Rainfall and River Flow Ensemble Verification: Phase 2

FINAL REPORT

February 2021

FOREWORD

This document is the Final Report for the "Rainfall and River Flow Ensemble Verification: Phase 2" (EnsVerP2) project forming part of the project "Improving confidence in Flood Guidance through verification of rainfall and river flow ensembles".

The EnsVerP2 project was carried out with contributions from the following teams:

UK Centre for Ecology & Hydrology: Seonaid R. Anderson, Robert J. Moore, Steven J. Cole

Met Office: Gabriella Csima, Sebastian Cole, Ric Crocker, Marion Mittermaier

Executive Summary

Forecasting the weather and floods is a challenging task and inherently uncertain. Acknowledging and accounting for the uncertainty in precipitation and flood forecasts has become increasingly important. This has partly been driven by the move of warning and guidance services to risk-based approaches that combine the likelihood of flooding with its potential impact on society and the environment. In the UK, such risk-based approaches underpin the National Severe Weather Warning Service delivered by the Met Office, and the Flood Guidance Statements produced by the Flood Forecasting Centre (FFC) and Scottish Flood Forecasting Service (SFFS).

A standard approach to accounting for forecast uncertainty is to use ensemble methods. For a number of years, FFC and SFFS have used precipitation ensembles coupled with the national Grid-to-Grid (G2G) model of river flow to underpin the Flood Guidance Statement. However, the performance of the overall end-to-end ensemble precipitation and river flow forecasting system is currently not verified routinely. This ensemble verification information and evidence is essential: its absence can limit end-user confidence and inhibit full exploitation for flood-risk guidance. In addition, the local flood forecasting systems - used by the Environment Agency (EA), Scottish Environment Protection Agency (SEPA) and Natural Resources Wales (NRW) - are planned to transition to ensemble forecasting and will have similar requirements for verification information.

A first step in addressing this operational gap has been to bring together existing expertise in meteorological and hydrological model performance assessment to design and develop a new, holistic **Ensemble Verification Framework**. Then to consider how this Framework could be used to develop an operational end-end interactive **Ensemble Forecast Visualisation and Verification System.** The Framework has been designed so that the operational system developed from it would help forecasters answer the following two key questions.

- How well has the ensemble precipitation and flood forecasting system performed in the (recent) past? Particularly for flood events of interest.
- What does this mean for interpreting today's forecast?

Forecasters could then make more informed decisions and increase their confidence in the use of ensembles for forecasting the severity and likelihood of precipitation and flooding.

To develop and test the potential verification approaches and operational displays, 16-months of precipitation and river flow ensemble forecasts have been processed and verified. Specific case-studies, identified with the help of stakeholders, have been used to prototype, demonstrate, assess and refine the verification tools. The Best Medium Range (BMR) precipitation ensemble is used as input to the national-scale G2G model of river flow across Great Britain and to a small selection of catchment-scale PDM local models of river flow. This approach has allowed rigorous scientific exploration of how to provide robust verification statistics of the ensemble precipitation inputs to the river flow modelling and of its ensemble river flow outputs.

The scientific analysis allowed identification of several points relevant to the underpinning verification methodology part of the Framework.

- Three different precipitation accumulation time-intervals were evaluated: 15 min (the temporal resolution of the river flow model and its precipitation inputs), hourly and daily. Daily precipitation accumulations appear to provide the best guidance in terms of rain volume for hydrological impacts. One reason for this may well be because it removes the impact of timing errors at the sub-daily scale. Sub-daily precipitation can be more closely related to river flow on an ensemble member-by-member basis.

- The source of observed precipitation (raingauge, radar or merged raingauge-radar) has an impact on the verification analyses and G2G river flow performance.
- The change in precipitation-intensity characteristics with lead-time between the STEPS, MOGREPS-UK and MOGREPS-G components of the BMR precipitation ensemble, are evident in both rainfall and river flow analysis.
- The length of period used for ensemble verification is an important factor: generally longer than two years is recommended if possible. The 16-month test period was sufficient for generating enough precipitation threshold-exceedances for the 95th percentile thresholds: but insufficient for higher thresholds and for considering river flow thresholds above one half the median annual maximum flood at sub-regional scales.
- New methods of presenting the precipitation forecast probabilities have been developed for precipitation thresholds that are hydrologically relevant. The verification of these Time-Window Probabilities (TWPs) has shown that the probabilities are larger, and also more reliable: so users can have greater confidence in using them.

For new real-time displays to be of value in operational settings, it is important that users (e.g. FFC hydrometeorologists or flood forecasting officers) find the displays understandable and easy to deploy in support of flood guidance and warning. Operational users have been engaged in co-design of the real-time forecast displays through the Project Board and a Workshop. These interactions have identified that the real-time displays need to be flexible and informative, with varying layers of detail. Viewing the precipitation and river flow together, however, is the most important ingredient along with using common methods for conveying information on both. Prototype joint rainfall and river flow displays have been created. Further co-design of interactive displays is recommended during future implementation and interactions should include operational users, researchers and system developers.

Case-studies have been used to highlight the potential benefits of these new real-time displays. They have demonstrated how the ensemble verification information can help users make more informed decisions when ensemble verification information is included. For example, knowing whether a forecast is over- or under-confident for different lead-times and severity-thresholds can be very helpful, particularly in marginal cases. That is if a forecast has a tendency to predict too high or low a probability of precipitation, or of river flow, exceeding a given level of severity.

Summary and key recommendation

Realising the benefit and value of probabilistic flood-risk information for decision-making was a key motivator for the “Rainfall and River Flow Ensemble Verification: Phase 2” project. The project succeeded in bringing together the meteorology and hydrology to define, test and demonstrate a joint Ensemble Verification Framework for ensemble precipitation and river flow forecasts. The outcomes of the project demonstrate how the subsequent verification information can be used to enhance the user’s perception and ability to deploy ensemble forecasts and derived probabilities in day-to-day flood risk decision-making.

Overall, the key finding is that joint precipitation and river flow ensemble verification is possible and useful. The primary recommendation is that an end-to-end interactive **Ensemble Forecast Visualisation and Verification System** for FFC (and SFFS) be implemented as soon as is practicable. The **Ensemble Verification Framework** provides the blueprint for the system and the **Joint Coding Framework** developed and applied here provides the basis for the algorithm and code. A detailed set of recommendations have been provided, including what is required for operational implementation. This also includes a priority list of recommendations for developing a minimum system.

The proposed system would address the current urgent operational gap in ensemble forecast verification capability for FFC and SFFS. It would mark a significant addition to the forecasters' toolkit by providing real-time displays that incorporate ensemble verification information for the first time, and in a usable form. In turn, this will facilitate enhanced and more informed decision-making at times of potential flood-risk. Local model systems have ensemble and probabilistic flood forecasting as an aspiration in their future plans. These systems would eventually benefit from the operationally urgent developments recommended here for the national-scale G2G model used by FFC and SFFS. Local model users could play an early and active part in system co-design as part of a staged implementation process for local model systems.

Acknowledgements

This research project was commissioned by the Flood Forecasting Centre (FFC) on behalf of the Environment Agency (EA), Scottish Environment Protection Agency (SEPA) and Natural Resources Wales (NRW). The research was formulated under the Flood and Coastal Erosion Risk Management (FCERM) R&D Programme entitled “Improving confidence in Flood Guidance through verification of rainfall and river flow ensembles” (SC150016). The Phase 1 study (Dey et al., 2019; Anderson et al., 2019) was co-funded by the FFC and SEPA, and carried out under the “Rainfall and River Flow Ensemble Verification” project (Phase 1) between September 2016 and July 2017. The Phase 2 study reported on here progressed during 2019 and 2020 with the addition of FCERM R&D Programme funds.

Thanks are due to the members of the Project Board for overseeing the Phase 2 project and contributing in various ways. For FFC: Dave Cox and Keith Fenwick. For EA: Tim Harrison, Susan Manson, Ian Perkins, Jenny Sansom. For SEPA: Jamie Rae. For NRW: Sam Mitchell. Participants from these agencies attending the User Feedback Workshop on Real-time Displays (incorporating ensemble verification information) and the Discovery Workshop “Ensemble Verification Beyond Phase 2” are thanked for their valuable interaction and feedback.

Adrian Wynn, Rob Cowling and Graeme Boyce of the FFC provided a “Statement of User Requirements” under the Phase 1 project. This served as a guide to the requirements of the FFC hydrometeorologists with regard to verification of rainfall and river flow ensembles.

Tim Harrison, Jamie Rae and Sam Mitchell are thanked for arranging provision of local PDM configuration information from their respective agencies. Paul Wass (JBA) is thanked for providing information on PDM workflows used in the EA′s new Incident Management Forecasting System (IMFS).

At the Met Office, Gavin Evans is thanked for information and advice on ensemble data access and Stephen Moseley for providing additional information about the Best Medium Range ensemble processing and how this has changed over time.

Table of contents

1 Introduction

Coupled with rainfall forecast ensembles, the national Grid-to-Grid (G2G) model forms a key element of the Flood Forecasting Centre's (FFC) capabilities and the operational flood guidance services it provides. However, the performance of the overall end-to-end ensemble flood forecasting system is currently not verified routinely. This can limit end-user confidence and inhibit full exploitation of the outputs. In turn, this imposes a constraint on the quality and effectiveness of the FFC's 5-day Flood Guidance Statement (FGS) relied upon by its users. These include Category 1 & 2 responders, along with local and central government, who use the FGS as a guide in taking appropriate action when preparing for and responding to flooding. It is therefore essential that the performance of the operational ensemble flood forecasting system is better understood, over a range of hydrometeorological situations and forecast lead-times. This understanding will provide end-users with the appropriate evidence-base required to improve operational and strategic decision-making at times of flood.

To address these needs, a research project was formulated under the Flood and Coastal Erosion Risk Management (FCERM) R&D Programme entitled "Improving confidence in Flood Guidance through verification of rainfall and river flow ensembles" (SC150016). Initial work, co-funded by FFC and SEPA, was carried out under the "Rainfall and River Flow Ensemble Verification" project (Phase 1) between September 2016 and July 2017. The Phase 1 Report "Rainfall and River Flow Ensemble Verification: Prototype Framework and Demonstration" (Dey et al., 2019) provided a foundation for the Phase 2 project. Phase 2 allowed the Phase 1 Report recommendations to be progressed during 2019 and 2020, with the addition of FCERM R&D Programme funds. A synthesis of aspects of the Phase 1 study was reported in Anderson et al. (2019). The present document and its Appendix is the Final Report of the Phase 2 study.

Phase 2 recognised new requirements for the Ensemble Verification Framework to support ensemble verification for local models - such as the PDM catchment model - employed in model networks configured to river basins. These local models are operated within the IMFS (Incident Management Forecasting System) by the EA, in FEWS Wales by NRW and in FEWS Scotland by SEPA. Greater use of probabilistic forecasting for these systems is a strategic aim and needs to be underpinned by an Ensemble Verification Framework.

The R&D reported here has brought together existing expertise in meteorological and hydrological model performance assessment to provide a new, holistic, end-to-end Ensemble Verification Framework. Application of the Framework, to both national and local flood forecasting models, is demonstrated here through analysis of case-studies and longer periods of ensemble rainfall forecasts: leading to recommendations on future operational implementation.

Phase 2 has introduced a much longer period for ensemble verification of ~16 months duration. The Phase 1 study was restricted to use of a very short and highly unusual period, with a focus on the month of December 2015 that encapsulated many high impact events associated with storms Desmond, Eve and Frank. Use of a longer period for verification analysis has provided a greater perspective on ensemble forecast performance under all conditions. It has also allowed further case-studies to be examined in greater detail.

The report is organised as follows. Section 2, summarising the scientific results, first outlines the verification methodology. The Ensemble Verification Framework is described along with work associated with the extended study period: including understanding concurrent weather model versions, accessing observation data, and selecting case studies. It also introduces the refined (for Phase 2) derivation of Time-Window Probabilities (TWPs) from the precipitation forecast ensemble. Also described, and new for Phase 2, is the derivation of catchment-based climatological precipitation distributions. This is followed by a summary of the scientific findings concerning verification, sensitivity to period length, impact of precipitation observation source, and joint verification considerations. Then developments in prototype real-time displays, introduced in Phase 1 and refined through a user-community workshop, are reported on. Findings from the case studies are reviewed. This is followed by a detailed look at the Joint Coding Framework supporting the Ensemble Verification Framework and its future operational use.

Section 3 provides the recommendations from the project in terms of implementation of the Joint Verification Framework and visualisation prototypes, focusing on the benefits for the hydrometeorological user.

A set of conclusions are provided in Section 4. The report Appendix has an extensive set of appendices which provide more detail on all the aspects summarised herein.

2 Summary of Scientific Results

To keep this Final Report concise, the summary of scientific results provided in this section is short and succinct. Further details are contained in the Science Reports of the separate Appendix.

2.1 Verification Methodology

Following a review of the Phase 1 analyses, several refinements to the Ensemble Verification Framework have been developed under Phase 2. These are summarised below and discussed in more detail in the sections indicated.

- Computation of something other than just the precipitation catchment-mean to represent precipitation over the catchment in the verification (Section 2.1.6, Appendix A.3).
- Creating a Joint Coding Framework (Section 2.5.1, Appendix).
- Adding the derivation of Time-Window Probabilities (TWPs) to the precipitation processing to mirror what is done for river flow verification, providing a more user-oriented view of overall precipitation ensemble usefulness and forecast performance. (Section 2.1.4, Appendix A.4).
- Deriving and examining a set of climatological thresholds to verify precipitation to be able to assess flooding potential more appropriately at the catchment level (Section 2.1.5, Appendices A.3 and A.5).
- Using and evaluating the raingauge-radar merged precipitation product with England & Wales coverage for both precipitation verification and as input for hydrological modelling (Sections 2.1.3, Appendices B.3 and B.5).

To ensure that this Final Report can be understood in isolation, an overview is first provided of the Joint Verification Framework developed under Phase 1 and extended under Phase 2 (Section 2.1.1). This is provided along with the data available for ensemble forecast verification (Section 2.1.2), the models of rainfall and river flow employed, and the data sources used (Section 2.1.3).

2.1.1 Joint Verification Framework

The Joint Verification Framework for ensemble verification of precipitation and river flow is outlined in Appendix A.1 as a set of selected metrics (scores and diagrams). It aims to give an overview of performance in general, first individually at each site and then over all sites, possibly split by catchment features (e.g. catchment size). This is where the standard ensemble verification scores are used, both in Numerical Weather Prediction (NWP) and Hydrological Forecasting. A summary of the key features of the verification metrics considered in the Joint Verification Framework, and detailed in Appendix A.1, is given in Table 2.1 for ease of reference.

Table 2.1 Overview of verification metrics used in the Joint Verification Framework.

	Verification metric	What the metric measures	Units	Performance indicator	
				Good	Poor
Verification score	**Continuous Ranked Probability Score (CRPS)**	Difference between the cumulative distribution estimated by the ensemble forecast, and the step-function cumulative density function of the observation	Units of the observation and ensemble forecasts	0	Large values
	Brier Score (BS)	Mean square probability error	Dimensionless	0	Large values
	Mean error (ME)	Measure of overall bias	Units of quantity being assessed	0	Large values
Verification Skill Score	**Continuous Ranked Probability Skill Score (CRPSS)**	CRPS compared to the ME of the observations over the verification period	Dimensionless	**1** indicate a perfect forecast	**0:** same value as climatological information only **<0:** less value than climatological information only
	Brier Skill Score (BSS)	BS compared to a reference given by the sample climatology			
	Relative Operating Characteristic Diagram and Area Under Curve Skill Score (ROCSS)	Area Under the ROC Curve (AUC) normalised with reference to a random forecast with no skill (an AUC equal to 0.5)			
Verification diagram	**Relative Economic Value (REV)**	Economic forecast value relative to a forecast based on climatological information			
	Rank Histogram	Reliability of the ensemble: that is, whether or not the ensemble and observations have been drawn from the same distribution.	N/A	Flat diagram	**U-shaped:** spread is too small **∩-shaped:** spread is too large **Asymmetric:** biased
	Reliability Diagram or Attributes Diagram	Reliability and Resolution of the probability forecasts		**Good Reliability and Resolution:** close to diagonal	**No Resolution:** horizontal line **Under forecasting:** above diagonal **Over forecasting:** below diagonal
	Relative Operating Characteristic (ROC) diagram	Potential skill of the ensemble: that is, the ensemble skill if ensemble probabilities were well-calibrated		Close to upper left corner	On diagonal

2.1.2 Verification periods and case studies

To produce verification analyses that are as relevant as possible to the current operational system for precipitation ensemble forecast production, a 16-month period was chosen (1 June 2017 to 30 September 2018) that followed the Unified Model (UM) PS39 upgrade (11 July 2017). This included a major resolution increase of MOGREPS-G to ~20 km, along with other major upgrades to data assimilation, and an improved UKV (United Kingdom Variable resolution) analysis upon which the MOGREPS-UK ensemble members are centred. To provide data for two summer seasons, June 2017 was also included even though it was before the PS39 implementation. This enables a comparison between summer 2017 and summer 2018.

The FFC, SEPA and NRW suggested an initial set of 23 case-studies which cover a variety of locations, synoptic conditions, Flood Risk Matrix severities, and range of impacts and model performance. In all, 14 case-studies have been considered in this Final Report. These are discussed in more detail in Section 2.4. A full list of the suggested case studies is given in Appendix A.2.

2.1.3 Overview of models and data sources used

Best Medium Range (BMR) precipitation ensemble forecasts of 15-minute precipitation accumulations are available with UK-wide coverage, extending out to over six days and issued four times a day with 24 ensemble members. The BMR precipitation ensemble forecast combines data from the STEPS extrapolation nowcasting system, the 2.2km grid-spacing convection-permitting MOGREPS-UK ensemble, and the ~20km grid spacing MOGREPS-G ensemble. All data are downscaled onto a fixed 2km grid over the UK, the British National Grid, as used by STEPS. To allow the latest forecast to be available to the FFC, the BMR forecasts are triggered based on the time when the required input data from the NWP model are available, as opposed to being clock-triggered at a fixed time. This results in forecast start-times which can vary by up to three hours. Forecasts are triggered around 4 hours after the driving MOGREPS-UK runs at 21:00, 03:00, 09:00, and 15:00. All forecasts terminate at 153 hours (6 days and 9 hours) after the driving MOGREPS-UK run (6.5 days after the driving MOGREPS-G run), giving forecast lengths which vary from 147 to 149 hours. For this study all forecast initiation times are considered, with forecasts combined based on the lead-time of the BMR forecasts. Further discussion of the effect of different triggering times is given in Section 2.1.7 and Appendix A.6.

River flow ensemble forecasts are analysed, both for the **distributed G2G model** with national coverage and catchment-specific **PDM local models,** with the BMR precipitation forecast used as input. The currently-operational configuration of the models in the FFC, SEPA, EA and NRW are used. When G2G is run over Scotland, this includes the use of the G2G Snow Hydrology module where snow is accounted for using air temperature as an additional input. Temperature data were recreated from the constituent UKV fields as the BMR temperature fields are not operationally stored on the Met Office MASS archive. Combining UKV temperature fields, there are four BMR temperature forecasts available per day as follows.

- 00:00 forecast using the 03:00 UKV data for T+3 to T+120
- 06:00 forecast using the 09:00 UKV data for T+3 to T+57 and 03:00 UKV data for T+58 to T+114
- 12:00 forecast using the 15:00 UKV data for T+3 to T+120
- 18:00 forecast using the 21:00 UKV data for T+3 to T+57 and 15:00 UKV data for T+58 to T+114.

The shorter length of the BMR temperature forecasts results in river flow BMR forecasts for Scotland which only extend out to a lead-time of around 4.5 days. As is done operationally, one single temperature forecast is used for all G2G ensemble member forecasts. For the other models considered (G2G for England & Wales and all PDM local models) the effects of snow are not considered. For all the hydrological models considered, Raingauge data are used up to the forecast start-time to produce the model's initial conditions.

River flow forecasts are verified against observations of **instantaneous river flow** available at **15-minute intervals.** Precipitation forecasts are verified against derived precipitation accumulations from **Raingauge, Radar, and Merged Raingauge and Radar** observation sources.

2.1.4 Derivation of precipitation Time-Window Probabilities (TWPs)

Phase 1 introduced the concept of the catchment areal mean precipitation as the unit for precipitation verification. This adaptation was applied alongside a conventional atmospheric model approach to precipitation forecast verification, exposing some weaknesses from the perspective of a hydrological user. In Phase 2, precipitation TWP-based verification analyses have been compared with the Phase 1 approach. The latter focused more on the quality of the precipitation used as input to G2G, by precisely matching the 15 min precipitation accumulations in the forecasts and observations for Days 1 to 6, rather than how a hydrometeorologist may use the precipitation ensemble forecasts to help in their decision-making process.

TWPs are intended to mirror what a hydrometeorologist does: scanning for the likelihood of a precipitation threshold exceedance at any time within a 24-, 48- or 72-hour period (time-window). By objectively deriving these probabilities, the forecast ensemble can also be objectively evaluated to assess the quality of this guidance. In essence, the use of TWPs removes, or at the very least, mitigates against the impact of timing errors in the precipitation forecast, increasing the detection of threshold-exceedance events.

Furthermore, TWPs are derived by examining the precipitation across *all* the grid-cells in a catchment, not just the catchment-mean or –median precipitation. In the process, a spatial coverage check ensures that any identified event is genuine. The probability for the time-window is then derived by counting the number of ensemble members that have *any* exceedances (that also meet the coverage criterion) at any time during the time-window. The derivation of TWPs is covered in detail in Appendix A.4.

2.1.5 Derivation of climatological precipitation thresholds over the UK

In Phase 1, fixed thresholds and percentile thresholds were used, tracking with the “rain of the day” or “rain of the hour”, and thereby trying to pick out the most interesting rain on any given day. The latter acted to remove the forecast bias but was highly unsatisfactory from a user perspective, given that the thresholds that were used remained often of little interest. Equally, the use of fixed thresholds applied nationwide are too rigid given the strong precipitation gradients seen across the UK, primarily from a wet west to a much drier east, but also from a wetter north to a drier south. Deriving a set of climatologically appropriate thresholds for each catchment means that each catchment can be evaluated against its own definition of “wetness”.

Considering the climatological distribution of precipitation for each catchment provides guidance on the variability of precipitation across the UK, and enables a threshold to be calculated that selects the “extreme” precipitation values for that catchment, noting that what is extreme for a low-rainfall catchment may be normal for a high-rainfall catchment. To create climatological catchment-precipitation thresholds, ten historical years of raingauge-rainfall data from 2007 to 2016 were used.

The hourly and daily precipitation values corresponding to the 90, 95 and 99th percentiles of their respective distributions were extracted and saved for future use. To increase the sample-size, and to investigate the overall precipitation distribution for each catchment, data were pooled over a number of days: firstly over all days-in-the-year to give an annual overview, and secondly over the 91 days centred upon each day in the year to give a rolling seasonal overview (varying by date-in-year). Both these methods have advantages: it can be argued that flood-events are dependent on a specific amount of precipitation, not on the time-of-year when this occurs; however, it can also be argued that differences in the precipitation characteristics at different times of year could be important, suggesting a seasonally-varying approach. To keep both options open for future investigation, both methods are used for the catchment-precipitation processing.

Appendix A.3 describes the data and process used to derive these precipitation thresholds and Appendix A.5 provides maps of annual and seasonal thresholds for the 90, 95 and 99th percentiles, for daily and hourly precipitation accumulations. These demonstrate the extremity of 4 mm h^{-1}, for example, at the hourly time-scale and why hourly precipitation is very difficult to evaluate from a flood potential perspective.

2.1.6 Precipitation catchment processing

Precipitation data were processed for all G2G catchments, and also selected PDM catchments provided by the EA, SEPA and NRW. To allow verification of 15-minute, hourly and daily precipitation accumulations, hourly and daily quantities were calculated by accumulating the 15-minute values. The precipitation accumulations were calculated as follows.

Observed data

Hourly accumulations ending on each whole-hour
Daily accumulations of all data in the previous 24h period ending on each whole hour

Forecast data

Hourly accumulations ending on whole-hour forecast lead-times (e.g. 1h, 2h, 3h …)
Daily accumulations ending on forecast lead-times 24h, 48h, 72h, 96h, 120h, 144h

For the precipitation catchment processing, all 15-minute and hourly data are converted to units of mm h^{-1}, and daily data are converted to mm d^{-1}. To give flexibility in the precipitation verification calculations, the following quantities were calculated and saved for future use.

- **Mean** (grid-cells falling fully in catchment and weighted mean)
- **Percentiles** of within-catchment distribution: 50, 75, 90, 95, 99
- **Fixed threshold** exceedance within catchment: 0.1, 1 and 4 mm h^{-1} (sub-daily accumulations), 0.1, 1, 4 and 8 mm d^{-1} (daily accumulations)
- **Climatological threshold** exceedance within catchment: 90, 95 and 99th percentile (Section 2.1.5)

Full details of the precipitation catchment-processing are given in Appendix A.3.

2.1.7 Aspects specific to forecast triggering and combining results

Appendix A.6 provides a brief overview of the triggering times for the Phase 2 study period. It was found that the trigger times were a non-issue, and indeed *should not* be or ever become an issue for forecasts that are of operationally critical importance, i.e. supporting vital services and decision-making. Any shifts or changes in the triggering should be reported and dealt with at the operational level and should not require action in this context.

2.2 Scientific Findings

This section presents an overview of the key findings of the overall precipitation and river flow verification analysis, presented in detail in Appendices B.1 to B.5. This includes consideration of the impact of different precipitation observation sources and observation uncertainty on ensemble verification. The key question of verification period length is also addressed, with discussion of sampling uncertainty and its relationship with the spatial areas and thresholds considered. Underpinning this is the question of what events are realistically verifiable, both from a river flow and precipitation perspective, given the availability of data and resources. For river flow, catchment-pooling techniques are employed to increase the sample size while maintaining the use of return-period based thresholds. For rainfall, the use of climatological thresholds (Section 2.1.5) provides an indication as to the severity of rainfall events that may be expected for a given catchment over a given time-period, and helps suggest appropriate threshold levels to use in ensemble precipitation verification.

2.2.1 Verification elements specifically relevant to precipitation

The findings from analysis of the Phase 1 dataset can be summarised as follows.

Observation uncertainty can have a significant impact on precipitation verification analyses (see Appendix B.3). This is important because forecast ensembles are used to provide spread (uncertainty) information about future predictions. Observations are imperfect and only provide an approximation of the true state, with an error which is generally partially or fully unknown. It had been anticipated that the uncertainty analysis using the Phase 1 dataset (Appendix B.3) would be repeated with the longer Phase 2 dataset, but this was not possible within the project timeframe. The initial estimates obtained from the Phase 1 analyses were not considered suitably robust to be applied under Phase 2. It would have been preferable to have observation uncertainty accounted for in the verification analyses presented here. However, the objective of this project was not to provide definitive verification analyses, but to demonstrate a framework and the principles and components required for real-time operational forecast monitoring. Should this framework be implemented, the methodology demonstrated using the Phase 1 dataset would need to be replicated, and observation uncertainty accounted for in the verification analyses, both for precipitation and river flow.

Verification of 15-min, hourly and daily precipitation accumulations. Under Phase 2, 15-minute precipitation accumulations were verified for the very first time, with the Phase 1 dataset evaluated to understand differences between these verification results and those from the hourly analyses. At longer lead-times, MOGREPS-G does not provide 15-min accumulations and the hourly totals are divided into four equal parts to form the 15-minute precipitation accumulations input to G2G. A comparison was made of the 15-min, hourly, and daily analyses to quantify the impact of accumulation length on skill and the impact of lead-time. These results are presented in Appendix B.4. It is worthwhile checking the precipitation input to G2G from a modelling perspective. 15-min and hourly precipitation accumulations tended to be small, with even the highest 15-min rates rarely sustained for long periods of time. River flooding is most often (the exception being very rapid response flash-flooding) the result of many successive 15-min or hourly accumulations which lead to a flood response. In this project, hourly precipitation accumulations were generally found to be sufficient when viewed alongside daily accumulations: the latter providing an understanding of overall rain volumes, the former providing some context as to how the precipitation accumulates and defining the details of the intensity-duration relationship.

Phase 2 introduced a new way of producing probabilistic forecasts from the precipitation ensemble using Time-Window Probabilities (TWPs). Figure 2.1 provides a brief snapshot of skill (as measured by the Brier Skill Score) achieved for TWPs of daily precipitation accumulation threshold-exceedance. This is for the Days 2-3 and Days 4-6 ensemble precipitation forecasts computed for the 95th annual precipitation (catchment-mean)

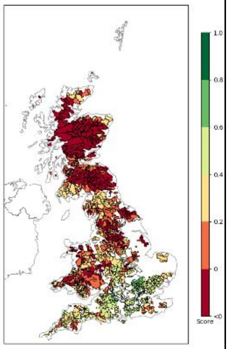
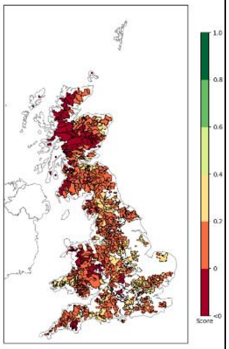
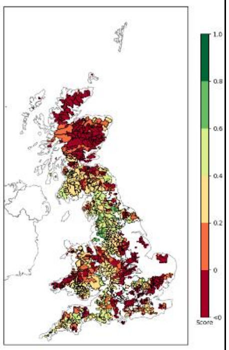
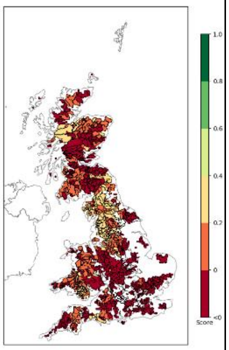
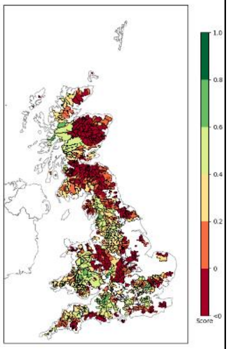
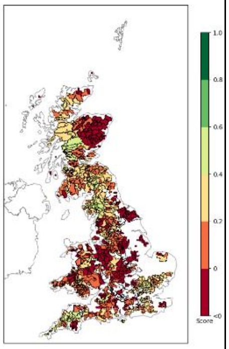
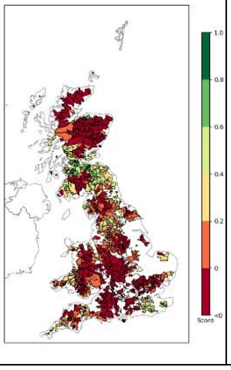
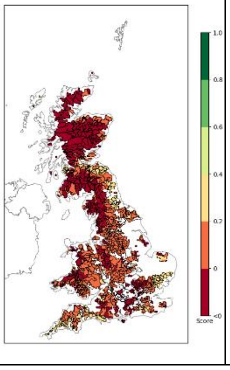
Monthly	Annual TWP 95th Days 2-3	Annual TWP 95th Days 4-6	Seasonal
Jul 2017			S1
Oct 2017			A1
Jan 2018			W1
Apr 2018			Sp2

Figure 2.1 Monthly Brier Skill Score (BSS) for daily precipitation accumulations for Days 2-3 forecasts (columns 1 and 2) and Seasonal BSS for Days 4-6 forecasts (columns 3 and 4). The observation source is gridded raingauge-rainfall. Scores are based on TWPs computed using the annual catchment-based precipitation climatologies. Seasonal 3-month periods are denoted: S1 = JJA 2017, A1 = SON 2017, W1 = DJF 2017/18 and Sp1 = MAM 2018

distribution percentile threshold. Appendix A.5 shows that for daily precipitation the 95th percentile is typically 20 mm or more for most catchments. Though month-by-month skill can vary considerably across the UK, many catchments show some modest skill even for the Days 2-3 precipitation ensemble forecasts and for such an extreme threshold. Scores and reliability are fairly robust even for month-long verification periods. More encouraging is the level of skill for the Days 4-6 forecast lead-time horizon, based on the seasonal (three-month) verification periods, which benefit from an increased sample size and greater stability. Catchments shaded dark red denote those where the ensemble precipitation forecast performs worse than the sample climatology. This is the case over large parts of Scotland where the signal persists for most months and lead-time horizons.

The precipitation assessment of the 16-month Phase 2 dataset can be summarised as follows (the full results are provided in Appendices B.1.1 and B.1.2).

User-relevant Time-Window Probabilities (TWPs) and their verification

- TWPs (considered in Appendix A.4) extract useful information content by removing or mitigating against timing errors in precipitation forecasts.
- TWPs allow the use and evaluation of higher thresholds, which better reflect the user needs for identifying and assessing potential flood risk from heavy precipitation.
- TWPs shifted the probability distribution to larger values which tend to me more reliable (with caveats).
- TWPs improve the sample size which helps to improve the stability and robustness of the verification scores for precipitation.
- TWPs computed using the seasonal and annual percentile precipitation thresholds show much better reliability and for higher thresholds. This is good news from the user perspective.
- There is little to choose between seasonal and annual precipitation thresholds. If a simple solution is sought then annual thresholds would do, though seasonal thresholds preserve more of the intra-annual variability in precipitation which exists across the UK.

Observation sources and biases

- Characteristics of the precipitation observations can have a strong influence on verification analyses, which can lead to the drawing of opposing conclusions about the performance of the same weather model, either over the same area (England & Wales) or over England & Wales and Scotland. The latter would seem to be unrealistic, especially if this is against the same precipitation observation source!
- Though an assessment of the bias would suggest that the precipitation forecasts are the least biased when assessed against the gridded raingauge-rainfall source (based on catchment means), this is considered somewhat misleading and gratuitous. At longer ranges the precipitation forecasts are strong under-estimates and the raingauge source is impervious to localised maxima, better reflected in the radar and merged rainfall products.
- The merged rainfall product would appear to be a good compromise for providing the texture that an interpolated raingauge analysis does not have whilst improving an inherent radar rainfall bias. This should be available right across the UK.
- Good precipitation observation QC is essential. The verification analyses over the Scottish Borders point at some kind of observation issue in the radar rainfall which is translated to the merged rainfall product.
- Physical biases in weather model ensembles can feed into the probability biases. The BMR ensemble is not seamless in time. Where different weather model configurations are joined together is evident in the verification analyses. From a flood forecasting perspective, the volume of water is of interest, highlighting the need and benefit of adjusting the forecast rainfall values.

2.2.2 Sensitivity to assessment period length

The assessment period length to be considered is directly related to the sample size used in the verification calculations. Thus, longer periods are needed when considering smaller spatial scales (e.g. the catchment-scale compared to regional- or national-scales) or higher thresholds. There is also a time-dependence associated with the occurrence of threshold-crossing or exceedance events. For example for river flow, more events were seen in the winter months - which are climatologically associated with higher flow values (and generally larger precipitation accumulations) - than in the summer months. This is demonstrated in Figure 2.2, which shows the number of times the observed river flow crossed the Q(2)/2 threshold for times corresponding to Day 1 BMR precipitation forecasts. (Here, Q(T) denotes the river flow of return period T years, with Q(2) the median annual maximum flood.) Whereas the 12-month periods show a reasonable distribution of threshold-crossings across the country, predominantly from the winter and spring seasons, the summer seasons show very few threshold-crossings.

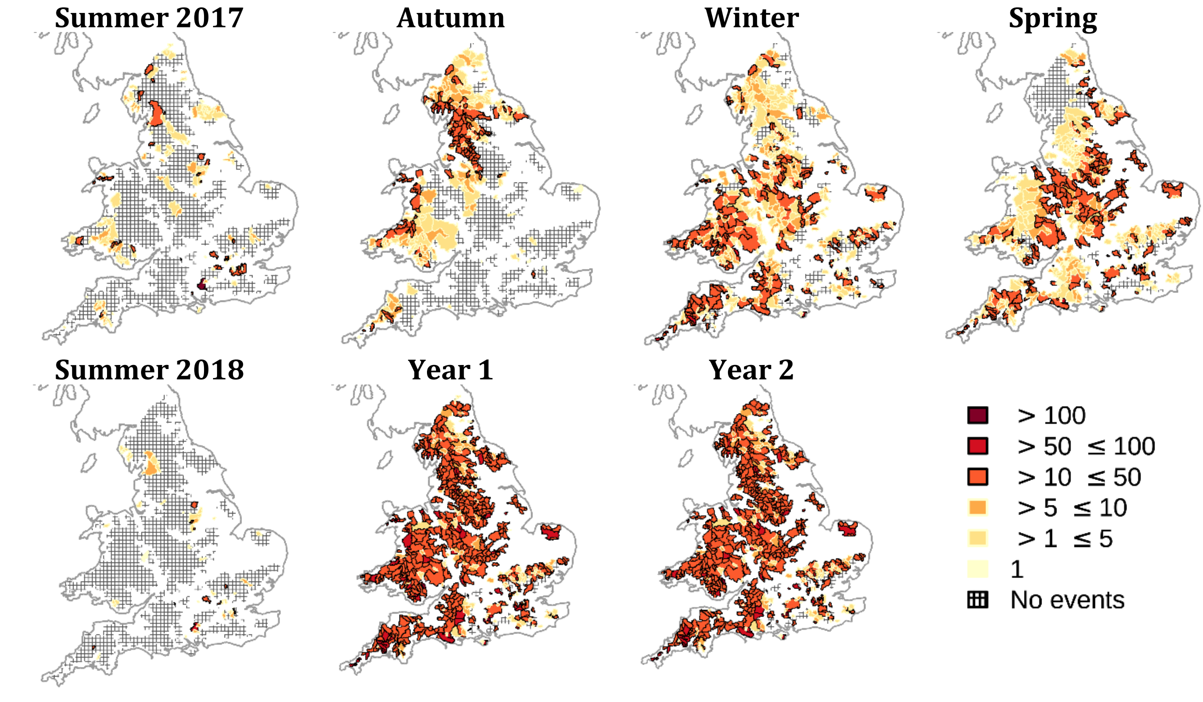


Figure 2.2 Number of river flow forecasts having observed threshold-crossings over different verification periods. For river flow threshold Q(2)/2 and time-periods corresponding to Day 1 forecasts.

Considering this, and the other river flow verification analyses (discussed fully in Appendix B.2), the following conclusions can be drawn.

1. For autumn, winter, and spring seasons, the number of river flow threshold-crossings can be sufficient to give meaningful verification analyses at the national-scale for the Q(2)/2 threshold. This depends on there being a large-enough pool of sites nationally and is, for example, not true for the spring season over Scotland with only 225 sites (compared to the 731 sites for England & Wales).
2. For the lower river flow thresholds (Q(2)/2 and Q(2)) a 12-month verification period can be sufficient to give meaningful verification analyses. If a rolling 12-month verification period were to be used, the analyses would be expected to be more sensitive to changes in the winter months as these contain the majority of threshold-crossing events.

3. For sub-regional scale analyses, sampling size and forecast skill are influenced by the time-window increasing in length with increasing lead-time.

It is informative to compare the 12-month Phase 2 period analyses with those obtained in Phase 1 for the abnormally wet December 2015 period. Figure 2.3 shows example Reliability Diagrams for Phase 2 and Phase 1 verification analyses (provided in full in the Phase 1 Report Figures 4.1 and 4.2). These suggest that the sampling issue for high thresholds is worse for the more-normal 2017 to 2018 12-month period of Phase 2 than was the case for the extremely wet December 2015 period of Phase 1. This is an important consideration for an operational verification system. Although a long and recent verification period is desired to capture up-to-date weather model behaviour, to capture extreme events it may be necessary to include a verification period longer in the past.

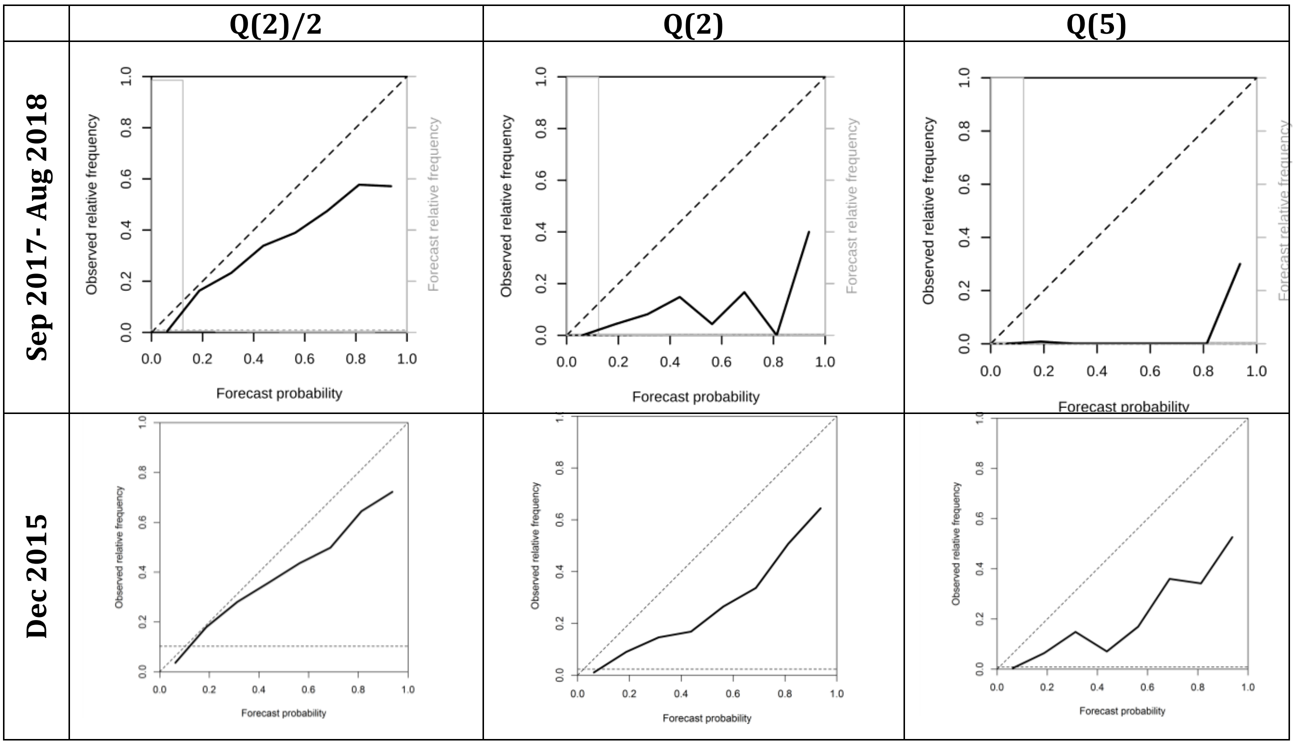


Figure 2.3 River flow Reliability Diagrams calculated using data pooled from all catchments for Scotland. For Day 1 forecasts and Q(2)/2, Q(2) and Q(5) thresholds over the Phase 1 period December 2015 (bottom) and Phase 2 12-month period September 2017 to August 2018 (top).

From a consideration of these analyses, and the other river flow verification diagrams of Phase 2 (Appendix B.2), the following conclusions can be drawn.

1. Apart from anomalously wet periods (e.g. December 2015 as used in Phase 1), the number of river flow threshold-crossings for Q(T) thresholds appropriate for flood forecasting (e.g. a minimum of Q(2)/2) is ***not sufficient*** for ensemble verification at ***sub-national scales*** when forecasts from only one season are considered.

Verification scores can be calculated for local models in the same manner as that shown for G2G. Local model verification analyses were completed for single catchments, so show high sampling uncertainties using 12-months of ensemble forecasts, even for the lowest threshold considered, Q(2)/2. This leads to the following conclusion.

2. In an operational system, the ***local model sample size*** would need to be ***increased*** through either ***multi-catchment pooling*** of analyses, consideration of a ***longer verification period***, or through using a ***fixed historical period known*** to have ***sufficient threshold-crossings***.

3. ***Changes to the model*** used (either river flow or precipitation) can have a significant impact on the verification analyses. In an operational system, the extent of these underlying model changes should inform the suitability of using existing verification analyses from a previous model until a sufficient dataset from current model runs are available for analysis. For far-reaching model changes or upgrades, it may be necessary to only use verification analyses from the current model.

The local model verification analyses can be used to inform prototype real-time displays (discussed in Sections 2.4 and 2.3.2). However, given the *high sampling uncertainties*, these should be considered as *demonstrative* rather than generally representative.

For **precipitation**, an evaluation of the length of the verification period found that forecast skill does vary from month-to-month and season-to-season, but regional differences are likely to be larger and more persistent. Annual analyses appear to be fairly robust, although those using the 99th percentile threshold is still somewhat sparse. Seasonal analyses are also stable.

If recent performance is of particular interest, a rolling 3-month window may well be very useful alongside something that tracks performance over 12 months or more. This ensures that the weather dependencies are better accounted for. Understanding forecast performance in the context of prevailing (dominant) weather patterns can be invaluable. Shorter verification periods will also respond to weather model changes more quickly and expose any specific or sudden changes in weather model behaviour. Such effects are masked in longer verification periods. Although individual monthly analyses can be useful, in terms of inferring continual performance some form of rolling performance information would be recommended.

2.2.3 Impact of precipitation observation source on river flow forecast performance

The effect on G2G modelled river flow of using different observation sources of rainfall was initially assessed for the period 1 March 2016 to 31 March 2017. With more-recent data becoming available later in Phase 2, this analysis was extended to include the following periods.

1 April 2016 to 31 March 2017 (using the most-recent observed river flow and raingauge data)
1 September 2017 to 31 August 2018 (Phase 2 12-month verification period “Year 2”)
1 October 2016 to 30 September 2018 (two full water years)
1 April 2016 to 30 September 2018 (full period of data available for comparison)

Four different sources of 15 minute rainfall accumulations on a 1km grid covering England & Wales are compared: (i) **raingauge** data from the EA and NRW raingauge networks over England and Wales, gridded by multiquadric interpolation with zero offset, as is currently used for maintaining G2G states, (ii) **radar** rainfall data from the Met Office RadarNet system, and (iii) & (iv) radar rainfall data from the Met Office RadarNet data **merged** with raingauge data, from a network of Met Office and EA/NRW gauges, using a Kriging with External Drift (KED) method (Jewell and Gaussiat, 2015). The merged rainfall data are available with a **1-hour delay** in real-time and **24-hour delay** in real-time, the latter allowing more raingauge data to be included and with better quality-control procedures applied.

Agreeing with the analyses of BMR ensemble verification in Section 2.2.2, considering a longer 2-year verification period gives clearer, less-noisy river flow verification analyses for higher thresholds, with more consistency both across regions and with the Q(2)/2 threshold analyses. This is due to a reduction in the sampling uncertainty. A reduction in the number of catchments available for analysis in the more-recent periods (from 898 to 730) was also found to make it harder to discern spatial patterns at the catchment-scale, especially if the selected catchments are unevenly distributed.

Agreeing with the initial analyses, those using the Bias, Correlation and R^2 Efficiency statistics, and the CSI (Critical Success Index) and False Alarm Ratio (FAR) skill scores, showed better performance when G2G employed Gauge rainfall as input. In contrast, the Probability Of Detection (POD) score was often better for the sources using radar rainfall (Radar and Merged) as input. Overall, the merged products perform better for more-recent periods (e.g. 1 September 2017 to 31 August 2018). In particular, the Bias in G2G simulated river flow is improved for the more recent verification period, and lies closer to that seen when Raingauge data are used as input, in contrast with earlier periods where the Bias is more-similar to that when Radar data are used as input. An example is given in Figure 2.4.

Further details of the comparison of G2G performance using different rainfall sources as input are given in Appendix B.5.

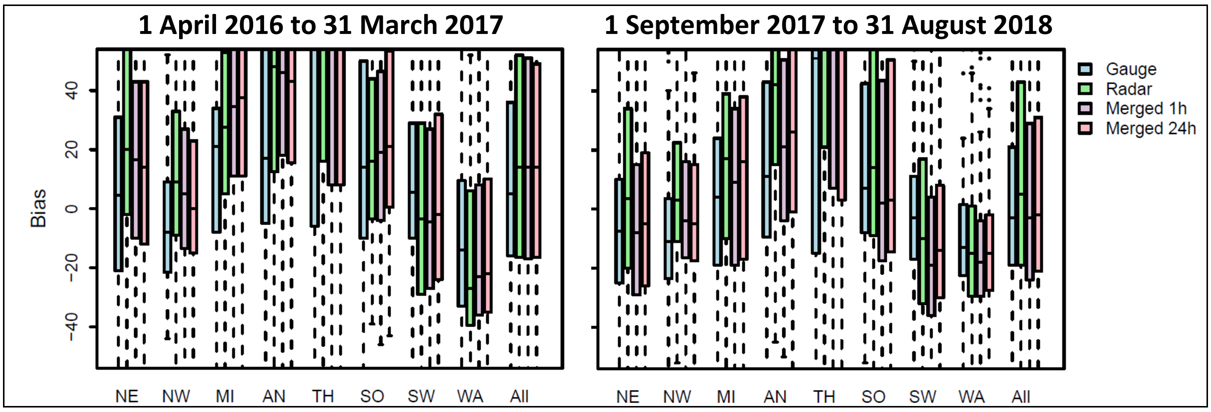


Figure 2.4 Box Plots comparing the performance of G2G river flow simulations using different observed precipitation sources as input as indicated by the bias. Results are shown for the periods 1 April 2016 to 31 March 2017 (left) and 1 September 2017 to 31 August 2018 (right). Bars are for each grouping of catchments considered: each region in England, for Wales, and for all catchments in England & Wales. Each set of bars contains (from left to right) results for G2G simulations using Gauge, Radar, Merged 1h and Merged 24h observed precipitation data as input. Each bar shows the median (solid line) and interquartile range (coloured box) of the distribution of statistics over the set of catchments. Dashed lines extend to 1.5 times the interquartile range from the box and indicate the typical range of the data. Outlying points are shown by black dots.

2.2.4 Overall joint verification assessment

There is great value in being able to view the precipitation and river flow ensemble forecasts *and* verification analyses side-by-side at the catchment-scale. This is also well illustrated via the case-studies, as will be shown in Section 2.3. The project has explored the similarities and differences that exist between the precipitation and river flow verification to find common ground in providing useful, relevant information to the user. The project has also investigated how best to implement this verification framework with recommended settings in a real-time operational context. These comparative considerations on precipitation and river flow verification are summarised in Table 2.2.

Table 2.2 Overall comparison of precipitation and river flow verification

	River flow	**Precipitation**
Score calculation		
Probability calculation	Threshold-exceedance within time-window	Conventional threshold-exceedance probabilities as well as threshold exceedance within time-window referred as time-window probabilities (TWPs)
Thresholds	***Return period*** severity thresholds appropriate to each G2G catchment were used for both G2G and PDM models.	***Fixed*** and ***climatological*** (10-year, Raingauge data based) thresholds were used and ***applied at the grid-scale*** to individual grid points within a ***given catchment boundary,*** and also to the ***catchment mean*** values.
Accumulation periods	Rainfall input 15-minute accumulations. Instantaneous river flow output every ***15 minutes***.	Analysis of hourly accumulations gives similar results to those for 15-minute accumulations, but with less processing time. Thresholds are exceedingly low for 15 min and hourly totals. ***Daily*** and ***hourly*** precipitation accumulations considered for seasonal and monthly verification periods, though ***only 24h accumulations provide sufficient evidence for event detection of interest*** to hydrological applications, even at the sub-daily scale.
Catchment processing	Outputs analysed at the ***outlet of Gauged catchments*** (within G2G model domain) and for ***site-specific PDM local models.***	***Catchment*** *mean* precipitation used for CRPSS, Rank Histogram and Mean Error calculations
Observation sources	Analysis of ensemble performance against ***observed river flow*** with initial conditions produced using ***Raingauge*** data. Analysis of ***simulation-mode*** performance using ***input Raingauge, Radar and Merged rainfall*** data sources	Analysis of ensemble performance against ***Raingauge, Radar*** **and** ***Merged rainfall*** data sources
Verification analyses		
Sampling	Sampling for **G2G** was Improved by ***regional-based pooling by catchment-size*** but still highly dependent on there being sufficient threshold-crossings in the selected verification period. For the **PDM** local models with no pooling, ***sampling uncertainty was too high to draw robust representative inferences***.	Sampling uncertainty was ***reduced through use of TWPs***, and allow application of higher thresholds

Time-of-year dependence	***Seasonal*** variations can be seen in threshold-crossings with more events occurring in the autumn and winter. ***Non-threshold scores*** show ***seasonal dependence*** (e.g. Rank Histograms).	**Monthly** variations are seen in skill. Results are surprisingly stable provided the sample size is adequate (i.e. the month has some rain). Extended dry periods are problematic. **Seasonal** analyses are stable and give useful information on weather dependencies.
Verification period length	Threshold-based ***seasonal*** analyses are meaningful at the national-scale only. ***Annual*** analyses can be meaningful at sub-national scales for Q(2)/2 and Q(2) thresholds.	A ***rolling 3-month window*** may well be very useful alongside something that tracks performance for ***12 months or more***. Shorter period verification provides insights into specific weather dependencies which are hidden in longer-term statistics.
Phase 1 period comparison (December 2015)	Overall, analyses from the Phase 2 verification periods were ***consistent*** with those from December 2015. Where seen, ***differences*** were often ***associated*** with regions with ***high sampling uncertainty.***	A ***direct comparison*** with the Phase 1 analyses is ***not possible*** due to the different treatment of thresholds and use of TWPs in Phase 2. Results are in broad agreement, given the limited and exceptional Phase 1 study period.
Bias	The bias of G2G and PDM ensemble forecasts was not assessed directly. From the case-study analysis some events showed the river flow peaks increasing with decreasing lead-time, whilst others showed the opposite.	Physical biases can feed into probability bias. Overall ***the catchment-mean precipitation*** is ***under-estimated*** with the under-estimation increasing with lead-time. This is due to use of different models to construct the ensemble forecast to span 6 days.
BMR rainfall ensemble configuration	For some case-study events it was clear from the river flow ensemble where the BMR rainfall ensemble models were joined together.	Where the rainfall ensemble models are joined together is evident in the precipitation verification analyses. Days 2-3 results tend to perform differently as this time horizon contains the model join.
Reliability	Overall, the G2G river flow ensemble is ***over-confident***, with this over-confidence increasing with increasing threshold.	Much better for higher thresholds and for higher probabilities when using TWPs. TWPs tend to shift the magnitude of the probabilities upwards, thus changing Reliability; TWP can switch a traditionally ***under-confident*** probability to one which is ***less under-confident, reliable or even over-confident.***
Calibration	***Reliability calibration*** would be useful. The methods of implementing this would need further research. ***Calibration of forecast probabilities*** is generally simpler than calibrating the precipitation input.	

Potential Skill	ROC curves suggested good potential skill if probabilities were well calibrated.	The potential skill was not analysed directly.
CRPSS	Little site-to-site variation was seen in the CRPSS for either river flow or precipitation (particularly hourly). For precipitation this is largely due to the atmospheric continuum. The choice of reference for the skill-score calculation could also be contributing	
Precipitation Observation source	Overall, ***G2G*** performance in simulation-mode was ***best when Raingauge data*** were used as input, and poorest when Radar data were used. ***Recent periods*** showed an ***improvement*** in the performance of the ***merged*** product over older periods, with the more-recent merged product data showing performance closer to that obtained using Raingauge-input than using Radar-input.	Observation characteristics can have a strong influence on verification analyses and lead to opposite conclusions about ensemble rainfall model performance. The ***merged*** product could be a good compromise because it incorporates the textural information of the spatial precipitation distribution available from radar with the relative accuracy of point-based gauge measurements. ***Quality control*** is essential in all cases (example of poor QC in radar rainfall over Scottish Borders).
Display of verification analyses		
Verification maps	Threshold-based score maps should be viewed alongside maps of the number of threshold-crossings	Threshold-based score maps for climatological thresholds should be viewed alongside maps of those thresholds to provide context.
Ensemble performance for threshold-crossings	Threshold-crossing performance ***can only be provided*** when there are ***sufficient threshold-crossings*** within the verification period	***Daily*** precipitation accumulation analyses generally provide more indication as to the occurrence of extreme precipitation, as hourly thresholds are not extreme enough.
Time-series	Hydrographs of the ***15-minute G2G or PDM forecasts*** allow the ensemble spread to be visualised, and individual ensemble members to be linked to the precipitation ensemble members (aided by use of a common colour-scheme).	A ***running-24h*** accumulation allows periods of high-accumulation precipitation to be identified. ***1h accumulation*** precipitation time-series allow individual ensemble members to be linked with the river flow ensemble members (aided by use of a common colour-scheme).

2.3 Visualisation and real-time displays

In this section the work on visualisation of ensemble verification information is discussed with a focus on real-time displays. This builds on the work done in Phase 1, and interaction with the user community through an interactive workshop (Section 2.3.1). Details are provided of the reasoning behind the display constituents and of their interpretation. Examples of real-time displays are presented later - in the context of the case-studies - in Section 2.4. A visualisation example is included also in the recommendations of Section 3.5.

2.3.1 Workshop with user community on real-time displays

For new real-time displays to be of value in operational settings, it is important that users (e.g. FFC hydrometeorologists or flood forecasting officers) find the displays understandable and easy to deploy in support of flood guidance and warning. To help achieve this, it is important that users have input into the prototype development process. Although the Project Board have representatives from the operating agencies and have provided continual input and feedback, it was recognised that it would be valuable to engage with a wider set of users that were not directly involved in the project.

On 29 April 2020, the Project Team held a feedback workshop as part of a Flood Forecasting Centre Operations Technical Team Meeting. This involved most of the operational FFC hydrometeorologists including the Chief Hydrometeorologist, Charlie Pilling. The workshop introduced the aims of the project to develop future real-time displays that pull-through information on past performance of the precipitation and river flow ensemble forecasts. Interaction revolved around the following three questions.

- What details need to be considered for the system to be useful?
- How can we focus on the flood-producing events of interest?
- So what does this mean for today’s forecast?

The workshop prompted some very useful discussion and the Project Team received positive feedback on the session from attendees such as “like the direction and distilling useful information” and “nice interactive session”. It should be noted that the FFC hydrometeorologists are familiar with visualising and using ensemble forecasts from rainfall (NWP), river flow (G2G) and coastal models. Some of the key points raised in the discussion were:

- seeing rainfall and river flow together is really useful
- seeing individual ensemble members is important (not just quantiles)
- careful choice of colours is needed for understanding and to be colour-blind friendly
- learn from coastal forecasting displays
- suggestions for display software functionality, clear keys/legends needed.

Several comments were general ones about viewing ensemble forecasts and some were specifically related to the verification information. The feedback received was incorporated in the updated prototype displays contained in this Phase 2 Report. It should be noted that this constructive interaction was a first step in designing operational displays and associated system functionality. Further co-design of displays and functionality between researchers, operational users, and system developers will be required during implementation.

2.3.2 Guide to river flow verification real-time displays

To be useful for real-time flood guidance and warning, there is a desire to view verification information that has already been interpreted and placed in context. For the verification of ensemble river flows, the performance of a specific forecast is assessed in three ways:

(i) analysing the ensemble hydrograph behaviour and threshold-exceedance at a given site,
(ii) placing the ensemble spread in the context of climatological spread for that site, and
(iii) analysing the threshold-exceedance from a regional perspective.

An example display for analysing the ensemble hydrograph behaviour and threshold-exceedance at a particular site is shown in Figure 2.5.

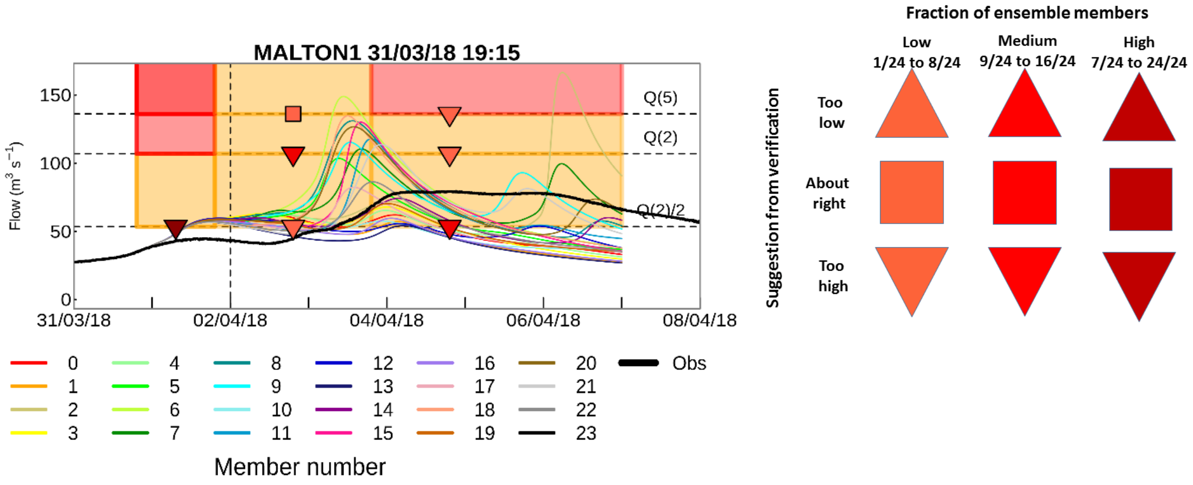


Figure 2.5 Example hydrograph display used to place the ensemble river-flow threshold-exceedance in the context of the ensemble verification information. The catchment shown is Derwent at Malton (Malton1, NE England) for a forecast time-origin of 19:15 31 March 2018.

For one forecast - selected, for example, to cover a specific time - the ensemble member hydrographs are plotted with one colour per ensemble member (with colours selected to match those used for the 24-member Storm Surge ensemble). If available (i.e. for post-event analysis), the observed flows are plotted in black to allow the ensemble performance to be visually assessed. Each flow threshold - of the set Q(2)/2, Q(2), Q(5) and Q(50) - appears as a horizontal black dashed line only when it is exceeded by at least one ensemble member or by the observations of river flow. If forecasts have been selected to analyse performance at a specific time of interest, this time is shown by a vertical black dashed line.

Ensemble probabilities of upward threshold-crossings are calculated for Day 1, Days 2-3 and Days 4-6 of the forecast. These are plotted at the relevant flow threshold, and the centre point of the lead-time range considered, with a coloured symbol indicating the probability of crossing each threshold. Light red indicates 0 to ⅓ of ensemble members crossed, medium red ⅓ to ⅔, and dark red ⅔ to 1. The symbol shape is used to indicate the direction of any correction suggested by the Reliability Diagram. An upper pointing triangle suggests a correction towards higher probabilities; a lower pointing triangle towards lower probabilities; and a square suggesting no correction. The suggested correction is calculated using a straight line of best-fit through the Reliability Diagram traces.

The background of the hydrograph is coloured according to the Overall Skill of the ensemble taken as the average of the BSS, CRPSS and ROCSS values calculated from the full Phase 2 Period (1 September 2017 to 31 August 2018). Here, the aim is to give a quick impression of the ensemble performance at the site of interest, and how this varies with threshold and lead-time**.** The Overall Skill has a transparent colour-scale defined as: dark red (very poor, worse than climatology) for values less than zero; red (poor) for values from 0 to 0.4; orange (satisfactory) for values from 0.4 to 0.6; green (good) for values from 0.6 to 1.0.

An example display for placing the river-flow ensemble dispersion in the context of climatological ensemble dispersion for a given site is shown in Figure 2.6. Here, the Coefficient of Variation (CV) is used as a dimensionless measure of ensemble dispersion. It is defined as the ratio of the ensemble Standard Deviation (spread), σ, to the ensemble mean, $\bar{y}$. The CV is calculated separately for each time-step in each forecast. For the individual forecast considered, the CV is plotted in red as a function of forecast lead-time in Figure 2.6. To calculate the climatological CV, the average is taken (separately at each forecast lead-time) of the CV values for all forecasts at the site of interest over the Phase 2 Period (1 September 2017 to 31 August 2018). This is plotted in black. Thus, when the red line in Figure 2.6 is above the black line, the individual ensemble forecast of river flow is more spread than the reference climatology.

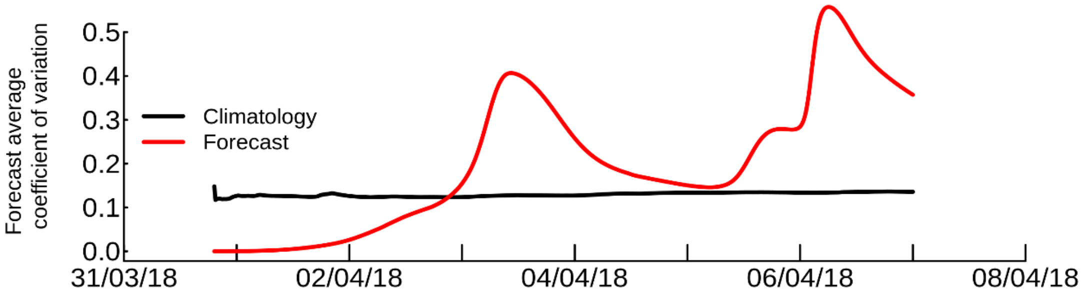


Figure 2.6 Example display of Coefficient of Variation of the ensemble forecast of river flow against forecast lead-time (given as the forecast time) for one ensemble forecast, placing the forecast ensemble spread in the context of climatological spread. The forecast time-origin is 19:15 31 March 2018.

To analyse the threshold-exceedance of the river-flow ensemble from a regional perspective, maps are drawn showing the threshold-exceedance for each site within a given region, for each threshold and lead-time range considered. An example for the North East of England is shown in Figure 2.7. The symbol at each site indicates the direction of any correction suggested by the Reliability Diagram, and the colour shows the fraction of ensemble members exceeding the threshold. The same symbols and colours are used as were used for the hydrographs in Figure 2.5.

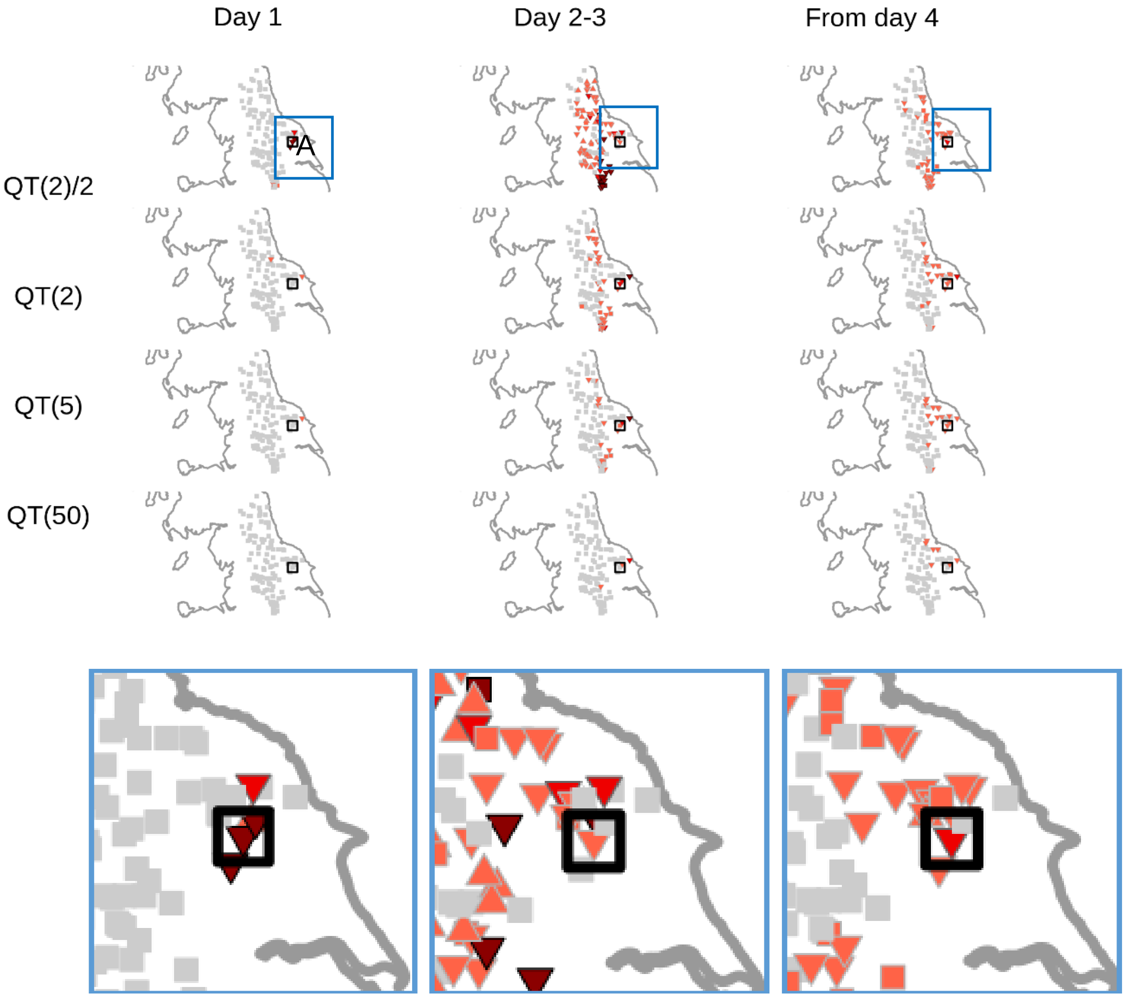


Figure 2.7 Example maps showing the variation in ensemble probability of river flow threshold-exceedance for a particular forecast over a given region (top panels) and zooming in on the catchment of interest (bottom panels). The time-origin of the forecast is 19:15 31 March 2018.

2.3.3 Guide to precipitation verification real-time displays

As discussed earlier in this report, hourly precipitation is often of lesser interest than daily precipitation from a flooding potential perspective but provides very useful context as to how any precipitation accumulates within a longer time-window. Hence, for the real-time displays it is useful to show something of how the precipitation accumulates over time. In Figure 2.8 a two-panel display of a single precipitation forecast initialisation for the Riccal at Nunnington shows both the hourly ***catchment mean*** totals (in the bottom panel) and a ***rolling 24-hour sum*** of hourly catchment means in the top panel. Ensemble members are shown in the same colours as in Figure 2.5; black denotes the catchment mean using the gridded raingauge-rainfall source. Similarly, the background conventions and information on the reliability of the ensemble forecast TWPs are provided in the same way as for the hydrographs. The symbols are shaded using the 99th percentile daily TWPs for this forecast initialisation. In addition to the fixed 0.5 and 8 mm d^{-1} precipitation thresholds, the annual 95th and 99th percentile precipitation thresholds for the Riccal at Nunnington are also shown, corresponding to 12 and 22 mm d^{-1} respectively. As a further guide, the maximum ensemble member accumulations for the 24-, 48- and 72-hour forecast-horizons are provided alongside the actual gauge-rainfall accumulations for the same period.

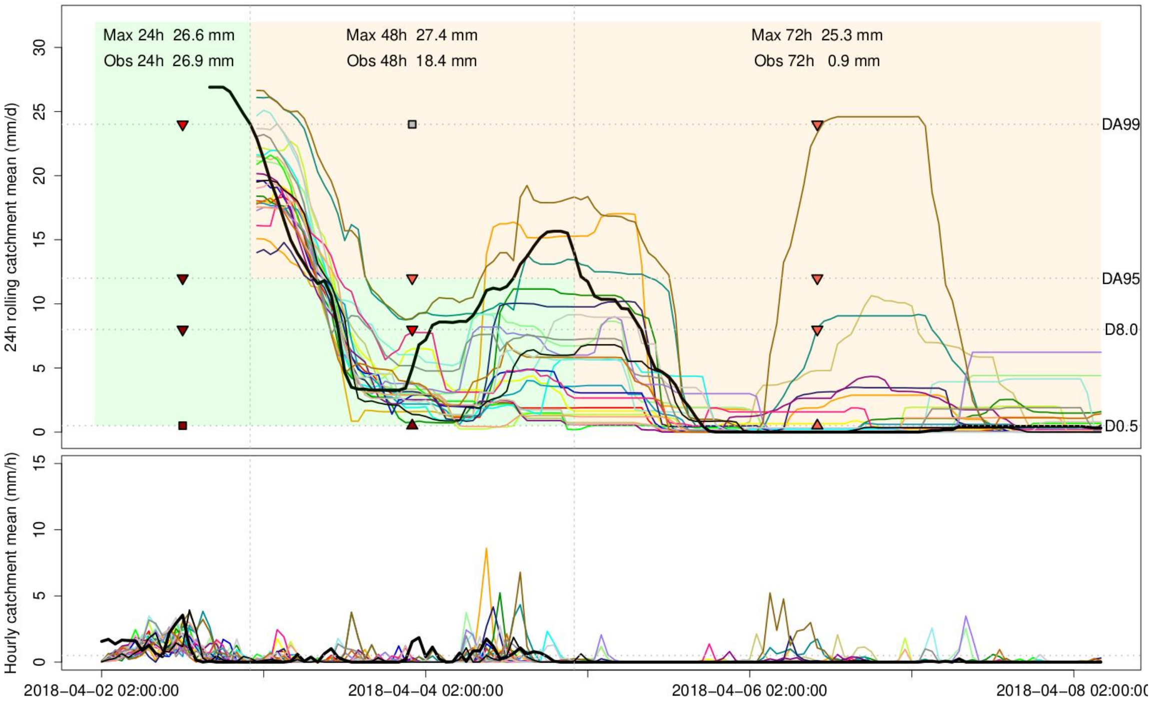


Figure 2.8 Example hyetographs showing the hourly catchment-mean accumulations (bottom panel) and a rolling 24h-window catchment-mean accumulation (top panel) for the ensemble precipitation forecast initialised at 01:00 2 April 2018 for the Riccal at Nunnington, NE England.

Spatial information is also important to provide the regional context that a catchment is in. It is known that precipitation forecasts may not have spatial accuracy at the catchment-scale. Whilst the percentile thresholds do vary in space, adjacent catchments are often similar so that when the TWPs are mapped, as shown in Figure 2.9, it provides an overview of where the 99th percentile thresholds are being exceeded in each catchment. As discussed in Appendix A.5, these are widely above 20 mm. Maps providing these percentile values, or having the ability to interactively hover over a catchment to obtain vital information for that catchment, are ways in which this information could be integrated for real-time use.

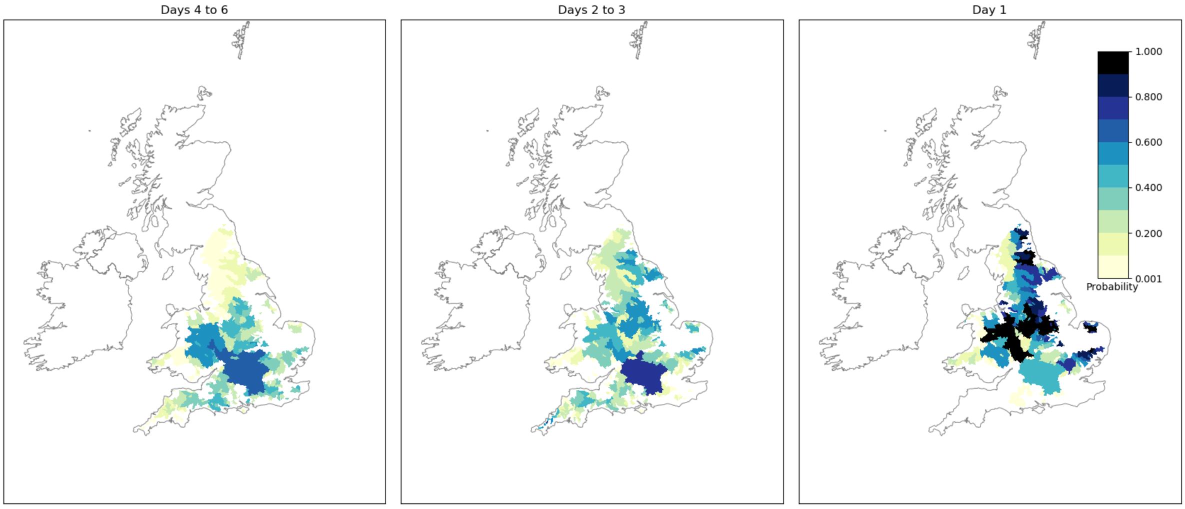


Figure 2.9 Maps of daily catchment TWPs exceeding the annual 99th percentile precipitation thresholds for the BMR forecast valid for the period ending 21:00 2 April 2018, demonstrating how the probabilities evolved between the different time-horizons (Days 4 to 6, Days 2 to 3, and Day 1).

2.4 Case Study Findings

The previous sections provided an overview of the main verification findings using the extended Phase 2 period of ensemble forecasts. Here, the river flow and precipitation BMR ensemble forecasts are compared directly for specific case-studies to provide greater insight into the ensemble behaviour for individual high river flow or high rainfall events. This allows direct comparison of member-to-member behaviour and the more-qualitative visual analysis of ensemble performance. In combination with the statistical verification analyses, which are displayed alongside the ensemble forecasts, this provides a comprehensive overview of the ensemble performance.

2.4.1 Case-study selection

Initial guidance on case-study selection was provide by the FFC, EA, NRW and SEPA (Section 2.1.2, Appendix A.2). The selection focussed attention on events where river flow or surface water flood impacts had been noted or forecast, for specific locations across the UK. From this list of case-studies, a subset were selected for joint hydrograph and hyetograph analysis based on observed river flow threshold-crossings. Catchments were selected based on the locations of observed or forecast impacts, and through analysis of the G2G simulated river flow hydrographs for the 2017 and 2018 water years. For river flow, both G2G and local model forecasts were analysed. These analyses are discussed in Section 2.4.2, with further details given in Appendix C.3.

A complementary analysis looked at all the suggested case-studies falling within the verification period June 2017 to September 2018, and analysed the national pattern of 24h raingauge catchment-precipitation accumulations. Locations of interest were identified as catchments with maximum precipitation accumulations, and compared with the locations with reported impacts. This analysis is presented in Section 2.4.3, with further details in Appendix C.2.

The catchments and case-studies considered for both forms of analysis are summarised in Table 2.3 (England & Wales) and Table 2.4 (Scotland).

2.4.2 Joint river flow and precipitation analysis of catchment time-series

Often a direct link can be made between the precipitation and river flow ensemble members for both PDM and G2G. Examples are given in Figure 2.10 for the Ricall at Nunnington catchment in NE England for the April 2018 case-study and Figure 2.11 for the Findhorn at Shenachie catchment in Scotland for the June 2017 case study. These figures show results for both G2G and PDM and are seen to be similar overall, with a direct correspondence between individual ensemble members. Full details of these case-studies are available in Appendix C.3.

It Is important to note how the PDM river flow forecasts do not show verification information above the Q(2)/2 threshold (Figure 2.11) despite many ensemble members crossing this threshold. As discussed in Section 2.2.2, the PDM verification scores are calculated for single-catchments only, and there are too few events above the Q(2)/2 threshold for score calculation. Of course, this case-study event in June 2017 does feature crossings above the Q(2)/2 threshold, and may allow scores to be calculated for higher verification thresholds if it were included in the overall verification period. However, as the verification period used in this instance is from 1 September 2017 to 31 August 2018, this is not the case. In a real-time forecasting system this might also be the case: there could be extreme river flow values which have not featured in the verification period used, and for which the verification can provide little guidance on ensemble forecast performance.

Table 2.3 Catchments used for case-studies over England & Wales. The catchments in the central column were analysed for both river flow and precipitation, whilst the catchments in the right-hand column were analysed for precipitation only. The exception to this is the 27 December 2017 case-study catchment Boyd at Bitton (53131 & PDM)* which was only analysed for river flow.

Case-study	Catchments selected based on river flow response	Catchment with peak 24h precipitation total
2017		
18 Jul		Walkham at Horrabridge (47118)
9 Aug		Gypsey Race at Boynton (Boyntn1)
23 Aug		Derwent (NE) at Low Marishes (MARISH1)
30 Sep		Kent at Sedgwick (730511)
21 Oct	Irwell at Irwell Vale (690140) Calder at Hebden Bridge (HEBDBR1) Calder at Todmorden (TODMDN1) Glaslyn at Beddgelert (065001_TG_1201 & PDM)	Glaslyn at Beddgelert (065001_TG_1201 & PDM)
3-4 Nov		Moors River at Hurn Court (43214)
22-23 Nov	Lune at Caton (724629) Wenning at Hornby (72452) Hindburn at Wray (724427) Wenning at Wennington (724326) Eden at Sheepmount (765512) Eden at Temple Sowerby (760502) Eden at Gt Musgrave Bridge (760112) Eden at Kirkby Stephen (760101) Glaslyn at Beddgelert (065001_TG_1201 & PDM)	Glaslyn at Beddgelert (065001_TG_1201 & PDM)
27 Dec	Boyd at Bitton (53131 & PDM)*	Wensum at Costessey Mill (E19862)
2018		
2-3 Jan		Derwent at Portinscale (751007)
12-14 Mar	Dove at Rocester Weir (4008) Dove at Hollinsclough (4033 & PDM) Torne at Auckley (4050)	Torne at Auckley (4050)
2-4 Apr	Derwent at Malton (Malton1) Derwent at Low Marishes (MARISH1) Riccal at Nunnington (Nunnington & PDM) Glaslyn at Beddgelert (065001_TG_1201 & PDM)	Glaslyn at Beddgelert (065001_TG_1201 & PDM)
20 Sep		Taff at Fiddlers Elbow (057007_TG_504)

Table 2.4 Catchments used for case-studies over Scotland

Case-study	Catchments selected based on river flow response	Catchment with peak 24h precipitation total
2017		
7 Jun	Lossie at Sheriffmills (234307) Mosset Burn at Wardend Bridge (234331 & PDM) Findhorn at Forres (234221) Findhorn at Shenachie (234306 & PDM) Divie at Dunphail (234206 & PDM) Nairn at Firhall (234218) Nairn at Balnafoich (234164)	Lossie at Sheriffmills (134307)
2018		
24 Jan	Tweed at Sprouston (15012) Ettrick Water at Lindean (14990) Ettrick Water at Brockhoperig (14987 & PDM Ettrick at Brockhoperig) Tima Water at Deephope (14986) Orchy at Glen Orchy (133087)	Orchy at Glen Orchy (133087)

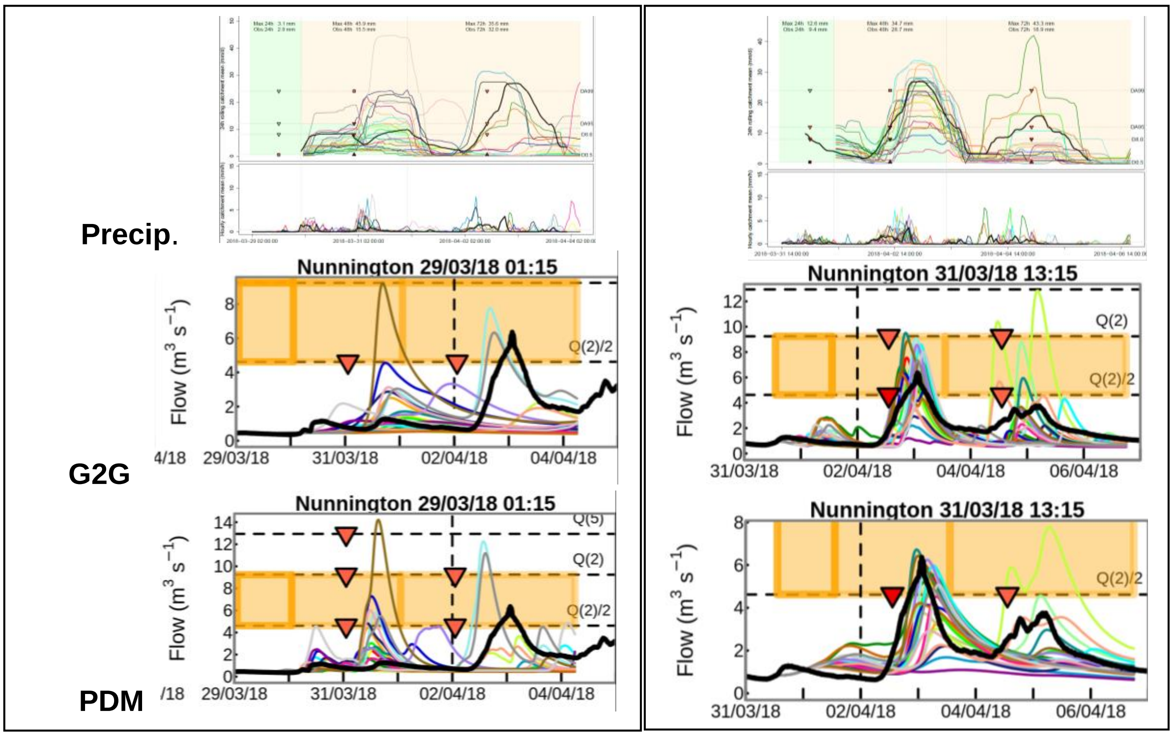


Figure 2.10 Example time-series of rainfall and G2G & PDM river flow (colours as discussed in Sections 2.3.2 and 2.3.3) for the BMR forecasts for Riccal at Nunnington initiated at 01:15 29 March (left) and 01:15 13:15 31 March (right) in 2018.

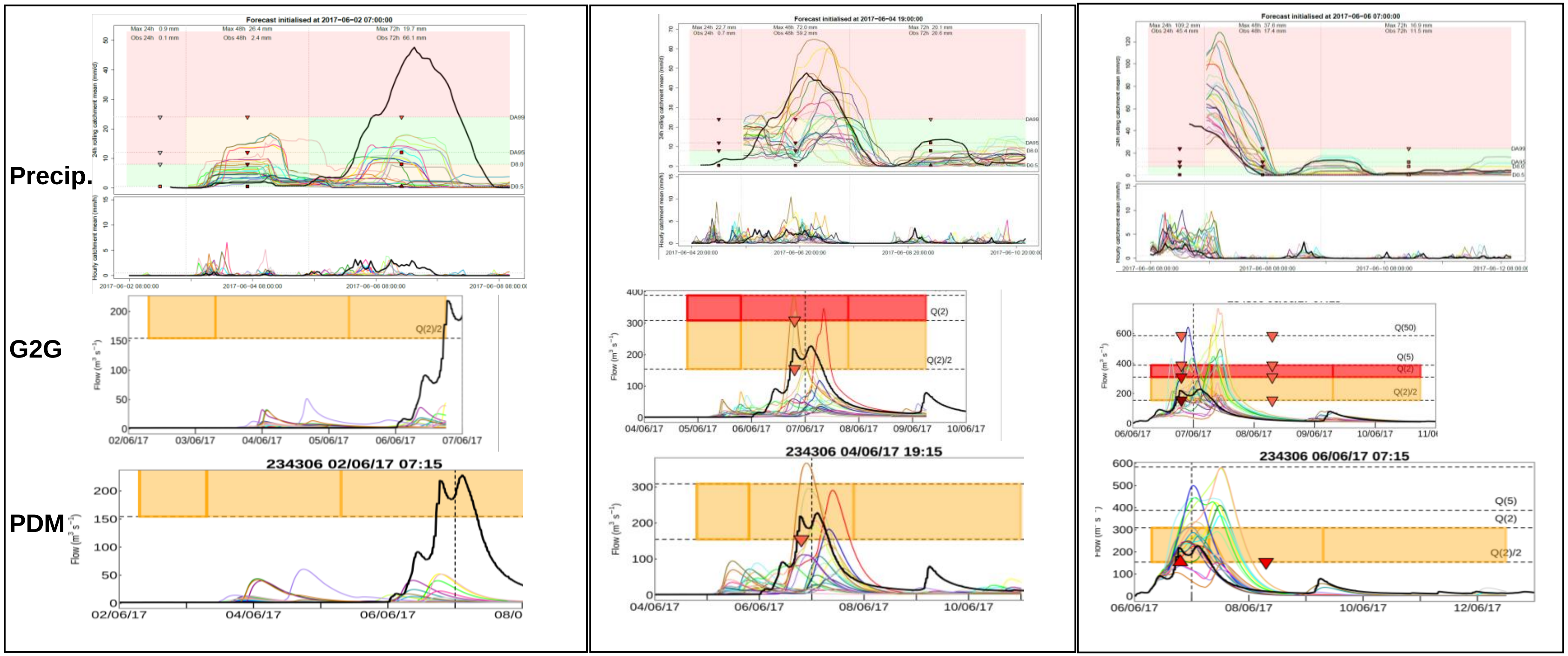


Figure 2.11 Example time-series of rainfall and G2G & PDM river flow (colours as discussed in Sections 2.3.2 and 2.3.3) for Findhorn at Shenachie (234306) BMR forecasts initiated at 07:15 2 June (left), 19:15 4 June (middle) and 07:15 6 June (right) in 2017.

For both rainfall and river flow, it was found that forecast performance does not necessarily improve with lead-time. For example, there are instances where longer lead-time forecasts perform better than those close to the event, or forecast performance varies between consecutive forecasts. This highlights the advantage of looking at multiple forecast-origins covering an event, not just the most-recent forecast.

Overall, it was concluded that useful information can be gained by viewing together the river flow and precipitation ensemble time-series. In general, for a given catchment, better performance is seen for PDM than for G2G. This is to be expected when comparing the performance of a countrywide distributed model to that provided by a set of catchment-calibrated local models.

2.4.3 Analysis of locations with the highest precipitation

The concept of extracting and visualising the spot maximum or 99th percentile within-catchment precipitation values was explored within the case-study assessment (Appendix C.2) and is considered to have some potential. For the case-studies, it was illustrated spatially with the daily precipitation accumulations for the three different observation sources. This served to highlight the differences between them illustrated in time-series form through the hyetographs. In this instance, only the raingauge observations are shown to avoid cluttering up the graphs unnecessarily. An example of the spatial map is provided in Figure 2.12. In a fully interactive real-time display, the user should be able to toggle between the maximum, 99th or catchment-mean precipitation forecast values at will for the spatial maps and the hyetographs. The observations should also be available retrospectively (at least for a time) to be able to review recent events.

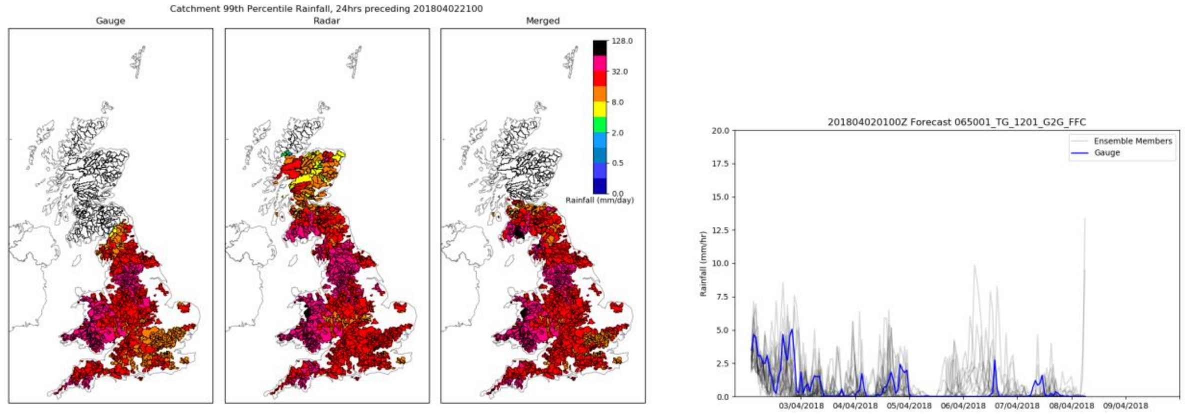


Figure 2.12 Example spatial maps (left) of the 99th percentile daily spot within-catchment precipitation accumulations (raingauge, radar, merged) for the 24h period ending 21:00 2 April 2018. Hyetographs (right) of ensemble members and raingauge observations showing the forecast evolution for the Glaslyn at Beddgelert (065001_TG_1201) catchment for the forecast at 01:00 2 April 2018.

2.5 Coding framework

2.5.1 Joint coding structures

The code structures used in Phase 1 were reviewed in detail by the Project Team to ensure that the code developed under the Project is robust, consistent across both river flow and precipitation, and structured appropriately to allow future flexibility and inform operational implementation. A detailed data flow diagram of the combined codes was created and is included in Appendix .

The following principles were used as a guide.

Code sharing. Where possible the same (identical) code will be used for river flow and precipitation processing and verification. Where this is not possible, the code will be consistent.

Saving data. Once calculated, all verification score data and products derived from the raw precipitation and river flow fields will be saved in a standard, simple, human-readable format. This gives the flexibility of reading in the data using different systems/coding-languages in the future. Processed river flow and precipitation data will be saved in daily or monthly blocks, giving flexibility when evaluating model performance over longer verification periods. In addition, certain scores can be calculated recursively (e.g. the Rank Histogram and CRPS) so values can be updated easily when new forecasts become available in real-time.

Plots and diagrams. The same file structures and naming conventions will be used to ensure these are easily comparable and identifiable for future systems. Where differences occur – for example, due to differences in the definition of thresholds - these will be clearly defined.

Parallelisation. Where possible, the elapsed time taken to complete a verification calculation will be reduced by splitting the geographical domains into smaller regions that can be run in parallel.

2.5.2 Considerations for operational implementation and use

When implementing the approach operationally, two modes for calculating the verification statistics will likely be needed; Mode 1: real-time calculation and update of statistics aligned with each operational forecast cycle, Mode 2: batch processing of long periods of ensemble data (this can include re-running the river flow forecasts).

Considerations for operational implementation can be divided into the following topics.

Data volumes and storage. Dealing with ensemble forecasts requires significant online storage. When calculating verification statistics in real-time (Mode 1) forecasts need to be dealt with when they are current i.e. as soon as the period covered by the forecast is in the past. After this, the forecast can be processed and deleted, or moved to long-term storage (access to this is needed for Mode 2).

Precipitation ensembles. This is the approach used for operational verification at the Met Office. The verification suites check every hour for new observations, and which forecast lead-times can now be verified. This means that online storage is actively managed and data volumes are kept stable with quick access to current information for as long as it is deemed necessary.

River flow ensembles. The same approach could be applied for flood forecasting systems. Long-term storage of the driving precipitation ensembles is required for Mode 2 batch processing.

Batch/bulk (Mode 2) vs keeping up with operational forecast cycle (Mode 1). Batch processing (Mode 2) is neither efficient nor is it fast, especially if a dedicated suite of maintained workflows has not already been created. This project has been a good example of this: back-processing 16 months of ensemble forecasts whilst developing the associated research workflows is a slow and laborious task which requires a lot of storage, compute power and time. However, at the start of any long-term monitoring this activity is unavoidable and provides the baseline. Beyond this point batch processing should be minimised where possible as it is far more efficient to keep up with operational forecast integrations close to real-time. However, there are occasions where batch processing is a necessity particularly on the flood forecasting side. For example, when a flood forecasting model has been upgraded, it would be desirable to batch calculate the verification statistics for that model over the agreed period of precipitation ensembles (e.g. since the last major weather model upgrade).

Minimising large data transfer. Where is the convergence of data streams? Observations and forecasts need to come together for visualisation: however, observations may or may not be available at the precipitation forecast generation end.

- Transferring ensemble forecast data in particular should only be done once.
- Ideally, all data-manipulation should be completed at source and only abstractions disseminated. Often this is not possible or desirable but transferring ensemble forecasts requires a “fat data pipe”, i.e. large data transfers need to be kept to a minimum.
- From a verification statistics perspective it is far more efficient to transfer verification statistics (over a “thin data pipe”) to a visualisation platform, with one caveat. If the forecasts are being transferred to a visualisation platform along with the observations for on-the-fly capability, there is no difference.

Should all of this be put in the cloud, then this would represent the perfect place for computing the verification too. Ideally, the hydrometeorologists should be given the best interactive experience possible and extract the relevant information that they need when they need it. There will undoubtedly come a time when the weather forecast models and ensembles will be run on cloud HPC computing. This is planned at the Met Office for HPC+1 or HPC+2 but timing is uncertain. Plans should be future proofed.

For reference, a summary is given below of the G2G data processing and volumes analysed for the England & Wales and Scotland domains over the period 1 June 2017 to 30 September 2018.

- 1,941 forecast-origins
- A total of over 20,000 days (equivalent to ~55 years) of G2G model simulations of river flow
- Large amount of computing power
- 487 dates with 1 job per date, 4 forecasts per day, taking ~20 hours to run all ensemble members and forecasts for England & Wales and ~12 hours for Scotland. Gives a total run-time of ~1.7 years if they were to be run in series.

Several large datasets created and backed-up at UKCEH.

- BMR rainfall ensemble SIDB database (GB coverage) ~1.1TB
- BMR deterministic temperature SIDB database for Scotland ~4.3GB
- Processed precipitation forecasts for rainfall verification ~138GB (once zipped)
- G2G ensemble forecasts (at gauged locations) for river flow verification ~10MB

What to do about weather model upgrades. The weather model is upgraded at least once a year. Any specific change may or may not have much impact for hydrometeorological applications, but the cumulative effect often does. Some changes will be more impactful than others, such as changes in spatial resolution. This poses an interesting verification dilemma. From a long-term monitoring perspective, it means the user has no clarity (and no immediate experience to draw on) as to how relevant the existing long-term verification information is when viewing forecasts from the new weather model configuration. However, the weather model rarely changes that radically. The following considerations are noted.

- The information that is needed has to be gathered through a more thorough end-to-end pre-operationalisation evaluation.
- The long-term trends will eventually adjust to the new configuration. This is why shorter verification windows can have value as they will track more closely with recent weather model changes (or weather dependencies).
- Access to monthly or seasonal behaviour is potentially very valuable.
- For hydrological modelling the use of canned data from some early weather model upgrade runs may be useful or the option for a period of parallel feeds.

It is also noted that reforecasts of past historical periods have long been recognised as beneficial but are extremely expensive to produce. Should the Met Office ever decide to pursue this, downstream applications such as hydrological modelling would also benefit. If only a few years could be re-run, it may also be beneficial to include some years having a reasonable number of high flow/flood events.

What to do about flood forecasting model upgrades? The national models used by FFC and SFFS are periodically updated at a national scale. Local models are updated, or new models added, on a rolling basis and can range from one catchment location to spanning across whole river basins. Often, the change in hydrological model performance can be marked between releases. So real-time calculations (Mode 1) should restart when the model is released. Secondly, historical ensemble runs could be created if computing capacity exists and for the period that a suitable version of the precipitation ensemble exists (e.g. since the last major weather model upgrade).

Computing and updating climatologies. To create climatological catchment-precipitation thresholds, ten historical years of raingauge-rainfall data from 2007 to 2016 were used (Section 2.1.5). To facilitate this task computationally, each date in the 10-year historical period was processed separately, and the distribution of precipitation values for all grid-cells falling within each catchment saved as a histogram. This was done separately for 15-minute, hourly (ending on the whole hour) and 24h (ending at 00:00) precipitation accumulations. These climatologies should be continually updated and expanded to the recommended 30 years over time.

Code ownership, maintenance, future development and overall responsibility*.* For an operational system the code would need to be associated with ownership. Code owners would be responsible for maintaining, upgrading the code, and fixing issues. For this to happen the code would need to conform to universal software quality assurance standards with respect to documentation, reviewing, etc. and be securely stored. This would require appropriate funding and be in partnership across the research side (UKCEH and Met Office) and operational user organisations.

3 Recommendations

Recommendations from the project are here grouped and numbered as follows.

- Ensemble Verification Framework overall considerations (Section 3.3)
- Calculation of ensemble verification statistics (Section 3.4)
    - Real-time incremental approach
    - Historical batch analysis over long runs
- Real-time displays and visualisation of ensemble verification statistics (Section 3.5)
- Improvements to the underlying approach (e.g. data, forecasting technique) (Section 3.6)

The following points have been considered, where appropriate, in forming the recommendations.

- Differences between FFC/SFFS (national model) requirements and EA/SEPA/NRW (local model) requirements, recognising the differences in current use of ensemble forecasts
- Relative priorities and what would be the minimum needed for an operational system, aimed at FFC/SFFS in the first instance
- Known constraints and dependencies (particularly those listed in Section 2.5.2)

For background, the motivation and current status are discussed in Section 3.1 along with an overview of the project rationale in Section 3.2.

3.1 Motivation and current status

This research was motivated primarily by the needs of the Flood Forecasting Centre (FFC) and Scottish Flood Forecasting Service (SFFS) to better understand the performance of the end-to-end ensemble flood forecasting systems that are used daily to underpin national flood guidance services. Although these systems have been in use for many years, their performance has not been routinely verified. Therefore, the research questions, from a practical FFC/SFFS user-perspective, can be summarised as follows.

- How well has the ensemble precipitation and flood forecasting system performed in the (recent) past? Particularly for flood events of interest.
- What does this mean for interpreting today's forecast?

Whilst the Phase 1 project focussed on an ensemble verification framework for the national-scale G2G river flow model, Phase 2 has seen an expansion of scope to also consider how local models – such as the PDM catchment model – could be included. These local models are employed in model networks configured to river basins and operate within FEWS-based forecasting systems at the EA, SEPA and NRW. Currently these local models are usually run to produce deterministic forecasts of river flow.

Whilst this was an R&D project using historical data and case-studies to illustrate approaches, the project was also tasked with considering **future operational implementation**. In this context, it is important to recognise that each agency is at a different point on the transition to ensemble flood forecasting and has differing needs and priorities. The FFC and SFFS have used national-scale ensemble precipitation and river flow (G2G) forecasts for many years: they have immediate needs for improved ensemble forecast verification. Whilst EA, NRW and SFFS do not yet routinely use ensemble precipitation forecasts as input to their local models, they are planning the transition to this in the future.

In the closing stages of the project, a Partner Workshop was organised in December 2020 by FFC with representatives from the SFFS, EA, SEPA and NRW. The purpose of the workshop is given below.

> *"Horizon scan across the partnerships to understand the future operational flood forecasting landscape and what this could mean for the application, in an operational setting, of rainfall and river flow ensemble verification information."*

A summary of the workshop is given in Appendix . Whilst the scope was wider than this project, it was useful to inform the recommendations reported here. It also reaffirmed that for many of the local forecasting processes, the question of how to move to probabilistic and ensemble forecasting was actively being considered. There was a clear aspiration to progress to probabilistic flood forecasting and appreciation of the future benefits from doing so in the medium- to long-term. There was acceptance across the group that ensemble rainfall and river flow verification would need to be integrated into any future local ensemble forecasting system.

3.2 Project rationale

Ensembles are necessary to capture the inherent uncertainties in precipitation forecasts and resulting river flow forecasts. These ensemble forecasts should be routinely verified but this information only becomes useful when it is provided to the users in a useable, digestible and meaningful form. Viewing and being able to interrogate verification information alongside the latest forecast is a powerful addition to the hydrometeorologists' toolkit for interpreting the information they have available and for subsequent decision-making.

This project has designed and developed an Ensemble Verification Framework suitable for hydrometeorological applications. This framework provides a useful foundation on which to build an operational long-term performance monitoring capability. Figure 3.1 gives an overview of an operational end-to-end interactive ensemble forecast visualisation and verification system built on the framework. Recent years have seen a significant investment being made in display and processing systems used by the operational flood forecasting agencies across Great Britain. The provision of ensemble and probabilistic forecast products and tools - which maximise useful information content alongside up-to-date verification information and observations - will support enhanced decision-making and increase confidence in the use of ensemble forecasts.

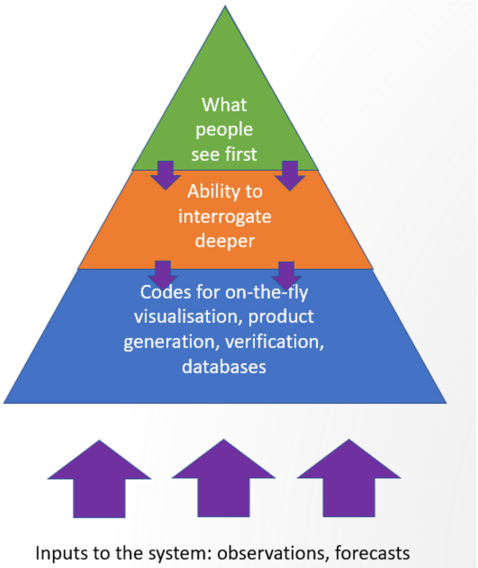

What people see first
Ability to interrogate deeper
Codes for on-the-fly visualisation, product generation, verification, databases
Inputs to the system: observations, forecasts

Figure 3.1 Hierarchical view of an end-to-end interactive forecast visualisation and verification platform.

The overview of the system, depicted in Figure 3.1, has succinct user-relevant displays at "the tip of the iceberg", i.e. what people see first. Examples of these high-level displays are provided in Section 2.4.2. These build on more detailed site-specific or regional real-time verification information that could be interactively viewed per site, or as national pictures of ensemble performance. Examples for the precipitation and river flow are given in Appendices B.1 and B.2 respectively. The key foundation to the system is a long period of ensemble forecasts and associated observation information, and verification analysis code for processing these.

This project has finally confirmed that hydrological applications require a different perspective to precipitation forecast verification. Thus far, this fact was only speculated on, and effectively ignored. Precipitation forecasts were used as input to river flow forecasting systems without having a specific, quantitative understanding of the quality of the input. The specific needs of the hydrological community were not being met.

Bringing the meteorology and the hydrology closer together has helped identify how precipitation forecasts can be presented and post-processed to support flood risk decision-making. For example, the creation of Time-Window Probabilities as a way of visualising precipitation forecasts has shown that these are computable for higher precipitation thresholds and provide larger probabilities which are more reliable and more skilful. Thus, the outcomes of this project should enhance and improve the decision-making processes involved in operational flood guidance and warning.

3.3 Recommendations: Ensemble Verification Framework overall considerations

- ***R1 Implementation of end-to-end interactive forecast visualisation and verification platform for FFC and SFFS.*** There is a long outstanding need for this from FFC so should be a priority for implementation. This report contains the outline for such a system. There are several options as to how advanced such a system should be, but recommendations **R2, R3, R4, R6, R8, R10 and R12** would provide a minimum system that could be advanced upon and provide a template for others. SFFS, as the other regular user of ensemble precipitation and river flow forecasts, could follow the same or similar steps. EA, SEPA and NRW local model users can feed-in to the process, or keep a watching brief, whilst developing their respective plans for probabilistic forecasting and warning with local models.
- ***R2 Creation of appropriate governance structure for the Ensemble Verification Framework, system and associated code.*** A life-cycle approach to managing and updating the Ensemble Verification Framework, system and associated code should be put in place. This would require appropriate funding and be a partnership across the research side (UKCEH and Met Office) and operational user organisations. The code management requirements are outlined in Section 2.5.2
- ***R3 The ensemble verification system should be part of the flood forecasting system.*** The precipitation and river flow ensemble verification system should form part of the flood forecasting system. As this is a key investment decision, reasons for this recommendation are given below.
    - All forecast and observation data required for the ensemble verification system are available in the flood forecasting system. Therefore it is much more efficient (time and costs) to verify the large data volumes (see Section 2.5.2) where the data are, rather than move large data to another verification platform.
    - Importantly, as part of the flood forecasting system, the verification system can update verification statistics continually in real-time (avoids need for periodic offline updates).

  - Verification information needs to be visualised and available in the flood forecasting system to support real-time decision-making (see recommendations in Section 3.5), so it is efficient to have the ensemble verification system as part of the flood forecasting system.
  - Creates an operational functionality in the flood forecasting organisations and robustness to the workflows. Can be re-used for historical period (e.g. when models change).
  - All flood forecasting systems used across the various operational agencies are FEWS-based so should make roll-out to other organisations and flood forecasting models easier in the future (e.g. inclusion of local models). In the first instance, this would be the Incident Management Forecasting System (IMFS) used by the FFC and EA, and, potentially, FEWS Scotland and FEWS Wales.
- ***R4 The Joint Coding Framework (Appendix) should be used as the basis for the operational implementation.*** Although modification may be required, it provides a detailed description of the calculation process based on lengthy discussions between the Met Office and UKCEH and aims to be computationally efficient, using common codes where possible.
- ***R5 Where possible, the period of ensemble data used for calculating verification statistics should be long enough (>2 years) to generate sufficient threshold-exceedances.*** The project used a period of 16-months which was sufficient for generating enough precipitation threshold-exceedances for the 95th percentile thresholds. But this period-length is insufficient for higher precipitation thresholds and especially for considering river flow thresholds above Q(2)/2 at sub-regional scales.

3.4 Recommendations: Calculation of ensemble verification statistics

Here, recommendations are primarily based on the discussion in Section 2.5.2. These ensemble verification statistics are the information that underpin the whole system so their creation in a future operational system needs to be considered carefully.

- ***R6 Real-time calculation of ensemble verification statistics in the flood forecasting system.*** This is essential for operational implementation of the Ensemble Verification Framework and should be part of the operational forecast cycle.
- ***R7 Batch calculation of ensemble verification statistics within the flood forecasting system for long periods of historical data.*** Although not essential for a minimal real-time verification system, it is strongly recommended that the system should be able to run periods of flood forecasts using archived ensemble precipitation forecasts and verifying observations, and calculate ensemble verification statistics from these. This would be required to accommodate flood forecasting model updates (e.g. G2G releases).
- ***R8 Initialisation of ensemble verification statistics for initial system.*** Depending when the system is first installed, a one-off offline creation of the starting verification statistics should be considered. The current recommendation would be to extend from the end of the current 16-month study period until the start of the system for the national models. This would be particularly beneficial for the river flow verification.
- ***R9 A process for identifying and handling major weather model upgrades to be discussed and agreed between Met Office and FFC/other flood forecasting agencies.*** See discussion in Section 2.5.2. When a weather model upgrade is expected to have a noticeable effect on the precipitation ensemble forecast characteristics, every effort should be made to provide flood forecasting agencies with a long period of ensemble forecasts ahead of the live datafeed switching over.

3.5 Recommendations: Real-time displays and visualisation of ensemble verification statistics

To support enhanced decision-making, the visualisation system needs to bring together ensemble and probabilistic forecast products which maximise useful information content alongside up-to-date verification information and observations. Such a system needs to be interactive and allow users to interrogate the system, displays and information to the required level of detail.

> ***R10 Implementation of new real-time forecast displays that incorporate ensemble verification information.*** There is a clear and pressing need for FFC and SFFS to have access, as soon as practicable, to new real-time forecast displays within the flood forecasting systems that incorporate ensemble verification information. Based on user interaction to date, a suggested starting point for this is to combine the time-series and spatial information presented for the case-studies in Sections 2.4.2 and 2.4.3: this is shown here in Figure 3.2. These two visualisations provide complementary information, allowing individual catchment values to be viewed in the context of the wider meteorological and hydrological situation.

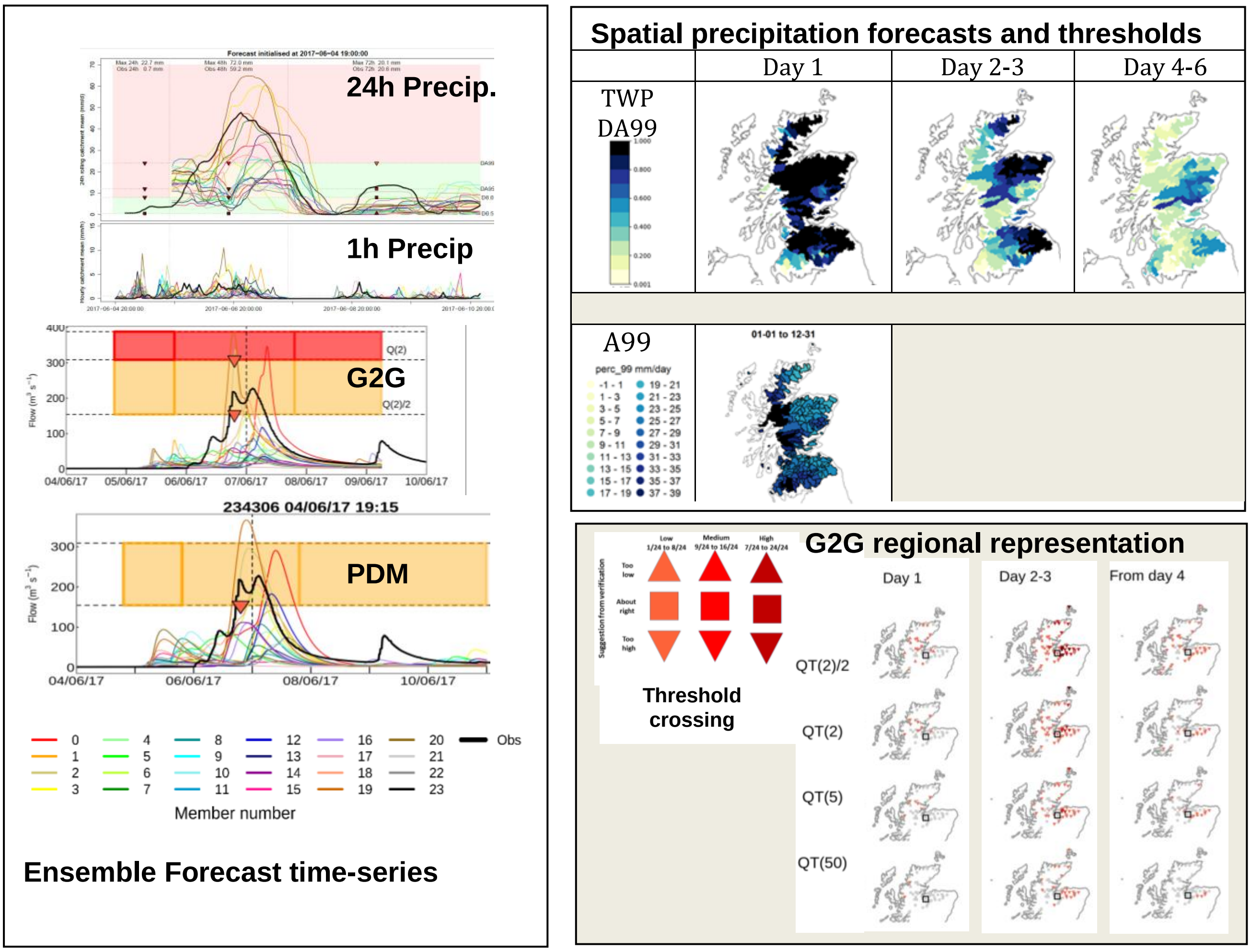


Figure 3.2 Example of a combined verification display for ensemble precipitation and river flow forecasts initiated at 19:15 4 June 2017, showing information as times-series and maps for the Findhorn at Shenachie (234306).

- ***R11 Implementation of a new site-specific and regional/national summary of forecast verification statistics.*** The underpinning verification statistics that are being updated in real-time should be available and interactively linked from the new real-time forecast displays. Example displays for the precipitation and river flow are given in Appendix B.1 and B.2 respectively.
- ***R12 A co-design approach for new displays is recommended.*** Further co-design of displays and interactive functionality between researchers, operational users, and system developers will be required during implementation. Future local model needs can be included as the plans for local ensemble forecasting evolves. Some potential ideas are to be able to view sequences of recent forecasts (Section 2.4.2), change precipitation thresholds interactively on the fly (Section 2.3.3), and inclusion of other relevant information such as antecedent soil moisture estimation via observations.
- ***R13 Multiple versions of ensemble forecasts and the associated verification statistics should be available.*** For example, comparison with a previous version of the ensembles or a parallel live feed of a new rainfall ensemble product. There should be no distinction between offline and online (real-time) visualisation capability. Introducing a common means of viewing and producing verification statistics will enable the smooth transition of model upgrades into operations, for example. Such capability should be seen as essential for future operational application. Providing an end-to-end testing platform whereby an existing operational processing-chain can be compared with pre-operational datafeeds is vital.

3.6 Recommendations for improvements to data and methods used in the Ensemble Verification Framework

This section lists potential improvements to forecast & observation data and methods used in the Ensemble Verification Framework not listed naturally above. These are not necessarily needed for the delivery of an initial operational ensemble verification system and likely be advanced through other R&D and operational activities. Nevertheless, these have been listed as they can be incorporated into future releases of the operational Ensemble Verification Framework.

- ***R14 The merged radar-raingauge precipitation product should be made available over the whole UK and include rain-snow discrimination.*** Whilst far from perfect, the merged product does provide a compromise between the radar and gauge-only rainfall products. The former has the spatial distribution detail whilst the latter has the spot accuracy of the raingauges. For Scotland, better rain-snow discrimination would also appear to be essential. New and emerging radar-precipitation products could provide useful additional information.
- ***R15 Precipitation ensembles used for flood forecasting should ideally come from a single weather model configuration.*** The construct of the BMR precipitation ensemble forecast - being the amalgamation of three different weather model configurations - means that the nature of the forecast changes more drastically with longer lead-times than if a given precipitation forecast was made up of a single model configuration. This is already likely under consideration and testing a 5-day MOGREPS-UK ensemble to replace the BMR ensemble at some point in the future is essential.
- ***R16 Ensemble Calibration should be explored as a future R&D priority***. This project has shown that the BMR forecast reliability could be improved with some calibration of the ensemble. In the first instance some form of reliability ensemble calibration would be recommended but more research work into physical calibration of precipitation for hydrological applications should also be considered. Post-processing of the river flow ensemble is another option to consider.

- ***R17 Utilise additional forecast precipitation products from IMPROVER.*** A number of more generic precipitation forecast products, such as being planned from the IMPROVER post-processing system, will also provide good additional guidance useful for decision-making. Their inclusion in future versions of the ensemble verification and visualisation system should be considered.
- ***R18 Explore best use of a longer multi-year period of ensembles.*** It is recommended to investigate best use of longer multi-year periods of ensemble forecasts as part of the life-cycle of the ensemble verification system. Being able to compile season-specific precipitation analyses over multiple years would be beneficial to capture known inter-annual variability. Also it would be useful to investigate if only using a subset of the forecast-origins each day (rather than all of them) loses any information. Such considerations would have implications for R7 and R8 and the computation time needed for batch processing.

4 Conclusions

The Flood Forecasting Centre (FFC) and Scottish Flood Forecasting Service (SFFS) have operated and used national-scale end-to-end ensemble flood forecasting systems for a number of years. However, the performance of these ensemble systems are not routinely assessed and verified. Allied to this, the local forecasting systems operated by the Environment Agency (EA), Scottish Environment Protection Agency (SEPA) and Natural Resources Wales (NRW) are planning wider use of ensemble and probabilistic forecasting systems in the future.

The operational research questions from a user perspective addressed in this project were:

- How well has the ensemble precipitation and flood forecasting system performed in the (recent) past? Particularly for flood events of interest.

- What does this mean for interpreting today's forecast?

In addition, the project also had to consider how any proposed solution could be implemented operationally. This had to recognise the immediate needs of FFC and SFFS, who use ensemble forecasts daily, and the longer-term needs of the EA, SEPA and NRW as they plan their transition to ensemble flood forecasting.

The project has addressed this challenge by designing and developing an **Ensemble Verification Framework.** It has also considered how this framework could be used to develop an operational end-end interactive **Ensemble Forecast Visualisation and Verification System**. This system aims to provide ensemble and probabilistic forecast products which maximise useful information content, alongside up-to-date verification information and observations to support enhanced decision-making: thereby increasing confidence in the use of ensemble forecasts.

To test and develop the potential verification approaches and operational displays, 16-months of precipitation and river flow ensemble forecasts have been processed and verified. Specific events have been identified with stakeholders for use in case-studies for evaluation and demonstration purposes. This has allowed rigorous scientific exploration of how to provide robust verification statistics of the ensemble precipitation inputs to the river flow modelling and of the ensemble river flows themselves. Section 2.1 summarises the Verification Methodology and Section 2.2 provides a summary of the scientific findings and evidence, complemented by the detail contained in the Appendix science reports.

Operational users have also been engaged in the design of real-time forecast displays through the Project Board and a Workshop. This interaction has identified that the real-time displays need to be flexible and informative, with varying layers of detail. Example real-time precipitation and river flow displays have been produced for a large number of case-studies using both the national G2G model and the local Probability Distributed Model (PDM) – a catchment rainfall-runoff model (Sections 2.3 and 2.4). Viewing the precipitation and river flow together, however, is the most important ingredient along with using common methods for conveying information on both. Prototype joint rainfall and river flow displays have been created (Section 3.5) but further co-design of interactive displays is recommended during implementation and should include operational users, researchers and system developers.

Overall, the key finding is that joint precipitation and river flow ensemble verification is possible and useful. The primary recommendation (R1, Section 3) is that an end-to-end interactive **Ensemble Forecast Visualisation and Verification System** for FFC (and SFFS) be implemented as soon as practicable. The **Ensemble Verification Framework** provides the blueprint for the system and the **Joint Coding Framework** (Section 2.5 and Appendix) developed here provides the basis for the algorithm and code. A detailed set of recommendations have been provided in Section 3, including what is required for operational implementation. This also includes a priority list of recommendations for developing a minimum system. A key recommendation is that the ensemble verification system is implemented in the downstream flood forecasting system rather than the upstream precipitation ensemble system.

The proposed system would address the current urgent operational gap for FFC and SFFS. It would mark a significant addition to the forecasters' toolkit by providing real-time displays that incorporate ensemble verification information for the first time, and in a usable form. In turn, this will facilitate enhanced and more informed decision-making at times of potential flood risk. Although local model systems are still planning for ensemble flood forecasting, these systems would eventually benefit from the FFC/SFFS developments and local model users could input into their co-design.

References

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. *J. Hydrol.*, **577**, 123966. https://doi.org/10.1016/j.jhydrol.2019.123966

Dey, S., Moore, R.J., Cole, S.J., Mittermaier, M., Csima, G., 2019. Rainfall and River Flow Ensemble Verification: Prototype Framework and Demonstration. Contract Report to FFC/SEPA/EA/NRW, Centre for Ecology & Hydrology and Met Office, V3.0, January 2019, Wallingford, UK, 171pp.

Jewell, S.A., Gaussiat, N., 2015. An assessment of kriging-based rain-gauge–radar merging techniques. *Q. J. R. Meteorol. Soc.*, **141**, 2300–2313. https://doi.org/10.1002/qj.2522

Appendix

Appendix A – Appendices for verification methodology and framework

A.1 – Joint Verification Framework

A.2 – Events for possible case-studies

A.3 – Catchment precipitation processing

A.4 – Definition of Time-Window Probabilities (TWPs)

A.5 – Climatological thresholds maps for England & Wales and Scotland

A.6 – Forecast triggering investigations

Appendix B – Appendices relating to scientific findings

B.1 – Overall precipitation verification summary
- **B.1.1 – Commentary on precipitation verification maps and plots**
- **B.1.2 – Precipitation verification maps and plots**

B.2 – Overall river flow verification summary
- **B.2.1 – Overall verification summary: river flow analyses**
- **B.2.2 – Supplementary plots (separate zip file)**

B.3 – Impact of observation uncertainty on verification metrics

B.4 – Fifteen-minute precipitation verification results and future plans

B.5 – Comparison of G2G river flows using different rainfall sources as input
- **B.5.1 – Comparison of G2G river flows using different rainfall sources as input**
- **B.5.2 – Supplementary plots (separate zip file)**

Appendix C – Appendices containing case-study analyses

C.1 – Evaluation and comparison of December 2015 case-study storms

C.2 – Precipitation assessment of case-studies

C.3 – Hydrograph analyses for flood-producing case studies
- **C.3.1 – Case study analysis: hydrological impacts, rainfall and river flow time-series**
- **C.3.2 – Supplementary plots (separate zip file)**

Appendix D – The Joint Coding Framework

Appendix E – Key findings from the Ensemble Verification Project Partner Workshop 17 Dec 2020