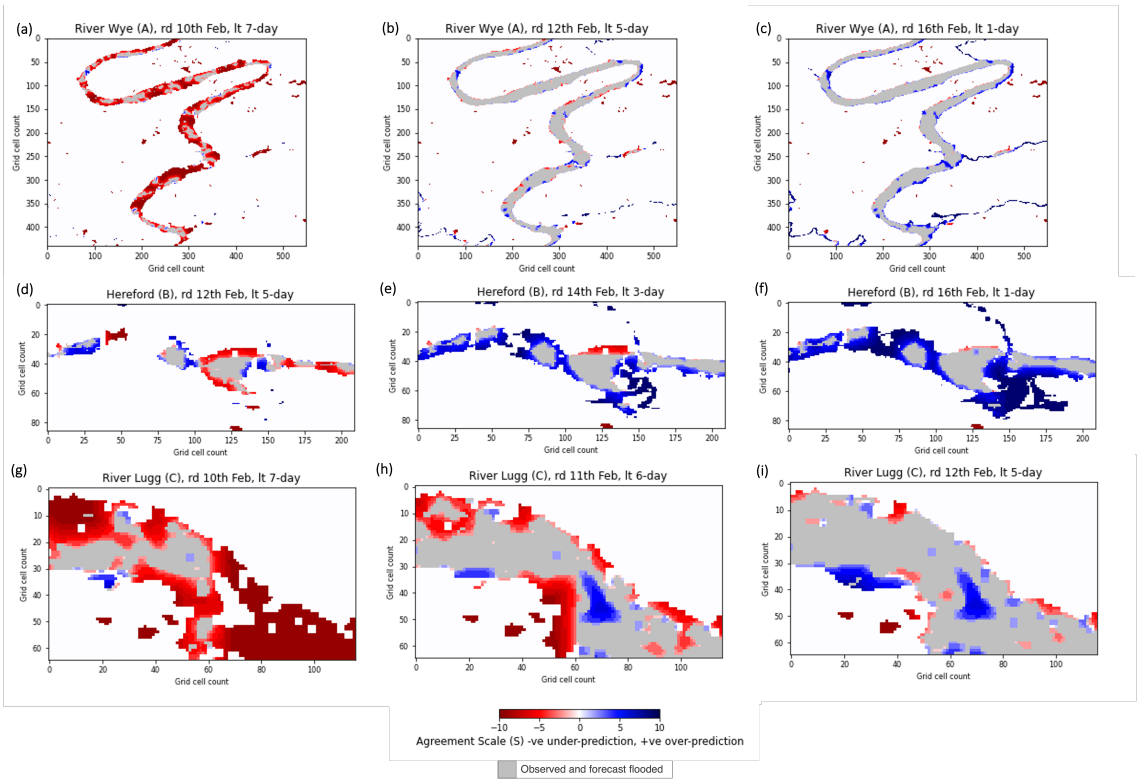


Figure 3.8: Categorical scale maps for each domain at various lead times (lt). Red indicates where the forecast flood extent is under-predicted, blue indicates over-prediction. The shading indicates the agreement scale, a measure of distance between the forecast and observed flood maps. Grey areas are correctly predicted flooded, white areas are correctly predicted unflooded. Each grid cell represents 25 m x 25 m for all domains. (Note: rd (forecast run date) varies between location, all dates have been evaluated and the most illustrative maps selected.)

The location-specific skilful scale varies with location and lead time as indicated on the categorical scale maps. For a 7-day lead time forecast for the River Wye (Fig. 3.8a), the model is indicating some flooding could occur, although under-estimating the total extent as show by the darkest red areas, which show the limits of the agreement scale have been reached. By 5-days lead time the forecast is in very close agreement with the observed flood at grid level (in grey) with larger agreement scales indicated by red/blue shading along some of the flood-edge locations (Fig. 3.8b) and a balance between under- and over-prediction. Over-prediction is more evident by 1-day lead time for the River

Wye (Fig. 3.8c) and flooding is also over-predicted along smaller tributaries. There are several detached areas of flooding observed remotely that are most likely due to ponding of surface water flooding, which were not predicted by the fluvial flood forecasting system.

The Hereford forecast is most skilful on the 12th February (Fig. 3.8d) with over-prediction, particularly towards the southwest at 3-day and 1-day lead times (Fig.3.8e and f). A small stream running southwards to the Wye, the Eign Brook, could be contributing to the over-prediction seen here. It is also worth mentioning that SAR will struggle to detect flood waters where buildings are closer together when the distance between them is less than the ground resolution of the SAR. Shadow and layover effects due to the side-looking nature of the SAR also mean flood detection is more difficult in urban areas (Mason et al., 2021a). This will likely only impact a small area of the Hereford domain but this observation uncertainty should be considered when interpreting these results. There is an area of under-prediction of the flood extent in the centre of the Hereford domain visible at all lead times. This could be due to surface water flooding, which most likely occurred due to the very high intensity rainfall observed. This combined with the urban area and steeply sloping gradient to the north of this area most likely contributed to rapid surface water runoff towards the river. Since Flood Foresight is a fluvial flooding forecast system we would not expect surface water flooding such as this to be predicted.

Flood Foresight selects multiple flood maps and stitches them together when the return period threshold is exceeded for a given area. The Hereford section of the Wye does not trigger a flood map selection until a 5-day lead time, this area also influences part of the River Lugg flood map and can be seen as a mismatch on the lower left hand side of Figure 3.8g and h. Once this is included the forecast flood map is in very good agreement from a 5-day lead time. There are areas that could be further improved, indicated by the lighter shading (Fig. 3.8i). An acceptable level of agreement scale could be determined for a given

situation, for example $n < 5$, and efforts made to understand/improve larger agreement scales at specific locations. These improvements might include changes to infrastructure included in the DTM used in the hydraulic modelling, for example.

3.6.4 SAR-derived flood map scale variation

In practice, particularly where a flood event is prolonged or the flood extent covers a wide area, there may be multiple sources of SAR data available for model evaluation, usually at higher spatial resolutions compared to the model grid size (e.g. ICEYE in spot mode at 1 m and strip mode at 3 m ground resolution). It is important to consider the impact of using observations at different spatial scales on the scale-selective approach results. By conducting a simulation experiment we address the question of how re-scaling and interpolating three higher spatial resolution SAR-derived flood maps (relative to the forecast flood maps) affects the scale selective skill scores and location-specific forecast skill. In order to simulate a range of observation spatial scales, SAR-derived flood maps are produced using method two described in Section 3.4.2 at spatial resolutions from 5 m to 25 m. These are re-scaled by 0-order spline interpolation (ndimage.zoom, 2021; Briand & Monasse, 2018) to match the model resolution (25 m) and compared to the forecast flood map for the River Lugg (5-day lead time). A comparison of the GEE flood map against the HASARD flood map, both at 20 m spatial scale produce almost identical verification scores for all performance measures for the River Lugg ($\Delta FSS < 0.01$).

The categorical scale maps for the comparison between the forecast flood map and the re-scaled simulated SAR-derived flood maps are shown in Figure 3.9. The resulting domain-averaged skill scores for the same forecast flood map against the four SAR-derived flood maps are displayed in Figure 3.10. The scores are calculated for the whole flood and the flood edge alone. In general, the categorical scale maps show similar regions of over and under-prediction but there are small location-specific variations in skilful scale. The

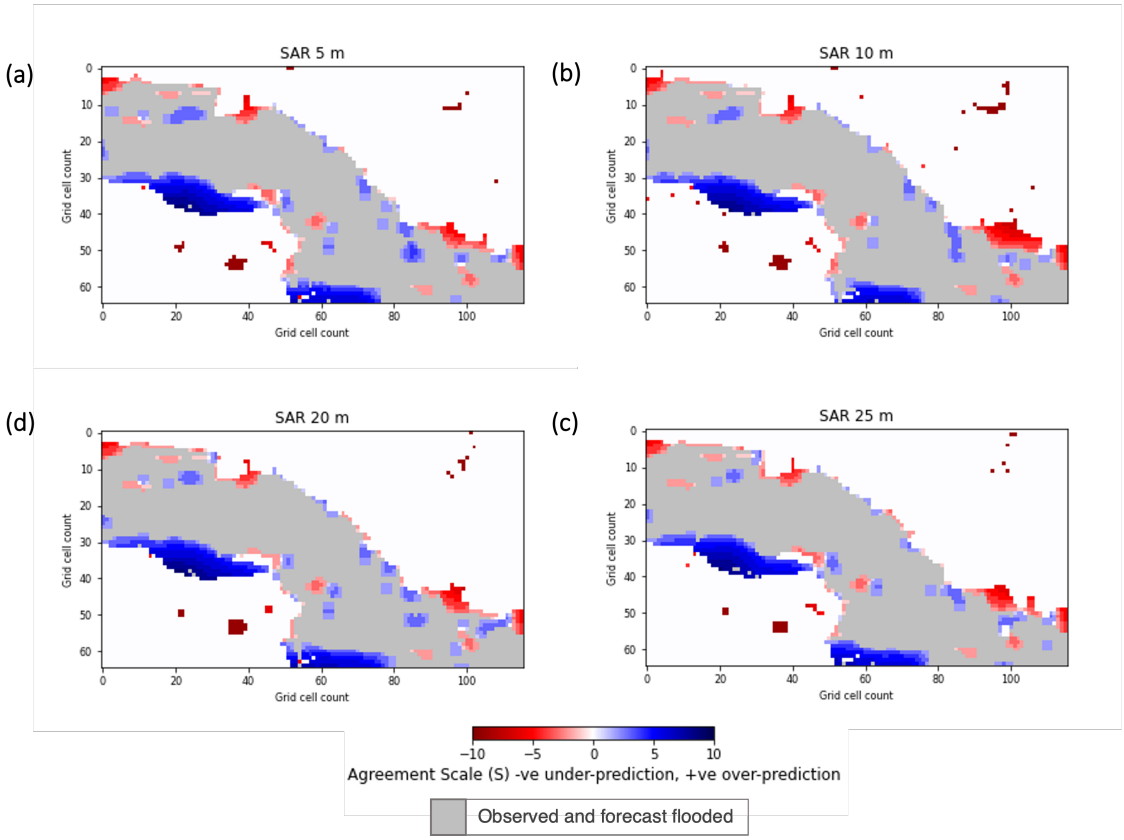


Figure 3.9: SAR-derived flood maps produced at different spatial resolutions (5 m to 25 m) are re-scaled to the model grid size (25 m) before categorical scale maps are calculated for the River Lugg (C), run date 12th Feb.

SAR-derived flood map at 25 m, the same spatial scale as the forecast flood maps, shows the best agreement away from the flood edge. This is also evident in the overall FSS score for the 25 m comparison, which marginally outperforms the evaluation after re-scaling finer observation flood maps (Fig. 3.10). The skilful scale determined for each observation comparison of the whole flood extent is $n = 1$ or at grid level, and for the flood edge is at $n = 5$.

Overall, based on the results from this simulation experiment, the scale-selective approach is not overly sensitive to the observation spatial scale and the skilful scale determined remains the same for each of the observed SAR-derived flood maps for both the

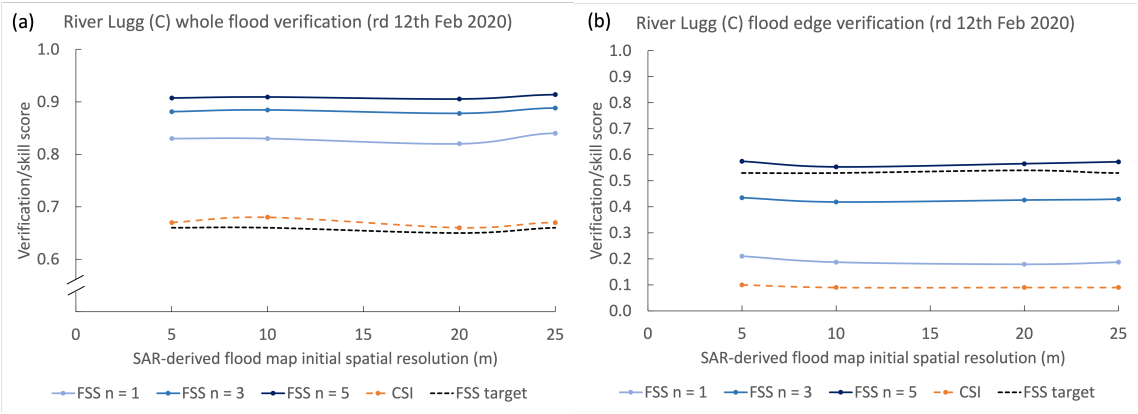


Figure 3.10: SAR-derived flood maps at different spatial resolutions (5 m to 25 m) are re-scaled to the model grid size (25 m) before verification scores are calculated for the whole flood (a) and the flood-edge (b). Note that axes in (a) and (b) are on different scales.

entire flood extent and the flood edge. Small errors are introduced by re-scaling and interpolating finer resolution observations to the model spatial scale which slightly reduce the skill score and change location-specific details on the categorical scale maps. Observation scale selection and re-scaling along with interpolation errors must be considered when evaluating model performance, particularly where model or observation scales vary in space and time, or where comparisons are made across different models.

3.7 Discussion and Conclusions

Overall, the aim of this paper was to introduce and apply a new scale-selective approach to forecast flood map evaluation with an emphasis on providing a physically meaningful verification of the flood-edge location. The skilful spatial scale for comparison of forecast flood inundation maps against SAR-derived observed flood extent has been evaluated by the application of the Fraction Skill Score: this provides a domain-averaged skilful scale. The verification measure has been applied to a forecast of an extreme flood event in the UK on the River Wye and the River Lugg following Storm Dennis in February 2020. Flood Foresight inundation predictions with lead times out to 10 days are evaluated against a

Sentinel-1 SAR-derived flood map captured close to the flood peak for three domains, each differing in hydrological characteristics. Conventional binary performance measures were calculated alongside the FSS for comparison. Flood-edge verification shows greater sensitivity to changes in forecast skill and spatial scale, relative to verification of the entire flood extent. The skilful scale determined is physically meaningful and can be used to estimate the average flood-edge discrepancy from the observed flood-edge. The observed flood map spatial resolution relative to the model scale is important and re-scaling and interpolation errors will impact the model verification scores. Ideally, the observed flood map should be derived at the same spatial scale as the forecast model to minimise these errors.

In operational practice the scale at which the forecast flood maps are presented to forecasters and decision makers should reflect the uncertainty within the forecast. Very high resolution flood maps can be presented where a detailed DTM is available. If this is presented as a deterministic forecast to flood risk management teams, it could lead to an over confidence in the forecast, or where the actual observed flood magnitude is different, the forecast may be devalued in the future (Savage et al., 2016; Speight et al., 2021). Application of a spatial-scale approach to forecast evaluation can determine the scale at which it is best to present the forecast flood map. Conversely, if the model is found to be skilful at grid level, there is scope to increase the flood map resolution adding more detail to the flood-edge location. Improvements made to hydrodynamic models, such as through data assimilation to improve inputs, initial conditions or model parameters may not improve the forecast flood-edge location at grid level. However, improvements may be evident through evaluation using FSS across a range of scales. Categorical scale maps are a useful evaluation and forecasting tool, adding location-specific detail. Model improvements can be spatially targeted and as improvements are made, the categorical scale map will highlight location-specific changes. For example, the categorical scale maps for

Hereford indicate the local infrastructure (in particular bridges) impact the movement of the flood wave, which suggests a digital surface model (DSM) would be beneficial in urban areas.

The verification approach is presented here in the context of an operational flood forecasting system. The skilful scale determined for each flooding scenario, lead time and at specific locations within a domain depends on the skill of the entire hydrometeorological chain of forecasting models from the meteorological inputs to the hydrodynamic model (run offline in the case study presented here) used to determine the inundation extent for a given river discharge. The scale-selective approach is equally applicable for the validation of flood maps from hydrodynamic models that are not part of an operational system. Here, we focus on the use of SAR-derived flood maps for validation, however the approach would apply to any remotely observed flooding such as from optical satellite data or UAV aerial imagery that can be converted into a gridded dataset. The FSS must be applied to binary data and for this reason it is very easily applicable to flood extent with grid cells categorised as flooded/unflooded. In operational forecasting, flood depth is also an important metric to verify and by applying a threshold (depths below/above a certain level or percentile), the depth data can be converted for application of FSS. The method for calculating categorical scale maps does not require binary data and so the depth values can be used directly in the calculations.

Ideally, in operational forecast systems, quantitative validation should run in tandem with the forecast system where observations are available. Over time, a catalogue of skilful scales, flood edge discrepancy distances and categorical scale maps could be built up. This catalogue would enable analysis of scale across different flood event type, season, meteorological scenario, forecast lead time and at specific locations within a catchment or sub-catchment. Such a verification library would enable forecasters to increase intuition

and expert judgement on the relevant scales for a given forecast. Based on this analysis and an increased understanding of the predictability of flood inundation, forecast flood maps could be presented at a variable scale. For example, a coarser scale at longer lead times becoming more detailed, closer to the flooding event. Coarse scales can appear jagged or with large steps along the edge and so ideally these would be converted to smooth contours, but with some indication (for example, lighter shading) that the flood edge lies somewhere within the width of the grid cell, rather than exactly at the contour edge. At shorter lead times, as forecast confidence is assumed to increase, the flood edge location would show more detail and a narrower band of uncertainty (grid cell width). This flood edge uncertainty information will prove invaluable for impact-based forecasting practice.

The spatial-scale approach will also prove a useful tool in multi-model performance comparisons where forecast flood maps are presented at different spatial resolutions or to evaluate the performance of an increase in model resolution. Evaluating a skilful scale for each model can be compared directly whereas the skill score values should not be compared across models with different spatial scales (Emerton et al., 2016). These methods will also benefit surface water flooding verification where the flood map is likely to be localised and discrete and accounting for variations in spatial skill more critical. An improved approach to evaluating forecast flood maps will result in improved accuracy in the predictions of flooding. Ultimately, this will benefit disaster management teams and those living in flood prone areas to enable future mitigation of flooding impacts.

3.8 Chapter summary

In this chapter, we described and applied a scale-selective approach as a new verification tool in flood inundation forecasting. The scale-selective approach overcomes issues with conventional binary performance measures such as spatial scale dependency and flood

magnitude biases. We found that verification of the flood edge gave more sensitive skill scores compared to verifying the whole flood extent. The resultant skilful scale can be converted into a discrepancy distance between the forecast and the observed flood edge location. Finally, a quantitative location specific agreement scale can be plotted on a categorical scale map, which also indicates whether the forecast is accurate, over- or under-predicting the flood extent at each grid cell location. This enables targeted improvements to be made in operational practice. In the next chapter, we develop the method further by considering the evaluation of the spatial spread-skill of an ensemble flood map forecast.

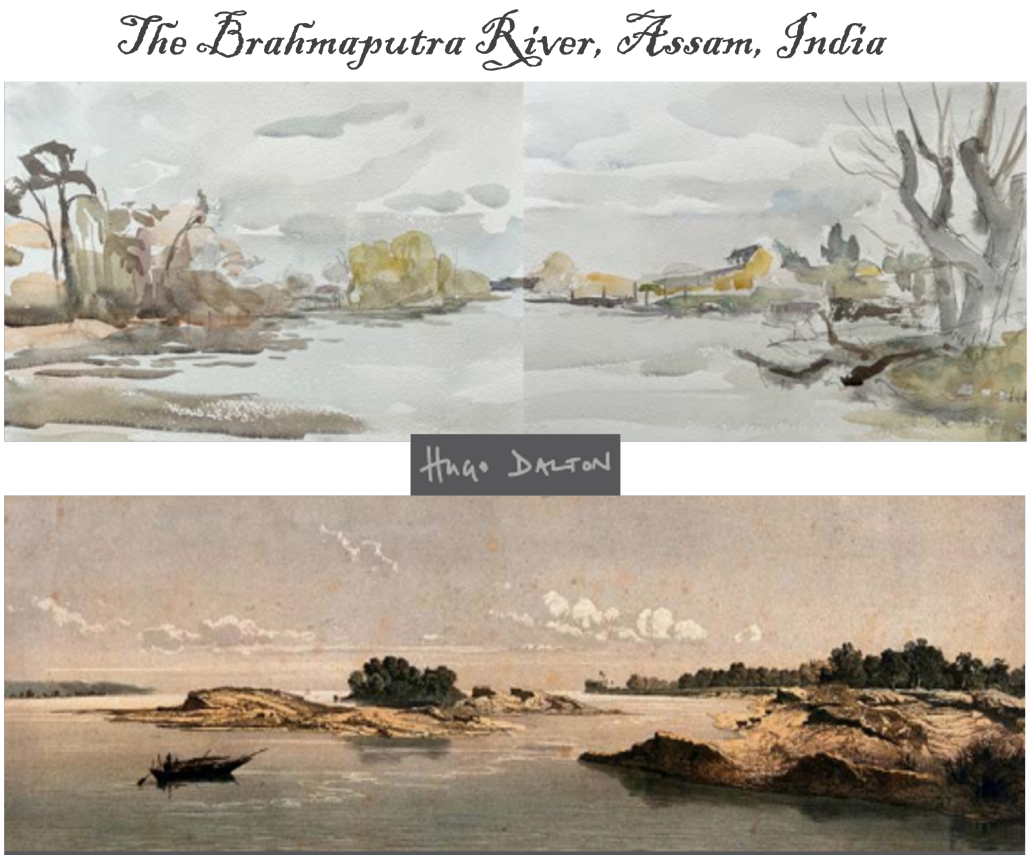
Chapter 4

Assessing the spatial spread-skill of ensemble flood maps with remote sensing observations

In this chapter we address the second research question outlined in Chapter 1; How skilfully does an ensemble of forecast flood maps represent the spatial uncertainty within the flood forecast?:

- How can we summarise the spatial predictability information in ensemble flood map forecasts?
- How can we evaluate the spatial spread-skill of an ensemble flood map forecast?
- How does the spatial spread-skill vary with location and how can this be presented?

The remainder of this chapter (except for the chapter summary, Section 4.7), has been published and is reproduced from (Hooker et al., 2023a).



4.1 Abstract

An ensemble of forecast flood inundation maps has the potential to represent the uncertainty in the flood forecast and provide a location specific likelihood of flooding. Ensemble flood map forecasts provide probabilistic information to flood forecasters, flood risk managers and insurers and will ultimately benefit people living in flood prone areas. Spatial verification of the ensemble flood map forecast against remotely observed flooding is important to understand both the skill of the ensemble forecast and the uncertainty represented in the variation or spread of the individual ensemble member flood maps. In atmospheric sciences, a scale-selective approach has been used to evaluate a convective precipitation ensemble forecast. This determines a skilful scale (agreement scale) of ensemble perfor-

mance by locally computing a skill metric across a range of length scales. By extending this approach through a new application, we evaluate the spatial predictability and the spatial spread-skill of an ensemble flood forecast across a domain of interest. The spatial spread-skill method computes an agreement scale at every grid cell between each unique pair of ensemble flood maps (ensemble spatial spread) and between each ensemble flood map with a SAR-derived flood map (ensemble spatial skill). These two are compared to produce the final spatial spread-skill performance. These methods are applied to the August 2017 flood event on the Brahmaputra River in the Assam region of India. Both the spatial-skill and spread-skill relationship vary with location and can be linked to the physical characteristics of the flooding event such as the location of heavy precipitation. During monitoring of flood inundation accuracy in operational forecasting systems, validation and mapping of the spatial spread-skill relationship would allow better quantification of forecast systematic biases and uncertainties. This would be particularly useful for ungauged catchments where forecast streamflows are uncalibrated and would enable targeted model improvements to be made across different parts of the forecast chain.

4.2 Introduction

Forecast flood maps indicating the extent and depth of fluvial flooding within an actionable lead time, are a useful tool for flood risk managers and emergency response teams prior to and during a flood event. Typically, forecast flood maps are presented as deterministic forecasts showing precisely where flooding will occur. This can lead to incidents of false alarms or missed warnings and subsequent recriminations causing mistrust in the system (Arnal et al., 2020; Savage et al., 2016). A timely prediction of exactly where and when fluvial flooding caused by intense or prolonged rainfall will occur is virtually impossible due to the chaotic nature of the atmosphere (Lorenz, 1969). The ensemble forecasting approach aims to address the sensitivity of the atmospheric dynamics to initial

conditions and through multiple model runs these initial condition uncertainties can be quantified (Leutbecher & Palmer, 2008). The ensemble forecast results in a probabilistic weather forecast that indicates the predictability of the atmosphere at a given space and time. State-of-the-art operational ensemble flood forecasting systems link together a chain of forecast models to produce probabilistic streamflow and flood inundation forecasts at national and global scales (Cloke & Pappenberger, 2009; Emerton et al., 2016; Wu et al., 2020). Ensemble Numerical Weather Prediction models provide meteorological inputs into land-surface, hydrological and hydraulic models, cascading the atmospheric uncertainty through to the flood forecast. Throughout this chain of models, multiple sources of uncertainty exist that have been investigated in numerous studies (Beven, 2016; Matthews et al., 2022; Pappenberger et al., 2005; Zappa et al., 2011). As discussed by Boelee et al. (2019), these uncertainties include those arising from meteorological inputs, measurements and observations, initial conditions, unresolved physics within the models and parameter estimates. A probabilistic flood inundation forecast should present a meaningful prediction of the likelihood of flooding so that there is confidence in the forecast, given the uncertainties represented in the system (Alfonso et al., 2016).

The accuracy of the location of flooding, predicted in advance, is defined as spatial predictability. The spatial predictability of ensemble forecasts of flood inundation could be verified by comparing with a remote observation of the flood from satellite or unmanned aerial vehicle (UAV) based sensors. Satellite-based optical and Synthetic Aperture Radar (SAR) sensors are well known for their flood detection capability (e.g., Horritt et al., 2001; Mason, Davenport, et al., 2012; Mason, Schumann, et al., 2012). SAR sensors are active, which enables them to scan the Earth through weather and clouds, and at night. The SAR backscatter intensity detected depends on the roughness of the surface, with unobstructed flooded areas and other surface water bodies appearing relatively smooth and returning low backscatter values. Dasgupta et al. (2018) detail some of the challenges along with ap-

proaches to solutions of flood detection using SAR, examples of these challenges include: roughening of the water surface by heavy rain and strong wind, emergent or partially submerged vegetation and flood detection in urban areas. Accurate flood detection in urban areas particularly due to surface water flooding has become increasingly important (Speight et al., 2021) and recent techniques have led to improved flood detection (Mason et al., 2018, 2021a, 2021b). Optical instruments rely on solar energy and cannot penetrate cloud, making them less useful during a flooding situation. Recent studies have investigated the flood detection benefits from combining both optical and SAR imagery (Konapala et al., 2021; Tavus et al., 2020). Improvements in the spatial-temporal resolution of SAR images and their open source availability mean that they are an increasingly valuable tool for hydraulic and hydrodynamic model improvements through calibration, validation and data assimilation (e.g., García-Pintado et al., 2015; Grimaldi et al., 2016; Cooper et al., 2018, 2019; Di Mauro et al., 2021; Dasgupta et al., 2018, 2021a, 2021b). The Global Flood Monitoring (GFM) product (EU Science Hub, 2021; GFM, 2021; Hostache et al., 2021) of the Copernicus Emergency Management Service (CEMS) (Copernicus Programme, 2021) produces SAR-derived flood inundation maps for every Sentinel-1 image detecting flooding. Three flood detection algorithms provide uncertainty estimation and population affected estimates within 8 hours of the image acquisition. The European Space Agency (ESA) Copernicus Programme have recently included the ICEYE constellation of small satellites into the fleet of missions contributing to Europe's Copernicus environmental monitoring programme (ESA, 2021). ICEYE captures very high resolution (spot mode ground range resolution = 1 m) SAR images which brings the potential for increased accuracy of flood detection, particularly in urban areas.

To evaluate the accuracy of an ensemble forecast, a number of verification measures have been proposed. Anderson et al. (2019) developed a joint verification framework for end-to-end assessment of the England and Wales Flood Forecasting Centre (FFC) ensem-

ble flood forecasting system. Anderson et al. (2019) describe verification metrics such as the continuous rank probability score (CRPS), rank histograms, Brier Skill Score (BSS) and the relative operative characteristics (ROC) diagrams that are commonly applied to assess the main ensemble attributes desirable in both precipitation and streamflow ensemble forecasts (e.g., Renner et al., 2009). These metrics refer to flooding events as part of a time series evaluated against a reference benchmark, such as climatology, to produce an average skill score. In contrast, here we consider ensemble *spatial* verification at a single time point. The verification of ensemble forecasts usually involves comparing the RMSE of the ensemble mean against an observed quantity to assess the *skill* of the forecast with the ensemble standard deviation used as a measure of *spread*. A perfect ensemble should encompass forecast uncertainties such that the ensemble spread is correlated to the RMSE of the forecast (Hopson, 2014). This *spread-skill* relationship was assessed by Buizza (1997) to investigate the predictability limits of the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS). This approach to ensemble verification is based on point values and makes the assumption that the ensemble mean is the forecast state with the highest probability and that the forecast distribution is Gaussian. Significant flooding events are, in their nature, a rare occurrence and in certain circumstances a few ensemble members can indicate a low probability of an extreme flood. Also, in particular atmospheric scenarios the ensemble forecast may result in a multi-modal forecast where two clusters of ensemble members are each equally likely (Galmiche et al., 2021). For example, both clusters may indicate flooding events but at different magnitudes. In both of these instances the individual ensemble member details are important and evaluation of the ensemble mean alone would not be meaningful. When mapping the flood extent prediction, the ensemble mean field alone does not retain the spatial detail of the individual member forecasts.

The spatial spread-skill of the ensemble forecast is determined by evaluating the full

ensemble against observations of flooding. For a flood map ensemble to be considered spatially well-spread, the spread or variation between ensemble members should equal the spatial predictability, or skill of the ensemble members (Dey et al. (2014), see Section 4.3). Presently, to the best of our knowledge, quantitative evaluation methods assessing the spatial spread-skill of ensemble forecast flood maps do not exist. However, previous work in numerical weather prediction by Ben Bouallègue and Theis (2014) investigated the application of spatial techniques to ensemble precipitation forecasts using a neighbourhood, or fuzzy approach that allowed comparisons at larger scales than grid level (native resolution). A location dependent approach to the spatial spread-skill evaluation of a convective precipitation ensemble forecast was developed by Dey, Roberts, et al. (2016). This method compares every ensemble member across a range of scales on a spatial field against an observation field to assess whether the ensemble forecast is spatially over-, under- or well-spread on average across a domain of interest (Chen et al., 2018). In a recent study, a scale-selective approach was developed and applied to evaluate a deterministic flood map forecast where comparisons were made against conventional binary performance measures (Hooker et al., 2022). A scale-selective approach to flood map evaluation was found to have several benefits over conventional binary performance measures. These include over-coming the double penalty impact problem when validating at higher spatial resolutions and accounting for the impact of the flood magnitude on the skill score. The work described here extends and applies this scale-selective approach to assess the spatial predictability and the spatial spread-skill of an ensemble flood map forecast.

In this paper we aim to address the following questions:

- How can we summarise the spatial predictability information in ensemble flood map forecasts?
- How can we evaluate and visualise the spatial spread-skill of an ensemble flood map

forecast?

- How does the spatial spread-skill vary with location and how can this be presented?

In Section 4.3 we present a new approach to the evaluation of spatial predictability and the spatial spread-skill of an ensemble flood map forecast by comparing against a remotely observed flood extent. We illustrate the features of the methods through an example case study of an extreme flooding event of the Brahmaputra River which impacted India and Bangladesh in August 2017, with focus on the Assam region of India. The flood event details are described in Section 4.4.1. The international ensemble version of the JBA Consulting Flood Foresight system provides forecast flood maps for the study and is described in Section 4.4.2. Observations of the flood are derived from satellite based SAR sensors and the method is explained in Section 4.4.3. The results including the Spatial spread-skill (SSS) map are discussed in Section 4.5. Our results show that individual ensemble member spatial predictions of flooding are meaningful and that the full ensemble spatial detail should be evaluated. We conclude in Section 4.6 that the spatial spread-skill of the ensemble forecast varies with location across the domain and can be linked to physical characteristics of the flooding event.

4.3 Ensemble flood map spatial predictability evaluation methods

In this Section we present new methods for evaluating and visualising the spatial-spread skill of an ensemble flood map forecast. Hooker et al. (2022) described and applied a new scale-selective approach to evaluate the spatial skill of a *deterministic* flood map forecast relative to an observed SAR-derived flood map. Here, we apply this same measure to evaluate different aspects of an *ensemble* forecast. The scale-selective Fraction Skill Score (FSS) method is outlined in Section 4.3.1. Agreement scale maps indicating forecast ac-

curacy are defined for location-specific comparisons between forecast and observed flood maps in Section 4.3.2. These are used to assess the spatial relationship between each unique pair of ensemble member flood maps (member-member) and between every ensemble member flood map and the observed SAR-derived flood map (member-SAR, Section 4.3.3). Visualisation methods of the spatial spread-skill relationship including the *Spatial Spread-Skill* (SSS) map are presented in Section 4.3.4.

4.3.1 Fraction Skill Score

The FSS is a scale-selective verification measure that can determine the skilful scale of a modelled flood map, when compared against a remotely sensed observation of flooding (Roberts & Lean, 2008; Hooker et al., 2022). We will call these flood maps the *model array* and the *observed array* respectively. For an ensemble forecast, the model array could be an individual ensemble member, or a summarised flood estimate derived from a combination of ensemble members such as a combined ensemble or the ensemble median (see Section 4.4.4). Both the model and observed arrays are converted into binary fields using a situation dependent threshold (e.g. depths greater than 0.2 m are labelled flooded). For this ensemble application of the FSS we evaluate the entire flood extent across the domain. Each grid cell is labelled as inundated (1) or dry (0). All grid cells are numbered according to their spatial locations (i, j), $i = 1 \dots N_x$ and $j = 1 \dots N_y$ where N_x is the number of columns and N_y is the number of rows. Surrounding each grid cell, a square of length n creates an $n \times n$ neighbourhood. The fraction of $1s$ (inundated cells) in the square neighbourhood area is calculated for every grid cell. This creates two arrays of fractions across the domain for both the observed O_{nij} and modelled M_{nij} data. The mean squared error (MSE) for the fraction arrays is calculated for the domain and a given neighborhood size, n:

$$MSE_n = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [O_{nij} - M_{nij}]^2. \tag{4.1}$$

A potential maximum $MSE_{n(ref)}$ depends on the fraction of flooding in the domain for the modelled and observed fields and is calculated as:

$$MSE_{n(ref)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [O_{nij}^2 + M_{nij}^2]. \tag{4.2}$$

Finally, the FSS is

$$FSS_n = 1 - \frac{MSE_n}{MSE_{n(ref)}}. \tag{4.3}$$

The FSS is initially calculated at grid level ($n = 1$) followed by the smallest neighbourhood size ($n = 3$) before increasingly larger neighbourhood sizes ($n = 5$, $n = 7$...) are considered. The FSS ranges between 0 (no skill) and 1 (perfect skill). Increasing the neighbourhood size typically leads to an improved FSS as the fractions are calculated over a larger area. Plotting FSS against the neighbourhood size can indicate a range of scales where the model is deemed to be the most skilful. A target FSS score (FSS_T) can be determined from the fraction of observed flooding across the whole domain (f_0):

$$FSS_T \geq 0.5 + \frac{f_o}{2}. \tag{4.4}$$

The point where the FSS_n exceeds FSS_T can be viewed as being equidistant between the skill of a random forecast and perfect skill (Roberts & Lean, 2008). A recent study by Skok and Roberts (2018) investigated the sensitivity of the calculated skilful scale to the constant value (0.5) in Eq. (4), and found that 0.5 gave meaningful results compared with the measured displacement. The magnitude of the observed flood, relative to the domain area, determines the value of FSS_T. This allows the comparison of the skilful scale (neighbourhood size) where FSS_T is reached across different domain sizes and floods of different magnitudes.

4.3.2 Location dependent agreement scales

The FSS (Section 4.3.1) gives a domain average measure of forecast performance and a minimum spatial scale at which the forecast is deemed skilful. To enable the spatial spread-skill of the full ensemble to be evaluated at specific locations, we first define an agreement scale (see Dey et al. (2014); Dey, Roberts, et al. (2016); Hooker et al. (2022) for full methodology). The agreement scale is calculated and mapped for every grid cell in the domain and shows a measure of similarity between two arrays of data. In contrast to the FSS method the arrays are not required to be thresholded. The agreement scale method can be applied to both binary flood extent maps as well as flood depth fields. These could both be ensemble member flood maps or an ensemble member flood map and an observed flood map. Two data arrays are compared F_{1ij} and F_{2ij} and the aim is to find a minimum neighbourhood size (or spatial scale) for every grid cell such that there is a predetermined acceptable minimum level of agreement between F_{1ij} and F_{2ij}. This is known as the agreement scale $S_{ij}^{A(F_1F_2)}$. (Note that the relationship between the agreement scale and the neighbourhood size described previously in section 4.3.1 is given by $S_{ij}^{A(F_1F_2)} = (n-1)/2$.) The agreement scale (now defined S for simplicity in the following equations) is determined individually for every grid cell by testing and meeting a chosen criteria.

A relative MSE, D_{ij}^S is calculated for all grid cells, initially at grid level, $S = 0$ ($n = 1$),

$$D_{ij}^S = \frac{(F_{1ij}^S - F_{2ij}^S)^2}{(F_{1ij}^S)^2 + (F_{2ij}^S)^2}. \tag{4.5}$$

If $F_{1ij} = 0$ and $F_{2ij} = 0$ (both dry) then $D_{ij}^S = 0$ (correct at grid level). The value of D_{ij}^S ranges between zero and 1. The arrays are deemed to be in agreement at the scale being tested if:

$$D_{ij}^S \leq D_{crit,ij}^S \quad \text{where} \quad D_{crit,ij}^S = \alpha + (1-\alpha)\frac{S}{S_{lim}} \tag{4.6}$$

The parameter value α indicates an acceptable bias at grid level such that $0 \leq \alpha \leq 1$. Additional historical forecast data of flood events is not available for the region in this study, so we assume there is no background bias between the forecast and the observations and set $\alpha = 0$. A fixed maximum scale S_{lim} is predetermined using human judgement considering the physical characteristics of the flood event. The value chosen for S_{lim} depends on the magnitude of the flood extent relative to the size of the sub-catchment. For the case study presented here, we set $S_{lim} = 80$ (2400 m), which is approximately $\frac{1}{4}$ to $\frac{1}{2}$ of the sub-catchment widths in the domain. If $D^S_{ij} \geq D^S_{crit,ij}$ then the next neighbourhood size up is considered ($S = 1$, $n = 3$, a 3 by 3 square) where F^1_{1ij} and F^1_{2ij} are arrays containing the average value of each neighbourhood surrounding the grid cell at position (i, j) for each array. The process continues by comparing increasingly larger neighbourhoods (e.g. $S = 2$, $n = 5$, a 5 by 5 square) until the agreement criterion:

$$S^{A(F_1F_2)}_{ij} \quad \text{or} \quad S_{lim} \quad \text{at} \quad D^S_{ij} \leq D^S_{crit,ij} \tag{4.7}$$

is met for every cell in the domain. The agreement scale at which the agreement criterion is met will usually vary from grid cell to grid cell and these values ($S = 0$, $S = 1$, $S = 2$ and so on up to S_{lim}), each specific to each grid cell location can be mapped onto the domain of interest to provide a location specific measure of agreement between the two data arrays that are compared. A small value for the agreement scale means that the two arrays being compared are very similar (spatially) at a specific location, whereas a large value for the agreement scale means that the two arrays being compared are dissimilar. Note that the skilful scale determined by the FSS (Section 4.3.1) differs from the agreement scale defined here. The former links directly with the spatial differences between objects e.g. Skok and Roberts (2018), whereas the latter reflects a pre-defined acceptable bias at different scales.

Validation of forecast flood maps against remotely observed flood extent is typically

carried out by labelling each grid cell using a contingency table with categories: correctly predicted flooded, under-prediction (miss), over-prediction (false alarm) and correctly predicted unflooded. In the contingency table under-predicted cells are set to +1, over-predicted cells are set to -1, correctly predicted flooded cells are assigned NaN and correctly predicted unflooded cells are set to 0. Mapping these categories creates a conventional contingency map, which combined (by element-wise array product) with an agreement scale map (Eq. (7)) creates a categorical scale map made by plotting the absolute agreement scale values coloured according to the contingency class. A categorical scale map shows a measure of spatial accuracy between two data arrays (Hooker et al., 2022). Categorical scale maps may be used as a basis for comparison between ensemble members and observations, as we illustrate with our case study in Section 4.5.3.

4.3.3 Ensemble spatial spread-skill evaluation

We assume that each ensemble forecast flood map represents an equally likely future scenario and the evaluation of the full ensemble is needed to quantify the uncertainty and to evaluate the spatial spread-skill relationship. The ensemble flood map spatial characteristics vary with location and in order to preserve the location dependent information, we utilise a method developed to evaluate a convective ensemble precipitation forecast (Dey, Roberts, et al., 2016; Dey, Plant, et al., 2016). Here, we outline the method and describe a new application to evaluate an ensemble forecast flood map.

A neighbourhood approach (Section 4.3.2) is used to assess the spatial agreement scale $S_{ij}^{A(F_1F_2)}$ or measure of similarity at each grid cell location (i, j) between each unique pair of ensemble flood maps. For an ensemble of M members, there are

$$M_p = \frac{M(M-1)}{2}, \tag{4.8}$$

unique pairs (e.g., 1275 pairs for a 51 member ensemble). For an ensemble, the skillful scale can be renamed as a *believable scale*, which is the scale where ensemble members become sufficiently similar to observations such that they are a useful prediction. Every paired ensemble agreement scale field is averaged at each grid cell to produce a mean field, from the agreement scale field defined in Eq. (7)

$$S_{ij}^{A(\overline{mm})} = \frac{1}{M_p} \sum_{F_1=1}^{M-1} \sum_{F_2=F_1+1}^{M} S_{ij}^{A(F_1F_2)} \tag{4.9}$$

indicating the location specific believable scales of the forecast flood map ensemble. Maps of $S_{ij}^{A(\overline{mm})}$ summarise the spatial spread of the full ensemble. Each of the agreement scale fields between the ensemble members and the observations are also averaged at each grid cell to give

$$S_{ij}^{A(\overline{mo})} = \frac{1}{M} \sum_{f=1}^{M} S_{ij}^{A(F_0)}. \tag{4.10}$$

A measure of the spatial spread-skill of the ensemble can be found by comparing the average agreement scale between the ensemble members $S_{ij}^{A(\overline{mm})}$ representing the ensemble *spread* with the average agreement scale between the ensemble members and the observed flood field $S_{ij}^{A(\overline{mo})}$ representing the ensemble *skill*.

4.3.4 Spatial spread-skill visualisation methods

To evaluate the spatial spread-skill relationship, $S_{ij}^{A(\overline{mm})}$ (representing the ensemble *spread*) must be compared in the same location as $S_{ij}^{A(\overline{mo})}$ (representing the ensemble *skill*). Data arrays can be visually compared using a binned scatter plot that averages across a selected bin of cells at the same location within the domain. Dey, Roberts, et al. (2016) demonstrated for an idealised example that by plotting $S_{ij}^{A(\overline{mm})}$ against $S_{ij}^{A(\overline{mo})}$ as a binned scatter plot in order to preserve the spatial location of the comparison (Fig. 4.1), the ensemble can be classified as over-, under- or well-spread. The ensemble is deemed to be *well-spread*

at a specific location in the domain of interest when the spread of the individual members represented at each grid cell by $S_{ij}^{A(\overline{mm})}$ equals the skill of the ensemble represented at each grid cell by $S_{ij}^{A(\overline{mo})}$, i.e. $S_{ij}^{A(\overline{mm})} - S_{ij}^{A(\overline{mo})} = 0$. The result would lie on a 1:1 line on the binned scatter plot. Where the spread between the ensemble members exceeds the skill of the ensemble forecast i.e. $S_{ij}^{A(\overline{mm})} > S_{ij}^{A(\overline{mo})}$ the ensemble is considered to be *over-spread* and the binned scatter plot will lie beneath the 1:1 line. The converse is true for an *under-spread* ensemble forecast where the agreement between members, the spread, is less than the agreement between the ensemble and the observations, the skill. Here, $S_{ij}^{A(\overline{mm})} < S_{ij}^{A(\overline{mo})}$ and the binned scatter plot would lie above the 1:1 line.

To summarise the spread-skill relationship we develop this visualisation further by plotting a hexagonal binned 2D histogram plot (an example hexbin plot is presented in Section 4.5.3). The domain is divided into a (pre-determined) number of hexagons. Hexagons minimize the perimeter to area ratio and therefore minimize the edge effects. The hexbin histogram plot colour shade represents the number of data points within each bin.

Whilst the hexbin plot is useful for gaining an understanding of the general spread-skill relationship of the ensemble flood map forecast, it does not tell us specifically where in the domain the ensemble spatial predictability is better or worse. The *Spatial Spread-Skill* (SSS) map plots $S_{ij}^{A(\overline{mm})} - S_{ij}^{A(\overline{mo})}$ at every grid cell location so that the spread-skill is mapped across the domain and can be linked directly to different sub-catchments and surface features such as tributaries, embankments, bridges and importantly the underlying topography or DTM, which influence the derivation of the ensemble flood maps. Regions on the SSS map where the ensemble is over-spread are positive with negative areas indicating where the ensemble is under-spread, zero values show a well-spread ensemble. Note that this does not necessarily mean that the entire ensemble is in agreement with

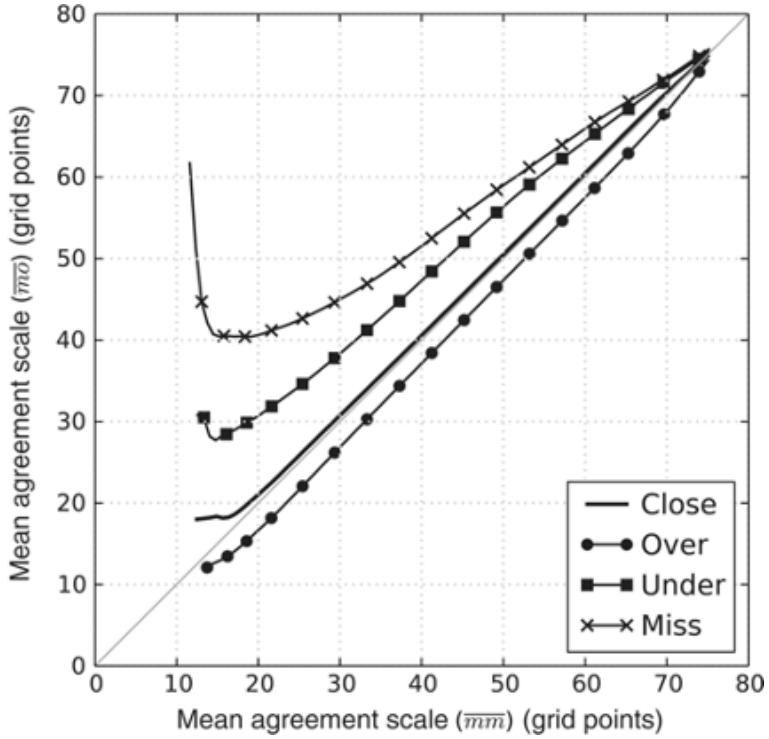


Figure 4.1: Figure reproduced with permission from Dey et al., (2016) showing results on a binned scatter plot from an idealised experiment indicating the spatial spread-skill relationship between an ensemble forecast and the observation.

observations at grid level, but that the agreement scales between $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are equal. (An example SSS map is presented in Section 4.5.3).

4.4 Ensemble forecasting flood event case study

In this section we describe an example flooding event used to demonstrate the application of the spatial spread-skill evaluation approach. We evaluate a 1-day lead time flood inundation 51 ensemble member forecast from the Flood Foresight system (Section 4.4.2) for the domain area against a satellite SAR-derived flood map (Section 4.4.3).

4.4.1 Brahmaputra flood, Assam India, August 2017

The origin of the Brahmaputra River (also known as the Yarlung Tsangpo in Tibetan, the Siang/Dihang River in Arunachali, Luit in Assamese, and the Jamuna River in Bangladesh) lies in the Himalayan Kailas Range of southwestern Tibet, China. Draining an area of 543,000 km^2, the Brahmaputra flows for 2000 km across the Tibetan Plateau and a further 1000 km parallel to the Himalayan foothills through the Assam Valley, India before entering Bangladesh where the Brahmaputra joins the Ganges River (Palash et al., 2020). The Brahmaputra baseflow originates from the upstream glacial snow melt, however the streamflow rates are dominated by the summer monsoon precipitation. The basin receives up to 95% of its annual rainfall during the pre-monsoon and monsoon season, which usually runs from April to September and causes annual flooding of the Brahmaputra. The Assam region typically records on average 2300 mm of annual rainfall and up to 5000 mm in the Himalayan foothills (Dhar & Nandargi, 2000, 2003).

For this example case we focus on the third wave of flooding that occurred during the monsoon season in August 2017, peaking around the 12^{th}. Figure 4.2 shows the location of the Brahmaputra and of a chosen domain centred upon some of the worst flooding that occurred. This area includes a confluence zone where the Subansiri River meets the Brahmaputra. The monsoon flooding impacted an estimated 40 million people across India and Bangladesh. Locally in the Assam region, the flooding in August affected over 3.3 million people and approximately 3200 villages, river embankments were damaged in 11 districts. Over 14,000 people were evacuated to one of around 700 relief camps that were also needed to house over 180,000 people relocated (Floodlist, 2017). The local Assam State Disaster Management Authority (ASDMA, 2017) flood early warning system issued a low warning alert (disasters that can be managed at the district level) on the 10^{th} August for the district.

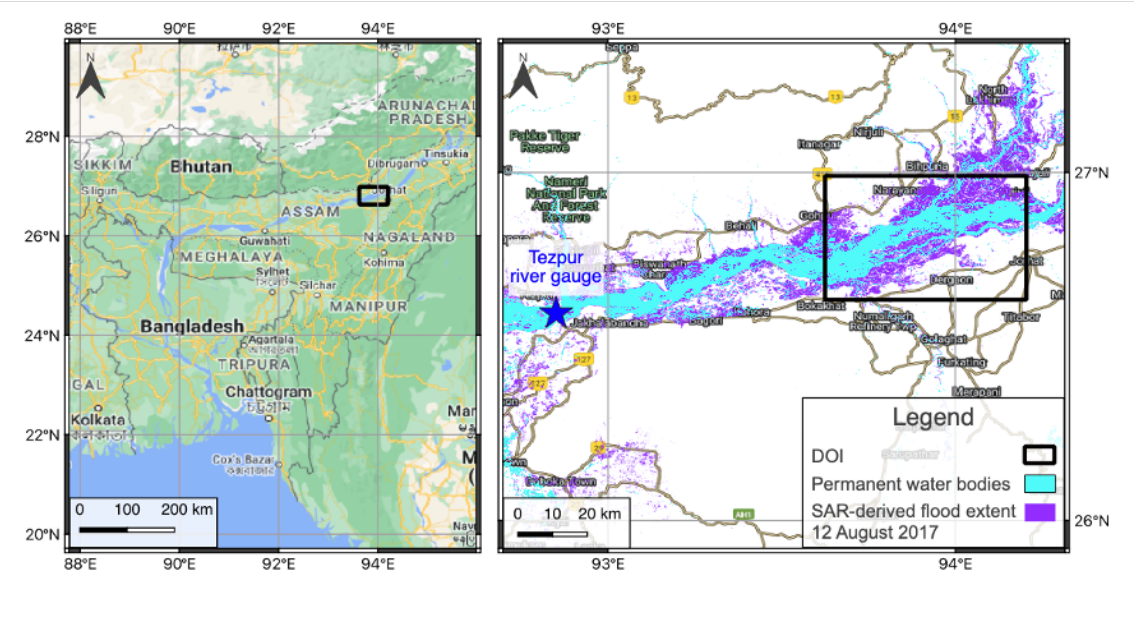


Figure 4.2: Left panel: domain location on the Brahmaputra River in the Assam region of India. Domain size is 57.5 km by 39.3 km. Right panel: Sentinel-1 SAR-derived flood map and permanent water bodies from the JRC Global Surface Water database for the domain of interest (DOI). Base map from ©Google Maps.

In 2017, the southwest monsoon season rainfalls were predicted to be *normal* by the South Asian Climate Outlook Forum (WMO, 2017). However, the pre-monsoon season began early in the year with heavy thunderstorms affecting the region from March onwards. In the Assam region, June and July were 60% wetter than the previous three years and during August more locally intense rainfall was recorded compared with historical observations (Palash et al., 2020). In higher latitude areas, 30 km to the north of the domain at North Lakhimpur, 215.8 mm rainfall was recorded in the three days prior to the flood peak (Floodlist, 2017; Hossain et al., 2021). An above normal flood situation is declared in India where the river water level exceeds the Warning Level, a severe flood occurs where the water level exceeds the Danger Level, and an extreme flood occurs where the previous Highest Flood Level is exceeded (Central Water Commission, 2023). The peak water level recorded downstream at Tezpur (Danger Level 65.23 m) on August 14$^{\text{th}}$ was 66.12 m. There are regional variations in maximum water levels reported, with upland

regions to the north of the Assam valley recording water levels that exceed the previous Highest Flood Level indicating an extreme flood level (Floodlist, 2017).

4.4.2 Ensemble flood forecasting system

The Flood Foresight system (Fig. 4.3), developed and operationally run by JBA Consulting, is a fluvial flood inundation mapping system that can be implemented at any river basin around the world. Flood Foresight utilises a simulation library approach to generate real-time and forecast flood inundation and water depth maps. The simulation library approach saves valuable computing time and allows the application of Flood Foresight in near continuous real-time at national and international scales. A pre-computed library of flood maps for a river basin or country are created using JFlow®(where a DTM is available), (Bradbrook, 2006) and RFlow (where a DTM is unavailable). JFlow uses a raster-based approach with a detailed underlying digital terrain model (DTM) and a diffusion wave approximation of the full 2D hydrodynamic shallow water flow equations. RFlow combines a 1D model based upon Normal Depth calculations, optimised for use on a Digital Surface Model (DSM, NEXTmap (2016)) with rapid 2D flood spreading (created by spreading Normal Depth from upstream to downstream) and is calibrated against JFlow. These equations capture the main controls of the flood routing for shallow, topographically driven flow. Six flood maps at 30 m resolution are created for 20, 50, 100, 200, 500 and 1500 year return period flood events (corresponding to annual exceedance probabilities (AEPs) of 5%, 2.5%, 1%, 0.5% and 0.2% and 0.07% respectively). Between each adjacent pair of modelled return period maps, five additional intermediate flood maps are created by linear interpolation of both flood depth and extent. An additional five flood maps are also created beneath the lowest return period flood map. This gives, in total, a library of 36 flood maps. Note that these flood maps are undefended and local temporary flood defences are not included. Flood foresight is set up for a region by dividing the river

basin into sub-catchments using the HydroBASINS data-set (level 12) (Lehner, 2014a). Flood Foresight takes gridded inputs of ensemble forecast streamflow and uses these to select the most appropriate flood map for each sub-catchment. These are mosaicked together and forecasts of ensemble flood maps are produced daily, out to ten days ahead.

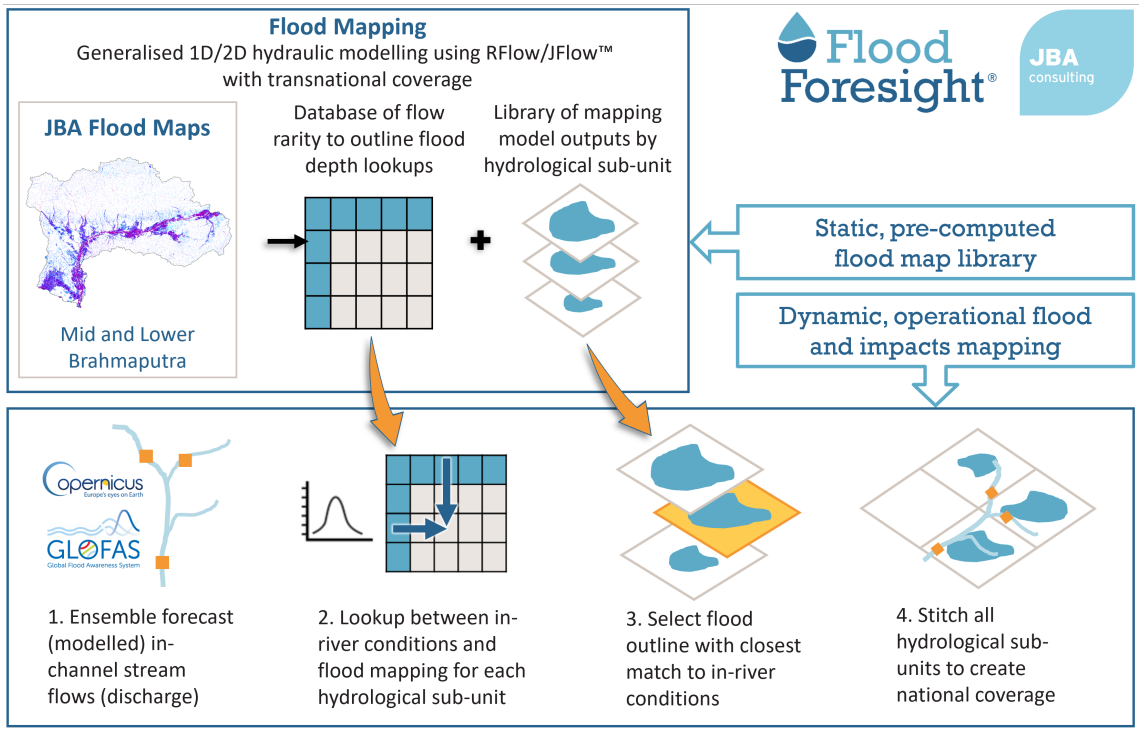


Figure 4.3: Flood Foresight ensemble forecast flood inundation and impact mapping work flow. Prepared by JBA Consulting.

The global (non UK and Ireland) configuration of Flood Foresight uses ensemble streamflow forecast data from the Global Flood Awareness System (GloFAS) (Alfieri et al., 2013; GloFAS, 2021). GloFAS was jointly developed by the European Commission and the European Centre for Medium-Range Weather Forecasts (ECMWF) and is composed of an integrated hydro-meteorological forecasting chain that couples state-of-the-art weather forecasts with a land surface and hydrological model. With its continental scale set-up, GloFAS provides downstream countries with forecasts of upstream river conditions

up to one month ahead as well as continental and global overviews for large world river basins. Meteorological forecast data are provided by the ECMWF Ensemble (IFS) model, the operational (51 member) ensemble weather forecasting product of the ECMWF. The meteorological forecast data provide inputs to the land surface module, HTESSEL (Hydrological Tiled ECMWF Scheme for Surface Exchange over Land). HTESSEL simulates the land surface response to the meteorological data, based on simulated interactions with soil conditions, idealised vegetation cover and land cover. From these simulations, HTESSEL outputs forecast global surface and sub-surface flows per grid cell. These simulated flows are then used by a simplified version of the hydrological model LISFLOOD, a 1D routing model which simulates the movement of the surface and sub-surface flows. The runoff data produced is routed through a representation of the river network using a double kinematic wave approach, which includes bankfull and over bankfull routing. The river network used is taken from the HydroSHEDS data-set (Lehner & Grill, 2013).

GloFAS outputs a gridded (approximately 10 km spatial resolution) ensemble forecast of river streamflow (Fig. 4.4). Each of the GloFAS grid cells are linked to the sub-catchments in the Flood Foresight system. The simulation library flood maps are selected when the forecast streamflow exceeds a return period threshold level within each sub-catchment. The RP threshold levels are calculated using ERA5 reanalysis data (Harrigan et al., 2020). Each ensemble member flood map forecast is created by aggregating the individual sub-catchment maps. In summary, the meteorological IFS 51 member ensemble input to the flood forecasting chain allows atmospheric evolution uncertainties to be represented within the ensemble streamflow forecast and the ensemble of inundation flood maps, thus creating a probabilistic flood map forecast, indicating the likelihood of flooding. Flood foresight produces daily ensemble flood depth and extent forecasts at 30 m spatial resolution out to 10-days.

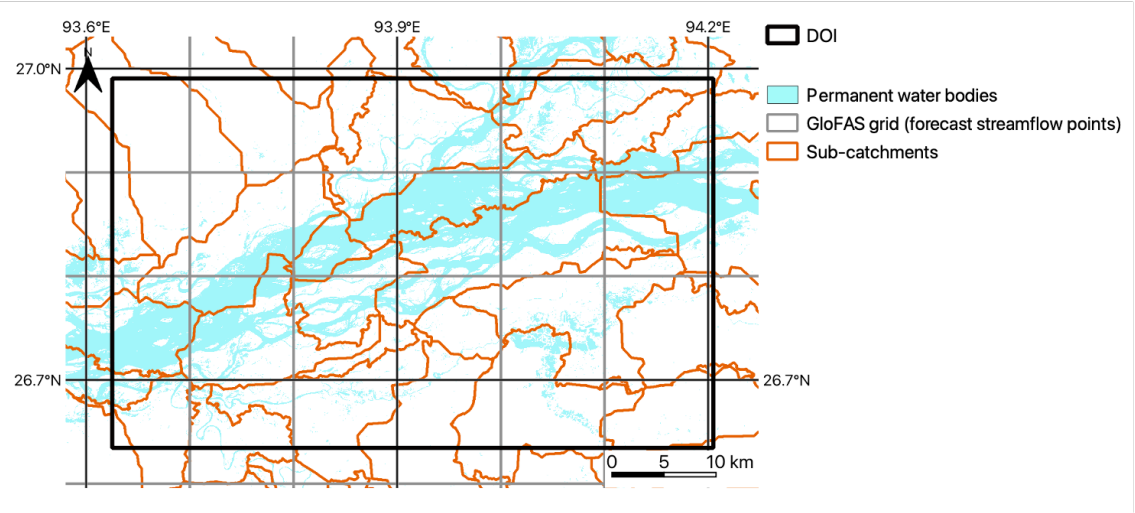


Figure 4.4: GloFAS grid, permanent water bodies and Flood Foresight sub-catchments for the domain of interest (DOI).

4.4.3 SAR-derived flood map

A Sentinel-1 (S1A) image was acquired in interferometric wide swath mode (swath width 250 km) around the time of the flood peak at 17:18 (IST) on the 12th August 2017. The ESA Grid Processing on Demand (GPOD) HASARD service (service terminated June 2021) was utilised to map the flooding. The flood mapping algorithm (Chini et al., 2017) uses an automated, statistical, hierarchical split-based approach to distinguish between two classes (background and flood) using a pre-flood and flood image. A pre-flood image (February 2017) from the same satellite sensor and track was used to derive the flood map (Fig. 4.2). Original SAR images (VV polarisation) were pre-proccesed, which involved: precise orbit correction, radiometric calibration, thermal noise removal, terrain correction, speckle reduction and re-projection to the WGS84 coordinate system. The HASARD mapping algorithm removes permanent water bodies that are detected on the pre-flood image, such as the unflooded river water, lakes and reservoirs by applying a thresholding approach. Flooded areas beneath vegetation, bridges and near to buildings will not be detected using this method. Flood Foresight forecast flood maps include the river channel

and exclude surface features such as vegetation and buildings. To smooth the HASARD flood maps and allow a fairer comparison we apply a morphological closing operation (without impacting the location of the flood extent) to flood fill vegetation and buildings. The wide and braided Brahmaputra River in the Assam region covers a significant area of the selected domain. In order to evaluate the flood prediction accuracy alone, the pre-flood occurrence of surface water using the JRC Global Surface Water database (Pekel et al., 2016) has been removed from the Flood Foresight forecast inundation maps. The observed flood extent derived from satellite based SAR data at 20 m grid size is re-scaled to match the forecast flood map grid size (30 m) using average aggregation. The closest available (cloud free) optical image available was a Sentinel-2 image on the 17$^{\text{th}}$ August 2017, 5 days after the SAR image acquisition. During this time the flood waters had receded from their peak, which makes this unsuitable for comparison with the SAR-derived flood map. Since no other validation sources are available, for the purposes of this study we assume that the SAR-derived observation of flooding represents the true flood extent. From October 2021, Sentinel-1 SAR images are processed by CEMS GFM (GFM, 2021) to derive flood extent and provide an uncertainty estimate of the grid cell classification. This means uncertainty information in the SAR-derived flood map could be accounted for in future evaluation studies by verifying across different levels of observation uncertainty. Additionally, a flood mask, indicating areas where flood detection using SAR data is not currently possible (at the Sentinel-1 spatial resolution) could be used to exclude areas from the evaluation process (note that this was not possible for this case study, since this information was not available in 2017).

4.4.4 Forecast data

Flood Foresight was set-up for the Brahmaputra basin in India and Bangladesh using the simulation library approach to flood mapping described in Section 4.4.2. Flood maps were pre-computed for the domain of interest (Fig. 4.2) using a DSM and RFlow. The forecast

data for the Brahmaputra flood event contains a 51 member ensemble of flood maps indicating flood extent, produced at a 1-day lead time. Vertically stacking each individual ensemble member flood map and adding vertically across every grid cell combines all ensemble members into a single flood map (all flooded grid cells are set to 1) showing where flooding is possible across all members (ens_{all}). A spatial median flood map is created (ens_{median}) where 26 members or more predict flooding at a particular grid cell location. Each of the ensemble member flood maps for the domain are plotted in Figure 4.5 along with ens_{all}, ens_{median} and the SAR-derived flood map.

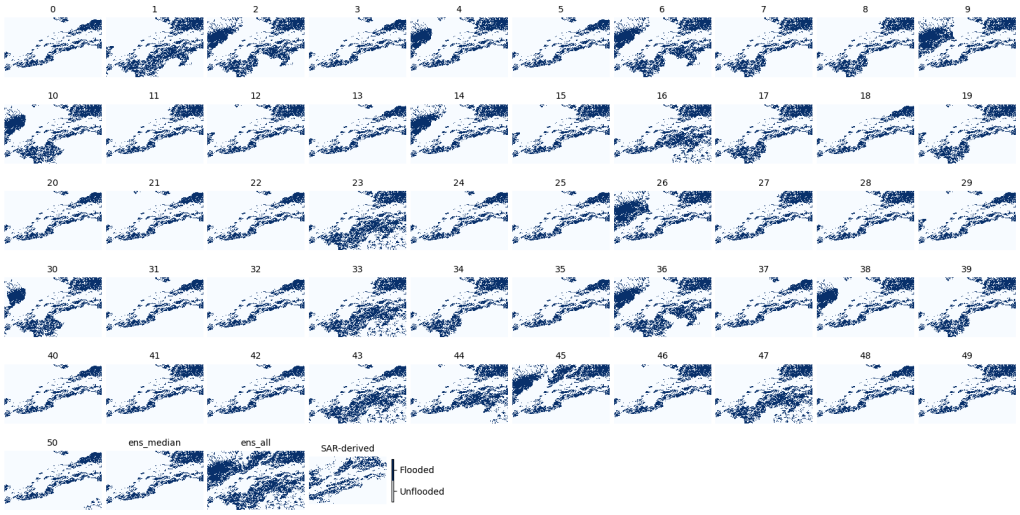


Figure 4.5: Brahmaputra River, Assam region, August 2017. 51 ensemble member forecast flood maps (labelled 0 to 50), ens_{median} and ens_{all} all at 1-day lead time and the Sentinel-1 SAR-derived flood map.

Figure 4.6 shows the amalgamated probabilistic ensemble forecast indicating the probability of flooding at each grid cell location. This was produced by vertically stacking each ensemble member flood map and adding vertically the number of flooded cells at each grid cell location across all ensemble members. The total is divided by 51 to calculate

the probability. Dark blue colours near to the central river channel indicate agreement between all ensemble members and 100% forecast probability of flooding, lighter colours to the north of the river indicate a low probability of flooding.

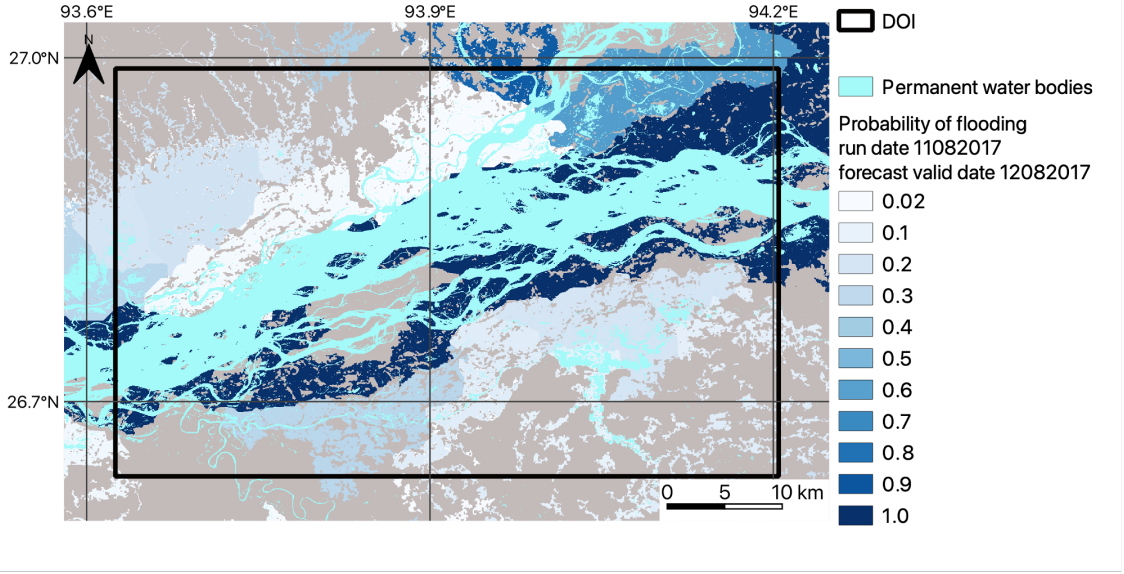


Figure 4.6: Brahmaputra River, Assam region, August 2017. Colour shading from white (low) to dark blue (high) indicate the forecast probability of flooding based on a 1-day lead time, 51 ensemble member flood map forecast for the Brahmaputra River in the Assam region, August 2017. (Note map background is grey)

4.5 Results and discussion

To demonstrate an application of the spatial scale approach to both ensemble forecast flood map evaluation of forecast skill and the spatial spread-skill relationship, we apply the methods outlined in Section 4.3 to the flooding case described in Section 4.4.1. First, in Section 4.5.1 we verify the full ensemble using a spatial scale approach to calculate a skilful scale of agreement between each ensemble member and the SAR-derived flood map (Fig. 4.2) along with the combined ensemble (ens_{all}) and the ensemble spatial median (ens_{median}). We evaluate the location specific spatial skill of the ensemble by calculating

categorical scale maps (Section 4.5.2) for ens_{all}, ens_{median} and a best and worst case ensemble members determined by the skilful scale calculated in Section 4.5.1. In Section 4.5.3 we evaluate the spatial predictability of the full ensemble and show this on the *Spatial Spread-Skill* (SSS) map, indicating regions where the ensemble is over-, under- or well-spread.

4.5.1 Ensemble spatial scale evaluation

Here we investigate how a scale-selective approach can be useful for extracting meaningful information from a flood map ensemble forecast where multiple forecast flood maps represent equally likely flooding scenarios (Fig. 4.5). A minimum skilful scale (where $FSS > FSS_T$) has been calculated for each individual member flood map, ens_{all} and ens_{median}. The results in Figure 4.7 show that individual ensemble member spatial skill varies considerably with FSS at grid level ranging from 0.35 to 0.59. One member ens_1, which would usually be disregarded as an outlier due to its low probability, outperformed all other members significantly with a skilful scale achieved at a neighbourhood size of $n = 3$. The combined ens_{all} showed more skill at grid level ($n = 1$) and smaller neighbourhood sizes compared with ens_{median}, both however exceeded FSS_T at $n = 41$, or 615 m. At neighbourhood sizes greater than $n = 41$, ens_{median} outperformed ens_{all}. There is a cluster of members showing similar skill to ens_{median} and ens_{all} and a second cluster, with more ensemble variation but indicating lower skill than the first cluster. The ens_{median} and ens_{all} flood maps outperform the second cluster, however there are individual members with a higher spatial skill score compared to ens_{median} and ens_{all}. These results show that all ensemble member flood maps, including outliers, should be considered individually as possible future flooding scenarios. Spatial variations across individual ensemble members (see Fig. 4.5 ens_1 compared to ens_{median}) indicate that it is not meaningful to consider only the ensemble median flood map to represent the information within the full ensemble.

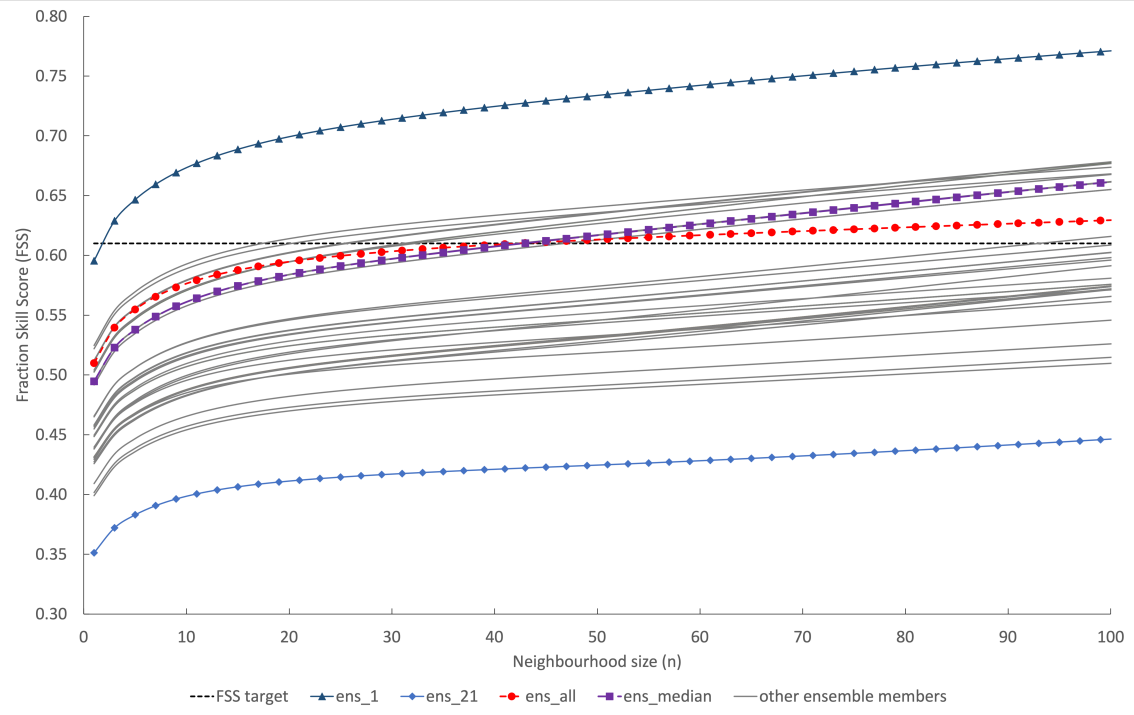


Figure 4.7: The spatial skill of each individual ensemble member forecast flood extent is evaluated along with the ens_{median} (a spatial median where 26 or more members predict flooding at a grid cell location) and ens_{all} (flooded grid cells from all ensemble members are combined). The FSS is calculated at increasing neighbourhood sizes to determine the scale at which the forecast becomes skilful at capturing the observed flood (FSS_T).

4.5.2 Ensemble spatial predictability

The scale-selective skill scores calculated for different aspects of the ensemble forecast give a domain-averaged score and skilful scale. To understand location specific spatial predictability of the ensemble forecast, categorical scale maps are calculated and presented in Figure 4.8. These show how the agreement scale (Section 4.3.2) varies with location for (a) ens_{all}, (b) ens_{median}, (c) ens_1, the 'best' performing ensemble member and (d) ens_{21}, the 'worst' performing ensemble member. The ensemble summary map, ens_{all} (Fig. 4.8 (a)) captures most of the observed flooding (in grey) with small regions of under-prediction (red). However, as you might expect to see by including every potential flooding realisation there are significant regions of over-prediction (blue) or false alarm. The region of over-

prediction to the south of the river is less evident in the ens_{median} categorical scale map (Fig. 4.8 (b)) which performs worse to the north by under-predicting flooding here. This flooding is captured well by ens_1 (Fig. 4.8 (c)) and in particular close to a confluence zone where the Subansiri River joins the Brahmaputra (grid cell location (1100, 250)). This ties in with the high rainfall totals accumulated just to the north of this region associated with localised very heavy rainfall (Floodlist, 2017). A region of under-prediction at grid cell location (750, 750) is missed by all members. In future work, a closer inspection of the DTM or surface features included/excluded in the hydraulic modelling, such as embankment heights, may indicate how this modelling could be improved. The 'worst' performing ensemble member ens_{21} (Fig. 4.8 (d)) accurately predicts flooding closer to the river channel, however under-prediction to the north along with over-prediction to the south show where the forecast was inaccurate. Categorical scale maps enable different ensemble flood map presentations to be evaluated so that the most useful presentation method can be determined for a particular flooding situation.

4.5.3 Ensemble spatial spread-skill

To evaluate the location specific skill of the full ensemble, one option would be to calculate 51 categorical scale maps from each individual member flood map (Fig. 4.5). This approach maintains the spatial detail held within each of the ensemble member flood maps, although does require multiple visual comparisons to be made by the flood forecaster or modeller, which takes time and effort. Making comparisons across the different ensemble member flood maps in Figure 4.5 provides a demonstration of these forecasting difficulties. Further, the categorical scale maps do not evaluate the ensemble spatial spread. To address this, we develop a Spatial Spread-Skill (SSS) map (derived from Fig. 4.9, presented in Fig. 4.10) showing the spread-skill of the full ensemble forecast and keeping the location specific detail. All ensemble members are included in this analysis which evaluates both the spatial skill and the ensemble spatial spread of the forecast against the remotely

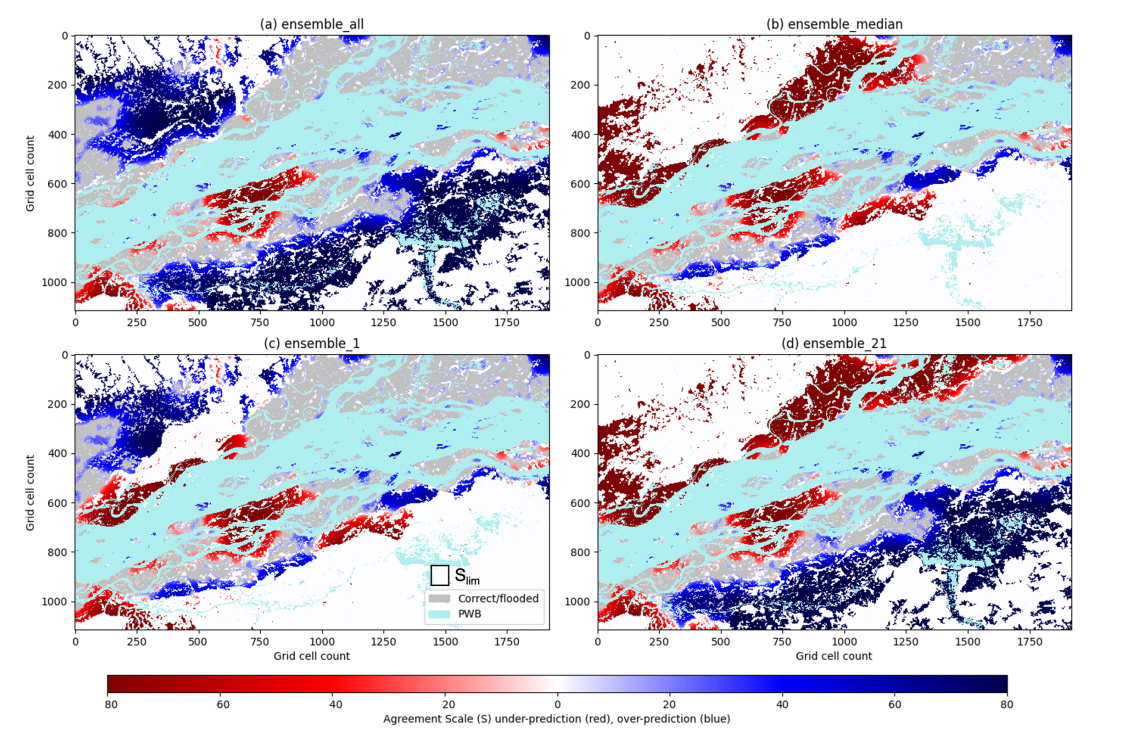


Figure 4.8: Brahmaputra River, Assam region, August 2017. Categorical scale maps for (a) ens_{all} (flooded grid cells from all ensemble members are combined), (b) ens_{median} (a spatial median where 26 or more members predict flooding at a grid cell location), (c) individual ensemble member 1 and (d) individual ensemble member 21. Red areas indicate where the forecast is under-predicted and blue regions represent over-prediction. The colour shade gives the scale of agreement (Eq. (7)) between the forecast and the observed flooding with lighter shading indicating a smaller agreement scale is required to reach the agreement criterion (Eq. (6)), a fixed maximum scale S_{lim} is drawn to scale (c). For georeferencing see Figure 4.6, each grid cell is 30 m x 30 m.

observed flood extent.

Figure 4.9 shows how the average-ensemble/ensemble-agreement scale in (a) $S_{ij}^{A(\overline{mm})}$ calculated at each grid cell (representing ensemble *spread*) compares with the average ensemble/observed scale in (b) $S_{ij}^{A(\overline{mo})}$ (representing ensemble *skill*) along with the hexbin scatter plot in (c) which compares (a) and (b) to indicate the spatial spread-skill of the forecast. The hexagonal tessellation is used so that the distances along the hexbin diagonal are on the same scale as those along the x and y-axis. For a perfect ensemble

forecast the average agreement scale between ensemble members should match the agreement scale between the ensemble forecast and observed flood map, i.e. they should align along the 1:1 line. The SSS map plots the difference between the ensemble/ensemble and the ensemble/observed average agreement scales at each grid cell (Fig. 4.10) and indicates where the spatial spread-skill is over-, under-, or well-spread. Three numbered areas (Fig 4.9(a)) identify three different ensemble spread-skill relationships. Area 1 shows that the agreement between ensemble members is close, but that they disagree with the observed flood extent. This is displayed in orange shades as an under-spread or miss region on the SSS map, Figure 4.10. This is the region close to the confluence area described in Section 4.5.2. Recall that in this region, most ensemble members did not predict the flooding that occurred with the exception of one ensemble member (ens_1). In area 2 on Figure 4.9, both (a) and (b) are in agreement at grid level, which indicates the ensemble is well-spread; these are shown in white on Figure 4.10. Away from the miss and well-spread regions in Figure 4.9, the overall visual impression is that the ensemble spread-skill lies below the 1:1 line and is over-spread, indicated by area 3. This corresponds to purple shading on the SSS map (Fig. 4.10). Overall Figure 4.9 tells us that the spread-skill relationship for this example case study is not uniform across the domain but is in fact location specific. The areas identified (1, 2 and 3) lie within different sub-catchments, which are linked to different GloFAS grid cells, driving the ensemble flood map selection for each sub-catchment. Inferences can be made about the spread-skill of the driving discharge data at sub-catchment level across the domain. Using the spatial spread-skill relationship shown on the ensemble SSS map we can infer how well the ensemble forecasting system encompasses the multiple sources of uncertainty and how meaningful the probabilistic ensemble forecast of flood inundation actually is. An ensemble flood map forecast that is well-spread suggests that the probabilistic forecast is meaningful. The SSS map is a useful evaluation tool for validating flood forecasts in un-gauged or partially gauged rivers. A simulation library approach, like the Flood Foresight maps used here, relies on the accuracy of the return period thresholds

set, the (ensemble) forecast streamflow and the accuracy of the flood inundation map for a given streamflow. The forecast evaluation approaches presented here enable these system attributes to be evaluated even where observed streamflow is limited or erroneous. The SSS map summarises the whole ensemble, which makes it useful for forecasters attempting to convey uncertainty information to decision makers, highlighting regions where there is high/low confidence in the forecast.

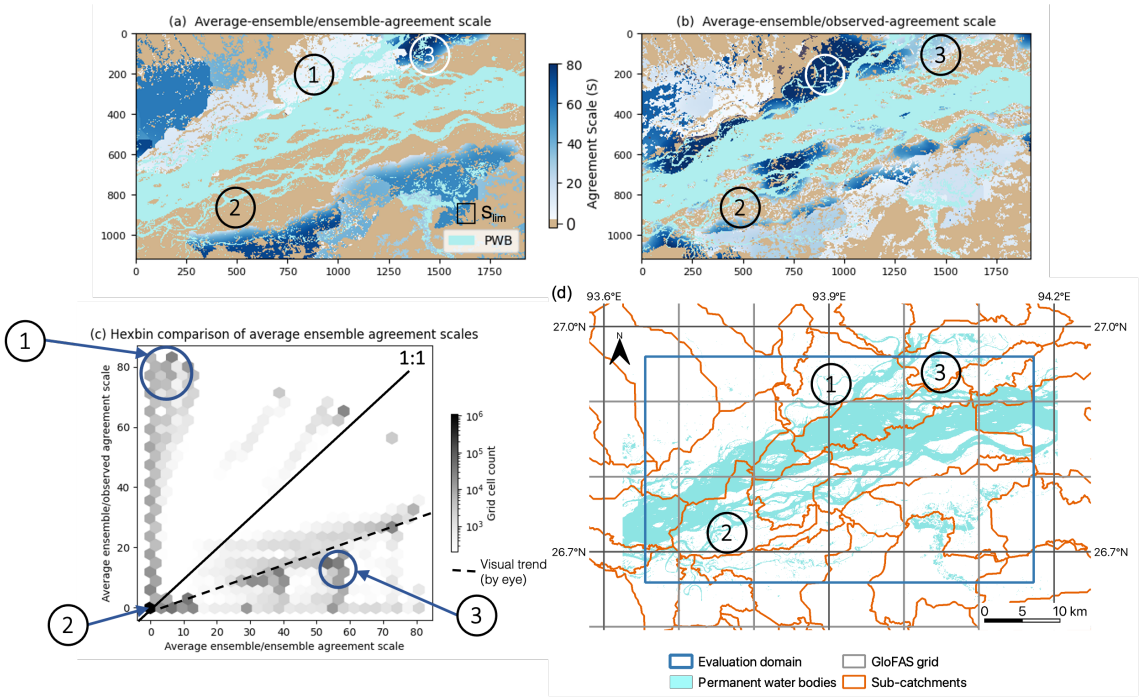


Figure 4.9: Brahmaputra River, Assam region, August 2017. (a) The average agreement scale map of each unique pair of forecast ensemble flood maps and (b) between each ensemble member compared against the observed SAR-derived flood map. (c) A binned histogram scatter plot compares (a) and (b) to indicate the spatial spread-skill of the forecast ensemble. (d) indicates the corresponding sub-catchment locations. Areas labelled (1, 2 and 3) are discussed in Section 4.5.3. A fixed maximum scale S_{lim} (Eq. (6)) is drawn to scale (a). Note PWB means permanent water bodies.

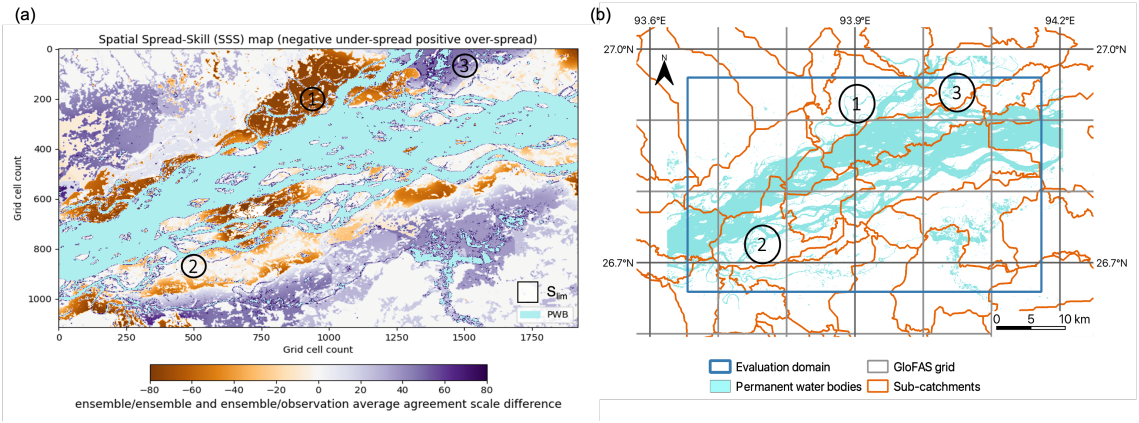


Figure 4.10: Brahmaputra River, Assam region, August 2017. (a) The Spatial Spread-Skill (SSS) map shows the difference between the ensemble/ensemble and the ensemble/observed average agreement scales at each grid cell. Negative values (orange) indicate where the ensemble is under-spread and positive values (purple) indicate where the ensemble is over-spread. White areas areas indicate where the average agreement scales match and indicate good spatial spread-skill. (d) Indicates the corresponding sub-catchment locations. Areas labelled (1, 2 and 3) are discussed in Section 4.5.3. A fixed maximum scale S_{lim} (Eq. (6)) is drawn to scale (a). Note PWB means permanent water bodies.

4.6 Conclusions

Differences between ensemble members in ensemble forecast flood map systems are mostly driven by initial condition perturbations at the top of the hydro-meteorological forecast chain within the numerical weather prediction system. Presently, there is limited understanding or evaluation of how these meteorological uncertainties link to mapped flooding predictability, which involves additional sources of uncertainty. An evaluation of the spatial predictability and the spread-skill relationship of the ensemble flood map forecast provides an improved understanding of the performance of the forecast system. Uncertainties in other parts of the forecast chain are not truly represented by the ensemble flood maps and evaluating the spatial spread-skill of the flood maps is important for understanding the likelihood of flooding that the ensemble flood maps capture. In this paper, we present a new scale-selective approach to assess the spatial predictability and spread-skill of an ensemble flood map forecast by comparing against a satellite SAR-derived observation of flood extent. By calculating a skilful scale at each grid cell for every unique ensemble member pair we can determine the ensemble *spatial spread*, and between every ensemble member and the SAR-derived flood map we can determine the ensemble *spatial skill*. The hexbin scatter plot summarises the spread-skill relationship so that a trend across the whole domain can be assessed. The difference between these skilful scales can be mapped onto the Spatial Spread Skill (SSS) map which shows for each specific location in the domain whether the ensemble is over-, under- or well-spread. The methods are applied to an example flooding event of the Brahmaputra River in the Assam region of India in August 2017.

In operational practice there are multiple options of ensemble flood map presentation type such as presenting the ensemble median or other exceedance probability for delivery to end-users and decision makers. An important aspect of developing an inundation flood

forecasting system is to determine the most useful way to present a spatial ensemble forecast. Using a scale-selective approach we have evaluated the performance of individual ensemble members, a combined total ensemble and the spatial ensemble median compared to a SAR-derived observation of flood extent. Other options could be to exclude ensemble member outliers, to spatially cluster similar ensemble members into groups of flood extent or to present a most likely, best and worst case ensemble flood map. Whichever presentation method is chosen, this should be fully explored using the spatial spread-skill methods described here to evaluate the ensemble performance of historical flooding events. We found for this example flooding event that one ensemble member significantly outperformed the combined and median flood maps and that potentially in some flood forecasting scenarios this member would have been excluded as an outlier. The categorical scale maps show the ensemble spatial median could miss vital flooding information and that all members should be considered as potential future flooding scenarios.

Through mapping the spatial-spread skill relationship, which varies with location, links can be made between the spatial variations in spread-skill and the physical characteristics of the flooding event. We found that one ensemble member outperformed all others in a region close to a confluence zone and nearby observed heavy rainfall. The region correlates to an area of under-spread ensemble members indicating that not enough members were predicting flooding here. Future studies could investigate the physical processes further using the methods presented here. The ensemble flood map spatial spread-skill could be investigated in the context of a particular physical process (such as rainfall intensity/location or an improved aspect of the hydrological model such as antecedent soil moisture) and how these uncertainties translate to the probabilistic flood map forecast. An understanding of the spatial predictability is particularly important for un-gauged catchments where the calibration of both forecast streamflow and return period thresholds (used to select the simulation library flood map) are rarely practiced routinely. Ideally, in operational prac-

tice, these spatial verification approaches including the categorical scale and SSS maps could be calculated and stored routinely as flooding events coincide with SAR-derived or other remotely observed flood maps to build up a verification catalogue/database. This database could then be used to investigate the spatial spread-skill model performance under different scenarios such as forecast lead time, month or season, or flood type. More locally, the impact of an improved DTM or the inclusion of a Digital Surface Model (DSM) or other surface features in the hydraulic model such as embankments could be considered. Over time, such a database would improve our understanding of the spatial predictability of an ensemble flood map system and how well the uncertainties present are represented by the ensemble forecast.

4.7 Chapter summary

In this chapter, we extend the scale-selective verification methods described in Chapter 3 with a new application to an ensemble flood map forecast. By considering the spatial spread between ensemble member flood maps, and between each ensemble member and a SAR-derived flood map, we evaluate the spread-skill of an ensemble flood map forecast. The spatial uncertainty held within multiple ensemble flood maps can be summarised onto a single map indicating the forecast performance at specific locations. For the case study evaluated, we found that individual ensemble member's forecasts can be more accurate than an aggregated forecast such as the ensemble median. This is an important consideration when choosing how to present ensemble flood map information in operational practice.

Chapter 5

A multi-system comparison of forecast flood extent using a scale-selective approach

In this chapter we address the third research question outlined in Chapter 1; How can a scale-selective approach be applied to evaluate multiple flood forecasting systems?:

- How can we evaluate the performance of flood forecasting systems predicting flood inundation extent at different spatial scales?
- What can we learn about the flood forecasting system performance and how does each compare?

The remainder of this chapter (except for the chapter summary, Section 5.9), has been published and is reproduced from (Hooker et al., 2023b).

5.1 Abstract

Fluvial flood forecasting systems increasingly couple river discharge to a flood map library or a real-time hydrodynamic model to provide forecast flood maps to disaster management teams and humanitarian agencies. The forecast flood maps can be linked to potential impacts to inform disaster risk reduction schemes, such as forecast-based financing. The success of forecast-based financing is dependent upon the accuracy of the forecast flood maps. We investigate a new application of scale-selective verification by evaluating the performance of three flood forecasting systems. Two simulation library systems, Flood Foresight (30 m) and GloFAS Rapid Flood Mapping (1000 m) and one hydrodynamically modelled system, the Bangladesh FFWC Super Model (300 m), all made predictions of flood extent at different spatial scales (grid lengths, shown in brackets) for the Jamuna River flood, Bangladesh, July 2020. These forecast flood maps are validated against Synthetic Aperture Radar-derived observations of flooding across four districts using a scale-selective approach that can compare directly across different spatial scales. Our results show that the simulation library system accuracy critically depends on the discharge return period threshold set to trigger a flood map selection and the number of hydrological model ensemble members that must exceed it. At short forecast lead times, the Super Model outperforms the other systems in three districts. Near to the Bangladesh border, the trans-boundary benefits of the two global systems are evident, with both outperforming the local model. We conclude that a scale-selective verification approach can quantify the skill of systems operating at different spatial scales so that the benefits and limitations can be evaluated. Multi-system comparison of flood maps is important for improving impact-based forecasts and ensuring funds and response activities are appropriately targeted.

HIGHLIGHTS:

- Through a new application of our scale-selective validation method we compare flood maps of different spatial scales (grid lengths) with SAR-derived observations of flooding.
- Three flood forecasting systems (two global ensemble and one local deterministic) operating during flooding in Bangladesh, July 2020 are evaluated.
- The return period threshold set and the number of ensemble members used to trigger a flood map from a simulation library proves crucial to the system performance.
- The results show the importance of accounting for spatial scale when interpreting skill scores in multi-system studies.

5.2 Introduction

Flood forecasting systems are increasingly used to improve preparedness ahead of a major flooding event (Stephens & Cloke, 2014a; Wu et al., 2020). One of the main action points from the recent Global Assessment Report (GAR2022) on Disaster Risk Reduction (DRR) is to 'design systems to factor in how human minds make decisions about risk' (UNDRR, 2022b). Whilst flood forecasting systems have improved significantly and continue to improve both globally and locally, the reliance on government departments and disaster managers to make the right decisions when faced with a potential crisis can result in inappropriate actions and unpreparedness (e.g., Fekete & Sandholz, 2021; Coughlan de Perez et al., 2022). The GAR2022 report shows that just 5.8% ($5.5 billion USD) of official development assistance contributes to disaster prevention and preparedness compared to 90.1% ($119.8 billion USD) for emergency response. Yet it has been demonstrated (for Europe) that financing for mitigation purposes such as flood forecasting systems can lead

to overall cost savings (Pappenberger et al., 2015).

Forecast-based financing (FbF) schemes can form a major element of DRR strategies and aim to directly link forecasts of extreme events to humanitarian actions (Coughlan de Perez et al., 2015, 2016). FbF schemes work by quantifying risks in advance of crises or disasters, prepositioning funds, and agreeing in advance how funds will be released based on forecasts, ahead of an event (OCHA, 2020). Anticipation and risk financing allows humanitarians to be better prepared by making important decisions before disaster strikes. These proactive decisions, directly linked to action (rapid fund release) remove the potential for reactive, incorrect decisions in the midst of a disaster. The success of the FbF system largely depends on the threshold triggers set and on the performance of the flood forecasting system at mapping the flood hazard.

Advances in flood forecasting systems both at global and local levels link together meteorological and hydrological forecasts to hydrodynamic models, simulating flood-wave propagation (Emerton et al., 2016; Wu et al., 2020; Apel et al., 2022). The resulting flood maps when directly linked to impacts can be used to inform DRR schemes. Multiple trade-offs exist in the development of such systems that inherently depend on observation data availability and computing power. These determine whether the maps can be modelled in real-time or are pre-calculated and form part of a simulation library; the spatial scale (grid size) of the forecast flood maps and whether the maps are deterministic or probabilistic (Savage et al., 2016). A recent review of flood inundation prediction (Bates, 2022) states that a key task to drive forward the development of better global hydraulic models will be more rigorous and comprehensive validation. Hoch and Trigg (2019) outline a Global Flood Model Validation Framework which includes a recommendation to routinely validate flood extent. Quantitative performance evaluation forms an important part of fitness-for-purpose assessment and continual system improvement. Currently, there

is limited quantitative validation of operational flood forecasting systems producing flood maps and operating in the same area. A recent advancement in flood map validation (Hooker et al., 2022) means that quantitative comparisons can be made across flood maps at different spatial scales, which makes a multi-system evaluation possible.

The accuracy of ensemble forecasts of flood extent can be verified by comparing with observations of flooding from unmanned aerial vehicles or satellite-based sensors. Satellite-based synthetic aperture radar (SAR) sensors are active, which means they can operate at night, through cloud and weather, and are well known for their flood detection capability (e.g., Horritt et al., 2001; Mason, Davenport, et al., 2012; Schumann et al., 2022). The SAR backscatter intensity depends on the smoothness of the surface, with unobstructed flooded areas returning low backscatter values. Recent techniques used to extract flood extent from SAR images have led to improved flood detection in urban areas (Mason et al., 2018, 2021a, 2021b). Since late 2021, SAR-derived flood maps are produced for every Sentinel-1 image detecting flooding around the world by the Global Flood Monitoring (GFM) service (EU Science Hub, 2021; GFM, 2021; Hostache et al., 2021), part of the Copernicus Emergency Management Service (CEMS) (Copernicus Programme, 2021). Within eight hours of the Sentinel-1 image acquisition, three flood detection algorithms are combined to give the flood class and uncertainty estimation per grid cell.

In this paper, through application of a new approach, we evaluate the inundation accuracy of three fluvial flood forecasting systems operating during severe flooding of the Jamuna River in Bangladesh, July 2020. The flood maps are compared against satellite SAR-derived flood extent. The systems are: Flood Foresight, a FbF system run by JBA Consulting working in partnership with the Start Network (Revilla-Romero et al., 2017); the Global Flood Awareness System (GloFAS) Rapid Flood Mapping (RFM) service (Alfieri et al., 2013; GloFAS, 2021) and the Bangladesh Flood Forecasting and Warning

Centre (FFWC) Super Model (BWDB, 2020), each producing forecast flood maps at different spatial scales (30 m, 1000 m and 300 m respectively). We investigate how a novel scale-selective spatial verification approach (Hooker et al., 2022) can be applied to multi-system studies where the forecast flood maps are presented at different spatial scales. Through applying this approach, we determine a skilful scale of each flood map that can be directly compared and discuss the benefits and limitations of each forecast system for flood mapping purposes that underpin the triggering of FbF schemes.

We describe the characteristics of the Jamuna River flood, July 2020 in Section 5.3. The three flood forecasting systems are described in Section 5.4 along with details of the SAR-derived observation of flooding. The scale-selective methods used to evaluate the forecast flood maps are outlined in Section 5.5. The performance of Flood Foresight with forecast lead time is presented in Section 5.6.1 and the multi-system flood map comparisons are presented in Section 5.6.2. In Section 5.6.3 we discuss the benefits and limitations of each system and conclude with recommendations in Section 5.7.

5.3 Flood event on the Jamuna River, Bangladesh July 2020

Bangladesh lies on the world's largest delta that drains the Tibetan plateau and the Himalayas to the north via the Brahmaputra and Ganges River systems. The river system capacity in Bangladesh is overwhelmed each monsoon season by the volume of water travelling through. Flooding is exacerbated in coastal regions where tidal surges from tropical cyclones impede the river drainage (Bernard et al., 2022). Due to its geographical location and the low lying, low slope nature of the land, Bangladesh is susceptible and vulnerable to flooding year after year and faces a worsening situation as a result of climate change and sea level rise (Hossain et al., 2021).

This study focuses on the Brahmaputra River (locally named the Jamuna River), which is characterised by braided, meandering channels that migrate continually due to frequent silting and erosion, particularly during flood events. The total length of the Brahmaputra is 2,900 km with a catchment area of around 583,000 km^2. Several flashy tributaries such as the Teesta join the main channel in the north from steep catchments in the southern Himalayas. The main distributary of the Jamuna River is the Old Brahmaputra, described by the Bangladesh Flood Forecasting and Warning Centre (FFWC) as a high flow spill river contributing largely to flooding, depending on the variations of siltation at the entry point (BWDB, 2020).

Bangladesh is divided locally into administrative districts, locally named zilas. Four of these that align along the Jamuna River have been chosen for the comparison of flood mapping skill, these are Kurigram, Gaibandha, Jamalpur and Sirajganj (Fig. 5.1).

Bangladesh experienced an active monsoon season during the summer of 2020 which brought severe and prolonged flooding in multiple spells. An unusually wet May following cyclone Amphan meant that water levels were already raised ahead of the monsoon season (Hossain, 2020). According to the Bangladesh Water Development Board (BWDB) the flooding had some remarkable characteristics. It began earlier than usual in late June and had a triple peak that had never been seen before. The flooding affected 40% of the country, inundating over 34,000 km^2. In 2020 this resulted in the second highest level of flooding since 1989 and the second longest flood duration since 1998. An estimated 5.5 million people were affected with 1 million houses waterlogged. Around 1.1 million people were displaced with almost 100,000 evacuated to over 1,500 shelters. Almost 1 million tube-wells and more than 100,000 latrines were damaged, 83,000 hectares of paddy fields were affected, and 257 people lost their lives.

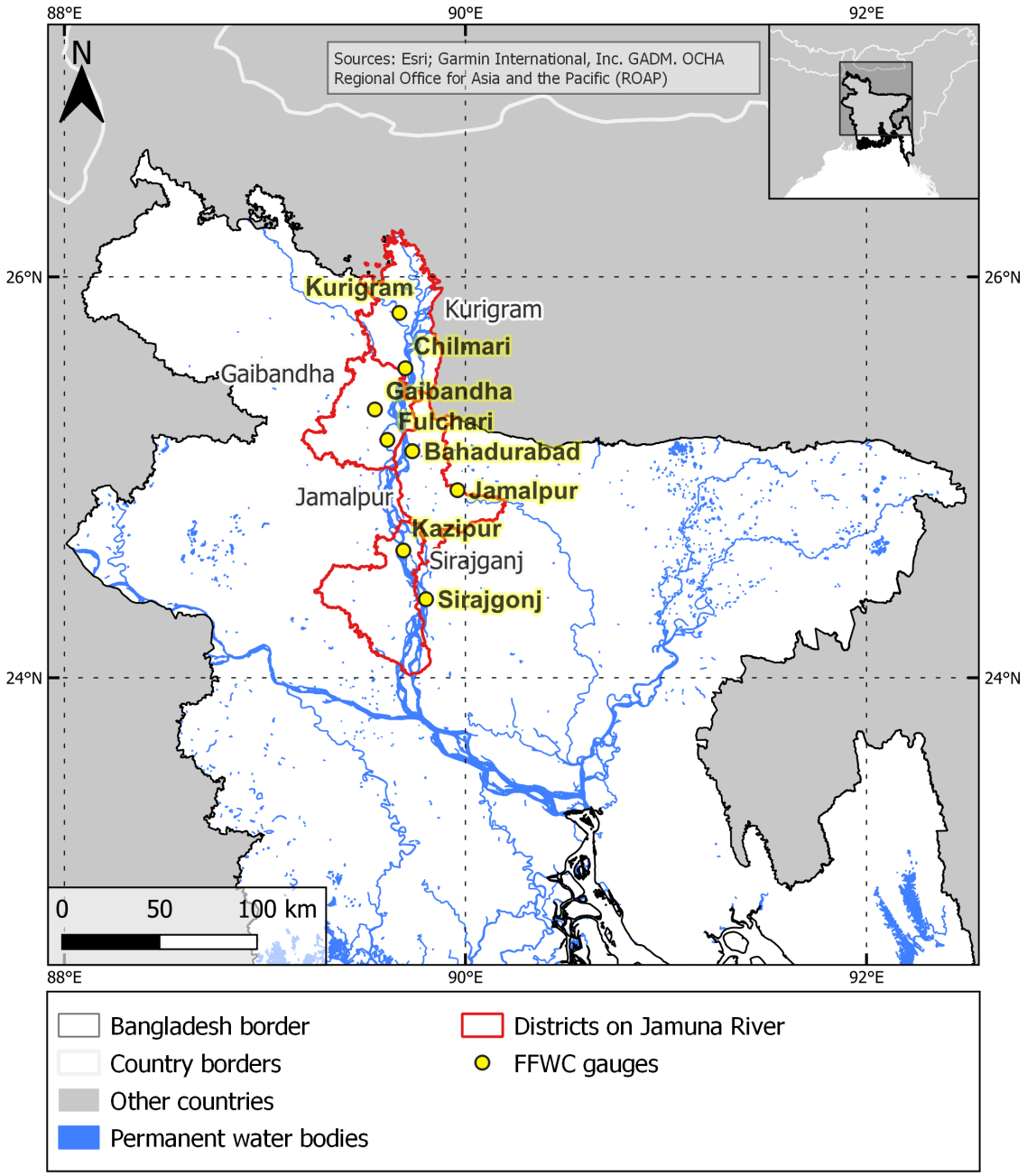


Figure 5.1: Four districts (zilas) of interest in the Jamuna catchment in northern Bangladesh.

The FFWC estimate that the Jamuna basin received 20% more rainfall in July than normal (BWDB, 2020). Gaibandha recorded the highest 1-day maximum rainfall across

the basin at 250 mm, with a 10-day consecutive maximum rainfall of 549.5 mm. The heavy monsoon rainfall in July caused two flood peaks in one month, the first peak around 15 July and the second around 25 July.

5.4 Flood forecasting systems and data

The focus of this multi-system comparison is to evaluate the performance of three systems at forecasting the flood inundation extent for the second flood peak on 25 July. In this section we briefly outline these three flood forecasting systems and summarise their similarities and differences in simulating flood inundation extent (Sections 5.4.1, 5.4.2 and 5.4.3). Table 5.1 details the main system attributes. The flood map data used for comparison from each system is described in Section 5.4.4.

Table 5.1: Flood forecasting system comparison. Any ensemble member forecast discharge value is defined as ens_{any}. The mean forecast discharge value of all ensemble members is defined as ens_{mean}.

Attribute	Flood Foresight	GloFAS Rapid Flood Mapping	FFWC
System application	Forecast-based financing for humanitarian early action	Global, broad scale medium range flooding prediction for large river basins	National flood forecasting and warning
System type	Ensemble simulation library (globally scalable)	Global ensemble simulation library (upstream drainage area >5000 km^2, river width >100 m)	Deterministic
Forecast type	Daily, 10-day lead time	Daily, maximum flood extent next 30 days	Daily, 5-day lead time
Meteorological model	ECMWF IFS	ECMWF IFS	BMD (WRF)
Hydrological model	LISFLOOD	LISFLOOD	MIKE II FF rainfall-runoff
Hydraulic model	RFLOW/JFLOW	CA2D	MIKE II GIS
Observed driving/input data	None	None	Rainfall and river water level
Grid length (m)	30	1000	300
DSM/DEM	NEXTMAP World30 DSM	SRTM (adjusted)	Survey of Bangladesh (>10 yrs old)
DSM/DEM grid resolution (m)	30	90 (re-scaled to 1000)	300
Modelled flood map return period thresholds (yrs)	20, 50, 100, 200, 500 and 1500 plus 30 interpolated flood maps	10, 25, 50, 100, 250, 500 and 1000	N/A
Flood map selection	$ens_{any} > 5yr$ return period threshold	$ens_{mean} > 10yr$ return period threshold	N/A
Defences included?	No	No	Yes

5.4.1 GloFAS Rapid Flood Mapping

GloFAS couples state-of-the-art numerical weather prediction (NWP) forecasts with a distributed hydrological model. With its continental scale set-up, it provides downstream countries with forecasts of upstream river conditions up to one month ahead as well as continental and global overviews for large river basins. As of version 3.1 (released in May 2021), this modelling chain is based on the full configuration of the LISFLOOD hydrological model, forced by an ensemble of meteorological inputs (GloFAS, 2021). The meteorological forecast data are provided to LISFLOOD by the ECMWF Integrated Forecasting System (IFS), the operational 51 ensemble member NWP system from ECMWF (Alfieri et al., 2013; GloFAS, 2021). The NWP data (precipitation, temperature, potential evapotranspiration, and evaporation rates for open water and bare soil surfaces) are taken as inputs into the hydrological model, LISFLOOD. LISFLOOD is a distributed hydrological rainfall-runoff model, simulating surface, groundwater and subsurface water flow and then routing the water to river channels and simulating the routing of the channel flow (LISFLOOD, 2022). LISFLOOD includes consideration of snow melt, infiltration, vegetation interactions (interception, evapotranspiration, water uptake) and exchange of soil moisture between a 3-layer soil water balance sub-model. The runoff data produced is routed through a representation of the river network using a double kinematic wave approach. The river network used is taken from the HydroSHEDS dataset (Lehner, 2014b). GloFAS is calibrated using historical streamflow records from selected stations worldwide (Hirpa et al., 2018). For Bangladesh, four river gauges have been used to calibrate GloFAS as reported on the GloFAS web viewer (GloFAS, 2022a). Two of these gauges are on river reaches that would impact the four Jamuna River districts: one at Bahadurabad on the main Jamuna River, and another upstream at Kaunia on the Teesta River, a tributary of the Jamuna River. The observation record at Kaunia is very short at around 7 years (1985-1992), while the record at Bahadurabad is over 35 years (1981-2015). Modified

Kling-Gupta Efficiency (KGE) is calculated for each station; a performance measure that indicates how well the model reanalysis (at day 0) replicates the flows observed, greater than 0.8 is very good and less than 0.2 is very poor. Both stations have KGE values above 0.7, indicating relatively good hydrological model performance.

GloFAS Rapid Flood Mapping (RFM) (GloFAS, 2022b) displays a maximum flood extent over the next 30 days by matching the return periods from the GloFAS streamflow forecast to a catalogue of modelled inundation extents. Flood maps are triggered for basins greater than 5000 km^2 and where the 10-year RP threshold is exceeded by the ensemble mean. The RP flood maps available are listed in Table 5.1. The flood maps were developed using the semi-inertial formulation of the CA2D hydraulic model which is a reduced complexity model based on the cellular automata approach and the diffusive wave equations, specifically designed to simulate flood inundation events involving wide areas (Dottori & Todini, 2011). Dottori et al. (2016) describe the methods used to derive the flood maps at specified return periods on a global scale using a vegetation corrected version of the global DEM SRTM (Farr et al., 2007). The hydraulic modelling was performed at 1 km grid resolution.

5.4.2 Flood Foresight

Start Network (Start Network, 2022) is a charity and network of over 80 humanitarian agencies and aims to develop locally led, early action by moving to a model of proactive funding to alleviate crises before they happen. JBA Consulting, in partnership with Start Network, have developed a Disaster Risk Financing (DRF) system for the Jamuna River that links a fluvial probabilistic flood inundation forecasting system, Flood Foresight (Revilla-Romero et al., 2017), to populations impacted by flooding (Fig. 5.2). The DRF system quantifies the flood risk to the population for the purposes of setting trigger threshold levels through a probabilistic global catastrophe risk model, FLY (Dunning,

2019).

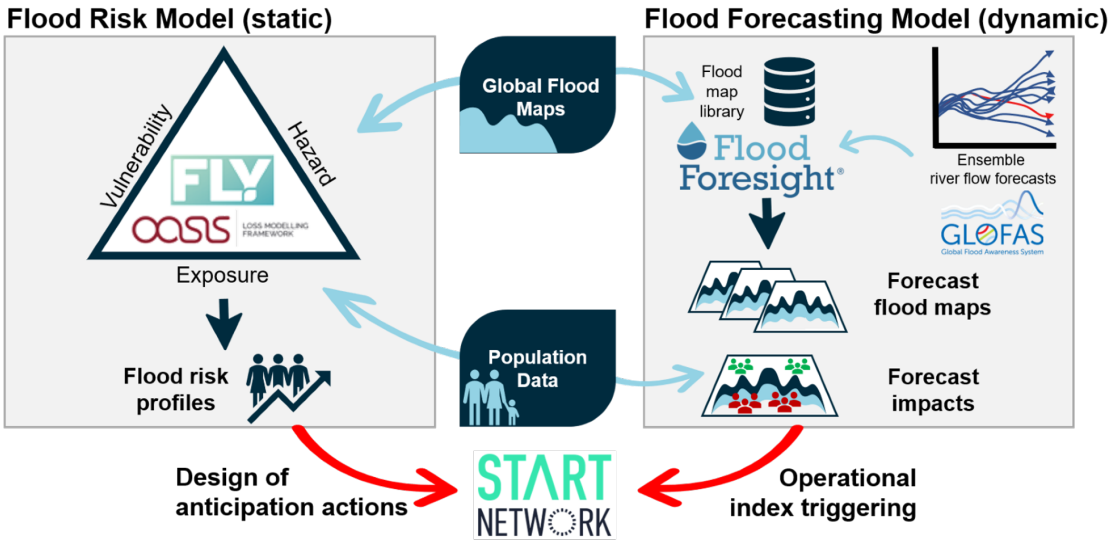


Figure 5.2: Flood Foresight/Start Network ensemble flood inundation forecast and population impacts work flow.

A domain of interest is divided into 'Impact Zones' (IZ) or sub-catchments using the HydroBASINS data-set (Lehner, 2014b). The Flood Foresight system links each IZ to GloFAS grid cells providing 51 ensemble member forecasts of river discharge (Section 5.4.1). Based on the forecast discharge for each IZ, a precomputed flood map is selected from a simulation library. The flood map library was hydrodynamically modelled using JFlow®, (Bradbrook, 2006) and RFlow using a detailed DSM at 30 m spatial scale at specified return period (RP) thresholds (detailed in Table 5.1). These were subsequently linearly interpolated at 5 intermediate intervals between each RP threshold and extrapolated between zero and the 20 year RP flood map (totalling 36 flood maps). The flood map selected is determined by the RP threshold exceeded within each IZ. An example forecast domain in Figure 5.3 shows neighbouring IZ trigger flood maps at different RP thresholds and the RP threshold is not exceeded in every IZ.

The simulation library approach enables rapid flood map selection so that the system

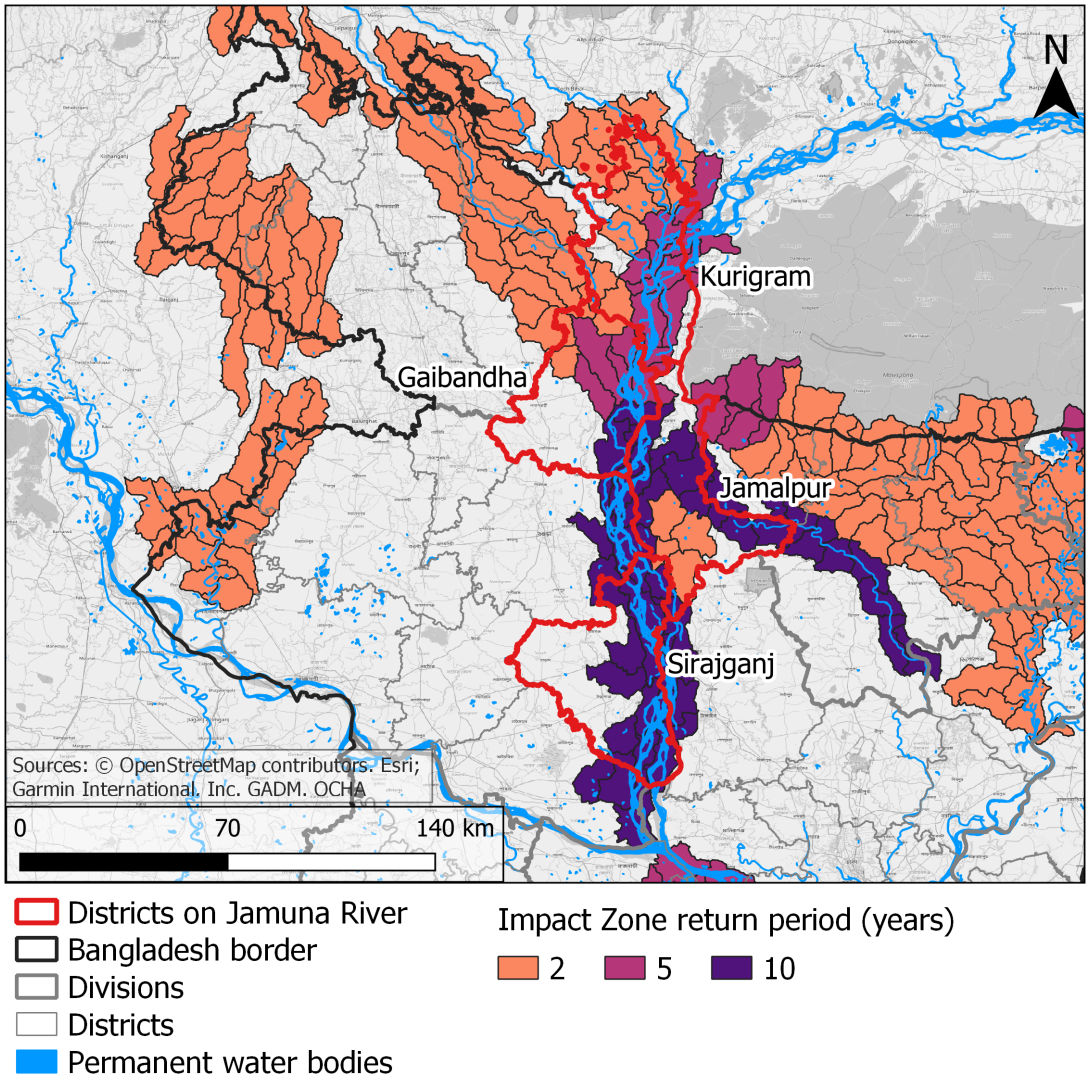


Figure 5.3: Example Flood Foresight forecast domain divided into Impact Zones (shown in colour with black outline where the 2-year return period threshold is exceeded), each linked to a GloFAS grid cell. The Impact Zone colour shows the corresponding return period threshold exceeded, determined by the GloFAS forecast discharge.

can be run in near real-time. Where the forecast discharge exceeds a 5-year RP threshold the probabilistic flood maps are triggered and linked to populations impacted. IZ linked to a 2-year RP threshold (Fig. 5.3) will not trigger a flood map due to low confidence in the RP threshold levels and the uncertain flood map interpolation process at low discharge

values. The system runs daily and produces 51 ensemble member flood extent and depth maps for forecast lead-times up to 10 days ahead. Hooker et al. (2023a) recommended consideration of all ensemble members as indicators of potential flooding so we evaluate each grid cell where any ensemble member indicates flooding (ens_{any}).

5.4.3 Bangladesh Flood Forecasting and Warning Centre

Historical catastrophic flooding in Bangladesh has led to well developed and forward-thinking flood forecasting and warning services. Under the Ministry of Water Resources, flood forecasting in Bangladesh is the responsibility of the Bangladesh Water Development Board (BWDB) following the BWDB Act-2000. The Flood Forecasting and Warning Centre (FFWC), established in 1972, is the lead organisation for flood forecasting and warning services. The FFWC act as coordinators between other Bangladesh agencies and ministries involved in flood disaster management. During the event in July 2020, FFWC provided a 5-day deterministic and a 10-day probabilistic flood forecast. FFWC have identified two main priority areas of improvement: to increase warning lead time and to make location specific flood forecasts (BWDB, 2020). Operationally, FFWC use real time hydrological data from water level and rainfall stations at 3-hourly intervals. Rainfall estimates are based on the preceding three days of rainfall along with analysis derived from NWP forecasts from the Bangladesh Meteorology Department (BMD, NCAR (2022)). Forecast flood bulletins are prepared daily and disseminated through various modes to multiple recipients.

During the monsoon flood season, the FFWC generate a daily hydrodynamically modelled flood inundation map for the Jamuna, Ganges and Meghna river basins. The flood maps are generated using output files from MIKE 11 FF Rainfall-Runoff hydrological model and hydrodynamic modelling simulations using a customized MIKE 11 GIS model (Havnø et al., 1995; Gourbesville, 1998). The Digital Elevation Model (DEM) used for the

hydrodynamic modelling has a 300 m spatial resolution that was collected by the Survey of Bangladesh (SoB) more than 10 years previously.

5.4.4 Observation data

The data described here are used to evaluate the three flood forecasting systems. Observations of river water level from eight river gauges across the four districts were provided by the FFWC for validation purposes. The gauge locations are shown on Figure 5.1 in yellow, five are located on the main Jamuna River channel and three are located on tributaries/distributaries of the Jamuna River. Three satellite SAR images from Sentinel-1 (S1A) acquired on 25 July 2020 and three pre-flood images from the same track from 7 June were used to derive a remotely observed flood map. The HASARD flood mapping algorithm (Chini et al., 2017) hosted on WASDI (WASDI, 2022) uses a statistical, hierarchical split-based approach to separate the two classes (flooded and unflooded) using the pre-flood and flood images. The HASARD mapping algorithm removes permanent water bodies, such as the river channel, reservoirs and lakes. Flooded areas beneath vegetation, near to buildings and under bridges will not be detected using this method. To smooth the HASARD flood maps and allow a fairer comparison we apply a morphological closing operation, without impacting the location of the flood extent, to flood fill buildings and vegetation. So that the flood prediction accuracy alone can be evaluated, the pre-flood occurrence of surface water using the JRC Global Surface Water database (Pekel et al., 2016) has been removed from the forecast inundation maps. The observed flood extent mosaic derived from the three SAR images at 20 m grid size was re-scaled to match the relevant forecast flood map grid lengths using majority (mode) aggregation.

5.5 Scale-selective evaluation methods

The flood forecasting systems detailed in Table 5.1 produce flood maps at 30 m, 300 m and 1000 m spatial scales (grid lengths). Validation of forecast flood maps against remotely observed flood extent is typically carried out by labelling each grid cell using a contingency table with categories: correctly predicted flooded, under-prediction (miss), over-prediction (false alarm) and correctly predicted unflooded. After labelling, conventional binary performance measures such as Critical Success Index (CSI) and Pierce Skill Score (PSS) are calculated and give a domain average skill score (Stephens et al., 2014). Comparing the binary performance measures of these flood maps at different spatial scales would not be meaningful as the higher resolution maps would be overly penalised due to the double penalty impact (Roberts & Lean, 2008; Hooker et al., 2022). Several commonly applied binary performance measures will be included here for demonstration and comparison purposes only and the details of these can be found in Appendix 5.8, Table 5.2.

To tackle the issue of validating across differing spatial scales, we apply a scale-selective approach to flood map evaluation. The scale-selective evaluation approach includes calculation of the Fraction Skill Score (FSS, Roberts and Lean (2008)) and location specific agreement scales (Dey, Roberts, et al., 2016), which are plotted on a Categorical Scale Map (CSM, Hooker et al. (2022, 2023a)). A brief summary of the method is given here. For full methodology please see Roberts and Lean (2008); Dey, Roberts, et al. (2016); Hooker et al. (2022, 2023a). The FSS calculates the accuracy of a forecast flood map by comparing against an observed flood map across a range of neighbourhood lengths (n, see Figure 5.4). First, every grid cell is compared ($n = 1$). Then, the next largest neighbourhood size, $n = 3$, surrounding each grid cell is compared and the process continues to $n = 5$, $n = 7$ and so on. The fraction flooded (number of flooded grid cells in the neighbourhood divided by the total number of grid cells in the neighbourhood) in each of the forecast and

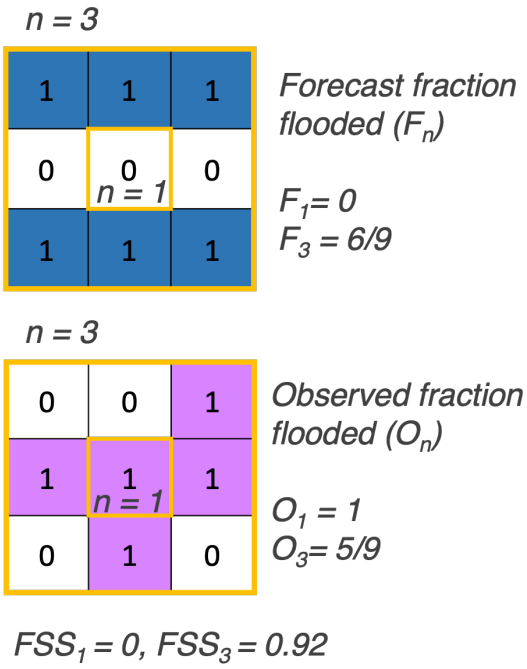


Figure 5.4: An example FSS calculation applied to forecast flood extent, 1 = flooded (in blue), 0 = unflooded (in white) compared to a remotely observed flood extent in pink. The FSS is calculated for two neighbourhood sizes, $n = 1$ (small gold box) and $n = 3$ (large gold box).

observed flood maps are used to calculate the FSS at each n. The FSS calculation is based on the Brier Skill Score. The FSS for each n is derived by calculating the mean squared error (MSE) between the forecast and observed fractions and dividing this by a reference MSE. This value is subtracted from 1 to give the FSS score. Increasingly larger neighbourhoods are compared until a target FSS score has been reached and exceeded at which point the skilful scale has been met (e.g. see Fig. 5.9). The target FSS score, given by $FSS_T \geq 0.5 + \frac{f_o}{2}$ depends on the fraction of observed flooding in the domain of interest, f_o.

The FSS gives a domain averaged skill score. We also calculate a local agreement scale (S) for each grid cell that can be mapped onto a Categorical Scale Map (CSM). The relationship between S and n is given by $S = (n-1)/2$. An acceptable level of

background bias between the forecast and observed flood maps can be pre-set. This is used to determine an agreement criterion. Like the FSS method, the comparison begins at each grid cell, if the agreement criterion is met, the grid cell is labelled with an agreement scale $S = 0$. Where the criterion is not met, a larger neighbourhood size is compared (e.g. $n = 3$). The fraction flooded in each of the forecast and observed flood maps are compared and if the criterion is met the agreement scale assigned would be $S = 1$. The process continues to larger neighbourhoods (e.g. $n = 5$, $S = 2$) until either the criterion is met or a predetermined limit is reached (S_{lim}, set to 9 for this application). The agreement scale at this limit would indicate a miss or false alarm for the grid cell. Combining a map of agreement scales with a conventional contingency map creates a CSM which shows the level of agreement (S) and whether the forecast is over or under-predicting the flood extent (e.g. see Fig. 5.8). The CSM shows a location specific skill score that can be linked to different aspects of the forecast system such as IZ and their associated river discharge forecast or return period thresholds, river channel bathymetry, the DSM or flood defences.

5.6 Results and discussion

Forecast flood maps for the Jamuna River flooding, July 2020 from each of the three systems described in Sections 5.4.2 - 5.4.3 are compared against SAR-derived flood maps (described in Section 5.4.4) using scale selective evaluation methods (as discussed in Section 5.5). First, in Section 5.6.1 the Flood Foresight system performance is evaluated against forecast lead time, where flood maps out to 10-days lead time are available. Performance evaluation of the other two systems with forecast lead time was not possible due to the availability of flood maps. In Section 5.6.2 we compare the forecast flood maps from each of the systems described in Section 5.4 and discuss their benefits and limitations in Section 5.6.3.

5.6.1 Flood Foresight Jamuna River case study

The performance of the Flood Foresight system with forecast lead time is evaluated here. Flood Foresight predicts the flood extent for second flood peak on 25 July at all lead times (out to 10-days). To evaluate the spatial accuracy within each district with forecast lead time, the absolute agreement scale score has been averaged across each district (solid lines, Fig. 5.5). An average agreement scale of zero would indicate the forecast and observed

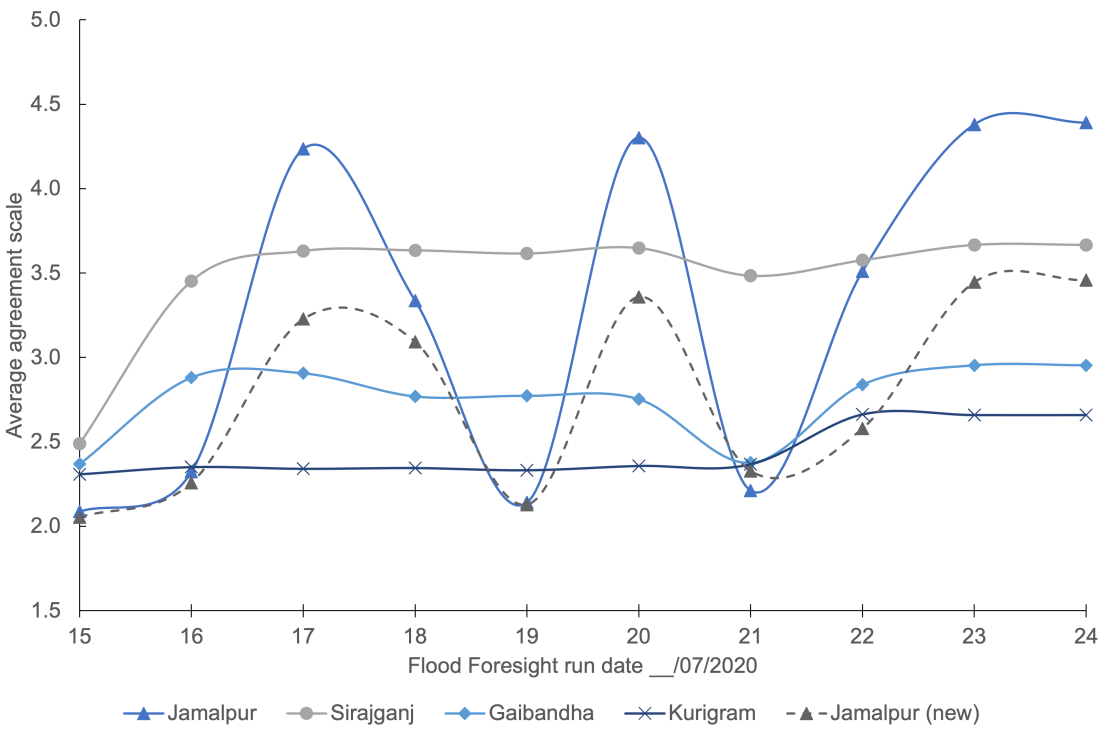


Figure 5.5: Flood Foresight average agreement scale against forecast lead time for each district and updated Jamalpur following reassociation of IZ with GloFAS grid cells, forecast valid for 25 July.

fields are in agreement at grid level, $S = 2$ means agreement is reached within a 5 by 5 neighbourhood. Across the 10-day forecast Flood Foresight performs best in Kurigram district in the north with consistently the worst performance in Sirajganj. Three of the districts (except Jamalpur) show a similar trend with forecast lead time with the best performance for all districts (smallest average agreement scale) occurring at a 10-day lead

time (2.31) with a second peak of performance at a 4-day lead time (2.61). The performance worsens from a 4-day lead time to a 1-day lead time across all districts, with an average agreement scale at a 1-day lead time of 3.42. The unusual variation in skill with forecast lead time for Jamalpur prompted further investigation into the driving data.

Section 5.4.2 describes how the domain is divided into IZ with each of these linked to the driving data, GloFAS river discharge (Fig. 5.3). The flood map selection within each IZ depends on the forecast discharge exceeding a particular RP threshold. A direct comparison of observed and modelled river conditions is not possible as observed data are river water levels, GloFAS provides forecast discharge, and the stage-discharge relationships are not available. The observed river water levels have been aligned using human judgment to the nearest GloFAS grid cell forecast discharge (1-day lead time control member) and compared for all river gauges. We aimed to match the trend in the two series whilst keeping the local station risk level close to the GloFAS 2 and 5-year RP threshold levels. Two of the eight river gauges are located in Jamalpur (Fig. 5.1). Bahadurabad is located on the main Jamuna channel and Jamalpur is on a distributary of the Jamuna River crossing the Jamalpur district (Fig. 5.6). (Hydrographs for gauges located outside of Jamalpur are plotted in Appendix 5.8, Fig. 5.10.) Bahadurabad (Fig. 5.6 (a)) is well calibrated in GloFAS (Section 5.4.1) and the overall forecast aligns well with the observed river water level; the second flood peak magnitude is slightly underestimated. The discharge forecast skill for Jamalpur station (Fig. 5.6 (b)) is very different to the observed river water level in terms of variation in water levels/discharge, timing and magnitude of flood peaks. A closer look at the river routing network in GloFAS shows that the Old Brahmaputra distributary in Jamalpur is not connected to the main river channel; this connection should occur at flood flows. To overcome the disconnection of the distributary in the river network, the IZ association to GloFAS grid cells in the Flood Foresight system design was manually updated. IZ aligning the Old Brahmaputra, crossing Jamalpur, were

reassociated to GloFAS grid cells located upstream on the main Jamuna channel. The new forecast hydrograph for Jamalpur (Fig. 5.6 (c)) shows an improved forecast compared to observed river water levels. However, the first flood peak arrives quicker than observed and the second flood peak magnitude is underestimated.

Following the gauge reassociation update to the Flood Foresight system, updated CSMs were reanalysed and this led to an improvement in the overall skill for the Jamalpur district (Fig. 5.5) but with variation in skill with forecast lead time. The variation is investigated by examining the CSMs for Jamalpur. Fig. 5.7(a) maps the original CSM for Jamalpur for run date 20 July. CSMs show a location specific (at each grid cell) agreement scale between the forecast flood map and the observed SAR-derived flood map. There is a large area of under-prediction around Jamalpur caused by the disconnection of the Old Brahmaputra distributary (as discussed above).

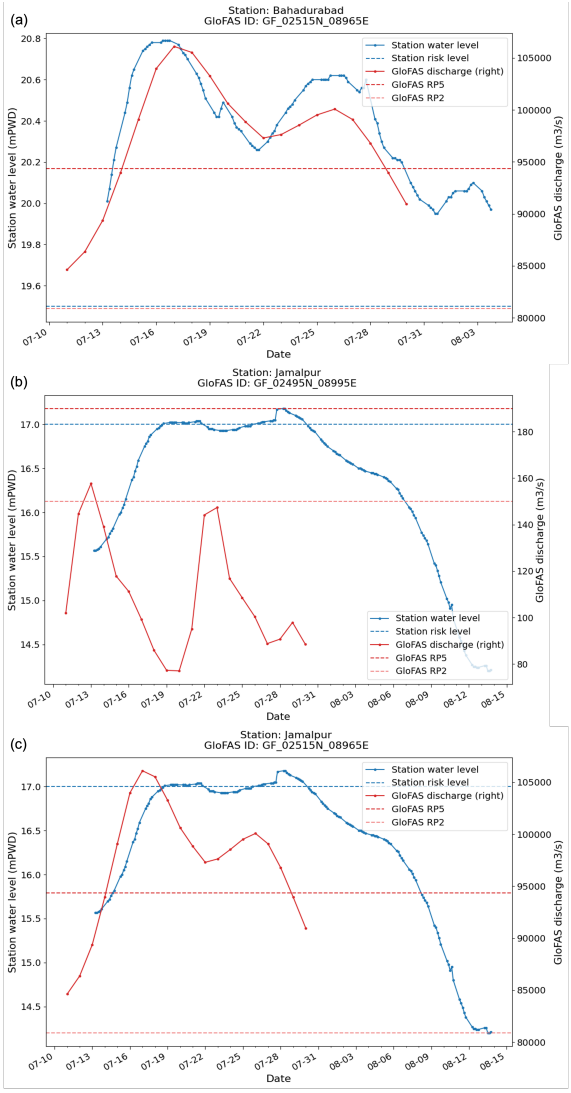


Figure 5.6: (a) GloFAS forecast discharge (control member, 1-day lead time) compared to FFWC observed river water level for the main Jamuna channel at Bahadurabad and (b) the old Brahmaputra distributary in the Jamalpur district. (c) The old Brahmaputra distributary forecast discharge following reassociation of IZ with GloFAS grid cells. The GloFAS RP threshold levels are taken from the nearest GloFAS grid cell to the gauge station location. Station risk levels are provided by FFWC.

The CSM change map (Fig. 5.7(b)) is calculated by taking the difference between the absolute CSM values, $|updated\ CSM| - |original\ CSM|$. The CSM change map for Jamalpur shows where the reassociation has impacted the flood map. A negative agreement scale change indicates an increase in skill (purple, smaller grid size), whereas a positive agreement scale change indicates a decrease in skill (orange, larger grid size). The CSM change values can highlight areas where the agreement scale has increased/decreased but have not reached $S = 0$ (agreement at grid level). Areas surrounding the Old Brahmaputra (in purple) show most of the increased skill following the reassociation. However, there are regions of Jamalpur not impacted by the reassociation that remain under-predicted where smaller distributaries run, that would not be captured/calibrated in GloFAS due to their small size (Fig. 5.7(b)). This results in some variation in skill with forecast lead time remaining after the reassociation (Fig. 5.5). The visualisation of incremental improvements at specific locations will benefit future flood map development work. The CSM change map shows more sensitivity to changes in skill that would not be visible on a conventional contingency map.

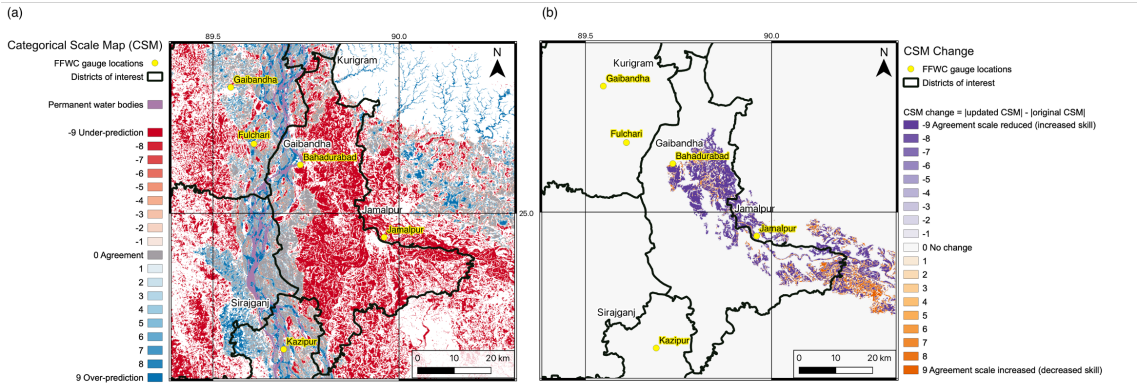


Figure 5.7: (a) Original CSM for Jamalpur. (b) CSM change (updated CSM - original CSM) for Jamalpur following reassociation of IZ with GloFAS grid cells. Run date 20 July, forecast valid for 25 July for (a) and (b).

5.6.2 Multi-system flood map comparison

GloFAS RFMs are displayed when the mean ensemble member exceeds the 10-year RP threshold within the next 30 days. By inspecting the reporting point hydrograph at Bahadurabad on each of the 10 days preceding the flood peak (25 July), we found that the 10-year RP threshold was exceeded just once by the ensemble mean on the 18 July (run date), which means only one forecast flood map was available from GloFAS RFM for evaluation. FFWC provided daily flood maps, based on most recent observations, valid for the same day. For the multi-system comparison, based on the availability of forecast flood maps, the following have been selected to compare in detail: Flood Foresight's flood map for the same run date as the available RFM forecast (18 July forecast date, valid date 25 July), the RFM (run date 18 July) and the FFWC flood map for 25 July (run date and forecast valid date) have each been compared to the SAR-derived flood map (25 July) and CSMs calculated (Fig. 5.8). CSMs have also been calculated for all available forecast lead times for the Flood Foresight system (not shown).

The large area of flood under-prediction (in red) to the northwest of Sirajganj on each CSM (Fig. 5.8) can be linked to observed very heavy rainfall and possible surface water flooding (SWF) detected by the SAR data. The recorded rainfall described earlier (Section 5.3) is confirmed by rainfall derived from satellite data, (Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) data (Funk et al., 2015)), which shows that the July rainfall anomaly in this area was 125 - 175 mm. The accumulated rainfall in GloFAS for the ensemble mean for the 20 days preceding 25 July amounts to a maximum of 300 mm, with up to 400 mm north of Bahadurabad. This is an under-prediction compared to observed rainfall (Gaibandha 1-day maximum rainfall 250 mm, 10-day consecutive maximum rainfall 549.5 mm), which explains the under-prediction seen in Fig. 5.8 (a) and (b). Note that each system forecasts fluvial flooding and does not account for SWF, which

was likely given the extreme rainfall accumulations recorded. This is a limitation of both Flood Foresight and RFM systems and demonstrates the need for combining fluvial and SWF forecasting systems or combining the forecast flood maps with SAR data using data assimilation (e.g., García-Pintado et al., 2015; Cooper et al., 2019).

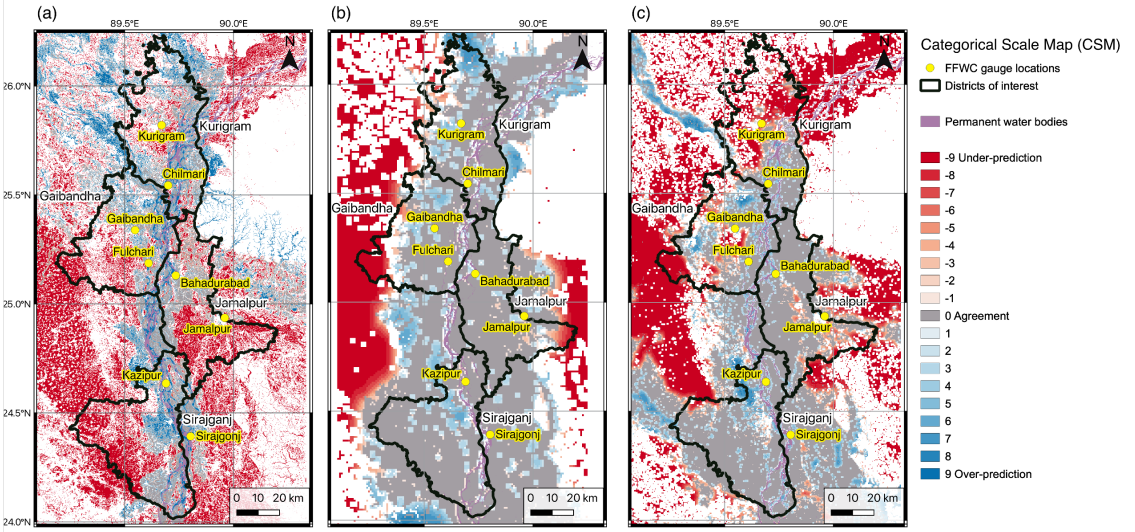


Figure 5.8: CSM for flood inundation forecasts from three forecast systems for flood peak 25 July compared to SAR-derived flooding. (a) Flood Foresight run date 18 July, (b) GloFAS RFM run date 18 July and (c) FFWC flood map run date 25 July.

The CSM for Flood Foresight (Fig. 5.8 (a)) shows areas of over-prediction (in blue) next to the Jamuna River. It is likely that more of the river channel has been removed from the SAR image during the flood mapping process compared to the 12-month occurrence of water in the Global Surface Water database. Also, the Jamuna River channel migration will also contribute to errors in this area as the DSM was acquired in 2016. To the west of Sirajganj, the FFWC correctly maps flooding associated with the Atrai River, a tributary of the Jamuna River that is not forecast by the other two systems. Multiple tributaries flow across Gaibandha and Sirajganj that are not currently resolved by GloFAS. This also impacts the performance of Flood Foresight when flood maps were not triggered above

the 5-year RP threshold.

Flood waters are (Fig. 5.8 (b)) spread further from the main river channel in the RFM, compared to the Flood Foresight flood map. This is due to the GloFAS configuration where clusters of cells are linked to the main channel reporting points. The RFM extent is also due to a smoother DEM created by re-scaling from 90 m to 1000 m. This will effectively remove flood barriers such as embankments and roads. An element of smoothing (albeit to a lesser extent compared to the RFM) would also occur in the Flood Foresight flood maps at 30 m grid length. The FFWC CSM shows good accuracy in Sirajganj and Jamalpur (Fig. 5.8 (c)). In Sirajganj there is a region of over-prediction on both the Flood Foresight and the RFM CSM not present on the FFWC CSM. The FFWC model includes flood defences and water level observations from Kazipur, both could contribute to the better performance seen here. FFWC maps perform less well in Kurigram where the flood extent is under-estimated. Kurigram is next to the northern border of Bangladesh where upstream water level data are unavailable for hydrological model calibration and validation. The benefits of the trans-boundary systems of GloFAS and Flood Foresight are evident here.

To quantify the district performance of each system the FSS has been calculated for neighbourhood sizes up to $n = 19$ or larger (not plotted) where the FSS target has not been reached (Fig. 5.9). The FSS gives a measure of spatial accuracy for each system flood map, however the score is not directly comparable across different spatial scales since the scores are calculated in terms of a neighbourhood size (Section 5.5). We can calculate a skilful scale, which is half of the neighbourhood size at which the FSS exceeds FSS_T, this accounts for the size of the grid cell and can be directly compared across each of the forecast systems. For example, in Figure 5.9(a) for Kurigram, at grid level ($n = 1$) the FFWC FSS (0.52) exceeds the Flood Foresight FSS (0.48). We saw on the CSM map (Fig.

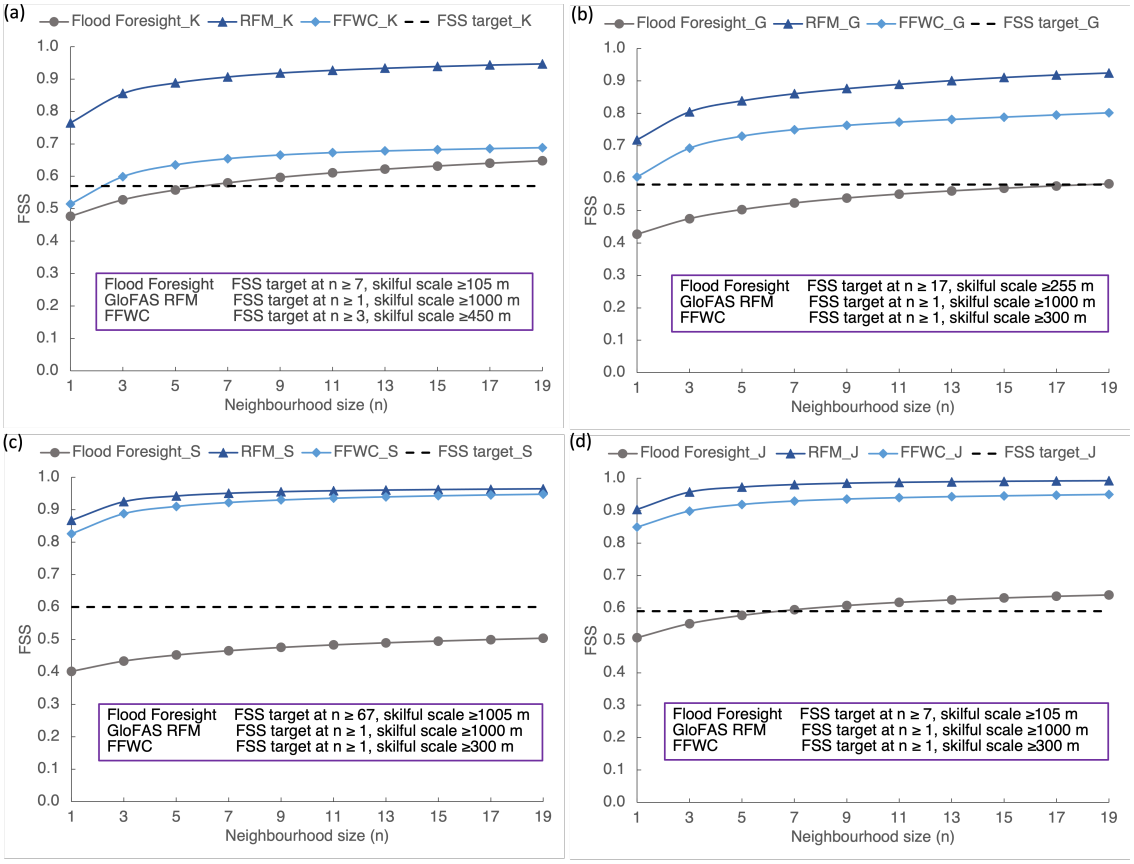


Figure 5.9: Average FSS plotted against neighbourhood size (n) for each forecast system (run dates as described in Figure 5.8) in Kurigram (a), Gaibandha (b), Sirajganj (c) and Jamalpur (d) and the target skill score for each district.

5.8) that the Flood Foresight map appeared to capture the observed flooding at Kurigram more accurately than the FFWC map. Despite this observation, the FFWC map has a higher CSI score (0.35, Fig. 5.11(c)) compared to the Flood Foresight CSI (0.31, Fig. 5.11(a)) because the grid size is not accounted for by the CSI. However, neither of the two systems flood maps have reached the target FSS_T at grid level (Fig. 5.9(a)). The FFWC map exceeds FSS_T at $n = 3$ and the Flood Foresight map exceeds FSS_T at $n = 7$. By accounting for the impact of the grid size, the skilful scale for Flood Foresight for Kurigram is 105 m ($\frac{1}{2}(7 \times 30)$) compared to 450 m ($\frac{1}{2}(3 \times 300)$) for the FFWC flood map indicating the Flood Foresight system is more accurate in Kurigram. In Gaibandha, the skilful scale

is similar for Flood Foresight and the FFWC maps (255 m verses 300 m, Fig. 5.9(b)). In Sirajganj, Flood Foresight requires a neighbourhood size similar to the grid size of the GloFAS RFM to exceed FSS_T at 1005 m (Fig. 5.9(c)). Following the reassociation of the IZ in Jamalpur, Flood Foresight has a skilful scale of 105 m here compared to a high FSS score for the FFWC model at 300 m grid level (Fig. 5.9(d)). Across all districts, GloFAS RFM exceeds the FSS_T at grid level with the best performance in Jamalpur and the worst performance in Gaibandha which can be linked to the under-estimation of flood extent seen on the CSM (Fig. 5.8(b)).

The scale selective approach is also useful for comparing performance above the FSS_T line where two systems exceed FSS_T at grid level ($n = 1$). For example, in Jamalpur (Fig. 5.9(d)) at $n = 1$ the RFM FSS is 0.90 compared to the FFWC FSS of 0.85. However, FFWC reaches the same score as the RFM (at $n = 1$, 0.90), at $n = 3$. In terms of spatial scale, the same FSS is reached by FFWC at $n = 3$ (450 m) as RFM at $n = 1$ (1000 m), indicating that the FFWC flood map is more accurate than RFM in Jamalpur. The same result is true in Sirajganj (Fig. 5.9(c)) where the FFWC FSS at $n = 3$ (0.89) exceeds the RFM FSS at $n = 1$ (0.87). In Gaibandha (Fig. 5.9(b)) the FFWC FSS at $n = 5$ (0.73, 750 m) exceeds the RFM FSS at $n = 1$ (0.72). Overall, by accounting for grid length, the FFWC provides the most accurate forecast flood map (albeit at a 0-day lead time) in Jamalpur, Sirajganj and Gaibandha. RFM is most skilful in Kurigram with Flood Foresight outperforming the FFWC model here.

5.6.3 Discussion

GloFAS RFMs are designed to give an early indication, up to a month in advance that severe flooding is possible for large rivers across the globe. The RFMs are triggered when the ensemble mean discharge exceeds the 10-year RP threshold over the next 30 days. The RP threshold levels are calculated using ERA5 reanalysis data (Harrigan et al., 2020).

GloFAS RFM is triggered for the Jamuna River only once at an 8-day lead time on 18 July (mapping maximum extent over the next 30 days) where the flood map shows a high level of skill across the districts of interest (Fig. 5.8), comparable to the local FFWC flood maps, which use local observations. Unfortunately, the RFM skill reduces closer to the event with no flood maps triggered after 18 July. We also see this impacting the Flood Foresight skill, which reduces closer to the flood peak (Fig. 5.5). The Flood Foresight system indicates flooding surrounding the Jamuna River at all forecast lead times. The flood extent is generally under-predicted. This is partly due to the discharge forecast not exceeding the 5-year RP threshold, which also links to unresolved/uncalibrated smaller tributaries/distributaries in the GloFAS river network.

Boelee (2022) finds (for Africa, based on GloFAS ensemble reforecast flows) that the forecast discharge exceedance above return period thresholds depends on both the forecast lead time and the number of ensemble members considered (the probability trigger set). Boelee found more flood occurrences were predicted at medium-range lead times, compared to short-range lead times. We infer that these results could also apply to Bangladesh and that they partly explain the RFMs best performance at 8 days lead time. The ERA5 reanalysis data is used to initialise the GloFAS forecast and determine the RP thresholds. Currently, the RP thresholds do not account for either forecast lead time or ensemble variability. The GloFAS hydrographs at Bahadurabad indicate a reanalysis discharge value of around the 2-year RP threshold for all lead times within 5 days of the flood peak. This is a significant underestimation compared to observed river levels (both in historical context and compared to FFWC danger levels and severe flood thresholds). High confidence is assigned to the initial conditions in the streamflow forecast and also in the short-term forecast before the ensembles show more variation at longer lead times. This leads to an overall under-estimation of the flood magnitude at shorter lead times. Zsoter et al. (2020) found that ensemble-reforecast-based thresholds would lead to an improved forecast at

lead times beyond a few days as they can account for variations in forecast skill with lead-time and ensemble variability. Ensemble-reforecast-based thresholds could improve the flood map selection in both the RFM and Flood Foresight systems, which presently are the main limitation of both systems.

The RFM uses the ensemble mean or 50% of ensemble members must exceed the RP threshold to trigger a flood map. However, Boelee (2022) found that the percentage of flood events exceeding the threshold for any return period dropped to less than 50% as soon as the required ensemble size was increased to two ensemble members or more, for all the lead times. Extreme flood event prediction, can lie in the ensemble member outliers (Hooker et al., 2023a). The Flood Foresight system uses information from all ensemble members for impact forecasts, which is a major benefit of this system. The automated probabilistic flood maps can be produced quickly in near real-time indicating a spread of possible conditions that could support the decision-making process. The RFM could see an improved forecast at more lead times if any ensemble member exceeding the RP threshold (rather than the mean) triggered the flood map selection. This would require further investigation in flood prone areas so that the number of ensemble members chosen can be optimised to avoid increasing false alarms. The RFM could also provide probabilistic information by combining flood maps from all ensemble members that exceed the RP threshold.

The FFWC model performs significantly better at shorter lead times compared to the other two systems, which is not surprising as locally observed river level and rainfall data are used as input data. The benefits of the FFWC model are that no RP thresholds need to be determined or exceeded to produce a flood map and there are no flood map interpolation uncertainties. The deterministic FFWC models forecast skill drops significantly with forecast lead time (BWDB, 2020), which reduces the usefulness for flood mitigation and FbF purposes at longer lead times. The FFWC model performs less well compared

to RFM and Flood Foresight near to the country's border where observations upstream are unavailable.

5.7 Conclusions

Forecast flood maps are increasingly sought to accurately link flooding hazard to populations impacted to inform FbF schemes. Humanitarian agencies in Bangladesh would like the impacts mapped in detail at Union level (4,571 Unions in Bangladesh) at long forecast lead times (out to 10 days) so that insurance funds can be locally targeted in good time. This creates a conflict between the detail or spatial scale (grid size) of the flood maps and the forecast skill of the flood forecasting system. Spatial validation of forecast flood maps from multiple systems has received little attention, partly due to the problem of comparing skill scores from maps at different spatial scales. Here, we applied a validation approach using scale-selective methods (Hooker et al., 2022) that determines a skilful scale, which can be directly compared across forecast systems.

We evaluated three flood forecasting systems, Flood Foresight (30 m), GloFAS RFM (1000 m) and the FFWC Super Model (300 m) each predicting flood extent at different spatial scales (shown in brackets) for the Jamuna River in July 2020. Each of the maps were compared against SAR-derived observations of flood extent. Evaluating Flood Foresight skill at all lead times out to ten days for four districts revealed issues with unresolved/uncalibrated tributaries/distributaries in GloFAS (used to input forecast discharge to Flood Foresight). Reconfiguring the Flood Foresight system led to an improved flood map forecast in one district (Jamalpur), but similar issues remained in other areas. This highlights one problem with trying to combine a gridded global hydrometeorological model with a detailed sub-catchment network and linking this to detailed flood maps. The flood mapping skill in the Jamuna basin is linked to the detail of the river network and whether

flood maps are triggered, which depends on exceeding the RP threshold set. Where Flood Foresight maps were triggered, such as in Kurigram, the flood map accuracy outperformed the local FFWC model. This is due to its location next to the border of Bangladesh and the lack of upstream observations. In other areas the FFWC model captures more detail in the river network and shows less under-prediction compared to the other systems. For FbF applications and humanitarian response in Bangladesh, a combination of Flood Foresight and the local FFWC model could produce flood inundation maps at a useful scale that can be linked to flooding impacts. Flood Foresight has the benefit of forecast skill at a longer lead times (up to 10 days) with probabilistic maps accounting for some of the forecast uncertainty. Flood Foresight could be supplemented or linked to driving data from the FFWC model at shorter lead times to incorporate local observations and a higher resolution river network. This would avoid regions of non-trigger where no flood map is selected from the Flood Foresight library due to the forecast discharge not exceeding the RP threshold. GloFAS RFM is designed as a deterministic 'heads-up' tool at a coarse resolution that would be difficult to link to impacts for FbF applications in its current configuration. All systems miss flooding across a wide area captured by the SAR data that is possibly due to surface water flooding (SWF). These fluvial flood forecasting systems are not designed to map SWF and we recommend combining the forecast flood maps with SAR-derived flood maps through data assimilation so that SWF can be accounted for in post event impact calculations for FbF schemes. Alternatively, a combined fluvial/pluvial flood forecasting system would be ideal, however pluvial flood forecasting practice is currently less developed in part due to the difficulties in observing SWF and accurately predicting convective rainfall (Speight et al., 2021).

For future spatial validation of flood maps using SAR data, we recommend making use of the Copernicus GFM product (GFM, 2021) which maps flooding detected from all Sentinel-1 images since October 2021. Importantly, an exclusion mask layer is available

which can be used to exclude areas where SAR is unable to detect flooding such as in dense urban areas, under vegetation and near steep topography. The forecast flood maps would no longer be penalised for over-prediction in regions where the SAR cannot reliably detect flooding. Another opportunity for improvement lies with the flood map library. Both Flood Foresight and RFM maps are currently undefended. However, both would likely be improved if flood defence information from FFWC could be incorporated, allowing areas benefiting from those defences to be discounted when flood conditions are at return periods lower than the standard of protection of the defence. Alternatively, a new high resolution DTM including local defence features, ideally through acquiring LiDAR data (where locally obtained), would improve the local accuracy of the flood maps. The maps held within the simulation library could be hydrodynamically precomputed at lower discharge return periods, which would avoid inaccuracies caused by flood map interpolation beneath the lowest RP level. This would increase the confidence in these flood maps so that a lower RP threshold could be used to trigger FbF allowing more people access to insurance funds.

Fortunately, some of the issues discussed here such as the river network detail will be partly resolved by the major upgrade to GloFAS with version 4.0 due in 2023 (Grimaldi, 2022). Significantly, the spatial resolution of GloFAS will increase 4 fold to approximately 5 km grid size. The river network will increase similarly, and more distributaries/tributaries will be included in Bangladesh. This upgrade along with the use of ensemble-reforecast-based RP thresholds should improve the flood map selection process used by both GloFAS RFM and Flood Foresight. Scale-selective validation methods will enable future system changes to be evaluated and compared meaningfully. Ideally, this would be automated and integrated into the flood forecasting system.

5.8 Appendix

Table 5.2: Binary performance measures and formulas. Contingency categories: correctly predicted flooded (A), over-prediction (B), under-prediction (C), correctly predicted unflooded (D)

Performance measure	Formula	Description [range min, range max, perfect score]
Bias	$\frac{A+B}{A+C}$	[0, ∞, 1] 1 implies forecast and observed flooded areas are equal > 1 indicates over-prediction, < 1 indicates under-prediction
Critical Success Index/Threat score $F^{<2>}$ (CSI)	$\frac{A}{A+B+C}$	[0, 1, 1] Fraction correct of observed and forecast flooded cells
$F^{<1>}$ Proportion correct	$\frac{A+D}{A+B+C+D}$	[0, 1, 1] Proportion correct (wet and dry) of total domain area
$F^{<3>}$	$\frac{A-C}{A+B+C}$	[-1, 1, 1] Score reduced by over-prediction
$F^{<4>}$	$\frac{A-B}{A+B+C}$	[-1, 1, 1] Score reduced by under-prediction
False Alarm Rate (FAR)	$\frac{B}{B+D}$	[0, 1, 0] Proportion of over-prediction of dry areas
Hit Rate (HR)	$\frac{A}{A+C}$	[0, 1, 1] Fraction correct of observed flooded area
Pierce Skill Score (PSS)	$HR - FAR$	[-1, 1, 1] Incorporates both under and over-prediction

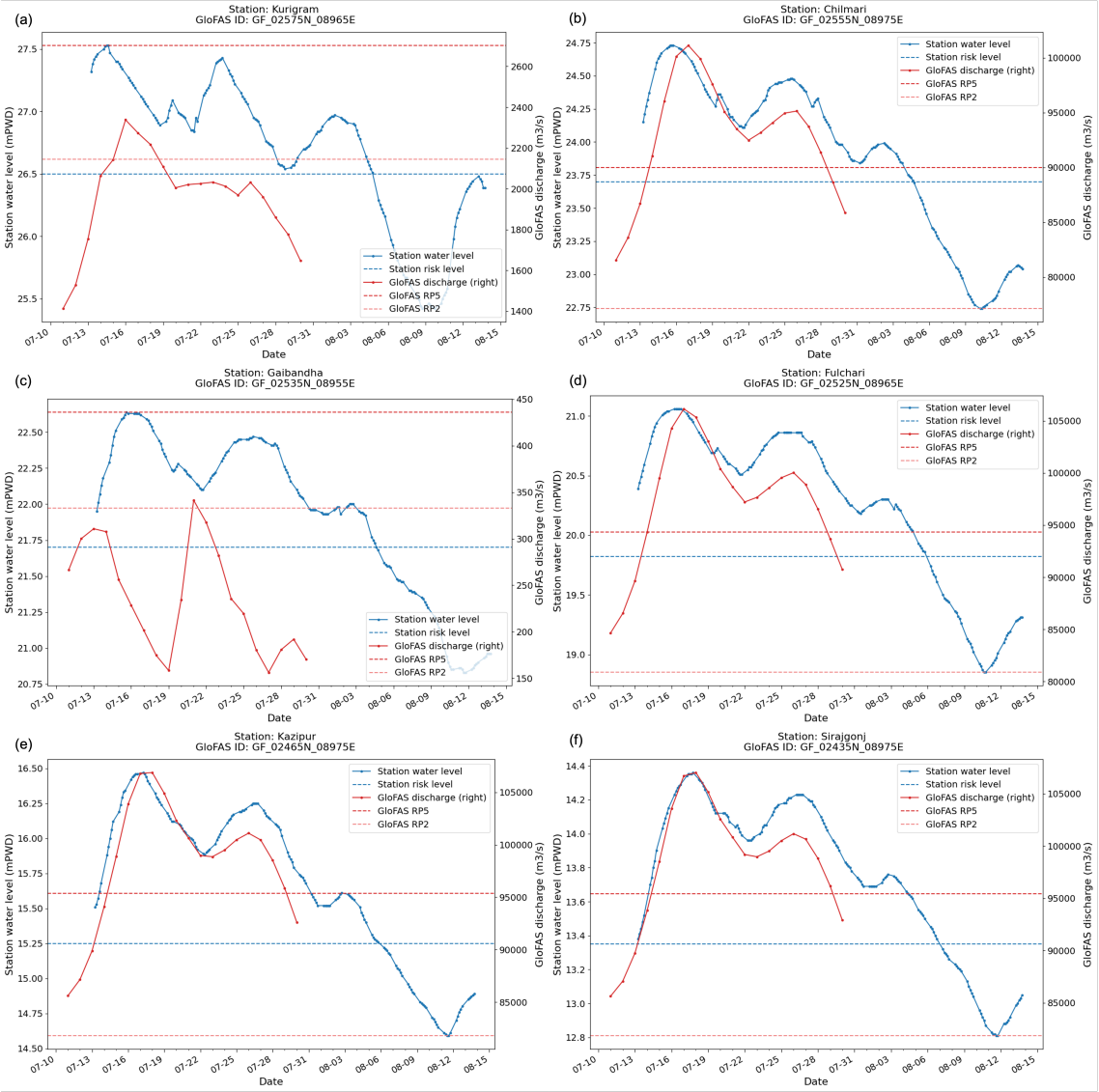


Figure 5.10: Six FFWC station water levels during July and August 2020 across the four districts compared with closest GloFAS grid cell discharge.

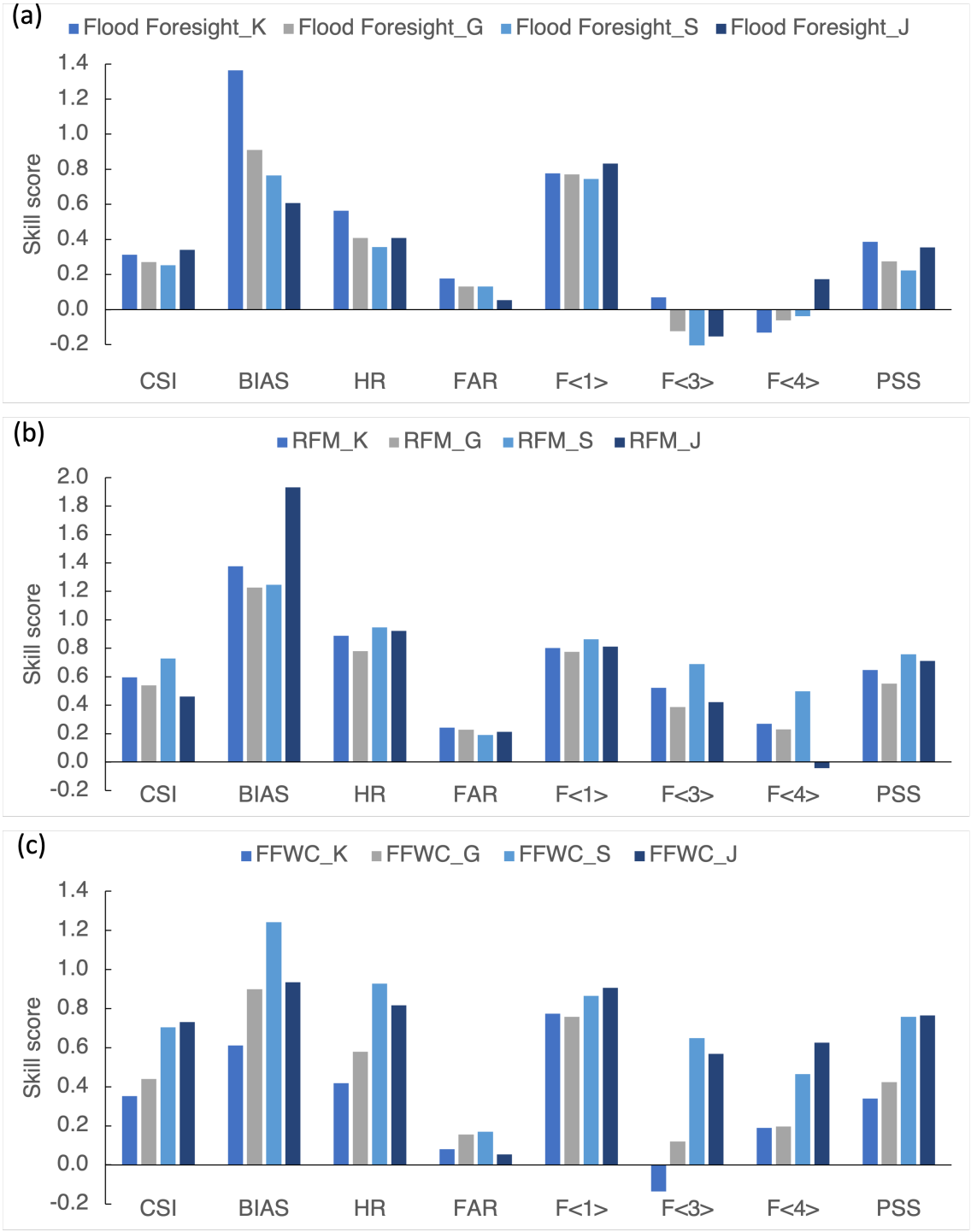


Figure 5.11: Binary performance measures for each district. Flood Foresight run date 18 July (a), GloFAS RFM 18 July (b) and FFWC flood map 25 July (c).

5.9 Chapter summary

In this chapter we apply the scale-selective methods presented in Chapter 3 to evaluate three flood forecasting systems, each presenting flood inundation forecasts at different spatial scales. Conventional binary performance measures would result in biases linked to spatial scale. We show how the scale-selective methods can be used to meaningfully compare across different flood forecasting systems. The benefits and limitations of the systems could be discussed as a result of the evaluation process. The issues identified here, that can lead to a non-trigger of flood maps in some sub-catchments, are addressed in Chapter 6. In Chapter 6 we will use the SAR data directly to improve the flood map selection for previously non-triggered flood maps in the Flood Foresight system.

Chapter 6

Updating simulation library flood map selection through assimilation of probabilistic SAR-derived flood extent

In this chapter we address the fourth research question outlined in Chapter 1; Does a data assimilation framework improve the analysis of flood inundation from a simulation library system?:

- Can we incorporate probabilistic information from remotely observed flood inundation into a data assimilation framework to improve the flood map selection within a simulation library flood forecasting system?
- How does the analysis flood map compare to independent validation data?


The Art of DA
Hugo Dalton

6.1 Abstract

Mitigating against the impacts of catastrophic flooding requires funding for the communities at risk, ahead of an event. Simulation library flood forecasting systems are being deployed for forecast-based financing (FbF) applications. FbF schemes use hydrometeorological predictions, linked to flood inundation maps that overlay local population (or other) impact maps to trigger a financial payment in advance of flooding. The FbF trigger is usually automated and relies on the accuracy of the flood inundation forecast. This reliance can lead to missed events that were forecast below the trigger threshold required or were not predicted at all. However, earth observation data from satellite-based synthetic aperture radar (SAR) sensors can reliably detect most large flooding events, although their use is limited in urban areas. Data assimilation (DA) combines forecast information with observation information to improve the current system state (the analysis). A new DA framework is presented to update the flood map selection from a simulation library system using SAR data, taking account of observation uncertainties. By utilising flood extent likelihood data derived from Sentinel-1 SAR images, we derive a new cost function that must be minimised. By iteration through the flood map library we optimise the flood map selection per sub-catchment. We have applied our DA method to the Pakistan 2022 flood. During this flood, the Indus River in the Sindh province downstream of the Sukkur barrage was not forecast to reach flood levels, which resulted in a non-trigger of the FbF scheme for this region. Our experiments have focused on three different scenarios: a large city area with limited SAR flood detection (due to a dense urban area); sub-catchments with mixed rural and urban areas; and flood edge sub-catchments. We found that the flood map selection could be triggered in four out of five sub-catchments tested, with the exception occurring in the dense urban area due to the simulation library flood map accuracy here. Thus, the analysis flood map, created by assimilating observations from SAR flood likelihood data, has potential to be used to trigger a secondary finance scheme

during a flood event and avoid missed financing opportunities for humanitarian action.

6.2 Introduction

Our warmer climate is increasing the frequency and intensity of extreme weather events and the exposure and vulnerability of communities and individuals (Pörtner et al., 2022). Large-scale flood forecasting systems predicting flood inundation extent are increasingly used for disaster risk reduction to improve preparedness ahead of a major flooding event (Stephens & Cloke, 2014a; Hooker et al., 2023b; Wu et al., 2020). An ensemble flood forecasting system creates probabilistic flood maps indicating the likelihood of flooding across a region or country. Flood impact risk factors such as population density, land-use types or vulnerable infrastructure can also be mapped for the same area. The forecast-flood-likelihood maps can be overlaid with impact maps and depending on the severity of the hazard and the level of impact, a risk profile can be determined. The flood risk profile can be used to inform forecast-based financing (FbF) schemes that enable the pre-release of funds based on the flood forecast, ahead of the flood event (Coughlan de Perez et al., 2015, 2016). Automation of FbF schemes is important for rapid action to take place to mitigate against flooding impacts. The skill of the flood forecasting system is key to *triggering* the FbF scheme. A *non-trigger* of FbF ahead of or during a flood event might prove catastrophic for those impacted.

Advances in flood forecasting both at global and local levels link together meteorological and hydrological forecasts of river discharge that drive the selection of pre-computed flood maps from a simulation library (Speight et al., 2021; Hooker et al., 2023a). The use of a simulation library obviates the need to run a hydrodynamic model as part of the forecast process, reducing computation time and allowing near real-time updating for large areas, which otherwise presents a significant challenge. The flood maps within the library

are at a relatively higher spatial resolution (e.g. 30 m) compared to the resolution of the driving global hydrological model (e.g. approximately 5 km). This mismatch in scales can lead to problems with flood map selection and can cause gaps where the minimum return period threshold has not been exceeded (a non-trigger) by the forecast discharge (Hooker et al., 2023b). The three main issues that cause this in the global scale model are the representation of river networks, the return period thresholds determined and the exclusion of dam operations. Rivers that are narrower than a particular width, or catchment areas smaller than a pre-determined size are not resolved by global scale models. In addition, the return period thresholds set may be poorly calibrated due to a lack of ground truth observational data such as river discharge or river water level (Boelee, 2022). These two limitations can lead to a non-trigger, i.e. no flood map is selected from the simulation library for a particular sub-catchment. Also, local dam operations such as diversions of river water for irrigation purposes or rapid releases of flood waters downstream, are not generally included in global scale models. This can lead to over- or under-prediction of forecast discharge, resulting in inaccurate or non-trigger of flood map selection in the forecast.

Satellite-derived observations of flooding have the potential to bring additional spatial information into flood inundation forecasts compared to in situ point gauging stations. These observations could be used to update and improve the FbF scheme either as part of a secondary finance payment following the acquisition of the satellite data or to improve the flood inundation forecasts going forwards as the flood event evolves. Synthetic aperture radar (SAR) sensors are particularly useful for remote flood detection, since they can see through cloud, most weather and are active both day and night (e.g. Mason, Davenport, et al., 2012; Schumann et al., 2022). Previously, SAR data have been used in several different ways to improve hydraulic models and flood prediction through data assimilation (DA). Data assimilation finds an optimal state (such as water level) and/or model

parameter values by accounting for the previous forecast, the observations available, and both of their associated uncertainties. The updated state (analysis) and/or parameter set are used to initiate the next forecast in a feedback loop or cycle. A review of approaches used to assimilate satellite-derived data into hydraulic models (from 2007 until 2015) can be found in Table 7 of Grimaldi et al. (2016) and Table 1 of Revilla-Romero et al. (2016).

Here, we summarise the different ways that observations of flood extent and water levels are used in various DA approaches and discuss some limitations. In order to extract flood extent from SAR data an image classification technique must be applied (see Section 3, Grimaldi et al. (2016)). Binary maps of flood extent were assimilated by Lai et al. (2014) using a 4D variational approach to successfully estimate the roughness parameter over the flood plain. More recent variational approaches combine in situ observations with high-resolution hydrometeorology and satellite altimetry data into a hydraulic–hydrological numerical model (Pujol et al., 2022). Other DA methods applied to flooding include filtering methods such as particle filtering (PF) and ensemble Kalman filter (EnKF) methods (van Leeuwen, 2009; Evensen, 1994). These methods have been used to assimilate SAR-derived water levels (WL), found by intersecting the edge of the binary flood map with a digital elevation model (DEM) (Mason et al., 2007; Mason, Davenport, et al., 2012). SAR-derived WL only provide information at the flood edge and rely on the spatial resolution and the vertical accuracy of the underlying DEM (Dasgupta et al., 2021a), which makes them difficult to obtain.

A probabilistic flood mapping procedure for SAR data was first introduced by Giustarini et al. (2016). This created the potential for flooding probabilities to be assimilated directly. More recently, observation uncertainty associated with classifying flood extent from SAR data is openly available through the Copernicus Emergency Management Service (CEMS) (Copernicus Programme, 2021). Probabilistic flood maps from SAR were used to quan-

tify observation uncertainty for assimilating flood extents by Hostache et al. (2018) using a PF approach. Hostache et al. found an improvement in forecast performance of WL compared to those found by García-Pintado et al. (2013); García-Pintado et al. (2015) who assimilated WL using a local ensemble transform Kalman filter (ETKF). Cooper et al. (2019) considered the direct assimilation of SAR backscatter values, avoiding the step of deriving an intermediate flood extent first. Cooper et al. assimilated pseudo-observed SAR backscatter values directly into a 2D hydrodynamic model using an ETKF approach and compared three different observation operators. The authors found a new backscatter observation operator performed well compared to more conventional options.

In general, filtering methods suffer from degeneracy and limited persistence of improvement following DA. Several attempts have been made to address this. By using a mutual information-based likelihood function, Dasgupta et al. (2021b) increased the persistence of the DA impact on improvement of both water depth and discharge. Di Mauro et al. (2021, 2022) aimed to overcome degeneracy issues by applying a tempered particle filter (TPF) and particle mutation so that variation across the ensemble is maintained. Comparisons against sequential importance sampling methods concluded that the TPF approach improved the persistence of the DA. More recent studies using EnKF methods include the assimilation of distributed WL derived from optical satellite data combined with WL from river gauges (Annis et al., 2022). Annis et al. found that improvements from the DA by including the satellite-derived WL data spread further across the domain compared to those without. However, these improvements also suffered from limited persistence.

One major assumption made when assimilating SAR data is that the observation errors are from a Gaussian distribution. Nguyen, Ricci, Piacentini, Fatras, et al. (2023) used a novel approach by assimilating flood extent data expressed as a wet surface ratio (Nguyen, Ricci, Piacentini, Simon, & Rodriguez-suquet, 2023) and Gaussian anamorpho-

sis (GA) to transform the non-Gaussian observation errors into Gaussian. They found this gave slightly better results compared to DA without GA but concluded that the sub-optimal assumption of Gaussian errors in EnKF may still give reliable and valid results.

All of these studies involved the assimilation of satellite-derived data to update hydrodynamic model states and/or parameter values by taking a data assimilation cycling approach. The assimilation provides updated initial conditions ahead of the next forecast cycle. Our new approach differs significantly as we aim to develop a DA framework to assimilate probabilistic flood maps into a simulation library flood inundation forecasting system. We aim to improve the flood map selection creating a new analysis flood map without a feedback loop by utilising spatially distributed flood likelihood information derived from SAR data. We test the DA framework using forecast and optical satellite observation data from a major flood event in Pakistan, August 2022 (Floodlist, 2022).

In this chapter, the flood forecasting system and derivation of the static simulation library along with satellite-derived observations of flood likelihood are outlined in Section 6.3. The development of the DA framework and verification methods are described in Section 6.4. Section 6.5 presents an overview of the 2022 Pakistan flood and details of the data used. The DA framework successfully triggered flood maps in 4 out of 5 sub-catchments tested as shown in our results, discussed in Section 6.6. We conclude in Section 6.7 with recommendations for future work to improve the use of SAR flood likelihood data to update a simulation library flood forecasting system through data assimilation.

6.3 Simulation library forecasting system and observation data

Flood Foresight, a simulation library flood inundation forecasting system and its application for disaster risk reduction through FbF is outlined in Section 6.3.1. Section 6.3.2 details the derivation of the flood likelihood data from SAR that will be used as observation data in the assimilation process. The extraction of flood extent information from optical images that will be used for validation is explained in Section 6.3.3.

6.3.1 Flood Foresight and Forecast-based-Financing

The Global Flood Awareness System (GloFAS) couples global ensemble weather forecasts with a hydrological model and provides daily ensemble forecast river discharge at approximately 10 km grid size (v3.2, GloFAS (2021)). The Flood Foresight system (Revilla-Romero et al., 2017; Hooker et al., 2023a) is a fluvial, probabilistic flood inundation forecasting system. Flood Foresight is set up by dividing the catchment into 'Impact Zones' (IZ) or sub-catchments using the HydroBASINS data set (Lehner, 2014b). Each IZ in Flood Foresight is linked to a GloFAS grid cell that provides a 51 ensemble member forecast of river discharge. Flood Foresight contains a simulation library of precomputed flood depth and extent maps. The flood map library was hydrodynamically modelled using JFlow®, (Bradbrook, 2006) and RFlow using a detailed 30 m digital surface model. The maps were modelled at specific return period (RP) thresholds (20, 50, 100, 200, 500 and 1500 years). Subsequently, these were linearly interpolated at 5 intermediate intervals between each RP threshold and extrapolated between zero and the 20 year RP flood map (totalling 36 flood maps). Depending on the forecast discharge from GloFAS for each IZ, a flood map is selected from the simulation library. The flood map selected is determined by the RP threshold exceeded within each IZ. The resultant forecast flood map is created by stitching together individual IZ flood maps (at various RP's) and is produced daily out

to 10 days ahead.

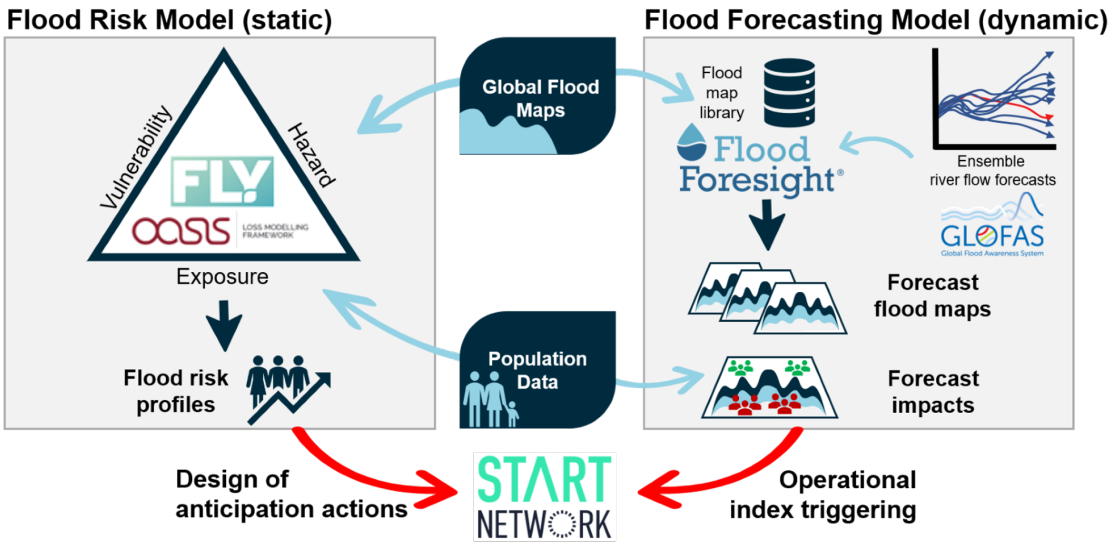


Figure 6.1: Flood Foresight/Start Network ensemble flood inundation forecast and population impacts work flow.

The charity Start Network (Start Network, 2022) brings together a group of over 80 humanitarian agencies and aims to develop local community-led, early action through a model of proactive funding to mitigate against the impacts of crises. JBA Consulting, in partnership with Start Network, have developed a Disaster Risk Financing (DRF) system for the Indus River basin in Pakistan that links Flood Foresight forecast flood maps to populations impacted by flooding (Fig. 6.1). For the purposes of setting FbF trigger threshold levels, the DRF system quantifies the flood risk to the population through a probabilistic global catastrophe risk model, FLY (Dunning, 2019). The analysis flood map produced here as a result of the DA relates to the dynamic operational index triggering.

6.3.2 Satellite-derived flood likelihood

The GFM service (GFM, 2021) combines the outputs of three different algorithms to extract flood extent and uncertainty information from Sentinel-1 SAR data. The process is automatic and runs continuously in near real-time (within 8 hours after image acquisition)

for every SAR image detecting flooding on a global-scale. The resulting flood information layers are openly available through open access. The mini-ensemble approach used for flood detection increases the confidence in the flood detecting capabilities of SAR. The first flood mapping algorithm developed by the Luxembourg Institute of Science and Technology (LIST) use a pair of SAR images (pre-flood and flood) and a hierarchical split-based change detection approach to classify permanent and flood waters (Chini et al., 2017). The classification uncertainty depends on the Bayesian inference classification probabilities. The second flood mapping algorithm by the German Aerospace Research Centre (DLR) uses a hierarchical tile-based thresholding approach and the optimization of the classification by combining various information sources using fuzzy-logic theory and region growing (Martinis et al., 2015; Twele et al., 2016). The uncertainty information depends on fuzzy memberships. The final algorithm from TU Wien uses the historical time series of the SAR backscatter values per pixel and classifies flood extent from the backscatter probability distribution (Wagner et al., 2020; Bauer-Marschallinger et al., 2022). The classification uncertainty is based on the Bayesian posterior probability. The output flood layer is derived using the mini-ensemble with a pixel classified as flooded where two out of three of the algorithms determines a flood class. The flood likelihood values are aggregated, first by converting each to lie in the same range [0, 100] before averaging the likelihood values. Regions where SAR is unable to detect flooding due to shadow or layover effects are removed from the classification process. This usually includes dense urban areas, densely vegetated areas, regions with steep slopes and regions that might appear flooded such as dry, sandy desert-like surfaces. The exclusion mask is available to download as an additional layer. Permanent and seasonal water bodies are classified separately as a reference water mask layer.

Krullikowski et al. (2023) applied and assessed the usefulness of GFM ensemble likelihood on two test sites in Myanmar and Somalia, both situated in challenging areas

for flood detection using SAR data. Krullikowski et al. found that the GFM ensemble likelihood layer resulted in a simplified appraisal of trust in the ensemble flood extent detection approach and provides more reliable and robust uncertainty information for detecting flooding compared to using a single algorithm only.

6.3.3 Optical Normalized Difference Water Index (NDWI)

Occasionally, optical satellite data can be useful for observations of flood extent. Flood detection from optical satellites depends on a near cloud free sky where the satellite acquisition coincides with the flood event. The Normalized Difference Water Index (NDWI) for flood and surface water detection is calculated with Sentinel-2 optical data using the green band (B03) and NIR band (B08) (Albertini et al., 2022). The NDWI is given by:

$$NDWI = \frac{B03 - B08}{B03 + B08}, \tag{6.1}$$

where positive values indicate water. Albertini et al. (2022) reviewed the performance of surface water and flood detection metrics using multispectral satellite data. They found that the average overall accuracy from previous flood studies applying NDWI to be 87.85% and for permanent surface water studies scored 94.41%. This included studies using data from different satellite sensors with spatial resolutions ranging from 10 m for Sentinel-2 to 500 m for Terra-Aqua MODIS.

6.4 Methods

6.4.1 Data assimilation framework

The aim of the DA framework is to update a previous forecast of flood inundation extent and depth from Flood Foresight (the background) where a non-trigger has occurred in the forecast system but where flooding was derived from concurrent satellite-based SAR data.

Using observation uncertainty information from the GFM flood likelihood layer (Section 6.3.2), we aim to improve the flood map selection for non-triggered IZ by minimising a cost function per IZ.

The data assimilation framework aims to update the state vector, $\mathbf{x} \in \mathbb{R}^n$ representing flood depths at each grid cell location. The total number of grid cells across an IZ is n, the total of observed unmasked grid cells is m. To find the optimum state accounting for observation uncertainty we define the observation *likelihood* term $P(\mathbf{y}|\mathbf{x})$ where observations, $\mathbf{y} \in \mathbb{R}^m$ have two possible binary outcomes, $y = 1$ flooded and $y = 0$ unflooded, following classification from SAR data. The likelihood term can be represented by a Bernoulli distribution (Lauritzen, 2023), defined as

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{m} L_i^{H(x_i)}(1 - L_i)^{1-H(x_i)}, \tag{6.2}$$

where L_i is the GFM flood likelihood value (see Section 6.3.2). The observation operator $\mathbf{H}(\mathbf{x})$ defined as

$$\mathbf{H}(\mathbf{x}) = \begin{cases} 1 & \text{flooded} \quad x_i > 0.2m \\ 0 & \text{unflooded} \quad \text{otherwise} \end{cases}, \tag{6.3}$$

acts to convert flood depths (state space) to a binary flood class (observation space) at unmasked grid cells. Thus, we exclude observation likelihood information for masked grid cells where SAR data cannot reliably detect flooding. To find the maximum posterior likelihood of the state variable, we take the negative log likelihood of $P(\mathbf{y}|\mathbf{x})$ and divide by the number of unmasked grid cells m to derive the cost function (averaged across unmasked grid cells per IZ)

$$J(\mathbf{x}) = -\frac{1}{m}\sum_{i=1}^{m}\big\{H(x_i)\ln(L_i) + (1 - H(x_i)\ln(1 - (L_i)\big\}. \tag{6.4}$$

The value of $J(\mathbf{x})$ is calculated per IZ by iterating through the flood map library (36 flood maps) and finding the flood map return period by minimising $J(\mathbf{x})$ and maximising the posterior likelihood, given the observation uncertainty data. To ensure that the minimum is reached across the flood map library, $J(\mathbf{x})$ is calculated for all 36 flood maps. Note that this is different to the standard data assimilation approach, minimisation would be accomplished via a gradient descent algorithm (Bannister, 2017). Following the assimilation process, replacing the non-triggered IZ with updated flood maps results in an analysis flood depth and extent map (the flood depth information is contained within the simulation library). This means that the analysis flood map remains consistent with the Flood Foresight system where the flood maps have been hydrodynamically modelled, i.e. they are physically realistic. Retaining the flood depth information is important for FbF applications for quantifying the risk of flood impacts. Since the observations have binary values (flooded/unflooded), we cannot distinguish between floods that have the same extent but different depths from the observation data. This property is inherited in the cost function.

6.4.2 Validation methods

The resulting analysis flood map, following assimilation of SAR-derived flood likelihood data, is validated by comparing against independent flood extent observation data derived from optical satellite data, the NDWI (Section 6.3.3). The results will be validated by calculating the Fraction Skill Score (FSS, Roberts and Lean (2008); Hooker et al. (2022)) and by mapping the performance on a Categorical Scale Map (CSM, Dey et al. (2014); Hooker et al. (2022, 2023a)). Both the FSS and the CSM avoid issues with the double penalty impact of conventional binary performance measures as well as the impact of flood magnitude on the skill score (Hooker et al., 2022). The FSS is based on the Brier Skill Score and uses a neighbourhood approach to determine the skillful spatial scale of the analysis flood map. The fraction of flooding within a given square neighbourhood size of length n is compared by calculating the mean-squared-error (MSE) between the analysis

and the validation flood maps to give

$$FSS_n = 1 - \frac{MSE_n}{MSE_{n(ref)}}, \tag{6.5}$$

where $MSE_{n(ref)}$ is a potential maximum MSE that depends on the fraction of flooding across the IZ on the analysis and the validation flood maps. A skilful scale is determined when $FSS \geq FSS_T$, the target FSS score, where $FSS_T \geq 0.5 + \frac{f_o}{2}$ depends on the fraction of observed flooding across the IZ, f_o. When the analysis and validation flood extents are equal in area across an IZ there is said to be no background bias and the maximum FSS is 1. Otherwise, the maximum asymptotic FSS (AFSS) is given by

$$AFSS = \frac{2f_o f_a}{f_o^2 + f_a^2}, \tag{6.6}$$

where f_a is the is the fraction of flooding on the analysis flood map per IZ.

The CSM plots a local agreement scale (S) at every grid cell. An overview of the method is presented here. Please see Chapter 3 or Dey et al. (2014); Hooker et al. (2022, 2023a) for full details of the methodology. A background bias between the analysis and verification flood maps that is deemed acceptable is predetermined. The pre-set bias is used to calculate an agreement criterion that must be reached by the flood map comparison calculation. The comparison begins at each grid cell $n = 1$, if the agreement criterion is met at grid level, the grid cell is labelled with an agreement scale $S = 0$. Where the criterion is not met, a larger neighbourhood size is compared (e.g. $n = 3$). The fraction flooded in each of the analysis and validation flood maps are compared and if the criterion is met, the agreement scale assigned would be $S = 1$. The process continues to larger neighbourhoods (e.g. $n = 7, S = 3$) until either the criterion is met or a predetermined limit is reached (S_{lim}, set to 9 for this application). The agreement scale at this limit would indicate a false alarm or miss for the grid cell. Note that the relationship between

n, the neighbourhood size used for the FSS, and S is given by $S = (n-1)/2$. The agreement scales are combined with data from a conventional contingency map (Stephens et al., 2014) and are mapped across an IZ. The CSM indicates a location-specific level of agreement and shows where the flood maps are over- or under-estimating flooding.

6.5 Pakistan flood 2022

6.5.1 Event overview

In Spring 2022, Pakistan saw a record-breaking heatwave with temperatures exceeding 50°C. The heat exacerbated upstream glacial snow melt feeding the Indus River basin, which runs over 3000 km across the length of Pakistan, draining the Himalayas to the Arabian Sea. An intense monsoon season followed in July and August, driving multiple flood-producing mechanisms including multi-day extreme precipitation that was the primary driver of floods (World Weather Attribution, 2022; Nanditha et al., 2023). Attribution studies indicate that the 5-day maximum rainfall over the provinces Sindh and Balochistan, which led to catastrophic flooding, was made 75% more intense by 1.2°C of global warming (World Weather Attribution, 2022). The northern Sindh province received an estimated 442.5 mm of rainfall in August, 784% more than usually recorded, causing inundation of 55,000 km^2 across the region (Floodlist, 2022). Despite early warnings of the potential for significant flooding from GloFAS, the unimaginable scale and magnitude of the flood impacted over 33 million people with over 1700 lives lost and costing more than $40 billion in economic damages (Floodlist, 2022; World Resources Institute, 2023).

6.5.2 Data

The DA framework (Section 6.4.1) was tested in the northern Sindh province where widespread flooding occurred during August 2022. Figure 6.2 maps the NDWI derived from Sentinel-2 optical data (Section 6.3.3) that was used to verify the resultant analysis

flood map following DA. Local reports and photographs of flooding were made in the cities of Sukkur and Larkana (DAWN, 2022; The Guardian, 2022; Sky News, 2022). The DA

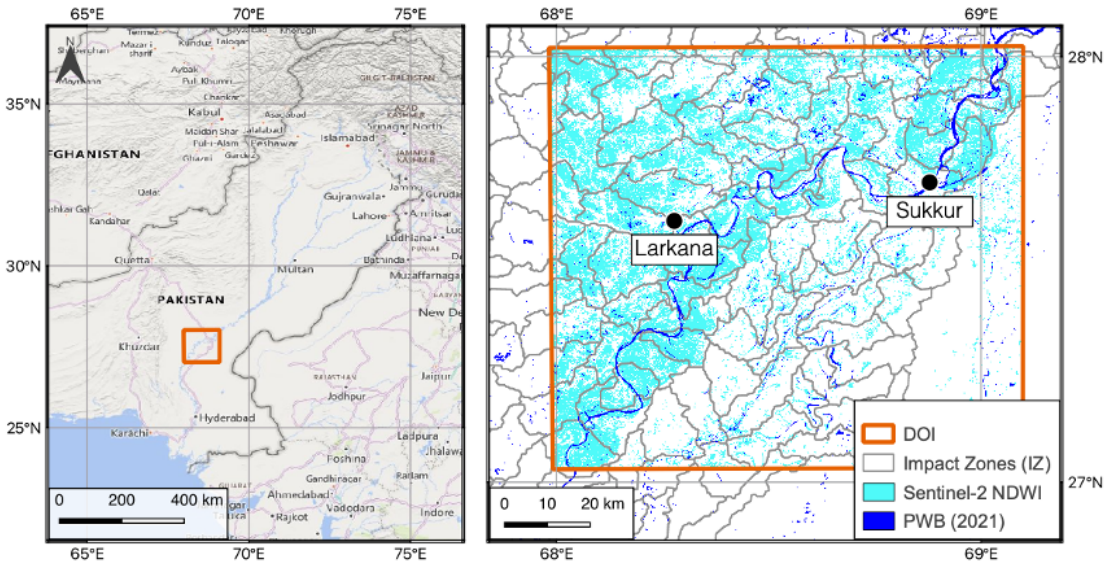


Figure 6.2: The domain of interest (DOI) is located on the Indus Basin, Sindh province, Pakistan (left). The region is divided into sub-catchments or Impact Zones (IZ) in Flood Foresight (right). Satellite-derived flooding (NDWI) from Sentinel-2 data (Section 6.3.3) from 31 August 2022 is highlighted along with permanent water bodies (PWB).

was applied to 5 IZ covering 3 different scenarios (Fig. 6.3): (1) One IZ where a large proportion of the IZ is a dense urban area and is masked (where the GFM product is currently unable to detect flooding), Sukkur (S), see Figure 6.3(a); (2) Two IZ with mixed urban and rural areas, Larkana north and south (LN, LS); and (3) two flood edge locations (FE1, FE2). The GFM flood likelihood data used to represent observation uncertainty in the DA is mapped in Figure 6.3(a) where darker shades of orange indicate a higher likelihood of flooding (Section 6.3.2). The forecast data from Flood Foresight is mapped in Figure 6.3(b) where the purple shades indicate the maximum return period flood map triggered by the system from 10 to 31 August, 2022. Each of the IZ selected were non-triggered IZ during this period. The driving forecast river discharge data from GloFAS did not reach the required threshold to trigger a flood map along the central Indus channel. This

is likely due to poor calibration of GloFAS due to a lack of observation of river stage or discharge. The Sukkur barrage operations for diversion and altering of river water flows are not currently included in the GloFAS system, which makes forecasts unreliable along this stretch of the Indus River.

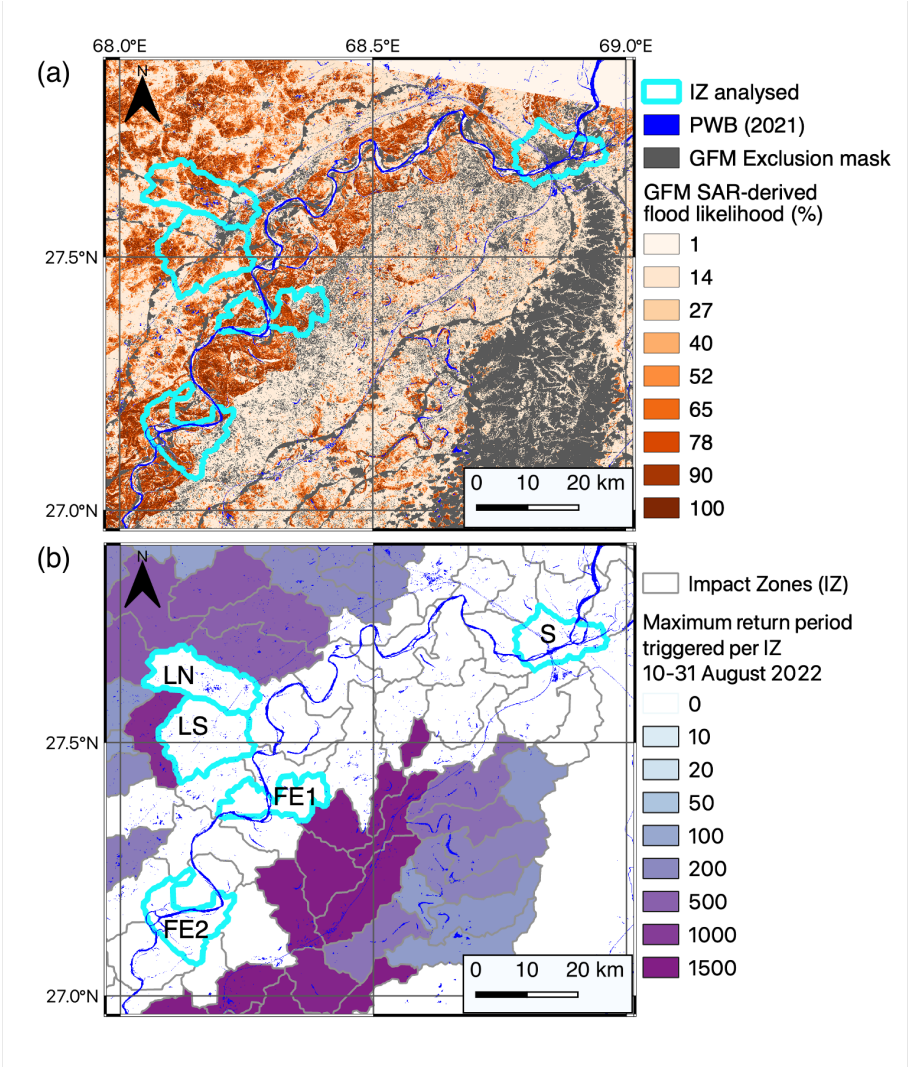


Figure 6.3: (a) GFM flood likelihood derived from Sentinel-1 SAR data, masked areas (grey) indicate where flooding cannot be reliably detected from SAR data. (b) The maximum return period threshold triggered per IZ by the Flood Foresight system during peak flooding 10-31 August, 2022. Five non-triggered IZ of interest labelled: S - Sukkur, LN - Larkana North, LS - Larkana South, FE1 - flood edge 1, FE2 - flood edge 2.

6.6 Results and discussion

Results are presented for the 3 scenarios tested in the following section. We discuss the benefits and limitations of the approach and how it could potentially be modified for improved performance.

6.6.1 Scenario 1

The DA framework was applied to 3 different scenarios totalling 5 IZ. The first scenario tested was an IZ centred on the city of Sukkur. Sukkur is located just south of a large barrage, used to control flood waters. Significant flooding was observed locally in the Sukkur region (Sky News, 2022), however the dense city centre means that flooding is difficult to detect using Sentinel-1 imagery at 20 m spatial resolution. Around one third of the IZ is masked by the GFM process (Fig. 6.4(c)) but high flood likelihood values are visible across some areas of the IZ (Fig. 6.3(a)). The aim is to test whether the DA framework can select a flood map from the simulation library based on limited usable SAR data.

The value of the cost function $J(\mathbf{x})$ from eqn. (4), Section 6.4.1 is plotted against the RP value of each flood map from the simulation library in Figure 6.4(a). The cost function was minimised at the lowest RP flood map (3 years) and we found that a 'no flood' map gave a slightly lower value of $J(\mathbf{x})$. In this instance, the lower RP flood maps over-estimated flooding in areas where low flood likelihood values were derived from SAR. The influence of the Sukkur barrage and river canals running across the IZ made the hydrodynamic modelling more difficult. Also, the flood maps do not include local defence information and the flood map interpolation process is highly uncertain at RP less than 20 years. The results mean that no flood map was triggered following the DA (Fig. 6.4(b and c)) with the CSM map (Fig. 6.4(d)) indicating where the flooding was underestimated,

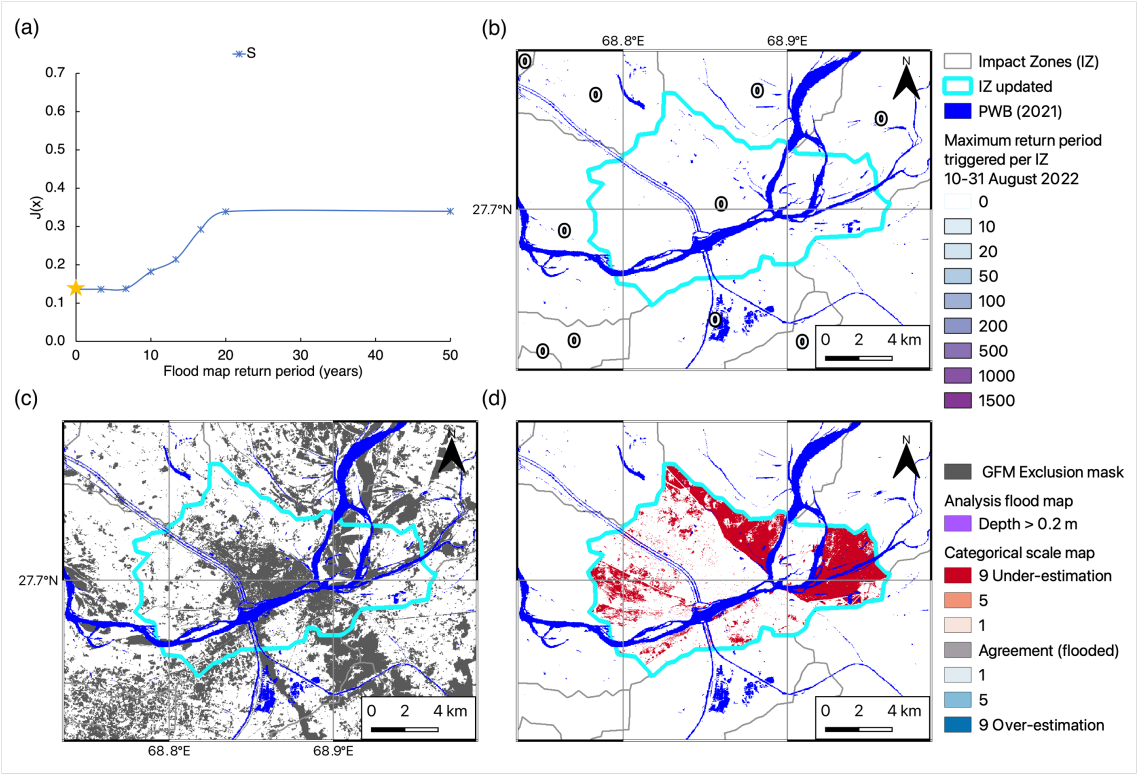


Figure 6.4: Scenario 1: Sukkur, a dense urban area. (a) The cost function (eqn. (4)) plotted against the RP value, the yellow star highlights the minimum value of $J(\mathbf{x})$. (b) The analysis RP triggered following DA. (c) The analysis flood extent map (note that no map was triggered for Sukkur) and (d) the CSM comparing the analysis flood extent map against Sentinel-2 NDWI.

particularly upstream of the Sukkur Barrage, with no flood map selected.

6.6.2 Scenario 2

The second scenario focused on a mixed urban and rural area with 2 IZ chosen around Larkana city. The dense urban area is again masked and is split across the 2 IZ (Fig. 6.3(a)), but there are large unmasked areas with high and low flood likelihood values. The assimilation results for scenario 2 are plotted in Figure 6.5(a) where the cost function minimum value is similar for both LN and LS with a 7 year RP flood map triggered for LN and a 13 year RP flood map triggered for LS (Fig. 6.5(b)). The resultant analysis flood maps selected (Fig. 6.5(c)) also indicate flooding within Larkana city, overlapping the

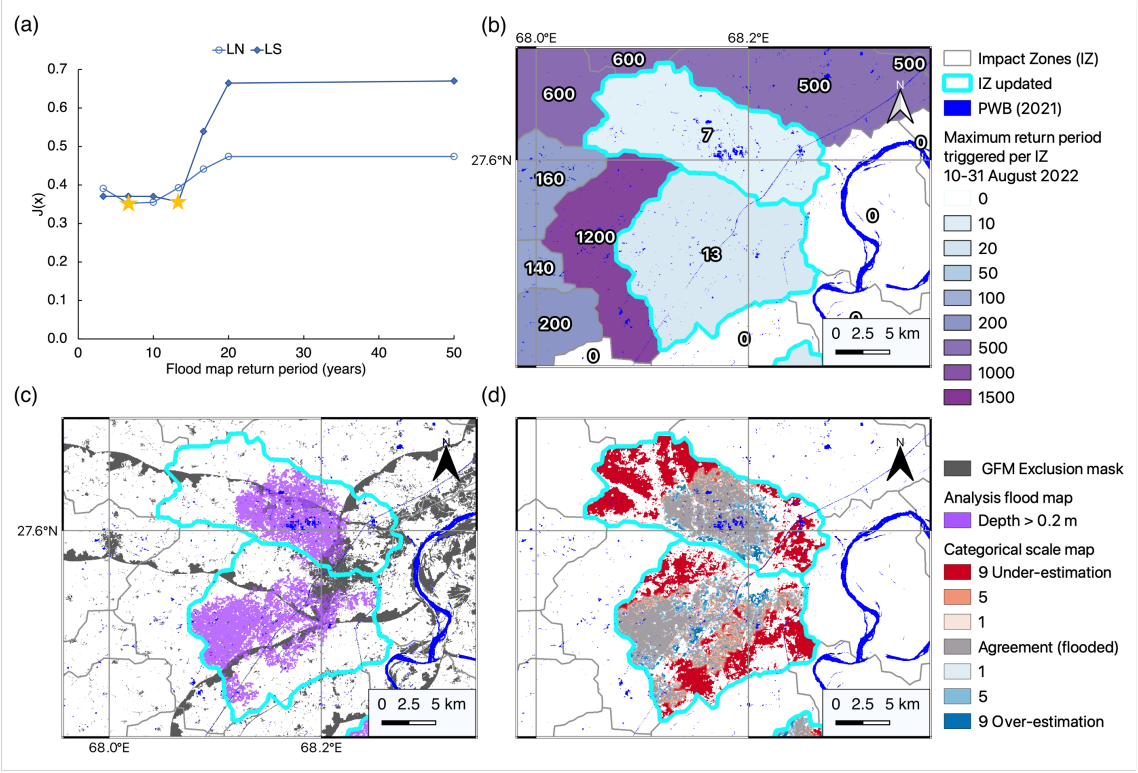


Figure 6.5: Scenario 2: Larkana, a mixed urban and rural area. (a) The cost function (eqn. (4)) plotted against the RP value, the yellow star highlights the minimum value of $J(\mathbf{x})$. (b) The analysis RP triggered following DA. (c) The analysis flood extent map and (d) the CSM comparing the analysis flood extent map against Sentinel-2 NDWI.

masked area. This is consistent with local observations and is important for population impact calculations for FbF schemes. The CSM indicates that whilst a large area is now correctly indicating flooding there are also large areas that are under-estimated by the analysis flood map (Fig. 6.5(d)). The neighbouring IZ that were triggered by the forecast system (Fig. 6.5(b)) are at much higher RP thresholds than the ones selected following the DA. By inspecting higher RP flood maps than those selected by the DA (for LN and LS) it became clear that these were over-estimating flooding in locations where low flood likelihood values were located causing $J(\mathbf{x})$ (Fig. 6.5(a)) to increase.

One potential solution to the inconsistency seen across the domain could be overcome

by including information from the forecast system. The assimilation process could be carried out across a region including multiple IZ at the same time, rather than considering individual IZ. Conditions could be imposed, such as a consistent flood depth across IZ boundaries away from the flood edge. One way to impose some smoothness would be through the use of a background error term. Note that the background error is the prior or forecast error. Information from neighbouring IZ could be spread across a domain by the background error covariance ($\mathbf{B}$) matrix used in variational DA (Bannister, 2008). The $\mathbf{B}$ matrix could be calculated offline using the content of the simulation library.

Once the entire IZ becomes inundated at a 20-year RP, $J(\mathbf{x})$ remains constant with increasing RP (Fig. 6.5(a)). Although the depth values are increasing, there are no significant changes in flood extent possible across the IZ, meaning the cost function cannot distinguish between flood maps over a 20-year RP. For the IZ tested here, the minimum has already occurred at lower RP, but it is possible that the minimum could occur where $J(\mathbf{x})$ is constant, meaning a range of potential RP are possible solutions. In order to distinguish between equally plausible flood maps, additional observation data would be required to measure flood depth. Flood depth data for sufficiently large floods could be derived from satellite altimetry data such as the Surface Water and Ocean Topography (SWOT) mission (Frasson et al., 2019; de Moraes Frasson et al., 2023).

6.6.3 Scenario 3

The final scenario investigates the impact of assimilating SAR-derived flood likelihood data on flood map selection where the flood edge lies within the IZ. The cost function value of $J(\mathbf{x})$ drops relatively sharply for FE1 in the north (Fig. 6.6(a)) from 0.56 at 3 years RP to a minimum of 0.34 at 17 years RP. Further south, $J(\mathbf{x})$ for FE2 is initially lower at 0.21 at 3 years RP, gradually decreasing to a minimum of 0.17 at 20 years RP. The shape of the cost function shows a smoother descent to a minimum compared to scenarios

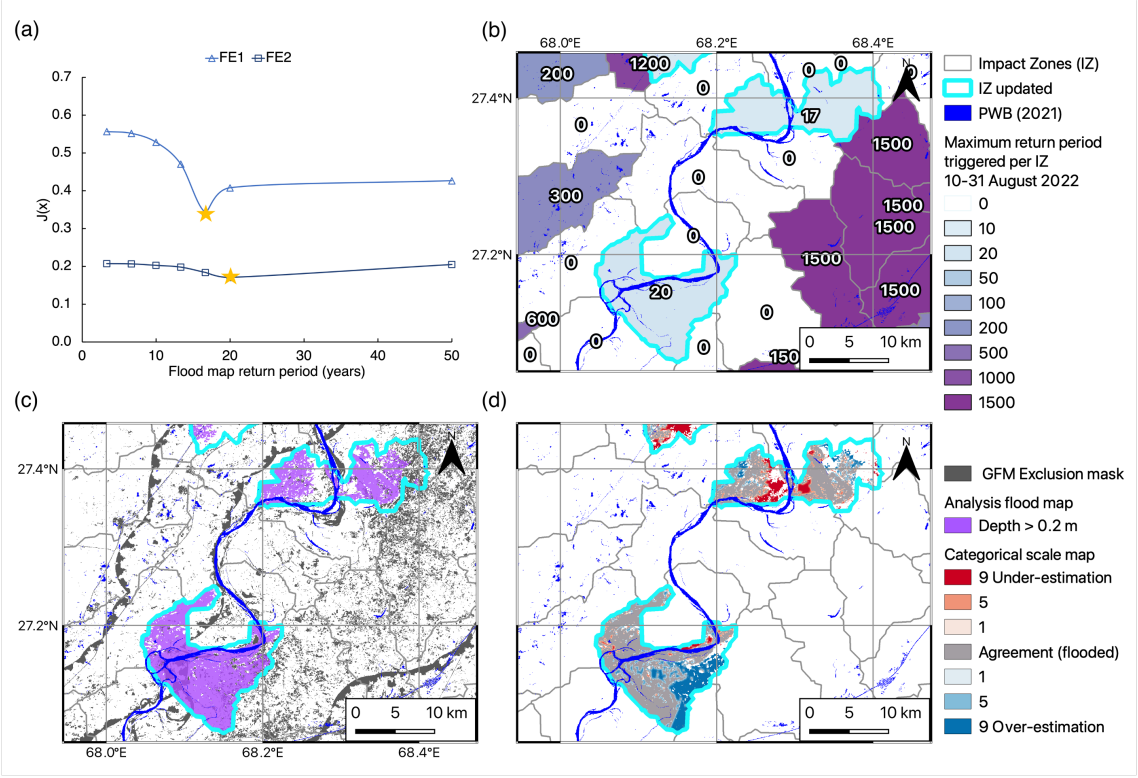


Figure 6.6: Scenario 3: Flood edge location. (a) The cost function (eqn. (4)) plotted against the RP value, the yellow star highlights the minimum value of $J(\mathbf{x})$. (b) The analysis RP triggered following DA. (c) The analysis flood extent map and (d) the CSM comparing the analysis flood extent map against Sentinel-2 NDWI.

1 and 2 as there is more variation in flood extent between flood maps at different RP at the flood edge location. In similarity to scenario 2 (Larkana), neighbouring IZ are again at very high RP levels (Fig. 6.6(b)). The analysis flood map for FE1 does not reach the edge of the IZ, whereas FE2 virtually flood fills the IZ (Fig. 6.6(c)). The CSM (Fig. 6.6(d)) shows that some flooding close to the main Indus River has been under-estimated in FE1 but with overall good accuracy and limited over-estimation. FE2 shows over-estimation in the west but again a large area in agreement with the flood extent derived from Sentinel-2 NDWI. Using flood extent observation likelihood data would be more useful where the flood edge stays within the IZ tested for the maximum RP flood map. There would be more chance of variation in the cost function value across the full range of RP flood maps.

In contrast, for a (near) flood filled IZ there would be less variation seen in $J(\mathbf{x})$ due to limited changes in flood extent. Future work could focus on assimilating flood edge IZ and sharing information with neighbouring IZ by including a background term to update regions closer to the river channel where the IZ are more likely to be flood filled (Section 6.6.2).

6.6.4 Analysis flood map validation

Table 6.1: Validation skill scores for each IZ analysis flood map compared against independent Sentinel-2 NDWI

IZ code	Analysis RP	FSS at $(n = 1)$	FSS_T	$AFSS$	n at FSS_T
S	0	0	n/a	n/a	n/a
LN	7	0.32	0.64	0.40	$AFSS < FSS_T$
LS	13	0.38	0.65	0.51	$AFSS < FSS_T$
FE1	17	0.55	0.66	0.71	n = 35 (525 m)
FE2	20	0.49	0.63	0.77	n = 15 (225 m)

In Table 6.1 the FSS (Section 6.4.2) has been calculated by comparing the analysis flood map selected per IZ with the corresponding Sentinel-2 NDWI representing observed flooding, with permanent water bodies excluded from the validation. The target skill score FSS_T and the asymptotic FSS $AFSS$ are also calculated. The FSS for scenario 1 (Sukkur) was 0 as no flood map was triggered. For scenario 2 (LN and LS) the FSS score at grid level $n = 1$ (0.32 and 0.38) is around half of FSS_T. Usually, by increasing the neighborhood size the value of FSS also increases, eventually exceeding FSS_T. In this case $AFSS$ (which is calculated using the fraction flooded across the IZ from both the analysis and observed flood maps) is less than FSS_T meaning the total differences in flood extent are too large for the FSS to reach or exceed FSS_T. The result of this is that there

is not a meaningful or skilful scale of the analysis flood maps for scenario 2. This is due to the under-estimation of flood extent seen on the CSM (Fig. 6.5(d)). For scenario 3, $FSS_T < AFSS$ which makes it possible to calculate a skilful scale where $FSS \geq FSS_T$. For FE1 this occurs when $n = 35$ or 525 m and FE2 at $n = 15$ or 225 m confirming that FE2 was the most accurate analysis. These results confirm that future work should focus on flood edge IZ first during the assimilation process.

6.7 Conclusions

In this chapter we introduced a new DA framework to update and improve the flood map selection within a flood forecasting system designed for FbF applications. Open access flood likelihood data derived from satellite-based SAR is used to update the flood map selection for previously non-triggered sub-catchments or IZ during a flood event. The framework is tested on the catastrophic flooding in Pakistan, August 2022 for 3 scenarios.

The first scenario tested an IZ where limited useful SAR data was available due to a dense urban area. This resulted in no flood map selection following the DA. The second scenario, where two IZ contained a mix of urban and rural areas did trigger flood maps but at low RP levels, relative to neighbouring IZ that were previously triggered. This resulted in under-estimation of the flood extent. However, the analysis flood map included flooding across parts of the city of Larkana. Information from the flood likelihood data from other areas of the IZ could select a flood map that included urban flooding. This is useful for FbF applications where population impacts are considered. The final scenario examined flood edge locations and these gave the best results as the variation in flood extent selected higher RP flood maps that were more closely matched to the validation data. The skilful scale of the analysis flood maps in the flood edge IZ was 225 to 525 m. Out of the 5 IZ tested, 4 resulted in a flood map selection with the dense urban area and

limited SAR coverage the exception. Each of these 4 IZ were non-triggered, and now they are, which is beneficial, even if the extents are not perfect.

The non-triggered flood maps could be updated quickly following the production of the GFM flood likelihood layer (approximately 8-hours after SAR acquisition). Although observed flood extent is used in the assimilation, the flood maps selected contain depth values that are already linked to a catastrophe risk model and population impact maps. Therefore the analysis is suitable to inform secondary financing schemes for flood response and recovery, during an event. Improvements could be made by the inclusion of prior information from the simulation library system. An additional background term in the data assimilation framework could improve the consistency of the flood maps selected across a region.

An additional benefit of our approach is that the analysis flood map could be used in a feedback loop to update the river discharge (e.g. in the associated GloFAS grid cell). This could be useful for hydrological model calibration or in updating the initial conditions for the next forecast.

Future options for optimising the simulation library flood maps (where remote flood depth observations are available) could use a conventional iterative approach (such as gradient descent methods) to step through depth values within each individual grid cell. The resultant analysis depth map would represent the best estimate of the *true* flood extent and could be used to inform secondary insurance payments. However, the analysis flood map would no longer be consistent with the simulation library system and it would be difficult to use this to update the hydrological model (i.e. in a feedback loop). The analysis will possibly include surface water flooding (SWF), which is likely for large relatively flat river basins where monsoon rainfall contributes to flooding, such as the Indus basin in Pakistan

and the Ganges-Brahmaputra-Meghna catchments of India and Bangladesh. The analysis is more likely to represent flooding as observed 'from the ground', which makes it more consistent with locally observed flooding. This combined SWF-fluvial analysis flood map would mean that secondary insurance payments are more fairly distributed as they do not depend on the type of flooding mapped within the fluvial simulation library. The inclusion of SWF in the analysis flood map causes additional inconsistencies with the fluvial flood forecasting system, which makes the analysis map less useful for updating the system in a feedback loop ahead of the next forecast.

In this application of the DA framework to selected IZ we were able to analyse the entire flood map library. For operational applications across a wider area, optimal iteration methods could be used to save computation time and storage.

6.8 Chapter summary

In this chapter we address some of the limitations of simulation library flood forecasting systems described in Chapter 5 by creating an assimilation framework using satellite flood likelihood data to improve the flood map selection for non-triggered flood maps. We applied the framework to three different scenarios where the flood maps were not previously triggered following flooding in Pakistan in 2022: a large city area with limited SAR flood detection (due to a dense urban area); sub-catchments with mixed rural and urban areas; and flood edge sub-catchments. The updated flood maps were evaluated against independent satellite data using the scale-selective verification methods introduced in Chapter 3.

Chapter 7

Conclusions

7.1 Main conclusions

In this thesis we have applied scale-selective verification metrics in a new application to evaluate several aspects of flood map forecasts from a simulation library system such as the spatial accuracy and the ensemble spatial spread-skill. The forecast maps were validated against SAR-derived observations of flooding for case studies both in the UK and internationally. We applied a new scale-selective verification approach to a multi-system comparison and addressed some of the limitations found by developing a new data assimilation (DA) framework to improve the flood map analysis. The main conclusions, which answer the four questions posed in Chapter 1, are:

1. **What are the skilful spatial scales in flood inundation forecasts made using a simulation library approach?**

    - In Chapter 3 we describe a scale-selective approach to evaluate forecast flood inundation maps. A skilful spatial scale for forecast flood maps was determined by calculating the Fraction Skill Score, a validation metric, found by comparing the forecast flood maps against a satellite SAR-derived observation of flooding

across a range of neighbourhood sizes. A target skill score was calculated and this depends on the magnitude of the observed flood. We found that the skilful spatial scale determined for the flood edge was more sensitive to changes in spatial accuracy and spatial scale compared to the skilful scale found by evaluating the entire flood extent. Categorical scale maps indicated that the skilful scale varies with location within a domain. Based on the results from evaluating a forecast of severe flooding on the River Wye and the River Lugg (UK) in February 2020, we found that the skilful scales depend on forecast inputs and can be linked to catchment characteristics such as flood plain topography, land use type and urban infrastructure. There are multiple operational uses of scale-selective evaluation of forecast flood maps such as model development and improvement and determining a meaningful spatial scale at which to present the forecast flood maps to end users.

2. **How skilfully does an ensemble of forecast flood maps represent the spatial uncertainty within the flood forecast?**

   - In Chapter 4 we presented a new scale-selective approach to assess the spatial predictability and spread-skill of an ensemble flood map forecast that accounts for the individual spatial prediction of flood extent held within each ensemble member flood map. The method determines, at specific locations within the domain, whether the ensemble forecast is over-, under- or well-spread. The spatial spread-skill relationship was mapped onto our new spatial spread-skill map. Results following the application of the method to an ensemble forecast of flooding on the Brahmaputra (Assam, India) in August 2017 show that the spatial spread-skill relationship is highly spatially variable and depends on multiple factors throughout the flood forecasting system chain. We found that one ensemble member flood map outperformed all others including summary

flood maps such as the ensemble median and a total combined ensemble.

3. **How useful are scale-selective evaluation approaches when applied to multiple flood forecasting systems?**

   - In Chapter 5 we investigated a new application of scale-selective verification by evaluating the performance of three flood forecasting systems. Two simulation library systems, Flood Foresight (30 m) and GloFAS Rapid Flood Mapping (1000 m) and one hydrodynamically modelled system, the Bangladesh FFWC Super Model (300 m), all made predictions of flood extent at different spatial scales (grid lengths, shown in brackets) for the Jamuna River flood, Bangladesh, July 2020. Our results show that the simulation library system accuracy critically depends on the discharge return period threshold set to trigger a flood map selection and the number of hydrological model ensemble members that must exceed it. At short forecast lead times, the Super Model outperforms the other systems in three out of four districts. Near to the Bangladesh border, the trans-boundary benefits of the two global systems are evident, with both outperforming the local model. We conclude that a scale-selective verification approach can quantify the skill of systems operating at different spatial scales so that their benefits and limitations can be evaluated.

4. **Does a data assimilation framework improve the analysis of flood inundation from a simulation library system?**

   - In Chapter 6 a DA framework was developed to integrate probabilistic flood extent maps from satellite-based SAR sensors into the simulation library flood map selection process. The method was tested on the severe flood event in Pakistan, 2022, where several sub-catchments resulted in a non-trigger of the forecast-based financing system deployed here, despite significant flooding evi-

dent from earth observation data. The DA successfully triggered flood maps in 4 out of 5 sub-catchments tested and we found that evaluating sub-catchments over the flood edge gave the best results. The analysis flood map remains consistent with the forecast-based financing system so could be easily linked to impacts to inform secondary finance payments during an event.

7.2 Thesis synthesis

The overall aim of this thesis was to improve flood inundation forecasts using satellite derived observations of flooding. It is important to understand meaningful length scales when comparing spatial observations with forecasts. Assimilating observations at grid level could lead to over-fitting or indicating over-confidence in the forecast. The forecast improvement may suffer from limited persistence through time. Finding meaningful or skilful spatial scales was the focus of Chapter 1. Visualising the skilful scale on an agreement scale map, brought additional benefits to the model evaluation, since improvements could be targeted to specific locations. The scale-selective evaluation approach overcomes issues such as the double penalty impact of high resolution verification and the impact of flood magnitude on skill scores. The method developed was used as an evaluation tool in Chapters 3 to 6.

In Chapter 4 the scale-selective verification approach developed in Chapter 3 was extended to an ensemble forecast system. The ensemble verification approach adds an extra dimension, the ensemble spread. The spread-skill relationship was evaluated for a simulation library system predicting flood inundation. The results showed that the most skilful member may lie in an ensemble outlier and the ensemble mean flood map may miss local detail. These are important considerations for interpreting ensemble flood maps in forecast applications.

The scale-selective verification approach has been applied to a simulation library system in Chapters 3 and 4. In Chapter 5 we also verify a local flood forecasting system along with two simulation library systems to show how the verification approach can be useful in multi-system evaluations where the flood maps are presented at different spatial scales. The skilful scale can be interpreted as a displacement distance, which can be directly compared across the systems. The simulation library system limitations found in Chapter 5 gave motivation to develop the data assimilation framework in Chapter 6.

In some forecast situations the forecast discharge driving the flood map selection may not exceed the return period threshold set to trigger the flood map selection. This can result in missed opportunities for forecast-based financing where flooding did occur but was not well forecast. The data assimilation approach developed in Chapter 6 aimed to overcome this by using satellite-derived flood likelihood data to improve the flood map selection from the simulation library, creating a new analysis flood map. The resultant analysis flood map was evaluated using the approaches developed earlier in Chapter 1.

7.3 Limitations

The satellite-derived observations of flooding from SAR data used for verification in Chapters 3, 4, and 5 make the assumption that the observed flooding extent is accurate. The limitations of SAR-derived flooding are discussed in each Chapter. However, new methods such as masking areas where SAR is unable to reliably detect flooding, described in Chapter 6 would have benefited the results in previous chapters. The inclusion of flood defences in the forecast flood maps would also give a better evaluation of model performance.

The flood inundation evaluation in Chapter 1 was developed on a conventional grid

across a region of interest. The neighbourhood approach may not be meaningful where the neighbourhood extends beyond a sub-catchment boundary, where flood waters are contained within each sub-catchment. The evaluation and data assimilation could be carried out on a sub-catchment level, rather than across a domain. We evaluated the spatial extent of flooding. Flood depth evaluation would add an extra dimension to the analysis, however this is complicated by the need for a detailed digital terrain model. Extrapolating water levels across a region from observations at the flood edge creates large uncertainties in the flood depth estimation. More recent techniques may reduce these uncertainties (Amitrano et al., 2024; Betterle & Salamon, 2024) so that agreement scales could be calculated for flood depths.

In Chapter 4 the spread-skill relationship was evaluated in detail for one forecast lead time. At the time, this was the only data available. It would be beneficial to evaluate the spread-skill at longer forecast lead times and for different flooding events. The multi-system comparison in Chapter 5 was limited by data availability, the study would benefit from additional data from the local model at more lead times to compare against the simulation library systems. The data assimilation framework in Chapter 6 uncovered occasions where the sub-catchments became flood filled and distinguishing between flood maps proved difficult. Recommendations were made in Chapter 7 and the next section to address this limitation.

7.4 Recommendations for future work

The application of scale-selective verification methods (Chapter 3) and the ensemble spatial-spread skill methods (Chapter 4) could be applied to other flood events. The case studies might include additional emphasis on the physical characteristics of the flooding or the driving meteorological situation, e.g., do the antecedent soil moisture conditions

impact the skilful scale of the forecast? This relies on the availability of a historical catalogue of events with flood inundation predictions. It would be interesting to investigate the impact of forecast lead time or an increased spatial resolution of the driving model, on the ensemble spatial-spread skill and the probabilistic skilful spatial scale (Necker et al., 2023).

An important aspect of developing an inundation flood forecasting system is to determine the most useful way to present both deterministic and ensemble flood map forecasts. Future work might include investigating the use of presenting forecasts to end-users using variable spatial scales, so that the scale reflects the forecast uncertainty. There is also the potential to smooth and contour the flood edge, particularly in rural areas, so that it is more representative of a flooding situation. Presentation options of ensemble forecasts could be investigated such as spatially clustering similar ensemble members into groups of flood extent or to present a most likely, best and worst case scenario ensemble flood map.

Several recommendations were made at the end of Chapter 6 for future data assimilation approaches applied to simulation library systems. An additional background term in the data assimilation framework could improve the consistency of the flood maps selected across a region. A recent paper assimilates precipitation features based on the FSS. Results based on synthetic forecast experiments showed that the proposed method improved the forecast accuracy (Otsuka et al., 2023). Future flood inundation assimilation could incorporate the skilful scale found using the scale-selective methods presented in Chapter 3. Assimilating at a coarser scale would save on computation time and reduce the risk of over-fitting. For example, if the flood maps were found to be skilful at $n = 5$, the fraction flooded at $n = 3$ could be used to improve the flood map selection from the simulation library.

References

Albertini, C., Gioia, A., Iacobellis, V., & Manfreda, S. (2022). Detection of Surface Water and Floods with Multispectral Satellites. Remote Sensing, 14(23). doi: 10.3390/rs14236005

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., & Pappenberger, F. (2013). GloFAS-global ensemble streamflow forecasting and flood early warning. Hydrology and Earth System Sciences, 17(3), 1161–1175. doi: 10.5194/hess-17-1161-2013

Alfonso, L., Mukolwe, M. M., & Di Baldassarre, G. (2016, feb). Probabilistic Flood Maps to support decision-making: Mapping the Value of Information. Water Resources Research, 52(2), 1026–1043. doi: 10.1002/2015WR017378

Amitrano, D., Di Martino, G., Di Simone, A., & Imperatore, P. (2024). Flood Detection with SAR: A Review of Techniques and Datasets. Remote Sensing, 16(4). Retrieved from https://www.mdpi.com/2072-4292/16/4/656 doi: 10.3390/rs16040656

Anderson, S. R., Csima, G., Moore, R. J., Mittermaier, M., & Cole, S. J. (2019). Towards operational joint river flow and precipitation ensemble verification: considerations and strategies given limited ensemble records. Journal of Hydrology, 577(March), 123966. Retrieved from https://doi.org/10.1016/j.jhydrol.2019.123966 doi: 10.1016/j.jhydrol.2019.123966

Annis, A., Nardi, F., & Castelli, F. (2022). Simultaneous assimilation of water levels from

river gauges and satellite flood maps for near-real-time flood mapping. Hydrology and Earth System Sciences, 26(4), 1019–1041. doi: 10.5194/hess-26-1019-2022

Apel, H., Vorogushyn, S., & Merz, B. (2022). Brief communication: Impact forecasting could substantially improve the emergency management of deadly floods: case study July 2021 floods in Germany. Natural Hazards and Earth System Sciences, 22(9), 3005–3014. doi: 10.5194/nhess-22-3005-2022

Arnal, L., Anspoks, L., Manson, S., Neumann, J., Norton, T., Stephens, E., ... Cloke, H. L. (2020). "Are We talking just a bit of water out of bank? or is it Armageddon?" Front line perspectives on transitioning to probabilistic fluvial flood forecasts in England. Geoscience Communication, 3(2), 203–232. doi: 10.5194/gc-3-203-2020

ASDMA. (2017). Assam state disaster management authority flood alert. `https://asdma.assam.gov.in/information-services/assam-flood-report`, last access 10th November 2021.

Bannister, R. N. (2008). Review A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. Q. J. R. Meteorol. Soc, 134, 1951–1970. Retrieved from `www.interscience.wiley.com` doi: 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(703), 607–633. doi: 10.1002/qj.2982

Bates, P. D. (2022). Flood Inundation Prediction. Annual Review of Fluid Mechanics, 54(1), 287–315. Retrieved from `https://doi.org/10.1146/annurev-fluid-030121-113138` doi: 10.1146/annurev-fluid-030121-113138

Bauer-Marschallinger, B., Cao, S., Tupas, M. E., Roth, F., Navacchi, C., Melzer, T., ... Wagner, W. (2022). Satellite-Based Flood Mapping through Bayesian Inference from a Sentinel-1 SAR Datacube. Remote Sensing, 14(15), 1–28. doi: 10.3390/rs14153673

Ben Bouallègue, Z., & Theis, S. E. (2014). Spatial techniques applied to precipitation en-

semble forecasts: From verification results to probabilistic products. Meteorological Applications, 21(4), 922–929. doi: 10.1002/met.1435

Bernard, A., Long, N., Becker, M., Khan, J., & Fanchette, S. (2022). Bangladesh's vulnerability to cyclonic coastal flooding. Natural Hazards and Earth System Sciences, 22(3), 729–751. Retrieved from https://nhess.copernicus.org/articles/22/729/2022/ doi: 10.5194/nhess-22-729-2022

Betterle, A., & Salamon, P. (2024). Water depth estimate and flood extent enhancement for satellite-based inundation maps. Natural Hazards and Earth System Sciences Discussions, 2024, 1–21. Retrieved from https://nhess.copernicus.org/preprints/nhess-2024-22/ doi: 10.5194/nhess-2024-22

Beven, K. (2016). Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. Hydrological Sciences Journal, 61(9), 1652–1665. Retrieved from http://dx.doi.org/10.1080/02626667.2015.1031761 doi: 10.1080/02626667.2015.1031761

Boelee, L. (2022). Evaluation of global flood forecasts in ungauged catchments (Doctoral dissertation, The Open University, UK). Retrieved from http://oro.open.ac.uk/84856/

Boelee, L., Lumbroso, D. M., Samuels, P. G., & Cloke, H. L. (2019). Estimation of uncertainty in flood forecasts—A comparison of methods. Journal of Flood Risk Management. doi: 10.1111/jfr3.12516

Bouttier, F., & Courtier, P. (2002). Data assimilation concepts and methods march 1999. Training.

Bradbrook, K. (2006). JFLOW: A multiscale two-dimensional dynamic flood model. Water and Environment Journal. doi: 10.1111/j.1747-6593.2005.00011.x

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. (1984). Classification and Regression Trees. Chapman and Hall/CRC.

Briand, T., & Monasse, P. (2018). Theory and practice of image B-spline interpolation.

Image Processing On Line, 8, 99–141. doi: 10.5201/ipol.2018.221

Briggs, W. M., & Levine, R. A. (1997). Wavelets and field forecast verification. Monthly Weather Review, 125(6), 1329–1373. doi: 10.1175/1520-0493(1997)125⟨1329:waffv⟩2.0.co;2

Buizza, R. (1997). Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. Monthly Weather Review, 125(1), 99–119. doi: 10.1175/1520-0493(1997)125⟨0099:PFSOEP⟩2.0.CO;2

BWDB. (2020). Bangladesh water development board annual flood report 2020. `http://ffwc.gov.bd/`, last access 26th October 2022.

Casati, B., & Wilson, L. J. (2007). A new spatial-scale decomposition of the brier score: Application to the verification of lightning probability forecasts. Monthly Weather Review, 135(9), 3052–3069. doi: 10.1175/MWR3442.1

Central Water Commission. (2023). Flood forecast dashboard. `https://cwc.gov.in/ffm_dashboard`, last access 23rd January 2023.

Chen, X., Yuan, H., & Xue, M. (2018). Spatial spread-skill relationship in terms of agreement scales for precipitation forecasts in a convection-allowing ensemble. Quarterly Journal of the Royal Meteorological Society, 144(710), 85–98. doi: 10.1002/qj.3186

Chini, M., Hostache, R., Giustarini, L., & Matgen, P. (2017). A hierarchical split-based approach for parametric thresholding of SAR images: Flood inundation as a test case. IEEE Transactions on Geoscience and Remote Sensing, 55(12), 6975–6988. doi: 10.1109/TGRS.2017.2737664

Cloke, H. L., & Pappenberger, F. (2008). Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures. Meteorological Applications, 15(1), 181-197. doi: https://doi.org/10.1002/met.58

Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. Journal of Hydrology, 375(3-4), 613–626. Retrieved from `http://dx.doi.org/10.1016/j.jhydrol.2009.06.005` doi: 10.1016/j.jhydrol.2009.06.005

Cooper, E. S., Dance, S. L., Garcia-Pintado, J., Nichols, N. K., & Smith, P. J. (2018, jun). Observation impact, domain length and parameter estimation in data assimilation for flood forecasting. Environmental Modelling & Software, 104, 199–214. doi: 10.1016/J.ENVSOFT.2018.03.013

Cooper, E. S., Dance, S. L., García-Pintado, J., Nichols, N. K., & Smith, P. J. (2019). Observation operators for assimilation of satellite observations in fluvial inundation forecasting. Hydrology and Earth System Sciences, 23(6), 2541–2559. doi: 10.5194/hess-23-2541-2019

Copernicus Programme. (2021). Copernicus Emergency Management Service. `https://emergency.copernicus.eu/`, last access 14 September 2021.

Coughlan de Perez, E., Berse, K. B., Depante, L. A. C., Easton-Calabria, E., Evidente, E. P. R., Ezike, T., ... Sant, C. V. (2022). Learning from the past in moving to the future: Invest in communication and response to weather early warnings to reduce death and damage. Climate Risk Management, 38, 100461. Retrieved from `https://doi.org/10.1016/j.crm.2022.100461` doi: 10.1016/j.crm.2022.100461

Coughlan de Perez, E., Van Den Hurk, B., Van Aalst, M. K., Amuron, I., Bamanya, D., Hauser, T., ... Zsoter, E. (2016). Action-based flood forecasting for triggering humanitarian action. Hydrology and Earth System Sciences, 20(9), 3549–3560. doi: 10.5194/hess-20-3549-2016

Coughlan de Perez, E., Van Den Hurk, B., Van Aalst, M. K., Jongman, B., Klose, T., & Suarez, P. (2015). Forecast-based financing: An approach for catalyzing humanitarian action based on extreme weather and climate forecasts. Natural Hazards and Earth System Sciences, 15(4), 895–904. doi: 10.5194/nhess-15-895-2015

Dance, S. L., Dalton, H., Carolin, C., Clark, J., Ferrando Jorge, N., Hooker, H., & Mason, D. (2022). Assimilated watercolours: Pop up art exhibitions in care homes, egu general assembly 2022, vienna, austria, 23–27 may 2022, egu22-11694. Retrieved from `https://doi.org/10.5194/egusphere-egu22-11694`

Dasgupta, A., Grimaldi, S., Ramsankaran, R. A., Pauwels, V. R., & Walker, J. P. (2018). Towards operational SAR-based flood mapping using neuro-fuzzy texture-based approaches. Remote Sensing of Environment, 215(June), 313–329. doi: 10.1016/j.rse.2018.06.019

Dasgupta, A., Grimaldi, S., Ramsankaran, R. A., Pauwels, V. R. N., Walker, J. P., Chini, M., ... Matgen, P. (2018). Flood mapping using synthetic aperture radar sensors from local to global scales. In Global flood hazard (p. 55-77). American Geophysical Union (AGU). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/9781119217886.ch4 doi: https://doi.org/10.1002/9781119217886.ch4

Dasgupta, A., Hostache, R., Ramasankaran, R., Schumann, G. J., Grimaldi, S., Pauwels, V. R. N., & Walker, J. P. (2021a). On the impacts of observation location, timing and frequency on flood extent assimilation performance. Water Resources Research. doi: 10.1029/2020wr028238

Dasgupta, A., Hostache, R., Ramsankaran, R. A., Schumann, G. J., Grimaldi, S., Pauwels, V. R., & Walker, J. P. (2021b). A Mutual Information-Based Likelihood Function for Particle Filter Flood Extent Assimilation. Water Resources Research, 57(2), 1–28. doi: 10.1029/2020WR027859

Davies, P. A., McCarthy, M., Christidis, N., Dunstone, N., Fereday, D., Kendon, M., ... Sexton, D. (2021). The wet and stormy uk winter of 2019/2020. Weather, 76(12), 396-402. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/wea.3955 doi: https://doi.org/10.1002/wea.3955

DAWN. (2022). Larkana faces threat of outbreak due to inundation. https://www.dawn.com/news/1707072, last access 15th June 2023.

de Moraes Frasson, R. P., Turmon, M. J., Durand, M. T., & David, C. H. (2023). Estimating the Relative Impact of Measurement, Parameter, and Flow Law Errors on Discharge from the Surface Water and Ocean Topography Mission. Journal of

Hydrometeorology, 24(3), 425–443. Retrieved from https://journals.ametsoc.org/view/journals/hydr/24/3/JHM-D-22-0078.1.xml doi: https://doi.org/10.1175/JHM-D-22-0078.1

Dey, S. R., Leoncini, G., Roberts, N. M., Plant, R. S., & Migliorini, S. (2014). A spatial view of ensemble spread in convection permitting ensembles. Monthly Weather Review. doi: 10.1175/MWR-D-14-00172.1

Dey, S. R., Plant, R. S., Roberts, N. M., & Migliorini, S. (2016). Assessing spatial precipitation uncertainties in a convective-scale ensemble. Quarterly Journal of the Royal Meteorological Society. doi: 10.1002/qj.2893

Dey, S. R., Roberts, N. M., Plant, R. S., & Migliorini, S. (2016). A new method for the characterization and verification of local spatial predictability for convective-scale ensembles. Quarterly Journal of the Royal Meteorological Society. doi: 10.1002/qj.2792

Dhar, O. N., & Nandargi, S. (2000). A study of floods in the Brahmaputra basin in India. International Journal of Climatology, 20(7), 771–781. doi: 10.1002/1097-0088(20000615)20:7⟨771::AID-JOC518⟩3.0.CO;2-Z

Dhar, O. N., & Nandargi, S. (2003). Hydrometeorological Aspects of Floods in India. In M. M. Q. Mirza, A. Dixit, & A. Nishat (Eds.), Flood problem and management in south asia (pp. 1–33). Dordrecht: Springer Netherlands. Retrieved from https://doi.org/10.1007/978-94-017-0137-2_1 doi: 10.1007/978-94-017-0137-2_1

Di Mauro, C., Hostache, R., Matgen, P., Pelich, R., Chini, M., Van Leeuwen, P. J., ... Blöschl, G. (2021). Assimilation of probabilistic flood maps from SAR data into a coupled hydrologic-hydraulic forecasting model: A proof of concept. Hydrology and Earth System Sciences, 25(7), 4081–4097. doi: 10.5194/hess-25-4081-2021

Di Mauro, C., Hostache, R., Matgen, P., Pelich, R., Chini, M., van Leeuwen, P. J., ... Blöschl, G. (2020). Assimilation of probabilistic flood maps from SAR data into a hydrologic-hydraulic forecasting model: a proof of concept. Hydrology and Earth

System Sciences Discussions(September), 1–24. doi: 10.5194/hess-2020-403

Di Mauro, C., Hostache, R., Matgen, P., Pelich, R., Chini, M., van Leeuwen, P. J., ... Blöschl, G. (2022, aug). A Tempered Particle Filter to Enhance the Assimilation of SAR-Derived Flood Extent Maps Into Flood Forecasting Models. Water Resources Research, 58(8). Retrieved from https://onlinelibrary.wiley.com/doi/10.1029/2022WR031940 doi: 10.1029/2022WR031940

Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., Hirpa, F. A., & Feyen, L. (2016). Development and evaluation of a framework for global flood hazard mapping. Advances in Water Resources, 94, 87–102. Retrieved from http://dx.doi.org/10.1016/j.advwatres.2016.05.002 doi: 10.1016/j.advwatres.2016.05.002

Dottori, F., & Todini, E. (2011). Developments of a flood inundation model based on the cellular automata approach: Testing different methods to improve model performance. Physics and Chemistry of the Earth, 36(7-8), 266–280. Retrieved from http://dx.doi.org/10.1016/j.pce.2011.02.004 doi: 10.1016/j.pce.2011.02.004

Dunning, P. (2019). Fly. https://www.jbarisk.com/news-blogs/fly-technology-revolutionising-the-world-of-catastrophe-modelling/, last access 20th January 2023.

ECMWF. (2022). Skill scores of forecasts of weather parameters by tigge centres. https://www.ecmwf.int/en/forecasts/charts, last access 20 April 2022.

Emerton, R. E., Stephens, E. M., Pappenberger, F., Pagano, T. C., Weerts, A. H., Wood, A. W., ... Cloke, H. L. (2016). Continental and global scale flood forecasting systems. Wiley Interdisciplinary Reviews: Water, 3(3), 391–418. doi: 10.1002/wat2.1137

Environment Agency. (2021). National LIDAR Programme. https://data.gov.uk/dataset/f0db0249-f17b-4036-9e65-309148c97ce4/national-lidar-programme, last access 29 April 2021.

ESA. (2021). ICEYE commercial satellites join the EU Copernicus programme. https://www.esa.int/Applications/Observing_the_Earth/Copernicus/ICEYE_commercial_satellites_join_the_EU_Copernicus_programme, last access 28 October 2021.

EU Science Hub. (2021). The JRC launches a revolutionary tool for monitoring ongoing floods worldwide as part of the copernicus emergency management service. https://ec.europa.eu/jrc/en/news/jrc-launches-revolutionary-tool-for-monitoring-floods-worldwide-part-copernicus-emergency-management-service, last access 28 October 2021.

Evensen, G. (1994, may). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research: Oceans, 99(C5), 10143–10162. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1029/94JC00572https://onlinelibrary.wiley.com/doi/abs/10.1029/94JC00572https://agupubs.onlinelibrary.wiley.com/doi/10.1029/94JC00572 doi: 10.1029/94JC00572

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., ... Alsdorf, D. (2007). The Shuttle Radar Topography Mission. Reviews of Geophysics, 45(2). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005RG000183 doi: https://doi.org/10.1029/2005RG000183

Fekete, A., & Sandholz, S. (2021). Here comes the flood, but not failure? Lessons to learn after the heavy rain and pluvial floods in germany 2021. Water (Switzerland), 13(21), 1–20. doi: 10.3390/w13213016

Floodlist. (2017). India – third wave of flooding hits assam, 2 million affected. http://floodlist.com/asia/india-assam-floods-august-2017, last access 10th November 2021.

Floodlist. (2022). Pakistan – almost 1,000 dead, 33 million affected in worst floods in a decade. https://floodlist.com/asia/pakistan-floods-update-august-2022,

last access 23rd May 2023.

Frasson, R. P. d. M., Schumann, G. J.-P., Kettner, A. J., Brakenridge, G. R., & Krajewski, W. F. (2019). Will the Surface Water and Ocean Topography (SWOT) Satellite Mission Observe Floods? Geophysical Research Letters, 46(17-18), 10435–10445. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL084686` doi: https://doi.org/10.1029/2019GL084686

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., ... Michaelsen, J. (2015). The climate hazards infrared precipitation with stations - A new environmental record for monitoring extremes. Scientific Data, 2, 1–21. doi: 10.1038/sdata.2015.66

Galmiche, N., Hauser, H., Spengler, T., Spensberger, C., Brun, M., & Blaser, N. (2021). Revealing Multimodality in Ensemble Weather Prediction. In D. Archambault, I. Nabney, & J. Peltonen (Eds.), Machine learning methods in visualisation for big data. The Eurographics Association. doi: 10.2312/mlvis.20211073

García-Pintado, J., Mason, D. C., Dance, S. L., Cloke, H. L., Neal, J. C., Freer, J., & Bates, P. D. (2015, apr). Satellite-supported flood forecasting in river networks: A real case study. Journal of Hydrology, 523, 706–724. doi: 10.1016/J.JHYDROL.2015.01.084

García-Pintado, J., Neal, J. C., Mason, D. C., Dance, S. L., & Bates, P. D. (2013). Scheduling satellite-based sar acquisition for sequential assimilation of water level observations into flood modelling. Journal of Hydrology, 495. doi: 10.1016/j.jhydrol.2013.03.050

GFM. (2021). GloFAS Global Flood Monitoring (GFM). `https://www.globalfloods.eu/technical-information/glofas-gfm/`, last access 28 October 2021.

Giustarini, L., Hostache, R., Kavetski, D., Chini, M., Corato, G., Schlaffer, S., & Matgen, P. (2016). Probabilistic Flood Mapping Using Synthetic Aperture Radar Data. IEEE Transactions on Geoscience and Remote Sensing, 54(12), 6958–6969. doi: 10.1109/TGRS.2016.2592951

GloFAS. (2021). Glofas methods. https://www.globalfloods.eu/general-information/glofas-methods/, last access 15th November 2021.

GloFAS. (2022a). Glofas map viewer. https://www.globalfloods.eu/, last access 8th December 2022.

GloFAS. (2022b). Glofas rapid flood mapping. https://confluence.ecmwf.int/display/CEMS/GloFAS+Rapid+Flood+Mapping+and+Rapid+Impact+Assessment, last access 4th November 2022.

Google Earth Engine CART. (2021). ee.Classifier.smileCart. https://developers.google.com/earth-engine/apidocs/ee-classifier-smilecart, last access 29 April 2021.

Google Earth Engine Catalog. (2021). Sentinel Collection. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S1_GRD, last access 4 August 2021.

Google Earth Engine Scale. (2021). Image Pyramids. https://developers.google.com/earth-engine/guides/scale, last access 16 September 2021.

Gourbesville, P. (1998, 09). Mike 11 gis: interest of gis technology for conception of flood protection systems. In Proceedings of the 3rd hydroinformatics conference - hydroinformatics'98. Copenhagen, Denmark.

Griffith, H. V., Wade, A. J., Lavers, D. A., & Watts, G. (2020). Atmospheric river orientation determines flood occurrence. Hydrological Processes, 34(23), 4547–4555. doi: 10.1002/hyp.13905

Grimaldi, S. (2022). Glofas v4.0 hydrological reanalysis. https://www.globalfloods.eu/news/113-glofas-v34-release-and-glofas-v40-hydrological-reanalysis-dataset-announcement/, last access 25th November 2022.

Grimaldi, S., Li, Y., Pauwels, V. R., & Walker, J. P. (2016). Remote Sensing-Derived Water Extent and Level to Constrain Hydraulic Flood Forecasting Models: Opportunities and Challenges. Surveys in Geophysics, 37(5), 977–1034. doi:

10.1007/s10712-016-9378-y

Hagen, A. (2003). Fuzzy set approach to assessing similarity of categorical maps. International Journal of Geographical Information Science, 17(3), 235–249. doi: 10.1080/13658810210157822

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., ... Pappenberger, F. (2020). GloFAS-ERA5 operational global river discharge reanalysis 1979–present. Earth System Science Data, 12(3), 2043–2060. Retrieved from https://essd.copernicus.org/articles/12/2043/2020/ doi: 10.5194/essd-12-2043-2020

Havnø, K., Madsen, M. N., & Dørge, J. (1995). Mike 11 - a generalized river modelling package. In V. P. Singh (Ed.), Computer models of watershed hydrology (p. 733-782). Colorado, USA: Water Resources Publications.

Hirpa, F. A., Salamon, P., Beck, H. E., Lorini, V., Alfieri, L., Zsoter, E., & Dadson, S. J. (2018). Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data. Journal of Hydrology, 566, 595–606. doi: 10.1016/j.jhydrol.2018.09.052

Hoch, J. M., & Trigg, M. A. (2019). Advancing global flood hazard simulations by improving comparability, benchmarking, and integration of global flood models. Environmental Research Letters, 14(3). doi: 10.1088/1748-9326/aaf3d3

Hooker, H., Dance, S. L., Mason, D. C., Bevington, J., & Shelton, K. (2022). Spatial scale evaluation of forecast flood inundation maps. Journal of Hydrology, 128170. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022169422007399 doi: https://doi.org/10.1016/j.jhydrol.2022.128170

Hooker, H., Dance, S. L., Mason, D. C., Bevington, J., & Shelton, K. (2023a). Assessing the spatial spread–skill of ensemble flood maps with remote-sensing observations. Natural Hazards and Earth System Sciences, 23(8), 2769–2785. Retrieved from https://nhess.copernicus.org/articles/23/2769/2023/ doi: 10.5194/nhess

-23-2769-2023

Hooker, H., Dance, S. L., Mason, D. C., Bevington, J., & Shelton, K. (2023b, sep). A multi-system comparison of forecast flooding extent using a scale-selective approach. Hydrology Research, 54(10), 1115–1133. Retrieved from `https://doi.org/10.2166/nh.2023.025` doi: 10.2166/nh.2023.025

Hopson, T. M. (2014). Assessing the ensemble spread-error relationship. Monthly Weather Review, 142(3), 1125–1142. doi: 10.1175/MWR-D-12-00111.1

Horritt, M. S., Mason, D. C., & Luckman, A. J. (2001). Flood boundary delineation from synthetic aperture radar imagery using a statistical active contour model. International Journal of Remote Sensing, 22(13), 2489–2507. doi: 10.1080/01431160116902

Hossain, S. (2020). Glofas case study: Bangladesh 2020. `https://www.globalfloods.eu/get-involved/case-study-bangladesh-2020/`, last access 18th October 2022.

Hossain, S., Cloke, H. L., Ficchì, A., Turner, A. G., & Stephens, E. M. (2021). Hydrometeorological drivers of flood characteristics in the Brahmaputra river basin in Bangladesh. Hydrology and Earth System Sciences Discussions, 2021, 1–28. Retrieved from `https://hess.copernicus.org/preprints/hess-2021-97/` doi: 10.5194/hess-2021-97

Hostache, R., Chini, M., Giustarini, L., Neal, J., Kavetski, D., Wood, M., ... Matgen, P. (2018). Near-Real-Time Assimilation of SAR-Derived Flood Maps for Improving Flood Forecasts. Water Resources Research, 54(8), 5516–5535. doi: 10.1029/2017WR022205

Hostache, R., Martinis, S., Bauer-Marschallinger, B., Chini, M., Chow, S., C. Cao, Pelich, R., ... Salamon, P. (2021). A first evaluation of the future CEMS systematic global flood monitoring product. `https://events.ecmwf.int/event/222/contributions/2274/attachments/1280/2347/Hydrological-WS-Hostache.pdf`, last access 4 August 2021.

Janjić, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., ... Weston, P. (2018). On the representation error in data assimilation. Quarterly Journal of the Royal Meteorological Society, 144(713), 1257–1278. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3130 doi: https://doi.org/10.1002/qj.3130

JBA. (2021). Storm Ciara, Dennis and Jorge. https://www.jbarisk.com/flood-services/event-response/storm-ciara-dennis-and-jorge/, last access 14 September 2021.

Kendon, M. (2020). Storm Dennis. https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/weather/learn-about/uk-past-events/interesting/2020/2020_03_storm_dennis.pdf, last access 29 April 2021.

Konapala, G., Kumar, S. V., & Khalique Ahmad, S. (2021, oct). Exploring Sentinel-1 and Sentinel-2 diversity for flood inundation mapping using deep learning. ISPRS Journal of Photogrammetry and Remote Sensing, 180, 163–173. doi: 10.1016/J.ISPRSJPRS.2021.08.016

Krullikowski, C., Chow, C., Wieland, M., Martinis, S., Bauer-Marschallinger, B., Roth, F., ... Salamon, P. (2023). Estimating Ensemble Likelihoods for the Sentinel-1-Based Global Flood Monitoring Product of the Copernicus Emergency Management Service. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16, 6917–6930. doi: 10.1109/JSTARS.2023.3292350

Lai, X., Liang, Q., Yesou, H., & Daillet, S. (2014). Variational assimilation of remotely sensed flood extents using a 2-D flood model. Hydrology and Earth System Sciences, 18(11), 4325–4339. doi: 10.5194/hess-18-4325-2014

Lauritzen, S. (2023). Fundamentals of Mathematical Statistics. doi: 10.1201/9781003272359

Lavers, D. A., Allan, R. P., Wood, E. F., Villarini, G., Brayshaw, D. J., & Wade, A. J. (2011). Winter floods in Britain are connected to atmospheric rivers. Geophysical

Research Letters, 38(23), 1–8. doi: 10.1029/2011GL049783

Lehner, B. (2014a). Hydrobasins global watershed boundaries and sub-basin delineations derived from hydrosheds data at 15 second resolution. https://www.hydrosheds.org/images/inpages/HydroBASINS_TechDoc_v1c.pdf, last access 5th November 2022.

Lehner, B. (2014b). Hydrobasins global watershed boundaries and sub-basin delineations derived from hydrosheds data at 15 second resolution. https://www.hydrosheds.org/images/inpages/HydroBASINS_TechDoc_v1c.pdf, last access 5th November 2022.

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrological Processes, 27(15), 2171-2186. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.9740 doi: https://doi.org/10.1002/hyp.9740

Leutbecher, M., & Palmer, T. N. (2008, mar). Ensemble forecasting. Journal of Computational Physics, 227(7), 3515–3539. doi: 10.1016/J.JCP.2007.02.014

LISFLOOD. (2022). Lisflood. https://ec-jrc.github.io/lisflood-model/1_1_introduction_LISFLOOD/, last access 4th November 2022.

Lorenc, A. C., Ballard, S. P., Bell, R. S., Ingleby, N. B., Andrews, P. L. F., Barker, D. M., ... Saunders, F. W. (2000). The met. office global three-dimensional variational data assimilation scheme. Quarterly Journal of the Royal Meteorological Society, 126(570), 2991-3012. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712657002 doi: https://doi.org/10.1002/qj.49712657002

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. Tellus, 21(3), 289–307. doi: 10.3402/tellusa.v21i3.10086

Martinis, S., Kersten, J., & Twele, A. (2015). A fully automated TerraSAR-X based flood service. ISPRS Journal of Photogrammetry and Remote Sensing, 104, 203–212. Retrieved from http://dx.doi.org/10.1016/j.isprsjprs.2014.07.014 doi: 10

.1016/j.isprsjprs.2014.07.014

Mason, D. C., Bevington, J., Dance, S. L., Revilla-Romero, B., Smith, R., Vetra-Carvalho, S., & Cloke, H. L. (2021b). Improving urban flood mapping by merging synthetic aperture radar-derived flood footprints with flood hazard maps. Water (Switzerland), 13(11). doi: 10.3390/w13111577

Mason, D. C., Dance, S. L., & Cloke, H. L. (2021a). Floodwater detection in urban areas using Sentinel-1 and WorldDEM data. Journal of Applied Remote Sensing, 15(03), 1–22. doi: 10.1117/1.jrs.15.032003

Mason, D. C., Dance, S. L., Vetra-Carvalho, S., & Cloke, H. L. (2018). Robust algorithm for detecting floodwater in urban areas using synthetic aperture radar images. Journal of Applied Remote Sensing, 12(04), 1. Retrieved from `http://centaur.reading.ac.uk/80110/8/045011{_}1.pdf` doi: 10.1117/1.jrs.12.045011

Mason, D. C., Davenport, I. J., Neal, J. C., Schumann, G. J., & Bates, P. D. (2012). Near real-time flood detection in urban and rural areas using high-resolution synthetic aperture radar images. IEEE Transactions on Geoscience and Remote Sensing. doi: 10.1109/TGRS.2011.2178030

Mason, D. C., Horritt, M. S., Dall'Amico, J. T., Scott, T. R., & Bates, P. D. (2007). Improving river flood extent delineation from synthetic aperture radar using airborne laser altimetry. IEEE Transactions on Geoscience and Remote Sensing, 45(12), 3932–3943. doi: 10.1109/TGRS.2007.901032

Mason, D. C., Schumann, G. J., Neal, J. C., Garcia-Pintado, J., & Bates, P. D. (2012). Automatic near real-time selection of flood water levels from high resolution Synthetic Aperture Radar images for assimilation into hydraulic models: A case study. Remote Sensing of Environment. doi: 10.1016/j.rse.2012.06.017

Matthews, G., Barnard, C., Cloke, H., Dance, S. L., Jurlina, T., Mazzetti, C., & Prudhomme, C. (2022). 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. Retrieved from https://hess.copernicus.org/articles/26/2939/2022/ doi: 10.5194/hess-26-2939-2022

Met Office. (2020). Record Breaking Rainfall. https://www.metoffice.gov.uk/about-us/press-office/news/weather-and-climate/302020/2020-winter-february-stats, last access 29 April 2021.

Met Office. (2022). Weather Observation Website (WOW). https://wow.metoffice.gov.uk/observations/details/20220503d5q856skshe63xibyyb96sm4oy, last access 4 May 2022.

Nanditha, J. S., Kushwaha, A. P., Singh, R., Malik, I., Solanki, H., Chuphal, D. S., ... Mishra, V. (2023). The Pakistan Flood of August 2022: Causes and Implications. Earth's Future, 11(3), e2022EF003230. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022EF003230 doi: https://doi.org/10.1029/2022EF003230

National River Flow Archive. (2021). NRFA. https://nrfa.ceh.ac.uk/data/station/info/55002, last access 29 April 2021.

NCAR. (2022). Weather research & forecasting model. https://www.mmm.ucar.edu/models/wrf, last access 14th November 2022.

ndimage.zoom. (2021). Scipy. https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.zoom.html, last access 21 September 2021.

Necker, T., Wolfgruber, L., Serafin, S., Dorninger, M., & Weissmann, M. (2023). How to use the fractions skill score for ensemble forecast verification. EMS Annual Meeting 2023, Bratislava, Slovakia, 4–8 Sep 2023, 20, 5194. Retrieved from https://meetingorganizer.copernicus.org/EMS2023/EMS2023-35.html

NEXTmap. (2016). Nextmap world30 dsm. https://www.intermap.com/nextmap, last access 24th January 2023.

Nguyen, T. H., Ricci, S., Piacentini, A., Fatras, C., Kettig, P., Blanchet, G., ... Baillarin, S. (2023, January). Assimilation of sar-derived flood extent observations for

improving fluvial flood forecast – a proof-of-concept. IOP Conference Series: Earth and Environmental Science, 1136(1), 012018. Retrieved from https://iopscience.iop.org/article/10.1088/1755-1315/1136/1/012018 doi: 10.1088/1755-1315/1136/1/012018

Nguyen, T. H., Ricci, S., Piacentini, A., Simon, E., & Rodriguez-suquet, R. (2023). Gaussian anamorphosis for ensemble kalman filter analysis of sar-derived wet surface ratio observations. , 1–19. Retrieved from www.opentelemac.org

OCHA. (2020). Bangladesh monsoon flooding 2020: anticipatory action pilot. https://www.unocha.org/our-work/humanitarian-financing/anticipatory-action/summary-bangladesh-pilot, last access 18th October 2022.

Otsuka, S., Awazu, T., Welzbacher, C. A., Potthast, R., & Miyoshi, T. (2023). Assimilating Precipitation Features Based on the Fractions Skill Score: An Idealized Study with an Intermediate AGCM. In S. K. Park (Ed.), Numerical weather prediction: East asian perspectives (pp. 283–294). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-031-40567-9_11 doi: 10.1007/978-3-031-40567-9_11

Palash, W., Akanda, A. S., & Islam, S. (2020). The record 2017 flood in South Asia: State of prediction and performance of a data-driven requisitely simple forecast model. Journal of Hydrology, 589(June). doi: 10.1016/j.jhydrol.2020.125190

Pappenberger, F., Beven, K. J., Hunter, N. M., Bates, P. D., Gouweleeuw, B. T., Thielen, J., & de Roo, A. P. (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(4), 381–393. doi: 10.5194/hess-9-381-2005

Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S., & Thielen, J. (2015). The monetary benefit of early flood warnings in Europe. Environmental

Science and Policy, 51, 278–291. Retrieved from http://dx.doi.org/10.1016/j.envsci.2015.04.016 doi: 10.1016/j.envsci.2015.04.016

Pappenberger, F., Frodsham, K., Beven, K., Romanowicz, R., & Matgen, P. (2007). Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations. Hydrology and Earth System Sciences, 11(2), 739–752. doi: 10.5194/hess-11-739-2007

Pekel, J. F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. Nature, 540(7633), 418–422. doi: 10.1038/nature20584

Pujol, L., Garambois, P.-A., & Monnier, J. (2022). Multi-dimensional hydrological–hydraulic model with variational data assimilation for river networks and floodplains. Geoscientific Model Development, 15(15), 6085–6113. Retrieved from https://gmd.copernicus.org/articles/15/6085/2022/ doi: 10.5194/gmd-15-6085-2022

Pörtner, H.-O., Roberts, D., Adams, H., Adelekan, I., Adler, C., Adrian, R., ... Ibrahim, Z. Z. (2022). Climate change 2022: Impacts, adaptation and vulnerability [Book]. Cambridge, UK and New York, USA: Cambridge University Press.

Renner, M., Werner, M. G., Rademacher, S., & Sprokkereef, E. (2009). Verification of ensemble flow forecasts for the River Rhine. Journal of Hydrology, 376(3-4), 463–475. Retrieved from http://dx.doi.org/10.1016/j.jhydrol.2009.07.059 doi: 10.1016/j.jhydrol.2009.07.059

Rentschler, J., Salhab, M., & Jafino, B. A. (2022). Flood exposure and poverty in 188 countries. Nature Communications, 13. doi: 10.1038/s41467-022-30727-4

Revilla-Romero, B., Shelton, K., Wood, E., Berry, R., Bevington, J., Hankin, B., ... Huyck, C. (2017, April). Flood Foresight: A near-real time flood monitoring and forecasting tool for rapid and predictive flood impact assessment. In Egu general assembly conference abstracts (p. 1230).

Revilla-Romero, B., Wanders, N., Burek, P., Salamon, P., & de Roo, A. (2016). Integrating

remotely sensed surface water extent into continental scale hydrology. Journal of Hydrology, 543, 659–670. Retrieved from http://dx.doi.org/10.1016/j.jhydrol.2016.10.041 doi: 10.1016/j.jhydrol.2016.10.041

riverlevels.uk. (2020). River levels, River Wye at Hereford Bridge. https://riverlevels.uk/herefordshire-hereford-old-wye-bridge-lvl#.YIFKt31KgUE, last access 29 April 2021.

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Monthly Weather Review. doi: 10.1175/2007MWR2123.1

Savage, J. T. S., Bates, P., Freer, J., Neal, J., & Aronica, G. (2016). When does spatial resolution become spurious in probabilistic flood inundation predictions? Hydrological Processes, 30(13), 2014–2032. doi: 10.1002/hyp.10749

Schumann, G. (2019). The need for scientific rigour and accountability in flood mapping to better support disaster response. Hydrological Processes, 33(24), 3138–3142. doi: 10.1002/hyp.13547

Schumann, G., Bates, P. D., Horritt, M. S., Matgen, P., & Pappenberger, F. (2009). Progress in integration of remote sensing-derived flood extent and stage data and hydraulic models. Reviews of Geophysics. doi: 10.1029/2008RG000274

Schumann, G., Giustarini, L., Tarpanelli, A., & Jarihani, B. (2022). Flood Modeling and Prediction Using Earth Observation Data. Surveys in Geophysics(0123456789). Retrieved from https://doi.org/10.1007/s10712-022-09751-y doi: 10.1007/s10712-022-09751-y

Sefton, C., Muchan, K., Parry, S., Matthews, B., Barker, L. J., Turner, S., & Hannaford, J. (2021). The 2019/2020 floods in the uk: a hydrological appraisal. Weather, 76(12), 378-384. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/wea.3993 doi: https://doi.org/10.1002/wea.3993

Skok, G., & Roberts, N. (2016). Analysis of Fractions Skill Score properties for ran-

dom precipitation fields and ECMWF forecasts. Quarterly Journal of the Royal Meteorological Society, 142(700), 2599–2610. doi: 10.1002/qj.2849

Skok, G., & Roberts, N. (2018). Estimating the displacement in precipitation forecasts using the Fractions Skill Score. Quarterly Journal of the Royal Meteorological Society, 144(711), 414–425. doi: 10.1002/qj.3212

Sky News. (2022). Pakistan floods: Satellite images and maps show scale of disaster. `https://news.sky.com/story/pakistan-floods-satellite-images-and-maps-show-scale-of-disaster-12685468`, last access 23 August 2023.

Speight, L. J., Cranston, M. D., White, C. J., & Kelly, L. (2021). Operational and emerging capabilities for surface water flood forecasting. Wiley Interdisciplinary Reviews: Water, 8(3), 1–24. doi: 10.1002/wat2.1517

Start Network. (2022). Anticipation and risk financing. `https://startnetwork.org/anticipation-and-risk-financing`, last access 25th October 2022.

Stein, J., & Stoop, F. (2019). Neighborhood-based contingency tables including errors compensation. Monthly Weather Review, 147(1), 329–344. doi: 10.1175/MWR-D-17-0288.1

Stephens, E., & Cloke, H. (2014a). Improving flood forecasts for better flood preparedness in the UK (and beyond). Geographical Journal, 180(4), 310–316. doi: 10.1111/geoj.12103

Stephens, E., Schumann, G., & Bates, P. (2014). Problems with binary pattern measures for flood model evaluation. Hydrological Processes. Retrieved from `https://doi.org/10.1002/hyp.9979` doi: 10.1002/hyp.9979

Tavus, B., Kocaman, S., Nefeslioglu, H. A., & Gokceoglu, C. (2020). A fusion approach for flood mapping using sentinel-1 and sentinel-2 datasets. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, 43(B3), 641–648. doi: 10.5194/isprs-archives-XLIII-B3-2020-641-2020

Tellman, B., Sullivan, J. A., Kuhn, C., Kettner, A. J., Doyle, C. S., Brakenridge, G. R., ...

Slayback, D. A. (2021). Satellite imaging reveals increased proportion of population exposed to floods. Nature, 596. Retrieved from http://dx.doi.org/10.1038/s41586-021-03695-w doi: 10.1038/s41586-021-03695-w

The Guardian. (2022). Pakistan floods cause devastation – in pictures. https://www.theguardian.com/world/gallery/2022/aug/30/pakistan-floods-cause-devastation-in-pictures, last access 15th June 2023.

Trepekli, K., Friborg, T., Balstrøm, T., Fog, B., Allotey, A., Kofie, R., & Møller-Jensen, L. (2021). UAV-LiDAR observations increase the precision of urban flood modelling in Accra by detecting critical micro-topographic features. EGU General Assembly online 19–30 Apr 2021. doi: 10.5194/egusphere-egu21-10457

Twele, A., Cao, W., Plank, S., & Martinis, S. (2016). Sentinel-1-based flood mapping: a fully automated processing chain. International Journal of Remote Sensing, 37(13), 2990–3004. Retrieved from http://dx.doi.org/10.1080/01431161.2016.1192304 doi: 10.1080/01431161.2016.1192304

UK Water Resources Portal. (2022). NRFA UKCEH UK Water Resources Portal. https://eip.ceh.ac.uk/hydrology/water-resources/, last access 4 May 2022.

Ulbrich, U., Leckebusch, G. C., & Pinto, J. G. (2009). Extra-tropical cyclones in the present and future climate: A review. Theoretical and Applied Climatology, 96(1-2), 117–131. doi: 10.1007/s00704-008-0083-8

UNDRR. (2022a). Early warnings for all the un global early warning initiative for the implementation of climate adaptation. https://library.wmo.int/idurl/4/58209, last access November 2023.

UNDRR. (2022b). Global assessment report on disaster risk reduction. https://www.undrr.org/gar2022-our-world-risk#container-downloads, last access 25th October 2022.

UNDRR. (2023). Gar special report 2023 mapping resilience for the sustainable development goals. https://www.undrr.org/gar/gar2023-special-report, last

access November 2023.

van Leeuwen, P. J. (2009). Particle Filtering in Geophysical Systems. Monthly Weather Review, 137(12), 4089–4114. Retrieved from https://journals.ametsoc.org/view/journals/mwre/137/12/2009mwr2835.1.xml doi: https://doi.org/10.1175/2009MWR2835.1

Wagner, W., Freeman, V., Cao, S., Matgen, P., Chini, M., Salamon, P., ... Briese, C. (2020). DATA PROCESSING ARCHITECTURES FOR MONITORING FLOODS USING SENTINEL-1. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-3-2020, 641–648. Retrieved from https://isprs-annals.copernicus.org/articles/V-3-2020/641/2020/ doi: 10.5194/isprs-annals-V-3-2020-641-2020

WASDI. (2022). Wasdi. https://www.wasdi.net/#!/marketplace, last access 4th November 2022.

WMO. (2017). South asian climate outlook forum held in bhutan. https://public.wmo.int/en/media/news/normal-rainfall-likely-much-of-south-asia-2017-southwest-monsoon, last access 23rd January 2023.

World Resources Institute. (2023). How floods in pakistan threaten global security. https://www.wri.org/insights/pakistan-floods-threaten-global-security#:~:text=In%20Pakistan%2C%20torrential%20rains%20and,and%20killed%20over%20900%2C000%20livestock., last access 23 August 2023.

World Weather Attribution. (2022). Climate change likely increased extreme monsoon rainfall, flooding highly vulnerable communities in pakistan. https://www.worldweatherattribution.org/climate-change-likely-increased-extreme-monsoon-rainfall-flooding-highly-vulnerable-communities-in-pakistan/, last access 23 August 2023.

Wu, W., Emerton, R., Duan, Q., Wood, A. W., Wetterhall, F., & Robertson, D. E. (2020). Ensemble flood forecasting: Current status and future opportunities. WIREs Water,

7(3), 1–32. doi: 10.1002/wat2.1432

Zappa, M., Jaun, S., Germann, U., Walser, A., & Fundel, F. (2011, may). Superposition of three sources of uncertainties in operational flood forecasting chains. Atmospheric Research, 100(2-3), 246–262. doi: 10.1016/J.ATMOSRES.2010.12.005

Zsoter, E., Prudhomme, C., Stephens, E., Pappenberger, F., & Cloke, H. (2020). Using ensemble reforecasts to generate flood thresholds for improved global flood forecasting. Journal of Flood Risk Management, 13(4), 1–14. doi: 10.1111/jfr3.12658