Article (refereed) - postprint

Anderson, Seonaid R.; Csima, Gabriella; Moore, Robert J.; Mittermaier, Marion; Cole, Steven J. 2019. **Towards operational joint river flow and precipitation ensemble verification: considerations and strategies given limited ensemble records**.

Towards operational joint river flow and precipitation ensemble verification: considerations and strategies given limited ensemble records

Seonaid R. Anderson*, Gabriella Csima[b], Robert J. Moore[a], Marion Mittermaier[b], Steven J. Cole[a]

[a] Centre for Ecology & Hydrology, Wallingford OX10 8BB, UK

[b] Met Office, Fitzroy Rd, Exeter EX1 3PB,UK

ABSTRACT

A framework for joint verification of river flow and precipitation ensembles is developed and demonstrated over Britain for eventual use in an operational flood forecasting setting. The river flow ensembles are obtained from a distributed hydrological model, the G2G model, using an ensemble of 15 minute precipitation accumulations as input on a 1 km grid. The precipitation ensemble consists of operational Numerical Weather Prediction (NWP) forecasts from the Met Office Unified Model. Both hourly and daily precipitation accumulations are verified, and the relevance of different accumulation periods discussed in the context of timing errors and hydrological response. The implications of precipitation observation error are investigated by comparing verification results from raingauge- and radar-derived precipitation estimates. Challenges of verification using only a limited record of precipitation ensembles, from a system only relatively recently made operational, are addressed. Methods of obtaining more robust verification statistics, given the available ensembles, are presented and demonstrated for an example period in December 2015. For precipitation, percentile thresholds are used to ensure a given number of threshold crossing events for analysis using a contingency table and derived skill scores. For river flow,

percentiles thresholds are of less relevance to operational flood guidance. Instead, exceedance of a flow threshold of given rarity (return-period) is used as a surrogate measure of flood severity. At the regional scale, both river flow and precipitation verification analyses are found to be dependent on the locations considered. This is linked to variations in precipitation amount. For river flows, catchment properties - and in particular catchment size - are found to be a key influence on verification. It is demonstrated how such behaviour can be used to obtain more-robust river flow verification statistics at sub-regional scales.

KEYWORDS

Verification, precipitation, river flow, ensemble, uncertainty, operational

1. Background and introduction

The development of hydrological ensemble systems is an active area of research, with investigations into the sources and drivers of uncertainty (e.g. Zappa et al., 2011; Brown et al., 2014a; Brown et al., 2014b; He et al., 2015), observation uncertainties (e.g. Caseri et al., 2016; Cecinati et al., 2017) and ensemble formulation (Marty et al., 2013; Davolio et al., 2013). Along with a prediction of the expected hydrological flows, these ensembles give a measure of the forecast uncertainty, allowing the forecasts to be interpreted probabilistically (Alfieri et al., 2011; Hardy et al., 2016; Alfonson et al., 2016). The Hydrologic Ensemble Prediction EXperiment (HEPEX), initiated in 2004, is a community of researchers and hydrological ensemble practitioners seeking to advance the science and practice of hydrological ensemble prediction (e.g. Schaake et al., 2007; Thielen et al., 2008).

Coupled river flow and Numerical Weather Prediction (NWP) ensembles are now used operationally across the world (e.g. Cloke and Pappenberger, 2009 and references therein; Alfieri et al., 2014; Demargne et al., 2014) to provide forecasts, guidance, and warnings of flooding. Over Britain, coupled with NWP ensemble precipitation forecasts from the Met Office, the Grid-to-Grid (G2G) distributed hydrological model (Moore et al., 2006; Bell et al.. 2009; Cole and Moore, 2009) configured at national-scale forms a key element of the operational flood guidance provided by the Flood Forecasting Centre (FFC) over England & Wales (Price et al., 2012) and the Scottish Flood Forecasting Service (SFFS) over Scotland (Cranston et al., 2012; Cranston and Tavendale 2012). These hydrological ensemble systems benefit from recent advances in NWP precipitation ensembles, which are now run at sufficiently high resolution that convection can be resolved (although not fully captured) by the model dynamics (e.g. Clark et al., 2016). These "convection permitting" NWP ensembles are produced operationally at a number of forecasting centres including the Met Office (Baldauf et al., 2011; Bouttier et al., 2012; Hagelin et al., 2017).

To fully benefit from these ensembles, it is necessary to understand their performance, behaviour, and drivers. Several recent studies have focussed on the verification of short- to medium-range hydrological ensemble forecasts, using example river basins (Addor et al., 2011; Zappa et al., 2013; Brown et al., 2014a; Brown et al., 2014b). Probabilistic forecasts from the European Flood Awareness System have also been recently assessed more generally, verifying against a reference simulation with observed fields as input (Alfieri et al., 2014).

At the FFC, the performance of the overall end-to-end ensemble flood forecasting system is currently not verified routinely. This paper reports first steps towards filling this operational-

critical knowledge gap. The aim is to develop a holistic, end-to-end ensemble verification framework, relevant to the ensembles' use in a flood-forecasting context. An operationally useful subset of existing metrics and methods are selected and discussed, along with the definition of thresholds, accumulation periods, and spatial scales relevant in this context. To this end, the focus here is on flood-producing thresholds when evaluating the river flow ensemble, and on precipitation thresholds that select the tail of the precipitation distribution when evaluating precipitation ensembles. Developing a verification framework appropriate for, and relevant to, flood forecasting is considered a necessary and important avenue of scientific investigation to enable the best use to be made of recent developments in verification metrics: namely, to "pull through" the science to the forecasting bench. Of course, many scientific challenges remain, including the most appropriate way of pairing precipitation events with the corresponding river flow events, both in terms of magnitude and time scale.

The development of the verification framework is first presented followed by a brief demonstration of its application with results, analysis and discussion. A 32-day example period (going beyond the often presented case-study approach) is used to put these considerations in context, and to demonstrate some of the overarching scientific, statistical, and technical challenges in this area. Such an analysis is an important precursor to a long time-period verification assessing the end-to-end forecasting system. Aligning with the operational setup at the FFC, a total of 898 catchments with gauged river flows within England and Wales are included here, omitting 225 catchments over Scotland for brevity of presentation.

The paper is structured as follows. First, the data, models and verification metrics are introduced in Section 2. Next, key verification considerations - including the use of thresholds, accumulation periods, and spatial scales - are discussed in the context of a short-range operational flood forecasting system in Section 3. Section 4 demonstrates the verification methods using December 2015 as an example study period. Further discussion in Section 5 is followed by the key conclusions in Section 6.

2. Framework for verification: data, models and metrics

2.1. Verification metrics considered

To provide an overview of ensemble performance, a range of well-established verification metrics and diagrams were selected and applied to verify both the continuous probabilistic forecasts, and the binary forecasts of event occurrence, in an operational flood forecasting context. The use of thresholds on river flow and on precipitation to define the binary forecasts of event occurrence is discussed in Section 3.2. The selected metrics and diagrams, summarised below for ease of reference, give an overview of the ensemble forecast accuracy. Of course, with an increasingly large selection of forecast verification metrics detailed in the hydrometeorological literature, other choices could have been made. This work focussed on using well established metrics giving a measure of the key ensemble forecast attributes: error of the full ensemble distribution, probability error, reliability, resolution, potential (calibrated) skill, discrimination and economic value. Each attribute is defined as the selected metrics are introduced. Some forecast attributes are evaluated by more than one metric, allowing the relative sensitivities of different metrics to precipitation observation uncertainty to be assessed.

To obtain a measure of the forecast skill in relative terms, a metric calculated from the ensemble forecasts can be compared to that calculated for a reference (benchmark) forecast. This gives the *Skill Score* for that metric. The choice of benchmark forecast depends on the aim of the verification, and must be considered in the interpretation of the Skill Score results: for example, see Pappenberger et al. (2015). Common benchmarks include climatology, persistence, and a random forecast.

2.1.1 Evaluation of continuous probabilistic forecasts

To assess the *error of the full ensemble distribution*, a continuous version of the Brier Score (Brier, 1950), the *Continuous Ranked Probability Score* (CRPS) was applied (Hersbach, 2000). The CRPS measures the difference between the cumulative distribution estimated by the ensemble forecast, and the step-function cumulative density function of the observation. As indicated in Hersbach (2000), the CRPS is commonly averaged over a number of cases. Here, the CRPS is averaged over all forecasts and all time periods contained within the forecast lead-time range being considered. A dimensionless skill score - the Continuous Ranked Probability Skill Score (CRPSS) - was formed by comparing to the CRPS calculated from the sample climatology. This benchmark was selected to be consistent with that used for the Brier Score, and the evaluation of longer lead-time river flow forecasts (not presented here). At best, the CRPSS takes a value of one and values less than zero indicate the forecast performs worse than the reference.

The *Rank Histogram* (Talagrand et al., 1997; Hamill, 2001) was used to assess the *reliability* of the ensemble: that is, whether or not the ensemble and observations have been drawn from the same distribution. A flat Rank Histogram suggests that the ensemble spread is an appropriate representation of the forecast uncertainty, whilst U- and domed-shaped Rank

Histograms indicate that the ensemble spread is too small and large overall respectively. An asymmetric Rank Histogram indicates that the ensemble is biased.

2.1.2 Evaluation of forecasts of binary events

The *Brier Skill Score* (BSS) was used to assess the probabilistic forecast skill. The BSS measures the proportional improvement in mean square *probability error* as defined by the Brier Score (Brier, 1950) with respect to a reference forecast in a standard manner (see Wilks, 2011). Here, the reference forecast is taken as the sample climatology of events occurring over the threshold of interest. In this context of binary events, the use of a sample climatology is preferred to a benchmark based on persistence. At best, BSS takes a value of one and values less than zero indicate the forecast performs worse than the reference.

The *Reliability or Attributes Diagram* (Wilks, 2011), plotting the forecast probability against the probability of the observation given the forecast, allows the *reliability* and *resolution* of the probability forecasts to be visually assessed. *Resolution* is an indication of how much the forecast deviates from the reference, with the forecast being more useful if resolution is larger (sharper, with smaller forecast spread), provided the ensemble is reliable. A forecast with no resolution has an observed relative frequency equal to the sample climatology, plotted as a horizontal line on the Reliability Diagram. In general it is considered good practice to include a *Sharpness Histogram* with the Reliability Diagram, showing the sample size for each probability bin. When this is the case it is referred to as an *Attributes Diagram*. Typically for a high threshold (precipitation or river flow), the low probability bins will be populated orders of magnitude more than the higher probability bins. This affects (and can skew) the interpretation.

The *Relative Operating Characteristic* (ROC) Diagram (see, for example, Jolliffe and Stephenson (2012)) plots, for a given threshold, paired values of Probability of Detection (POD) and False Alarm Rate (*F*) of ensemble forecasts for different probability of exceedance. The ROC Diagram measures the *discrimination* of the forecasts: the ability of the forecasts to distinguish between observed events and non-events. This diagram was also used to assess the *potential skill* of the ensemble: that is, the ensemble skill if forecast probabilities were well calibrated. A Skill Score based on the Area under the ROC curve, the ROC Skill Score (ROCSS), is defined as the Area Under the ROC Curve (AUC) normalised with reference to a random forecast with no skill (an AUC equal to 0.5). This reference was used to relate directly to the ROC Diagram. A ROCSS of one indicates a perfect forecast and if above zero the forecast has a skill better than a random forecast.

The economic benefit of a forecasting system depends on the cost-loss ratio of a particular user. To assess the *economic value* of forecasts, the *Relative Economic Value* (REV) statistic (Murphy, 1977; Richardson, 2000; Wilks, 2001; Zhu et al., 2002) was used. The REV is widely used in the verification of both hydrological and meteorological forecasts (e.g. Roulin, 2006; Magnusson et al., 2014) and uses information derived from a contingency table (e.g. Wilks, 2011) to calculate the economic value relative to a forecast based on climatological information. A cost-loss decision making model is assumed to define the cost for taking action (irrespective of whether or not the event occurs), and the loss incurred when the event occurs but no action was taken. The *REV Diagram* presents the REV for different cost-loss ratios. The REV has a maximum value of one for a perfect forecasting system, a value of zero for forecasts having the same value as climatological information only, and is negative for forecasts which have less value than using only climatological information.

2.2. Meteorological and hydrological models used

To focus on the considerations and strategies for joint river flow and precipitation ensemble verification, this paper restricts attention to the first 24 hours of river flow ensemble forecasts, and the corresponding precipitation ensembles. The precipitation ensemble consists of nowcasts from the Short-Term Ensemble Prediction System STEPS (Bowler et al., 2006), which merges a radar extrapolation nowcast with a spatially downscaled NWP forecast, and forecasts from the Met Office Global and Regional Ensemble Prediction System, MOGREPS (Hagelin et al., 2017; Bowler et al., 2008), run using the operational Met Office Unified Model (UM) (Davies et al., 2005; Tang et al., 2012). STEPS aims to account for uncertainty in the motion and evolution of radar-based precipitation fields; MOGREPS-UK aims to account for uncertainties in meteorological initial and boundary conditions and sub-grid processes in the NWP model. The river flow ensemble is generated using the G2G distributed hydrological model (Moore et al., 2006; Bell et al., 2009; Cole and Moore, 2009).

2.2.1. Best Medium Range precipitation ensemble

Best Medium Range (Best MR) ensemble forecasts of 15 minute precipitation accumulations (mm per 15 minutes) are produced with UK coverage, extending out to over 6 days and issued four times a day. These forecasts use the "best available" precipitation estimates. For the lead-time period considered in this study, the ensemble forecasts are a blend of the 2 km resolution STEPS extrapolation nowcasting system and the convection-permitting, 2.2 km MOGREPS-UK for the first ~7 hours (depending on forecast triggering, see below), and MOGREPS-UK beyond. For the operational forecasts considered here (December 2015), MOGREPS-UK extends out to 36 hours. To produce the Best MR product, all MOGREPS

forecasts are downscaled onto a fixed 2 km grid over the UK, the British National Grid, as used by STEPS.

To allow the latest forecast to be available to the FFC, the Best MR 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 vary by up to three hours. To simplify interpretation, only Best MR forecasts issued at 0100, 0700, 1300 and 1900 - each four hours after the associated MOGREPS-UK run - have been used. Forecasts issued at these times correspond to around 65% of the total number of forecasts issued. The operational Best MR forecasts are archived from 25 November 2015 to present. This archive reflects the biannual upgrades to the UM; past forecasts are not re-run for a new model version and no hindcast archives are available. The best use of such an archive is a key consideration for the operational implementation of an ensemble verification system of flood events, and is discussed further throughout this paper.

2.2.2. River flow ensemble forecasts using the G2G distributed hydrological model

G2G is a physical-conceptual distributed hydrological model developed by the Centre for Ecology & Hydrology (CEH) to forecast river flow and surface water flooding (Moore et al., 2006; Bell et al., 2009; Cole and Moore, 2009). G2G takes account of the effects on grid-cell runoff production of land-cover and soil/geology properties, along with antecedent wetness conditions. With water flows routed from cell to cell, G2G is formulated to represent spatial variability in river flow response to precipitation across a landscape with catchment, river basin and countrywide coverage. G2G can make full use of spatially-distributed precipitation data derived from observation networks of weather radars and raingauges, as well as precipitation forecasts from nowcasts, NWP models and their combination.

G2G is in operational use as a countrywide flood forecasting system by both the FFC over England and Wales (Price et al., 2012) and by the Scottish Flood Forecasting Service (SFFS) across Scotland (Cranston et al., 2012). Five-day outlook forecasts from G2G are used in preparing the Flood Guidance Statements issued by these operational bodies. For this study, the operational G2G configuration on a 1 km grid and for a 15 minute time-step is employed (Price et al., 2012), with the period January to March 2008 used for calibration. A raingauge-based precipitation truth (see Section 2.3.1) is used in the calibration of G2G, and is also used to obtain the initial conditions for each G2G forecast. Spatial datasets (e.g. terrain, soil/geology, and land-cover) are used to support its configuration and parameterisation, lessening the need for extensive calibration. Data assimilation of river flow observations helps to maintain realistic model states, which are then used to initialise forecasts of river flow. Flow-insertion is applied to correct flows to those observed at gauged river locations (and thereby improve the flows propagated downstream). A conservative form of empirical state-updating uses the observed flow to gradually adjust the G2G water storage upstream of gauged river locations.

For this work, the river flow ensemble forecasts were re-run using the operational G2G configuration and the 15-minute accumulation Best MR ensemble rainfall forecasts as input. Instantaneous river flows (m^3s^{-1}) were output every 15 minutes for the 898 river gauging station locations used operationally in G2G over England & Wales. Observed river flows are available from the Environment Agency and Natural Resources Wales for these sites (Section 2.3.2). No additional uncertainties are incorporated in the river flow ensemble river flow ensemble: the ensemble only accounts for uncertainty in the input precipitation.

2.3. Verification data sources

2.3.1. Precipitation

To consider the effect of observation uncertainties on forecast verification, two precipitation truth types are used: raingauge-based and radar-based. The raingauge-based precipitation truth uses data from the raingauge network across England and Wales operated by the Environment Agency (EA) and Natural Resources Wales (NRW). A gridded 1 km raingauge-based truth is then calculated by fitting a multiquadric surface with zero offset to the point raingauge observations accumulated to a 15 minute interval (Moore et al., 1994; Cole and Moore, 2008). Raingauge data have been quality-controlled at CEH using the methods presented in Howard et al. (2012). The radar-based precipitation truth is generated using the Met Office RadarNet system (Harrison et al., 2012) which combines 5 minute scan data from individual radars and includes data quality-control. Radar data processing includes a raingauge-based mean-field adjustment (constant over the domain of a single radar) that uses data from the Met Office Raingauge network, applied over a time period dependent on the number of recent raingauge-radar pairs available. As the radar rainfall data does not relate directly to the 15-minute raingauge data used for the raingauge-composite precipitation truth, these datasets are considered suitably independent observation sources for verification.

2.3.2. River flow

Data on river flow at 15 minute intervals were obtained for the 898 EA and NRW river gauging stations used operationally in G2G over England & Wales. Of these catchments around half are less than 100 km^2, with 75% less than 250 km^2 and 90% less than 700 km^2. The catchment response times range from less than an hour to a few days. The river flow data have been quality-controlled at CEH including visual checks to identify periods of

erroneous data. Consideration of uncertainties in the river flow data in a forecast verification context was beyond the scope of the current study, but is recognised as an important avenue of future investigation.

2.4. Ensemble verification period

Ensemble verification was undertaken over the 32-day period from 25 November to 26 December 2015, hereafter referred to as the study period. Throughout this paper, the term *sample climatology* refers to average values calculated over this period. This winter period was very wet, as revealed by the December 2015 precipitation anomaly from the 1981-2010 average which exceeded 200% for much of, and 300% for parts of, Wales and Northern England (Fig. 3 of McCarthy et al. (2016)). Many record-breaking precipitation totals were seen over this period including the highest 24-hour total ever recorded, and the second highest rain-day rainfall reliably recorded in the British Isles (Burt, 2016). Flooding associated with three Met Office and Met Éireann named storms occurred over this period: 5 to 6 December from Storm Desmond, 24 to 26 December after Storm Eva, and 29 to 30 December from Storm Frank. Precipitation from these storms fell onto already saturated ground following three storms the previous month, resulting in very high river flows. Flow records for nine catchments held in the National River Flow Archive set new data-era peak flows during this period, a number of flow events were assessed to have return periods greater than 100 years , and widespread flood damage was suffered across large parts of Britain (Barker et al., 2016; Marsh et al., 2016).

3. Approach to verification

To obtain an overview of ensemble performance, national and regional scales must be considered alongside the performance for individual catchments. This paper focusses on ensemble verification over England and Wales, with national-scale verification undertaken using data for all 898 catchments for which river flows are used operationally in G2G. Eight catchment groups (defined based on aggregated river drainage basins aligned to Wales, and Environment Agency regions over England) are used to verify forecasts at the regional-scale. These catchment groups are shown in Fig. 1.

< Figure 1 here please >

To facilitate a joint verification of river flow and precipitation, it is necessary to match, as closely as possible, the precipitation and the hydrological response. To this end, all precipitation verification reported here employs catchment-average precipitation, calculated from the gridded precipitation ensemble output. This differs from the conventional meteorological, and currently operational, verification of precipitation at either individual observation locations or using a gridded radar product (e.g. Mittermaier et al., 2013; Mittermaier, 2014; Mittermaier and Csima, 2017), and is essential for meaningful hydrological comparison. 0ne of the primary challenges for joint precipitation and river flow verification is how best to link precipitation events with the corresponding river-flow events, in terms of both magnitude and time-scale. In the sections that follow, methods of choosing accumulation periods and thresholds for precipitation and river flow are discussed which go some way towards addressing this challenge, with the aim of developing a meaningful joint verification framework in an operational flood forecasting context. These choices are then discussed further, and put into context in Section 4 when the verification framework is demonstrated.

3.1. Thresholds

To obtain ensemble verification results that are useful and relevant in an operational flood forecasting context, it is necessary to define river flow thresholds which select the flooding events of interest, and precipitation thresholds which select relevant precipitation values. Using these thresholds the observed time-series of river flow and precipitation are converted to binary time-series (where a value of one indicates an *observed event*) and time-series of ensemble forecast probabilities (whose values indicate the *forecast probability* of an event occurring). These binary and forecast probability time-series are then used to calculate the threshold-based metrics and diagrams described in Section 2.1: namely the BSS, Attributes Diagram, ROC Diagram, ROCSS and REV Diagram.

Traditionally, thresholds for forecast verification are treated differently between hydrological and meteorological communities. It could be argued that only precipitation that leads directly to a flood response in the river flow is of interest. Unfortunately this view is too simplistic as, for example, the same precipitation totals over the same catchment may not lead to the same river flow response. The river flow response is determined by multiple factors which interact non-linearly and non-systematically. Thus it is necessary to consider separately the calculation of precipitation and river flow thresholds. In this paper we consider two methods of calculating precipitation thresholds, and one method for calculating river flow thresholds.

For the verification of precipitation, thresholds are usually specified as either fixed values (e.g. 4 mm h^{-1}) or as percentiles of the total precipitation in the verification domain, for example over England and Wales, at a specific time. Thus, in this method, *percentiles of the spatial precipitation distribution* are used as precipitation thresholds (hereafter referred to

as "*spatial percentile thresholds"*), with a new threshold calculated every time the forecasts and observations are compared (i.e. the *spatial percentile thresholds* vary temporally). This method for precipitation is discussed further in Section 3.1.2. An alternative method of calculating precipitation thresholds, appropriate at the catchment scale, takes *percentiles of the temporal distribution* of catchment-average precipitation values for each catchment. Thus, this second method uses percentiles of the *sample climatology for each catchment* as time-invariant thresholds that vary from catchment to catchment. This thresholding methodology, denoted "*temporal percentile thresholds*" is discussed in Section 3.1.3. The method used for defining river flow thresholds based on return-periods is discussed first in Section 3.1.1.

In addition to the need for different methods of calculating thresholds for river flow and precipitation, it is also necessary to consider how best to apply those thresholds. For precipitation, where high (but not necessarily increasing) values are relevant for hydrological response, it is appropriate to consider threshold exceedance: any precipitation values over the threshold are considered to be "events". However, for river flow flood forecasting interest is focussed on the rising flows: the start of a potential flood event. To focus on these times, *upward threshold crossings* (i.e. the point at which rising river flows first cross a given threshold value) are used in this paper to define hydrological events. This is discussed further in Section 3.1.1.

3.1.1. Return-period based thresholds for river flow

Here, a hydrological threshold is selected as the river flow corresponding to a specific return period, where a return period of *n* years means that there is, on average, a 1 in *n* chance of a flood of at least that magnitude occurring in one particular year. Thus, higher return

periods correspond to more extreme events and higher flow thresholds. Return period threshold values from the Flood Estimate Handbook (FEH, Institute of Hydrology (1999)) were scaled to match the G2G median flood (equal to *Q*(2), the flow *Q* of return period 2 years) calculated over the water years 2007 to 2015 as done operationally for G2G forecasts.

It is noted that the median flood has a close association with the bankfull discharge for natural rivers. For flood guidance purposes, the FFC use a 1km grid of Q(T) values, for a range of T values (return period in years), as a nationally consistent indicator of flood severity when referenced against the G2G flows. This approach complements flood thresholds for specific sites associated with actual flooding and used with local models.

In this study, a forecast hydrological event is defined for each ensemble member as at least one upward crossing of a threshold (*upward threshold crossing*) occurring anywhere within the forecast lead-time range of 0 to 24 hours. Upward threshold crossings are calculated from the instantaneous river flow data (m^3s^{-1}) at two consecutive time-steps (i.e. separated by 15 minutes). Note that timing uncertainties up to 24 hours in the river flow forecasts are tolerated in this analysis. For flood guidance purposes, there is interest in a given threshold being crossed at any time within the next 24 hours. This approach is preferred to verifying 24-hour precipitation accumulations as it retains the shape and magnitude of the 15 minute time-step flood hydrograph. Accommodating timing uncertainties can be particularly important in this analysis where instantaneous river flows are verified (Section 2.2.2).

Hydrological forecast probabilities are calculated by taking the average number of events across the ensemble. An observed hydrological event is defined as an upward threshold crossing occurring anywhere within the corresponding 24 hour observation period. An

upward threshold crossing is used in preference to a threshold exceedance for river flow to focus on the start of a potential flooding event, and accommodating timing errors within a 24-hour period. Using a threshold exceedance would be particularly inappropriate for less extreme thresholds which may be exceeded for a number of days. In an operational verification system, several verification thresholds would be used, spanning a range of return-periods. This would allow the ensemble performance at different points in the flood hydrograph to be monitored. For meaningful verification statistics it is necessary to consider a large number of events: this becomes difficult for thresholds of high return-period. For this reason, the focus here is on one return-period threshold: ½Q(2). This corresponds to half the river flow of the bankfull level (which can be related to the Q(2) threshold), and is considered to be the minimum threshold of interest in a flood-forecasting context. It will be shown in Section 4 that, even for this low threshold, sampling uncertainties impact on the verification results.

3.1.2. Spatial percentile thresholds for precipitation

As discussed above, *spatial percentile precipitation thresholds* are calculated from the spatial distribution of all precipitation in the verification area at a particular time. For this study where catchment average precipitation values are used, the spatial distribution is formed of the catchment-average precipitation values of all catchments in the verification area. Thus, the spatial percentile thresholds focus on the tail of the precipitation distribution for every forecast. Of course percentile-based results are dominated by the use of more modest precipitation accumulation thresholds, but at least they allow for higher thresholds to be included when they do occur (e.g. Mittermaier et al., 2013). Spatial precipitation thresholds also allow for a consistent interpretation across the different catchments within

the verification area. By definition spatial percentile thresholds only select a small number of events leading to large sampling uncertainties. However, this compromise is necessary to focus on the catchment-average precipitation relevant in a hydrological context. For the study period used here, the 95th percentile was found to be a good compromise given these considerations, and is used for all spatial percentile precipitation threshold results. Thus, after application of this threshold to a particular ensemble member precipitation forecast, or to the precipitation observations, 5% of the catchments considered will be denoted as having an event (and allocated the value one) and the remaining catchments will have no event (allocated a value of zero).

In this paper, the group of catchments used to calculate the spatial percentile thresholds depends on the scale of the analysis being conducted. For the national-scale analysis, all 898 catchments in England & Wales are used; for the regional analysis, only catchments within the region of interest are used. Thus for a particular time, the regional analysis will use a different threshold value for each region in Fig. 1. Note that these regional-analysis spatial percentile threshold values will also differ from that used for the national scale analysis at this time. For analysis at the catchment scale, spatial percentiles thresholds calculated from the regional analysis are applied. Fig. 2 shows how the 95th percentile threshold regional analysis values relate to catchment-average precipitation values for two example regions. The North West region of England (predominantly upland) shows very heavy precipitation at the 95th percentile, up to 10 mm h^{-1} with four days having over 100 mm of precipitation. This highlights the unusual nature of the selected verification period, and why this period is relevant for understanding model performance in a flood warning context. Sample size tends to restrict the computation of verification statistics for fixed thresholds exceeding 4 mm h^{-1} unless computed over a very long time period. The other example region in Fig. 2,

Anglian in lowland eastern England, shows much lower precipitation thresholds, and is much more representative of conditions more generally.

< Figure 2 here please >

3.1.3. Temporal percentile thresholds for precipitation

As discussed above, *temporal percentile thresholds* are calculated separately for each catchment by taking percentiles of the temporal distribution of catchment-average precipitation values. Thus, unlike the spatial percentile threshold approach discussed in Section 3.1.1, results using temporal percentile thresholds cannot be consistently compared across different geographical areas. In this study the use of temporal percentile precipitation thresholds is limited to that at the catchment scale. Of course, with knowledge of the precipitation values corresponding to each catchment-threshold, useful insight can still be gained when comparing individual catchment performance. Maps of the time-invariant temporal percentile threshold values, calculated from the full 32-day study period (25 November to 26 December 2015 – see Section 2.4) are shown in Fig. 3. Consistent with the spatial percentile thresholds, the 95th percentile value is used here. Maps are shown for both the raingauge and radar data (centre and right) and also for an example ensemble member (other members lead to similar conclusions). The spatial distribution of thresholds is similar for all three. Agreeing with the results of Fig.2, Fig. 3 shows lower precipitation thresholds to the southeast, and higher precipitation values in Wales and the northwest of England. However, the most extreme values shown in Fig. 2 are lost from the analysis when using these sample climatology thresholds, with maximum totals of 7.9 mm h^{-1} and 129 mm d^{-1} occurring for the radar data (the values for raingauge and ensemble precipitation

are slightly lower). Thus, the chances of capturing and evaluating the characteristics of the very highest precipitation totals (isolated in time, and localised in space) are reduced.

< Figure 3 here please >

3.2. Accumulation periods

From a NWP perspective it is desirable to consider different precipitation accumulation periods, as longer accumulation periods have higher forecast skill (Duc et al., 2013). The intensity-duration relationship is highly non-linear. Longer accumulation periods do not necessarily imply that it rained for longer periods, but larger time windows have the ability of blurring or mitigating against the impact of timing errors. The precise definition of the accumulation window can be important though, as from a hydrological perspective, it can affect the time-delay in any catchment flow response to precipitation. For example, the river flow at the outlet of a large slow-response catchment will be linked to precipitation falling further in the past than for a small rapid-response catchment. Here, both 24-hour and one-hour (daily and hourly) precipitation accumulations are considered for verification for all metrics considered (i.e. both threshold-based and non-threshold-based). Thus, for each forecast origin, there is one comparison of the forecasts and observations when considering a 24-hour accumulation (with units mm d^{-1}), or 24 comparisons when considering one-hour accumulations (with units mm h^{-1}).

Of course it is still desirable to consider directly the precipitation at the temporal resolution that is used as input to the G2G model. Although this like-for-like correspondence between river flow and precipitation time-interval is desirable, it is considered more important that the precipitation verification should be appropriate to catchment response, and meaningful

in a flood forecasting perspective. Future work will extend this study to consider 15 minute precipitation accumulations. This is a very stringent timing test for the forecasts. Even for 24-hour accumulation precipitation forecasts, any skill will likely be attributable to a lack of timing errors as it is expected that timing errors will be the dominant error source.

As discussed in Section 3.1, an event for river flow threshold-based verification metrics is defined for each ensemble member as at least one upward threshold crossing occurring anywhere within the forecast lead-time range of 0 to 24 hours. Thus, although the threshold crossings are evaluated using the 15-minute river flow data (units m^3s^{-1}), the consideration of timing uncertainties is comparable to that of the analysis of daily precipitation accumulations. For the calculation of non-threshold-based metrics (e.g. CRPSS, Rank Histogram) two methods are applied to account for timing uncertainty in the river flows: firstly taking the mean flow over the 24-hour period (units $m^3 s^{-1}$), and secondly taking the maximum value over the 24-hour period (units $m^3 s^{-1}$). The former evaluates the 24-hour river flow volume and the latter focusses on the highest points in the hydrograph, of interest in a flood forecasting context. Additionally, for completeness, the 15 minute river flows are evaluated directly.

3.3. Summary of verification approach

When comparing any of the methods for precipitation forecast verification with those used for the hydrological ensemble forecasts, there are some key differences deserving of further discussion. In particular, the following points are noted.

- For the hydrological forecasts, *upward threshold crossings* are used as this is what is of operational interest for flood guidance and warning systems. For precipitation

forecasts, threshold *exceedance* is used instead, as the hydrological response is determined by high (but not necessarily rising) precipitation values.

- For hydrological forecasts, events are defined when a threshold is crossed *at any 15-minute time-step* in the forecast period of interest (here 24 hours) whereas, for precipitation, *each accumulation period is treated separately*. Thus, in terms of the time-period considered, the performance of the daily precipitation accumulation ensemble links more directly with the river flow ensemble performance. However, the performance of the hourly precipitation accumulation ensemble relates more directly to the catchment runoff response, as the hydrological ensemble is driven by 15 minute precipitation accumulations and run at a 15 minute time-step.
- For hydrological forecasts *return period thresholds* are used to select flooding events of interest. For precipitation two methods of using percentile thresholds are used: *spatial percentile thresholds* (calculated from the spatial distribution of catchment-average precipitation values, varying in time) and *temporal percentile thresholds* (time-invariant and calculated separately from the temporal distribution of precipitation values for each catchment).
- For metrics calculated using the full ensemble distribution (not threshold, e.g. CRPSS, Rank Histogram) both *daily and hourly precipitation accumulations* are evaluated. For these metrics the *15 minute river flow data* are evaluated alongside the *daily mean and daily maximum* river flows. This allows the effect of timing uncertainties to be investigated, and links made between these metrics and those using thresholds.

4. Demonstration of verification framework: results for example period

4.1. Overall analyses

To give an overview of ensemble performance, forecasts from all catchments in England & Wales are first considered together for the calculation of verification statistics and associated diagrams. Fig. 4 shows the Reliability, ROC and REV diagrams (with bootstrap confidence intervals at the 75th, 90th and 99th percentile in grey shading) and Rank Histograms for this overall verification.

< Figure 4 here please >

Considering the Reliability Diagrams (and associated Sharpness Histograms showing the sample size for each probability bin), it can be seen that the river flow ensemble is over-forecasting (the probabilities are too high) and also over-confident (larger probabilities are more over-forecast). This is also seen for the hourly precipitation accumulation ensemble, suggesting that the over-confidence in the input precipitation ensemble is contributing to the over-confidence in the river flow ensemble. In contrast, the daily precipitation accumulation ensemble shows good reliability for forecast probabilities up to 0.8. By considering daily accumulations the effects of timing errors are reduced: this gives an upper band on the ensemble performance. For probabilities above 0.8, both the hourly and daily precipitation accumulation ensembles show an increased over-confidence. Thus, the raingauge-based precipitation used as truth for Fig. 4 is not capturing the highest precipitation values as frequently as they are forecast, possibly due to the extreme precipitation values not occurring at raingauge locations (signalling observation error in the form of raingauge representativity).

All three Sharpness Histograms show a higher forecast relative frequency for low forecast probabilities as expected: there is generally a low chance of the threshold being crossed

(river flow) or exceeded (precipitation). For river flow, there is also a slight increase in the forecast relative frequency for the highest Sharpness Histogram bin. Thus at times when it is likely that a threshold will be crossed, it is more-common for the majority of ensemble members to predict this event than for the ensemble to be split between members that do and do-not capture the event. It is possible that, for the short and abnormally wet study period considered here (Section 2.4), the river flow Sharpness Histogram is influenced by flooding events with flows rising much higher than the ½Q(2) threshold. Note that, as the Reliability Diagram shows these high probability forecasts to be forecast too frequently (the ensemble is over-confident and over-spread) the observed increase in sharpness does not lead to higher forecast accuracy.

The effects of sampling uncertainty can be seen in both the river flow and the daily precipitation accumulation results, with the larger bootstrap uncertainties seen for 24h accumulations. This difference is thought to be due to the different approaches to thresholding the river flow and precipitation accumulations. In particular, the choice of absolute-value river flow thresholds indicating possible flood events results in a much smaller sample of river flow threshold-crossings and higher sampling uncertainties. It may perhaps be surprising to see that the daily precipitation confidence intervals are wider than those for hourly precipitation. This is primarily because daily precipitation accumulations can span a wider range of values (mm d^{-1}); the range of hourly precipitation accumulation values (mm h^{-1}) is generally much less (the exception being precipitation that leads to flash floods on short time-scales). The larger sample size available for analysing hourly precipitation accumulations will also contribute to the narrower uncertainty bands. Like those for river flow, precipitation probabilities tend to be over-confident especially for larger probabilities.

The ROC Diagrams indicate high potential skill for both the river flow and precipitation ensembles. Agreeing with the Reliability Diagrams, the highest potential skill is seen for the daily precipitation accumulations. The river flow ensemble shows higher potential skill than the hourly precipitation accumulation ensemble, suggesting that re-calibration of the river flow forecast probabilities in particular could lead to improved performance. For all three ensembles, the REV Diagrams show positive REV over a range of different cost-loss ratios, with higher probability thresholds having lower REV values, but over a larger range of cost-loss ratios. Comparing the river flow and precipitation REV Diagrams it is seen that, overall, the river flow ensemble has a narrower envelope of cost-loss ratios with positive REV than for precipitation. The ½*Q*(2) threshold river flow ensemble forecasts have comparable economic value to the daily 95th percentile threshold precipitation forecasts, though the daily precipitation forecasts show somewhat higher REV for high cost-loss ratios. The hourly precipitation forecasts show a smaller envelope of positive REV with a lower peak and for a smaller range of cost-loss ratios. Interestingly this difference was not seen when comparing the ROC curves: it occurs only when considering the cost-loss ratio. The lower potential skill (indicated by the ROC Diagram) and lower REV for hourly accumulations is tied to the spatial constraints applied in this analysis, where the precipitation is expected to occur in the right place (catchment) at the right time. Even though the precipitation forecast is an ensemble, any mismatches in space and/or time are accentuated for shorter accumulation periods. For higher precipitation thresholds (not shown) the REV curves are more similar to those for the ½*Q*(2) river flow threshold, suggesting that these differences may also be due to differences in the thresholding methods. This highlights the need for a thorough understanding of both river flow and precipitation verification methods for a meaningful joint verification.

Thresholding differences do not exist for the Rank Histograms, which are calculated from the full ensemble distribution. Rank Histograms show larger differences between river flow and precipitation ensemble performance. The river flow ensemble Rank Histograms show the observations falling in the lowest bin (the ensemble over-predicting river flow) over 40% of the time, and into the highest bin (the ensemble under-predicting river flow) around 30% of the time when instantaneous 15-minute river flows are used. Similar results are obtained using the daily maximum and daily mean instantaneous river flows (shown by the hashed bars in Fig. 4), although the relative population of the highest and lowest bins changes slightly. This suggests that timing uncertainties from the use of instantaneous river flows are not causing the strong under-dispersion seen in the river flow Rank Histograms. Instead the under-dispersion is thought to relate to several different factors. As discussed in Section 2.2.2, the river flow ensemble only takes account of rainfall uncertainty. Other forms of uncertainty, such as model uncertainty in representing the hydrological processes, may be important to accurately capture the ensemble dispersion (e.g. Brown et al. (2014b)). It is possible that the unusual nature of the verification period (Section 2.4) also acts to highlight the river flow ensemble under-dispersion. Additionally, although the ensemble is not reproducing the range of observed values and so is, in an overall sense, under-spread, a similar effect could also be caused by conditional biases in the ensemble forecasts. For example, if an ensemble had a high bias for half of the forecasts evaluated, and a low bias for the other half, the Rank Histogram from all evaluated forecasts would show higher populations for both the lowest bin (from the high-biased forecasts) and the highest bin (from the low-biased forecasts). It is interesting that the ensemble under- and over-predicts the flow values, given that the forecast *probabilities* were seen from the Reliability Diagrams to be overestimated only. This suggests that, although the high flows (irrespective of

absolute magnitude) are overestimated by the ensemble (giving an over-confidence in predicting threshold crossings), the low flows are underestimated: that is, the ensemble flows are too "peaky". Of course, the overall quality of the hydrological simulation also impacts the ensemble performance. As a physically-based model which conserves water balance, G2G does not contain a bias correction term, and the ensemble forecasts for some sites will have high/low bias. This would also show in the Rank Histograms as a conditional bias, contributing to a U-shaped Rank Histogram. Future work will investigate the use of post-processing to bias-correct the river flow ensemble members. Even when river flow data assimilation is used, and the input rainfall ensemble is reliable, recent studies (e.g. Bourgin et al. (2014)) have shown that post-processing is needed to obtain reliable river flow ensembles.

Although more uniform than the river flow Rank Histograms, those for precipitation also show observations falling too frequently in the ensemble extremities: at the high-end of the ensemble for hourly accumulations, and at the low-end of the ensemble for daily accumulations. However, these differences in precipitation Rank Histograms were found to be highly sensitive to the precipitation observation type, and should hence be treated with caution. Rank Histograms in particular are known for being sensitive to observation uncertainty (e.g. Hamill (2001)), but other diagnostics which can be related to the distribution can also be affected.

Fig. 5 shows the equivalent precipitation ensemble verification diagrams as Fig. 4 but using a radar-based precipitation truth. As the river flow results are not evaluated for different precipitation truths, they are unchanged from Fig. 4 and are not repeated in Fig. 5 for brevity of presentation. Overall, the radar-based Reliability Diagrams give a similar message

to those with a raingauge-truth: the ensembles are over-forecasting and overconfident. However, there are differences, particularly for the daily precipitation accumulations which show much poorer reliability when a radar-based truth is used. For a radar-based truth, performance is similar across the full range of probabilities. For both truth types, the relationship between hourly and daily precipitation Reliability Diagrams is similar. Only subtle differences are seen in the ROC and REV. The radar-based daily precipitation accumulation Rank Histograms in Fig. 5 suggest the smallest accumulations occur more frequently compared to the raingauge-based Rank Histogram in Fig. 4, though both are suggesting a dry bias and insufficient spread. For the hourly accumulations the shape of the Rank Histogram changes more dramatically, looking fairly well spread based on the radar-rainfall accumulations in Fig. 5, whilst for gauge-rainfall in Fig. 4 it shows that observations fall in the largest accumulation bin more frequently. That is, the same forecast against a different observation has a wet bias, under-forecasting lighter precipitation according to the raingauge observations. This suggests an interesting dynamic between hourly and daily precipitation. The differences may simply be due to the temporal granularity, as an hourly accumulation may not be a good fit when it comes to defining events, whereas on the daily time-scale, events are generally less susceptible to timing errors, unless it is a very long duration event: that is, less likely to straddle adjacent time periods with detrimental impact. These results highlight both the benefit of considering multiple verification diagrams, and also the importance of considering observation uncertainty.

< Figure 5 here please >

4.2. Regional-scale verification

The ensemble performance is found to vary considerably at the regional scale. Fig. 6 shows Reliability and ROC diagrams stratified by region for river flow and hourly precipitation accumulation, with shading showing bootstrap confidence intervals at the 99th percentile. Daily precipitation results were found to be similar to those for hourly accumulations and are not included here. Overall, higher reliability is seen for precipitation than for river flow. Although it is expected that 15 minute precipitation accumulations (not considered here) would have lower reliability than daily or hourly ones, it is not thought that this would fully account for these differences. For river flow, there is more regional variation in reliability than for precipitation, with regions to the south east of the country having much lower reliability when considering river flow. This is partly explained by the large sampling uncertainties seen for these regions, as shown by the bootstrap confidence intervals. For these regions there are so few river flow threshold crossings, even for the ½Q(2) threshold, that the range of possible reliability goes from very poor to good. Hydrological differences between the regions are also thought to contribute to the greater regional variation for the river flow ensemble performance. Comparing the individual region performance, there is not a direct correspondence between the river flow and precipitation reliability. For example, the North East and North West of England regions perform best for river flow, but Midlands and Wales perform best for precipitation. Anglian performs worst for both river flow and precipitation. These differences highlight again the dependence of the river flow ensemble behaviour on hydrological processes controlling runoff production, water storage and translation. The river flow ensemble is not a simple transformation of the rainfall ensemble members, and the river flow ensemble performance cannot be directly estimated in a simple manner from that of the rainfall ensemble. The ROC Diagrams stratified by region show higher potential skill for river flow than for hourly rainfall accumulations across

the majority of regions, agreeing with the national-scale analyses (Section 4.1). Exceptions to this are the Thames and Anglian regions where the river flow analyses are dominated by sampling uncertainties with rivers being generally less responsive to precipitation. A similar regional dependence is seen in the ROC and Reliability diagrams.

< Figure 6 here please >

4.3. Verification at the catchment scale: the CRPSS

As fewer catchments are used to calculate threshold-based verification statistics, the effective sample size decreases and sampling uncertainties increase. In this section, the spatial distribution of the CRPSS scores calculated for individual sites are considered. As the CRPSS is calculated from the full ensemble distribution, it is less targeted to the flood-forecasting context than the threshold-based metrics. However, a brief analysis of the CRPSS serves as useful complement to the threshold-based scores, through giving a better understanding of the ensemble performance as a whole. Additionally, as the CRPSS is calculated from the full ensemble distribution, for all forecasts in the verification period, the CRPSS is less influenced by sampling issues and the catchment-scale performance can be better-evaluated. This is important in an operational context, where flooding cases may cover only a small number of catchments.

Fig. 7 shows maps of CRPSS for instantaneous 15-minute river flows, and for hourly and daily precipitation accumulations. Results for daily mean and daily maximum river flows are very similar to those for the instantaneous river flows, leading to the same discussion and conclusions, and are not included here for brevity of presentation. All CRPSS values were formed by comparing CRPS calculated from all ensemble forecasts in the sampling period to the CRPS calculated from the sample climatology (as defined in Section 2.1.1). For the

majority of catchments, the ensemble forecasts are more skilful than the sample climatology, with positive CRPSS values. There are a few exceptions for river flow, predominantly in the Midlands region (Fig. 1) and corresponding to catchments with unnatural flow regimes (artificial influences such as abstractions, discharges and reservoirs) which are not represented in detail in G2G.

< Figure 7 here please >

For river flow, little consistent variation is seen in CRPSS values across England & Wales. This suggests that the influence of non-location-specific catchment properties, such as catchment size, are influencing the CRPSS more than locally consistent catchment properties. In contrast, the precipitation CRPSS values show clear spatial variations which can be linked to the distribution of precipitation accumulations (e.g. poorer skill in the south and east of England where smaller accumulations were experienced during the winter period and there was less deviation from the sample-climatological values). Thus, the spatial variations seen in the precipitation CRPSS scores is related to the use of a reference based on the sample climatology. This relationship will be investigated further when the verification framework is applied to longer study periods, using longer climatological references. Overall, precipitation skill is spatially more coherent purely because the atmosphere is a continuum and inhomogeneous catchment properties (aside from relief) remain irrelevant until the precipitation has reached the ground. Overall, similar results are obtained from raingauge- and radar-based truths. Any differences in the northwest and southwest corners of both England and of Wales can be linked to the extent and quality of radar coverage in these areas.

Without differences in thresholding methods, a more-direct comparison can also be made between river flow and precipitation ensemble CRPSS values. The river flow CRPSS values are calculated from G2G modelled flows using 15 minute precipitation data as input. Ideally, these would be compared with 15 minute accumulation precipitation CRPSS values: however this was not possible in this study due to data processing constraints. Instead, the river flow CRPSS values are compared to those calculated from both precipitation hourly and daily accumulations. This gives an indication of the dependence of the precipitation CRPSS on the temporal resolution used, and allows for an informed comparison with the river flow verification.

From scatter plots of the CRPSS for river flow against raingauge-based precipitation shown in Fig. 8, it is seen that a smaller range of CRPSS values are obtained for precipitation than for river flow. Thus the river flow ensemble is being influenced by hydrological effects in addition to the precipitation uncertainty input through the precipitation ensemble. Moving from daily to hourly accumulations, the standard deviation of the precipitation CRPSS values decreases slightly. This narrowing of the range of CRPSS precipitation scores for the shorter accumulation period is initially somewhat unexpected, but is related to the CRPS being in the same units as the variable that is being verified. Daily precipitation totals have a larger range compared to hourly ones, giving a larger range of CRPSS magnitudes. As the CRPS calculated from the ensemble forecasts will vary more between catchments than the CRPS of the sample-climatology reference, this larger precipitation range for the daily accumulations will also result in a larger range of CRPSS values. The hourly scores are lower, which is expected, given that timing errors will have a larger impact at this temporal granularity. Also shown in Fig. 8 are the q-q (quantile-quantile) plots of the CRPSS for river flow against that for raingauge-based precipitation. By considering only quantiles of the

CRPSS distributions, attention is focused at the regional scale: the relationship between individual catchments is no longer preserved. The scores are presented in an ordered or ranked fashion and show the range of values for a region for both the precipitation and the river flow. The q-q plots show, particularly for higher CRPSS values (indicating higher skill), an almost linear relationship between the river flow and precipitation CRPSS values. For lower CRPSS values this relationship is less clear: a range of river flow CRPSS values are seen for a given precipitation CRPSS value. Similar results (not shown) are obtained when comparing against a radar-based precipitation truth.

< Figure 8 here please >

4.4. Pooling of river flow data by catchment size

Given the dependence of the river flow CRPSS on non-location-specific catchment properties (as shown in Fig. 7 and discussed above), the relationship between various catchment properties (e.g. catchment size, terrain slope and sub-catchment properties) and threshold-based verification scores was investigated. Fig. 9 shows river flow Reliability and ROC diagrams, calculated using all catchments in England and Wales pooled by catchment size (5 pooling groups each containing around 180 catchments). This catchment property was found to relate directly to the threshold-based verification scores. In particular, the Reliability Diagram shows a clear trend of reliability decreasing with decreasing catchment size. For forecast probabilities up to 0.4, these differences are larger than the 90^{th} percentile bootstrap sampling uncertainties. This trend agrees with that found in Alfieri et al. (2014). A similar trend is seen in the ROC Diagram for the four largest catchment size groups; however, the group of smallest catchments does not follow this trend and shows performance similar to the group of largest catchments. Similar conclusions were found

using more catchment groups: best performance was seen on the ROC Diagram for the smallest and largest catchments, with middle-sized catchments performing worse. The REV curves partitioned by catchment size (not shown) also lead to similar conclusions. This feature is unexpected (generally larger catchments are expected to perform better than smaller ones), and may be related to the unusual verification period considered here, with many cases of large-scale heavy precipitation and extreme flood events (Section 2.4). Future work using a longer verification period will aim to disentangle this.

< Figure 9 here please >

4.5. Verification at the catchment-scale: threshold-based scores

Sampling uncertainties at individual sites are expected to be large, particularly for river flows using flood-producing thresholds. Hence, in addition to considering threshold-based river flow results at individual sites, the relationship between catchment size and river flow ensemble performance (Section 4.4) is exploited by pooling the data from several catchments within a given geographic region. The aim is to exploit the clear relationship between river flow ensemble performance and catchment area shown in Fig. 9 to reduce the river flow sampling uncertainties for the calculation of threshold-based verification metrics at sub-regional scales. As similarly sized catchments have similar ensemble performance for river flow, data from similarly sized catchments can be sensibly combined to calculate verification statistics, thus reducing sampling uncertainty.

As the river flow ensemble takes precipitation input which varies coherently across geographic regions (as discussed in Section 4.3 with reference to the precipitation CRPSS maps), it is also important to retain some regional variability in the river flow verification statistics. Here, the aggregated river drainage basins (regions) shown in Fig. 1 are used. Use

of regions based on hydro-climate, or broad-scale landscape features, are other options for regional pooling that might be considered in future work. Within a given region, the catchments are ranked by catchment size. For each catchment within that region, data from a fixed number of other catchments with closely ranking areas are also used to calculate the verification scores. Thus, a moving catchment-size bin is used, centred upon the size of the catchment of interest. The width of this moving-bin is defined based on a fixed number of catchments – for example, 15 sites larger and 15 sites smaller than the catchment of interest – irrespective of the distribution of catchment sizes. This option is used here as it ensures the same amount of data is used for each calculation. Another option would be to fix the range of catchment sizes to include in the moving bin: for example, to consider all sites within 200 km of the catchment of interest. Due to the long-tailed distribution of catchment sizes, this would result in a large number of catchments falling within the moving catchment-bin for small catchments, and only one or two catchments falling within the moving catchment-bin for large catchments. Hence, this is not a sensible option to use.

The number of sites to use for each pool of data is a compromise between retaining all the catchment-scale information but not reducing sampling uncertainty at all, and losing all the catchment-scale information (i.e. evaluating at the regional scale) but reducing the sampling uncertainty. To investigate the effects of pool size, moving-bins consisting of 11, 21, 31, 41 and 61 sites were considered. By comparing the ROCSS and BSS values for all these options, it was found that large differences were seen when moving from a 1 to 11, 11 to 21 and 21 to 31 site pool. However, when increasing the pool size further, the BSS and ROCSS results converged and smaller differences were seen. Hence, a pool of 31 sites was considered appropriate. In Fig. 10 river flow BSS and ROCSS results are mapped both for individual sites, and using a moving catchment-size pool of 31 sites within each region. Note that for

precipitation accumulations, local properties dominate, and pooling by catchment size is not sensible. Instead, it is necessary to consider individual site scores in the context of the scores obtained for neighbouring sites.

< Figure 10 here please >

From Fig. 10 it can be seen that, when using individual site data only, it is not possible to calculate (marked by hatching) the ROCSS and BSS at a large number of sites, due to a small sample size. By applying an area-based pooling, the values are more consistent between neighbouring locations. Note that the pooling does not affect the overall interpretation of the maps at national and regional scales. For example, differences seen between the ROCSS and BSS maps are consistent between the no-pooling and pooling results in Fig. 10. This is also true when a different number of catchments (e.g. 11, 21) are used in the catchment-size pool (not shown for brevity of presentation). Of course, area-based pooling only has an effect if there are threshold crossings at some sites within a given area-pool. Hence, although pooling has been found to be a useful method of obtaining more meaningful information from a small sample of hydrological events, it is still reliant on having *some* hydrological events. In this study the BSS and ROCSS were calculated at all sites (or pools of sites) where any events occurred (i.e. when there was at least one event in the verification period). This method was used to retain the maximum possible amount of information, from as many sites as possible. Of course, it also results in some scores being calculated from a very small number of events, particularly in regions to the southeast of England. In future work with the benefit of a longer verification period, the use of a higher sample-size cut-off for scores calculation will be investigated.

Precipitation accumulation BSS and ROCSS results calculated using spatial percentile thresholds, for daily and hourly accumulations, are shown in Fig. 11. Both raingauge- and radar-based precipitation are considered as truth. Sampling effects are responsible for the lack of scores for many catchments, as indicated by the hatching. As discussed in Section 3.1.2, this resulted from the thresholds being calculated from the catchment-average values in a region such that, by definition, only the catchments with mean precipitation totals in the top 5% are verified. The exceptionally wet period meant that these thresholds can be locally very high. However, on a larger multi-catchment scale the precipitation lacked the spatial uniformity of intensity to exceed such thresholds over multiple catchments. The net result is that many catchments did not receive the required amount of precipitation to have a sufficient sample for verification. The BSS is overwhelmingly negative (skill worse than sample climatology) with a scattering of catchments with small positive skill scores. The ROCSS score shows more regions with skill, with higher levels of skill for the daily precipitation accumulations. Overall there are strong similarities in the signals from daily and hourly accumulations for Day 1, with the daily showing some additional skill. Results between radar- and gauge-rainfall accumulations are surprising similar, in terms of spatial patterns, with some notable exceptions, e.g. mid-Wales.

< Figure 11 here please >

BSS and ROCSS results using the second thresholding methodology for precipitation accumulations at the catchment-scale – using temporal percentile precipitation thresholds (Section 3.1.3) – are shown in Fig. 12. Here, as the number of threshold exceedances is fixed for each catchment, the effects of sampling uncertainty are less, and scores can be calculated for all catchments. Both the BSS and ROCSS are seen to vary smoothly across the country, with more skill (higher scores) seen to the northwest, and lower scores to the

southeast. Scores are much higher than those using spatial percentile thresholds (Fig. 11), reflecting the less extreme values of precipitation that are being verified. Thus, for lower precipitation thresholds, further from the tail of the precipitation distribution, the precipitation ensemble performs better. However, to understand the precipitation ensemble performance in a flood forecasting context it *is* necessary to focus on the tail of the precipitation distribution. Thus, despite the presence of sampling uncertainties, it is necessary to use other methods, such as the spatial percentile thresholds used in Fig. 11, in this context. Of course, to reduce sampling uncertainties, verification should be performed over a longer verification period. However, given the extreme nature of precipitation and river flow events of interest in a flood forecasting context, it is likely that sampling uncertainty, and methods of its reduction, will remain an important consideration in ensemble verification.

< Figure 12 here please >

For both river flow and precipitation (using either thresholding methodology), higher values of ROCSS than BSS are obtained suggesting that biases in the forecast probabilities are reducing the overall ensemble skill. It is therefore conceivable that some post-processing of the precipitation totals and probabilities could be of benefit. In general, the BSS values for river flow are more similar to those for precipitation when calculated with spatial percentile thresholds (for catchments where the scores could be calculated) than with sample-climatology based thresholds. This is expected as the spatial threshold method evaluates more directly situations where flooding is more likely to occur. For the ROCSS this is less clear, with the river flow ROCSS values generally sitting between those calculated for the two precipitation thresholding methodologies: higher than those for spatial percentile thresholds, but lower than for temporal percentile thresholds.

5. Discussion

Using a sample one month case-study period (December 2015), this paper has demonstrated important considerations for a joint verification framework for hydrological and meteorological ensemble forecasting systems. The sample represents a very short and atypical period: good for getting enhanced sampling of flood-producing rain but not generally representative. The focus has been on verification information relevant in an operational flood forecasting context. To gather meaningful statistics that are more representative of all possible scenarios the verification will need to be run over extended (operational) periods before overall conclusions on true skill can be drawn.

Due to the infrequency of flood events, sampling uncertainties are inherent in the verification of flood forecasting ensembles, for example as discussed by Cloke and Pappenberger (2009). By considering lower thresholds more robust statistics can be obtained: but such results may be of lesser practical relevance to those using flood-forecasting ensembles for operational decision-making. Individual case-studies of flooding events for specific catchments can help to provide some confidence in forecast performance; however, in a probabilistic sense, no meaningful evaluation of individual case-study performance is possible.

In this paper, a short verification period of 32 days was used, moving beyond the individual case-study approach. This very wet period was selected so as to contain a number of flood events, and thus to enable ensemble verification to be examined in a flood forecasting context. Of course, by definition, this means that it is not a “representative period”, and model performance over this period will not be indicative of the model performance in less-

extreme cases. Results were presented for the lowest threshold considered appropriate in a flood-forecasting context, equivalent to half the flow of the two-year flood event.

Methods of increasing the sampling size were discussed throughout this paper. In particular, the consideration of a larger number of catchments (grouped nationally and regionally), the use of percentile thresholds for precipitation, and the pooling of river flow results based on catchment size, were presented as possible methods of reducing sampling uncertainties to a manageable level. The performance of precipitation forecasts is very strongly dependent on the choice of thresholds used in the calculation of skill. Spatial percentile thresholds can often lead to skill metrics being dominated by non-extreme events. They pick up what happens in a particular accumulation period, which most of the time is not extreme. Based on the ROC and Reliability diagrams and skill scores used here, the skill and performance of the precipitation ensemble is good. However, as the exceptionally wet period used in this paper shows, when the sampling does fall in the tails of the distribution, the performance of the precipitation forecasts shows certain weaknesses. For the shorter accumulation periods the lack of skill could well be dominated by timing errors, which are not accounted for here. Therefore, various sources of sampling uncertainty will always feature as a key consideration when interpreting verification results from flood-forecasting ensemble systems. This is particularly true for operational NWP ensemble systems, such as the MOGREPS, where long-term hindcasts are not produced and weather models are in continuous development.

For forecast performance to be better understood, it is vital that the associated sampling uncertainties are fully and comprehensively conveyed to operational forecasters and decision-makers. Some form of post-processing of the precipitation accumulations (bias

correction) in the first instance, and potentially also a subsequent probability calibration, could add benefit. However, post-processing precipitation can be challenging due to the non-Gaussian nature of its distribution (e.g. Scheuerer and Hamill (2015), Ben Bouallègue (2013), Bentzien and Friedrichs (2012)).

For flood forecasting to benefit from a joint hydrological-meteorological ensemble verification, relevant and physically meaningful time-scales must be considered, and sources of uncertainty identified. To this end, precipitation results were presented using both hourly and daily accumulations to encompass the effects of catchments with different hydrological response-times. Each precipitation accumulation period was considered separately in the verification. For hydrological ensembles, to focus on hydrological threshold-crossings in a flood-warning context, hydrological events were defined when a river flow threshold was crossed at any 15-minute time-step in the 24-hour forecast period of interest. For precipitation, the two truth types – raingauge-based and radar-based – were used to quantify the effects of precipitation observation uncertainty on the verification results. The Rank Histogram was found to be particularly sensitive to this observation uncertainty. Reliability Diagrams also showed sensitivity to the precipitation truth-type, with a raingauge-truth suggesting high probabilities were particularly underestimated. Although not seen for the hourly precipitation results, daily accumulations appeared to be much more reliable when comparing to raingauge data. As only one set of river flow observations are available, consideration of river flow observation uncertainty is a more-involved process, beyond the scope of this current study but an important topic for future work.

For the short verification period used in this study, probabilistic forecasts derived from both the river flow and the precipitation accumulation ensembles tended to be over-confident,

with over-confidence increasing with forecast probability. This is related to the lack of reliability in the forecast probabilities. The river flow ensemble was found to be more severely under-spread than the precipitation accumulation ensemble according to the Rank Histogram. This suggests that unaccounted-for uncertainties in the hydrological modelling process (the river flow ensemble considers precipitation uncertainty only) may be important for forecast accuracy, in agreement with the conclusions of Brown et al. (2014b). Hydrological uncertainties will be considered in future work.

6. Conclusions

From the investigations and analyses presented above, the following key conclusions can be drawn.

- For the full evaluation of operational flood forecasting ensembles, it is necessary to consider both precipitation and river flow ensembles in a joined-up manner. Differences in the physical nature of precipitation and river flow require consideration, and lead to different verification solutions and interpretations. Examples have been presented of pooling river flow verification analyses (but *not* precipitation analyses) based on catchment properties, and using percentile thresholds when verifying precipitation accumulations.
- Different spatial scales (e.g. national, regional, and sub-regional where possible) should be considered to give informed and physically relevant information about the ensemble's performance. This study has highlighted the varying effects of sampling uncertainty at different scales, and the information that can be gained from a multi-scale analysis.

- To obtain a representative and unbiased view of ensemble performance, it is necessary to use a range of metrics for both river flow and precipitation verification. This is particularly important given the differing sensitivities of verification metrics to observation error. In this study, sensitivity to precipitation observation error was exemplified by the Rank Histogram.

These conclusions will form the basis of future work seeking an end-to-end ensemble verification framework relevant to operational flood forecasting ensembles. In particular, the results of this initial detailed case-study over a 32-day period will allow appropriate choices to be made when considering longer datasets, and different hydrological ensemble systems.

Acknowledgements

The authors thank the two anonymous reviewers, and Associate Editor Maria-Helena Ramos, for their detailed comments that helped improve the quality and clarity of this paper. The work reported on here formed part of the "Rainfall and River Flow Ensemble Verification" project commissioned by the Flood Forecasting Centre on behalf of the Environment Agency, Scottish Environment Protection Agency and Natural Resources Wales. Preparation of this paper was made possible through Centre for Ecology & Hydrology and Met Office science funding.

References

Addor, N., Jaun, S., Fundel, F., Zappa, M., 2011. An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. Hydrol. Earth System Sci. 15(7), 2327–2347.

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., 2014. Evaluation of ensemble streamflow predictions in Europe. J. Hydrol. 517, 913–922. https://doi.org/10.1016/j.jhydrol.2014.06.035.

Alfieri, L., Velasco, D., Thielen, J., 2011. Flash flood detection through a multi-stage probabilistic warning system for heavy precipitation events. Adv. Geosci. 29(1), 69–75. https://www.adv-geosci.net/29/69/2011/adgeo-29-69-2011.pdf.

Alfonson, L., Mukolwe, M.M., Di Baldassarre, G., 2016. Probabilistic flood maps to support decision-making: mapping the value of information. Water Resour. Res. 52, 1026 – 1043.

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., Reinhardt, T., 2011. Operational convective-scale numerical weather prediction with the COSMO model: description and sensitivities. Mon. Wea. Rev. 139(12), 3887–3905.

Barker, L., Hannaford, J., Muchan, K., Turner, S., Parry, S., 2016. The winter 2015/2016 floods in the UK: a hydrological appraisal. Weather 71(12), 324–333.

Bell, V.A., Kay, A.L., Jones, R.G., Moore, R.J., Reynard, N.S., 2009. Use of soil data in a grid-based hydrological model to estimate spatial variation in changing flood risk across the UK. J. Hydrol. 377(3–4), 335–350. https://doi.org/10.1016/j.jhydrol.2009.08.031.

Ben Bouallègue, Z., 2013. Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. Wea. Forecasting 28, 515–524. https://doi.org/10.1175/WAF-D-12-00062.1

Bentzien, S., Friederichs, P., 2012. Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. Wea. Forecasting 27, 988–1002. https://doi.org/10.1175/WAF-D-11-00101.1

Bourgin, F., Ramos, M.H., Thirel, G. Andréassian, V., 2014. Investigating the interactions between data assimilation and post-processing in hydrological ensemble forecasting. J. Hydrol. 519, 2775–2784. https://doi.org/10.1016/j.jhydrol.2014.07.054.

Bouttier, F., Vié, B, Nuissier, O., Raynaud, L., 2012. Impact of stochastic physics in a convection-permitting ensemble. Mon. Wea. Rev. 140(11), 3706–3721. https://doi.org/10.1175/MWR-D-12-00031.1

Bowler, N.E., Arribas, A., Mylne, M.R., Robertson, K.B., Beare, S.E., 2008. The MOGREPS short-range ensemble prediction system. Q. J. R. Meteorol. Soc. 134, 703–722.

Bowler ,N.E., Pierce, C.E., Seed, A.W., 2006. STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. Q. J. R. Meteorol. Soc. 132(620), 2127–2155.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review 78(1), 1–3.

Brown, J.D., Wu, L., He, M., Regonda, S., Lee, H., Seo, D-J., 2014a. Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification. J. Hydrol. 519, 2869–2889. https://doi.org/10.1016/j.jhydrol.2014.05.028.

Brown, J.D. He, M., Regonda, S., Wu, L., Lee, H., Seo, D-J., 2014b. Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic

Ensemble Forecast Service (HEFS): 2. Streamflow verification. J. Hydrol. 519, 2847–2868. https://doi.org/10.1016/j.jhydrol.2014.05.030.

Burt, S., 2016. New extreme monthly rainfall totals for the United Kingdom and Ireland: December 2015. Weather 71(12), 333–338.

Caseri, A., Javelle, P., Ramos, M.H., Leblois, E., 2016. Generating precipitation ensembles for flood alert and risk management. J. Flood Risk Manage. 9(4), 402–415.

Cecinati, F., Rico-Ramirez, M.A., Heuvelink, G.B.M., Han, D., 2017. Representing radar rainfall uncertainty with ensembles based on a time-variant geostatistical error modelling approach. J. Hydrol. 548, 391–405. https://doi.org/10.1016/j.jhydrol.2017.02.053.

Clark, P., Roberts, N., Lean, H., Ballard, S., Charlton-Perez., C., 2016. Convection-permitting models: a step-change in rainfall forecasting. Meteorol. Appl. 23, 165–181. https://doi.org/10.1002/met.1538.

Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: A review. J. Hydrol. 375(3–4), 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005.

Cole, S.J., Moore, R.J., 2008. Hydrological modelling using raingauge- and radar-based estimators of areal rainfall. J. Hydrol., 358, 159-181.

Cole, S.J., Moore, R.J., 2009. Distributed hydrological modelling using weather radar in gauged and ungauged basins. Adv. Water Res. 32, 1107-1120.

Cranston, M., Maxey, R., Tavendale, A., Buchanan, P., Motion, A., Cole, S., Robson, A., Moore, R.J., Minett, A. 2012. Countrywide flood forecasting in Scotland: challenges for hydrometeorological model uncertainty and prediction. In: Weather Radar and

Hydrology (ed. by R.J Moore, S.J. Cole, A.J. Illingworth) (Proc. Exeter Symp., April 2011), IAHS Publ. no. 351, 538-543. http://nora.nerc.ac.uk/19637/.

Cranston, M.D., Tavendale, A.C.W., 2012. Advances in operational flood forecasting in Scotland. Water Management 165(2), 79–87. https://doi.org/10.1680/wama.2012.165.2.79.

Davies, T., Cullen, M.J.P., Malcolm, A.J., Mawson, M.H., Staniforth, A.A, White, A.A., Wood, N., 2005. A new dynamical core for the Met Office's global and regional modelling of the atmosphere. Q. J. R. Meteorol. Soc. 131, 1759–1782.

Davolio, S., Miglietta, M.M., Diomede, T., Marsigli, C., Montani, A., 2013. A flood episode in northern Italy: multi-model and single-model mesoscale meteorological ensembles for hydrological predictions. Hydrol.Earth Syst. Sci. 17(6), 2107–2120.

Demargne, J. et al., 2014. The science of NOAA's operational hydrologic ensemble forecast service. Bull. Amer. Meteor. Soc. 95(1), 79–98.

Duc, L., Saito, K., Seko, H., 2013. Spatial-temporal fractions verification for high-resolution ensemble forecasts. *Tellus, Series A: Dynamic Meteorology and Oceanography*, 65(1).

Golding, B.W. et al., 2014. Forecasting capabilities for the London 2012 Olympics. Bull. Amer. Meteor. Soc. 95(6), 883–896.

Hagelin, S., Son, J., Swinbank, R, McCabe, A., Roberts, N., Tennant, W., 2017. The Met Office convective-scale ensemble, MOGREPS-UK. Q. J. R. Meteorol. Soc. 143, 2846–2861.

Hamill, T.M., 2001. Interpretation of rank histograms for verifying ensemble forecasts. Mon. Wea. Rev. 129(3), pp.550–560.

Hardy, J., Gourley, J.J, Kirstetter, P-E., Hong, Y, Kong, F., Flamig, Z.L., 2016. A method for probabilistic flash flood forecasting. J. Hydrol. 541, 480–494. https://doi.org/10.1016/j.jhydrol.2016.04.007.

Harrison, D.L., Norman, K., Pierce, C., Gaussiat, N., 2012. Radar products for hydrological applications in the UK. Water Management 165(2), 89–103.

He, X., Højberg, A.L., Jørgensen, F., Refsgaard, J.C., 2015. Assessing hydrological model predictive uncertainty using stochastically generated geological models. Hydrol. Process. 29(19), 4293–4311.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Wea. Forecasting 15(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Howard, P.J., Cole, S.J., Robson, A.J., Moore. R.J., 2012. Raingauge quality-control algorithms and the potential benefits for radar-based hydrological modelling. In: Weather Radar and Hydrology (ed. by R.J Moore, S.J. Cole, A.J. Illingworth) (Proc. Exeter Symp., April 2011), IAHS Publ. no. 351, 219-224.

Institute of Hydrology, 1999. Flood Estimation Handbook. 5 volume set. Available from Centre for Ecology & Hydrology, Wallingford, UK.

Jolliffe, I.T., Stephenson, D.B. 2012. Forecast Verification: A Practitioner's Guide in Atmospheric Science. 2nd Edition, John Wiley & Sons Ltd, Chichester, UK, 274pp.

Magnusson, L., Haiden, T., Richardson, D., 2014. Verification of extreme weather events: discrete predictands. ECMWF Tech. Memo. No. 731, . European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, UK, 27pp.

Marsh, T.J., Kirby, C., Muchan, K., Barker, L., Henderson, E., Hannaford, J., 2016. The winter floods of 2015 / 2016 in the UK - a review. Centre for Ecology & Hydrology, Wallingford, UK. 37 pp

Marty, R., Zin, I., Obled, C., 2013. Sensitivity of hydrological ensemble forecasts to different sources and temporal resolutions of probabilistic quantitative precipitation forecasts: Flash flood case studies in the Cévennes-Vivarais region (Southern France). *Hydrol. Proces.* 27(1), 33–44.

McCarthy, M. Spillane, S., Walsh, S., Kenton, M., 2016. The meteorology of the exceptional winter of 2015/2016 across the UK and Ireland. Weather 71(12), 305–313.

Mittermaier, M.P., 2014. A strategy for verifying near-convection-resolving model forecasts at observing sites. Wea. Forecasting 29(2), 185–204. https://doi.org/10.1175/WAF-D-12-00075.1.

Mittermaier, M.P., Csima, G., 2017. Ensemble versus deterministic performance at the kilometer scale. Wea. Forecasting 32, 1697–1709.

Mittermaier, M.P., Roberts, N., Thompson, S., 2013. A long-term assessment of precipitation forecast skill using the fractions skill score. Meteor. Appl. 20, 176–186.

Moore, R.J., Cole, S.J., Bell, V.A., Jones, D.A., 2006. Issues in flood forecasting: ungauged basins, extreme floods and uncertainty. In: I. Tchiguirinskaia, K. N. N. Thein, P. Hubert (eds.), Frontiers in Flood Research, 8th Kovacs Colloquium, UNESCO, Paris, June/July 2006, IAHS Publ. 305, 103-122.

Moore, R.J., May, B.C., Jones, D.A., Black, K.B., 1994. Local calibration of weather radar over London. In: M.E. Almeida-Teixeira, R. Fantechi, R. Moore and V.M. Silva (eds), Advances

in Radar Hydrology. Proc. Int. Workshop, Lisbon, Portugal, 11-13 November 1991, Report EUR 14334 EN, European Commission, 186-195.

Murphy, A.H. 1977. The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. Mon. Wea. Rev. 105(7), 803–816.

Price, D., Hudson, K., Boyce, G., Schellekens, J., Moore, R.J., Clark, P., Harrison, T., Connolly, E., Pilling, C., 2012. Operational use of a grid-based model for flood forecasting. Water Management 165(2), 65–77. https://doi.org/10.1680/wama.2012.165.2.65.

Richardson, D.S. 2000. Skill and relative economic value of the ECMWF ensemble prediction system. Q.J. R. Meteorol. Soc. 126, 649-667.

Roulin, E., 2006. Skill and relative economic value of medium-range hydrological ensemble predictions. Hydrol. Earth Syst. Sci. 11, 725-737 .

Schaake, J.C. , Hamill, T.M., Buizza, R., Clark, M., 2007. HEPEX: The Hydrological Ensemble Prediction Experiment. Bull. Amer. Meteor. Soc. 88(10), 1541–1547.

Scheuerer, M., Hamill, T.M., 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. Mon. Wea. Rev. 143, 4578–4596. https://doi.org/10.1175/MWR-D-15-0061.1.

Talagrand, O., Vautart, R., Strauss, B., 1997. Evaluation of probabilistic prediction systems. In Proc. ECMWF Workshop and Predictability. ECMWF, 1–25.

Tang, Y., Lean, H.W., Bornemann, J., 2012. The benefits of the Met Office variable resolution NWP model for forecasting convection. Meteorol. Appl. 20, 417–426.

Thielen, J., Schaake, J., Hartman, R., Buizza, R., 2008. Aims, challenges and progress of the Hydrological Ensemble Prediction Experiment (HEPEX) following the third HEPEX workshop held in Stresa 27 to 29 June 2007. Atmos. Sci. Let. 9, 29–35.

Wilks, D.S., 2001. A skill score based on economic value for probability forecasts. Meteor. Appl. 8, 209–219.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. 3rd Edition, Academic Press, 704pp.

Zappa, M., Jaun, S., Germann, U., Walser, A., Fundel, F., 2011. Superposition of three sources of uncertainties in operational flood forecasting chains. Atmos. Res. 100(2–3), 246–262. https://doi.org/10.1016/j.atmosres.2010.12.005.

Zappa, M., Fundel, F., Jaun, S., 2013. A "Peak-Box" approach for supporting interpretation and verification of operational ensemble peak-flow forecasts. Hydrol. Process. 27(1), 117–131.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., Mylne, K., 2002. The economic value of ensemble-based weather forecasts. Bull. Amer. Meteor. Soc. 83(1), 73-83.

A framework for joint verification of river flow and precipitation ensembles is developed and demonstrated over Britain for eventual use in an operational flood forecasting setting. The river flow ensembles are obtained from a distributed hydrological model, the G2G model, using an ensemble of 15 minute precipitation accumulations as input on a 1 km grid. The precipitation ensemble consists of operational Numerical Weather Prediction (NWP) forecasts from the Met Office Unified Model. Both hourly and daily precipitation accumulations are verified, and the relevance of different accumulation periods discussed in

the context of timing errors and hydrological response. The implications of precipitation observation error are investigated by comparing verification results from raingauge- and radar-derived precipitation estimates. Challenges of verification using only a limited record of precipitation ensembles, from a system only relatively recently made operational, are addressed. Methods of obtaining more robust verification statistics, given the available ensembles, are presented and demonstrated for an example period in December 2015. For precipitation, percentile thresholds are used to ensure a given number of threshold crossing events for analysis using a contingency table and derived skill scores. For river flow, percentiles thresholds are of less relevance to operational flood guidance. Instead, exceedance of a flow threshold of given rarity (return-period) is used as a surrogate measure of flood severity. At the regional scale, both river flow and precipitation verification analyses are found to be dependent on the locations considered. This is linked to variations in precipitation amount. For river flows, catchment properties - and in particular catchment size - are found to be a key influence on verification. It is demonstrated how such behaviour can be used to obtain more-robust river flow verification statistics at sub-regional scales.

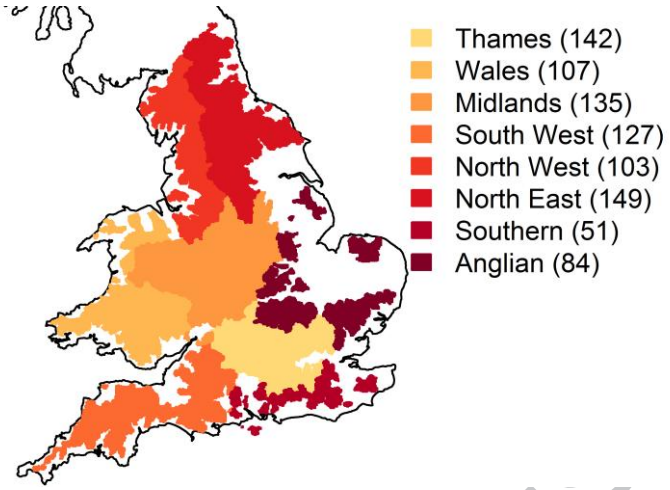

Fig. 1. The eight catchment groups used for regional-scale ensemble verification. The catchment groups are defined based on aggregated river drainage basins aligned to Wales, and Environment Agency regions. The bracketed numbers give the number of catchments in each region. COLOUR FOR ONLINE ONLY

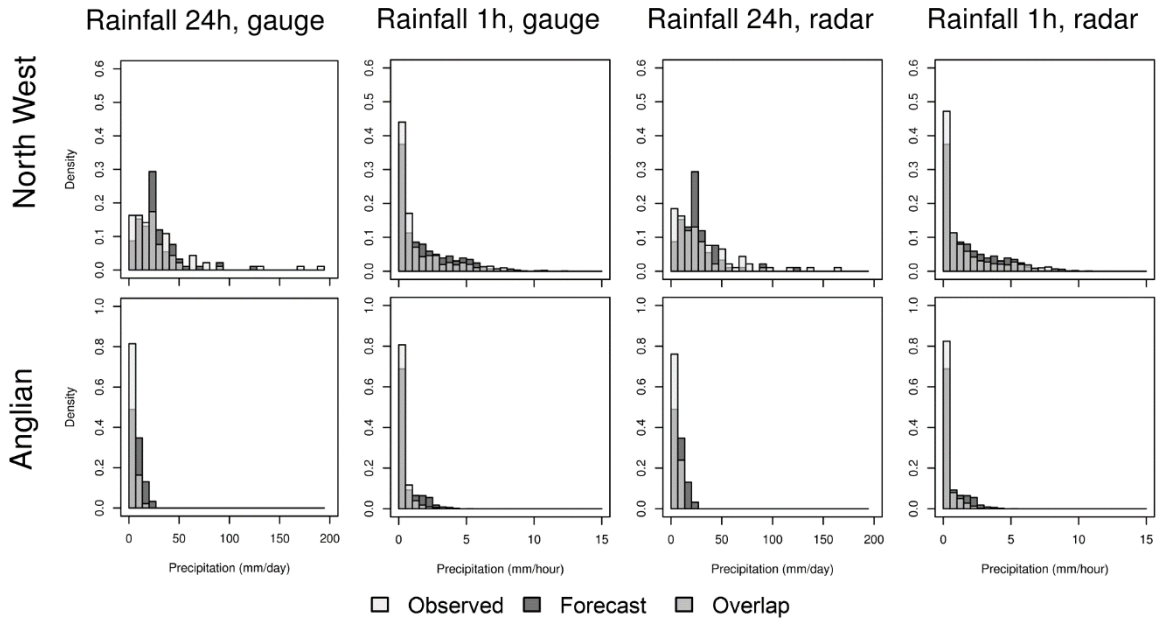

Fig. 2. Frequency histograms (daily and hourly) of observed (raingauge and radar) and forecast precipitation accumulation (daily in mm d^{-1}, hourly in mm h^{-1}) corresponding to the spatial 95th percentile of catchment-average precipitation for all catchments in the North West (top) and Anglian (bottom) regions.

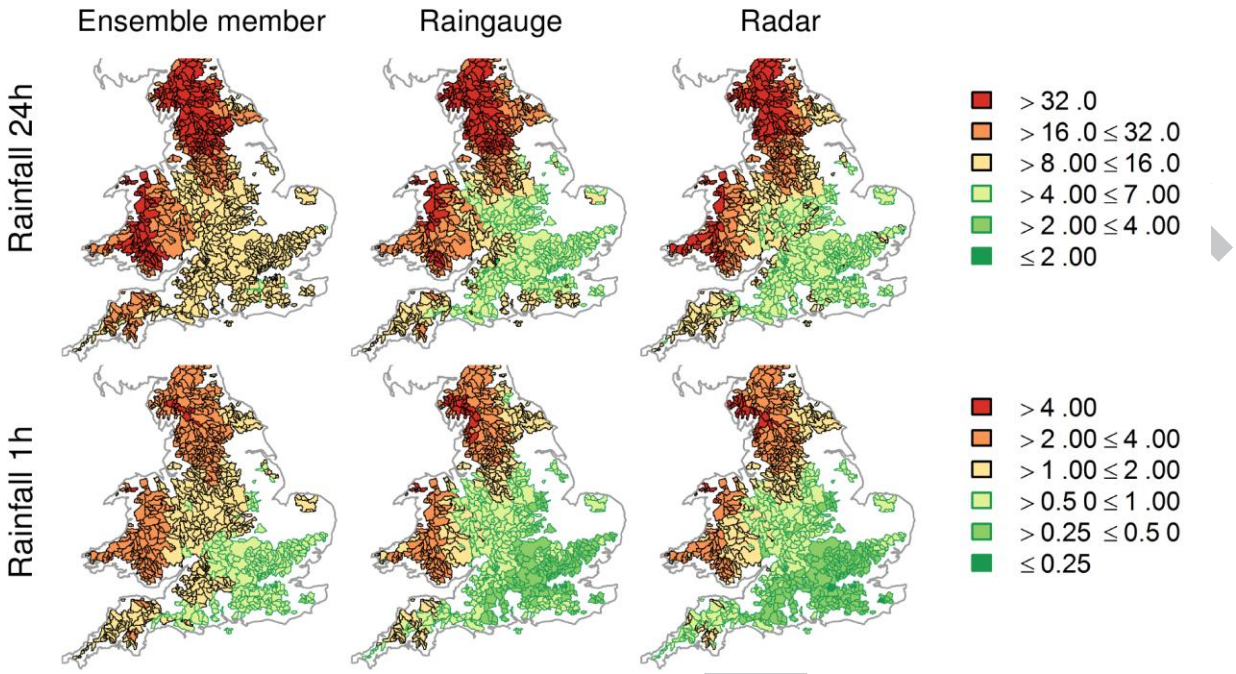

Fig. 3. Temporal 95th percentile precipitation values for daily (top) and hourly (bottom) totals (daily in mm d^{-1}, hourly in mm h^{-1}), calculated from the full 32-day study period 25 November to 26 December 2015. Results are shown for 24h (top) and 1h (bottom) precipitation accumulations from an example ensemble member (left), raingauge data (middle) and radar data (right). COLOUR FOR ONLINE ONLY

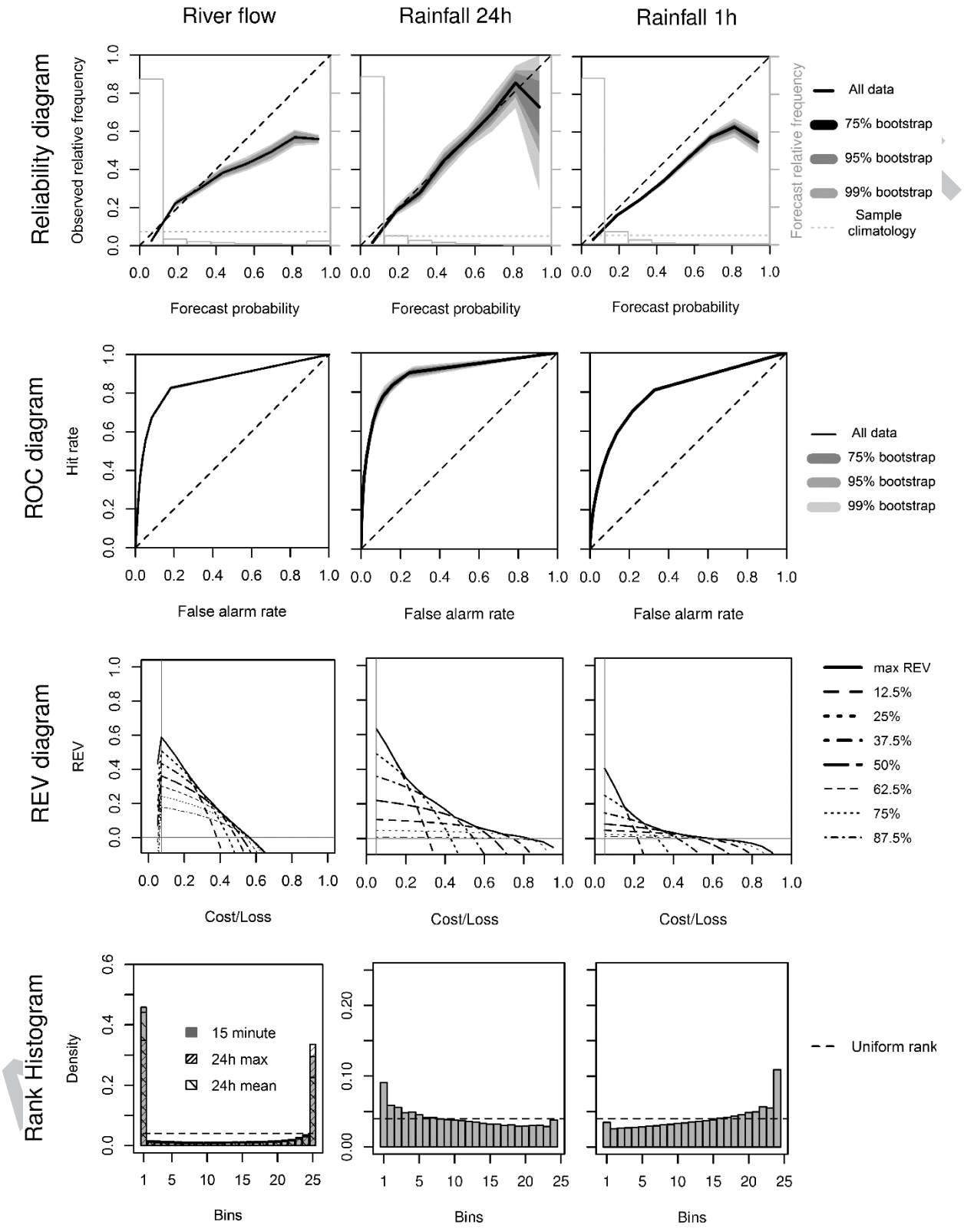

Fig. 4. Verification diagrams calculated using data pooled from all catchments in England and Wales. From top to bottom: Reliability, ROC and REV diagrams, and Rank Histogram. Results are shown for river flow (left) and 24h (middle) and 1h (right) precipitation accumulations (verified against raingauge data). For the Reliability, ROC and REV diagrams the ½Q(2) threshold was used for river flow, and the spatial 95th percentile threshold for precipitation accumulations.

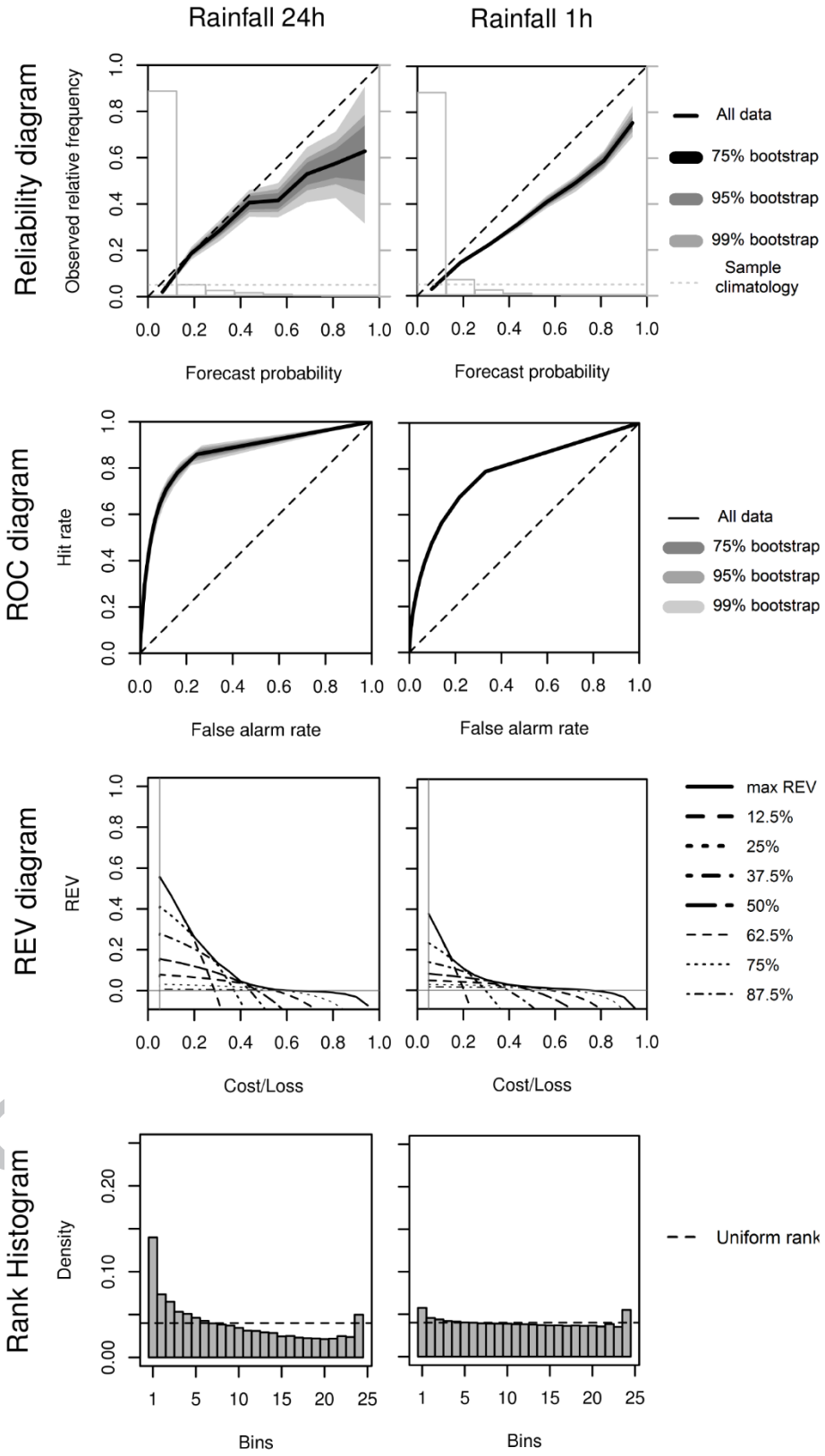

Fig. 5. Verification diagrams calculated using data pooled from all catchments in England and Wales. From top to bottom: Reliability, ROC and REV diagrams, and Rank Histogram. Results are shown for 24h (left) and 1h (right) precipitation accumulations (verified against radar data). For the Reliability, ROC and REV diagrams the spatial 95th percentile threshold was used.

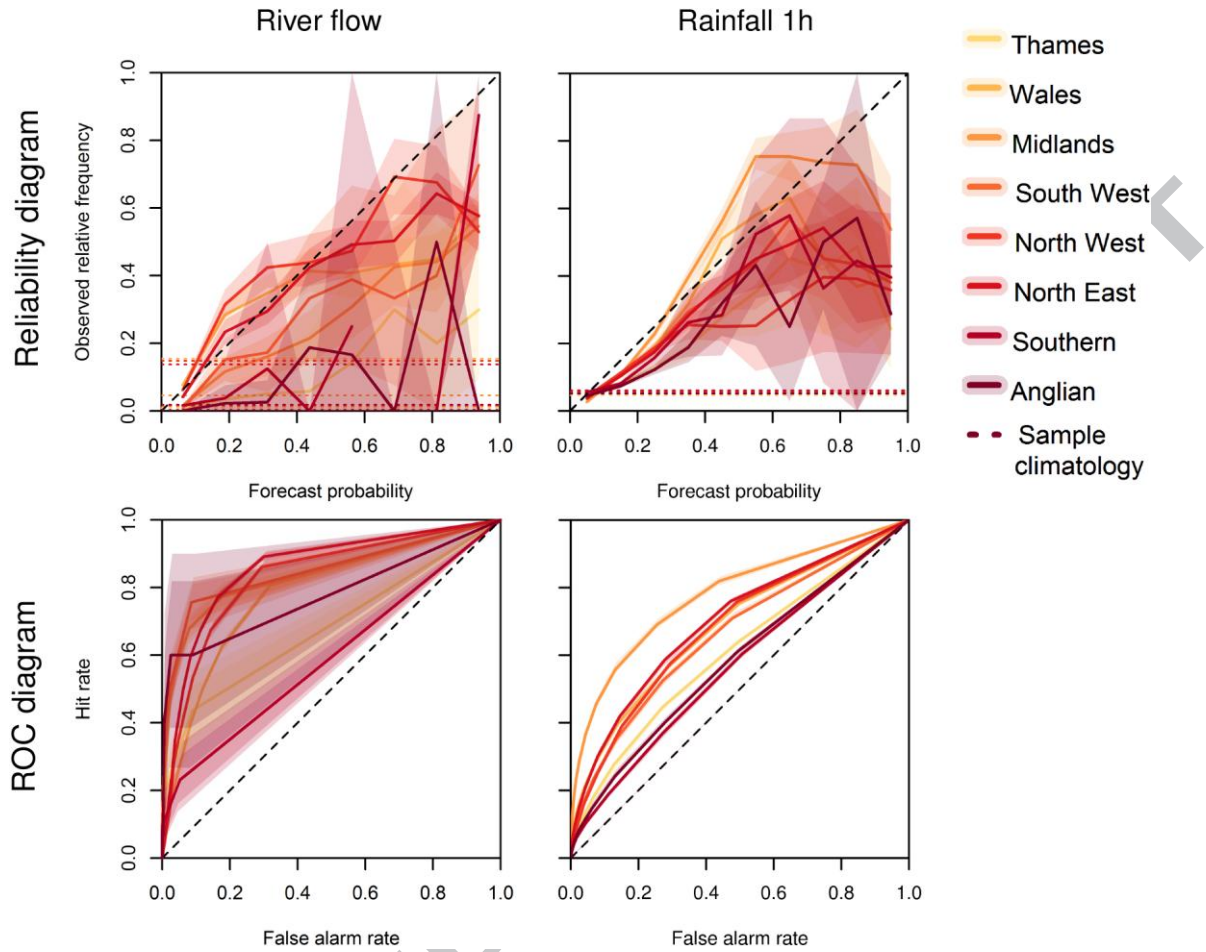

Fig. 6. Verification diagrams for catchments pooled by region. Reliability (top) and ROC (bottom) diagrams are shown for river flow (left) and 1h precipitation accumulations verified against raingauge data (right). The ½Q(2) threshold was used for river flow, and the spatial 95th percentile threshold for precipitation accumulations. Shaded areas around each line show the 99th percentile bootstrap uncertainty, and horizontal dashed lines show the sample climatology. The use of 24h precipitation accumulations, or verification against radar data, lead to similar results.

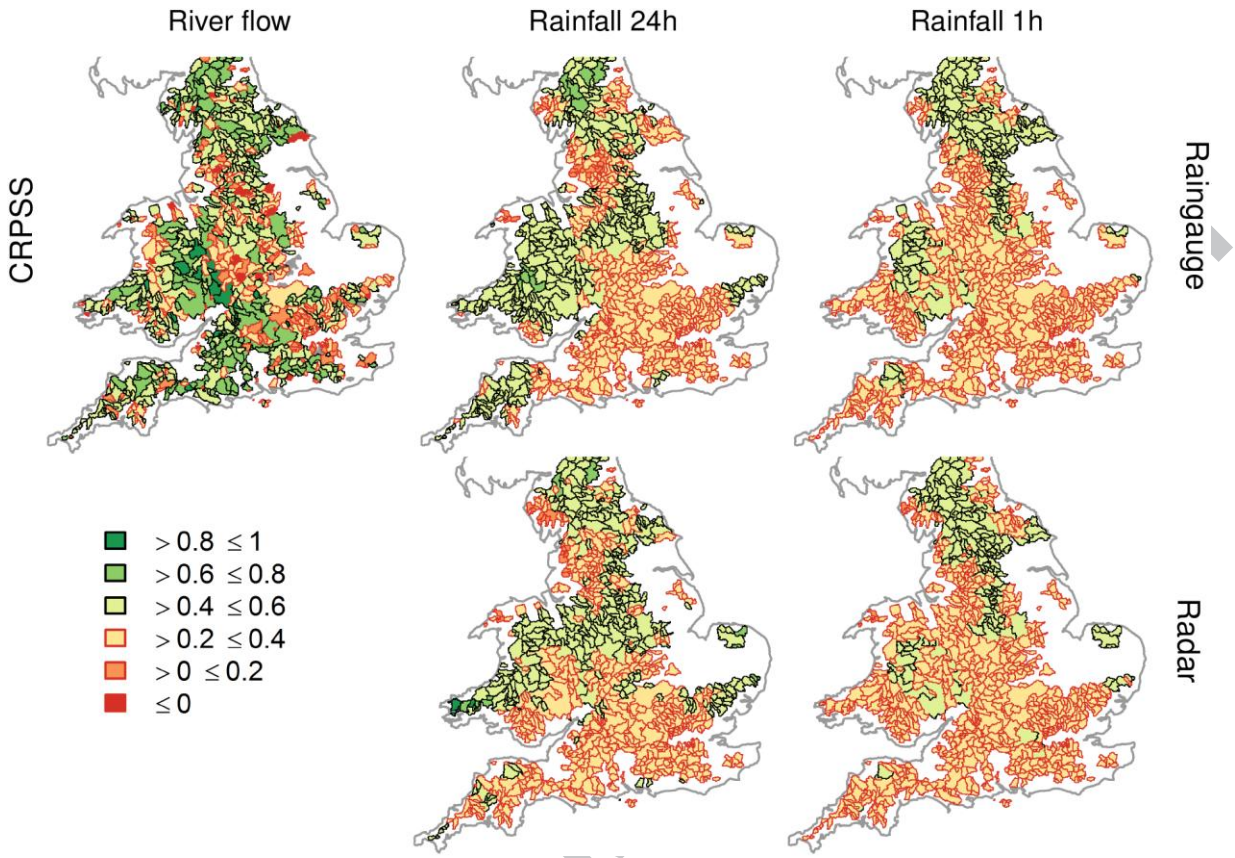

Fig. 7. Maps of the Continuous Rank Probability Skill Score (CRPSS) calculated for individual catchments in England and Wales. Results are shown for river flow (left), and 24h (middle) and 1h (right) precipitation accumulations verified against raingauge (top) and radar (bottom) data. COLOUR FOR ONLINE ONLY

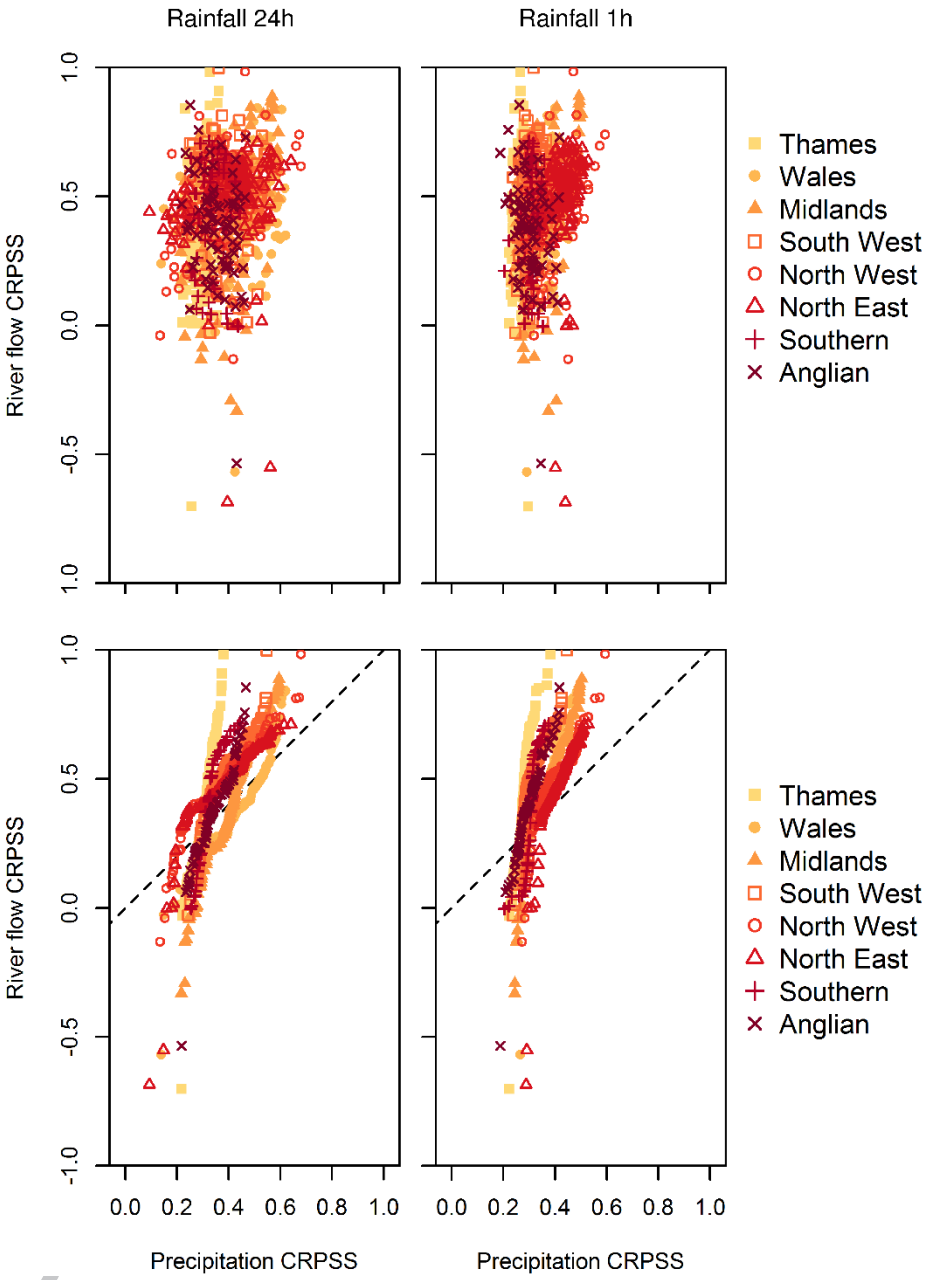

Fig. 8. Scatter (top) and quantile-quantile (bottom) plots of the instantaneous 15-minute river flow CRPSS against the 24h (left) and 1h (right) precipitation accumulation CRPSS, with precipitation verified against raingauge data. For each q-q plot the quantiles of the river flow CRPSS distribution are plotted against the corresponding quantiles of the CRPSS precipitation accumulation distribution. Catchments are pooled by region. COLOUR FOR ONLINE ONLY

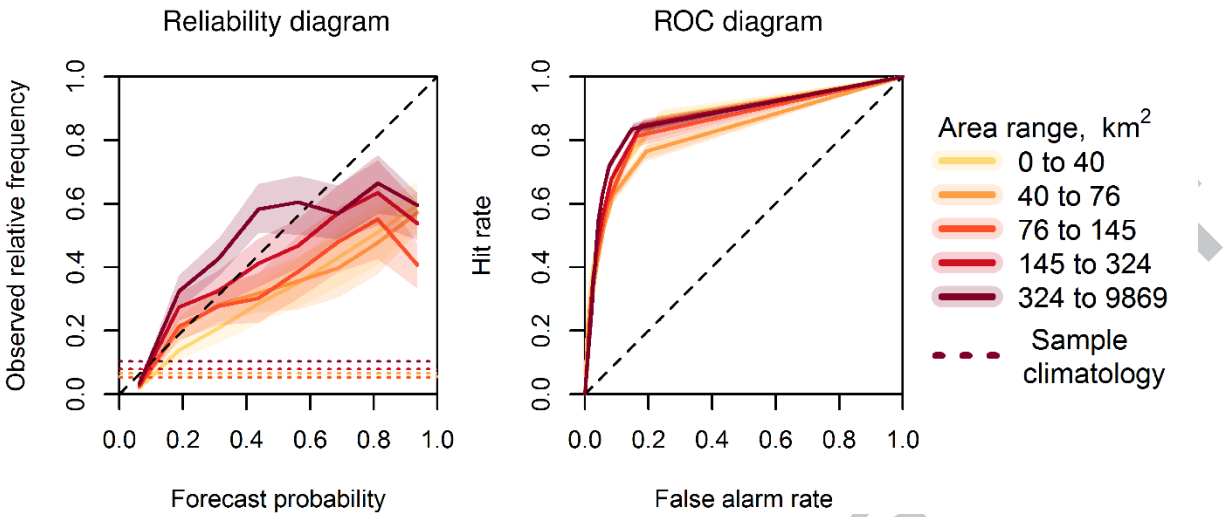

Fig. 9. River flow verification pooled by catchment area using Reliability (left) and ROC (right) diagrams with the ½Q(2) threshold. Five catchment area pooling groups are used, giving around 180 catchments per group. Shaded areas around each line show the 99th percentile bootstrap uncertainty. COLOUR FOR ONLINE ONLY

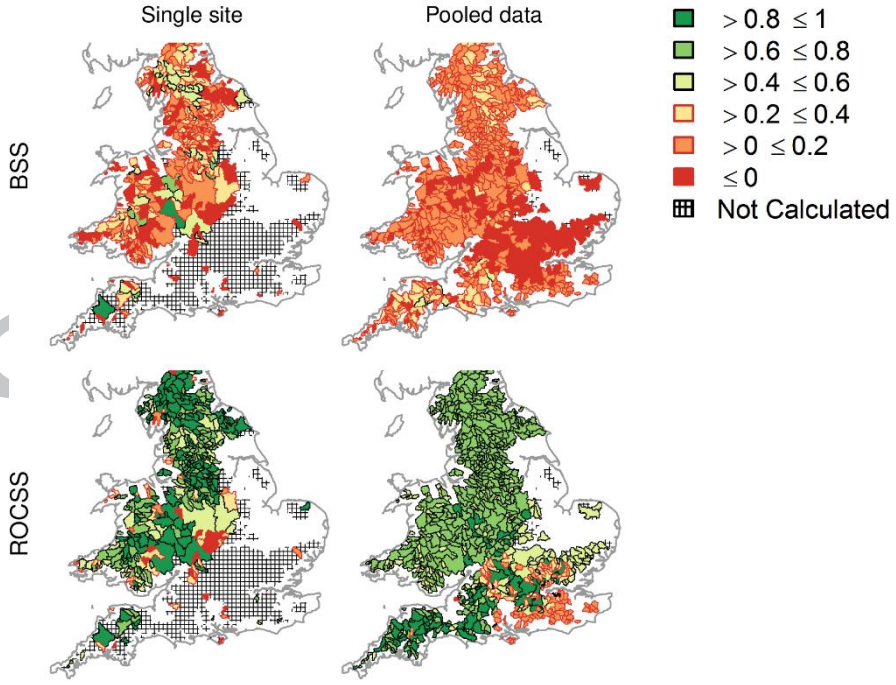

Fig. 10. Maps of the Brier Skill Score (top) and Relative Operating Characteristic Skill Score (bottom) calculated for river flow for all catchments in England and Wales using the ½Q(2) threshold. Results are shown calculated using individual catchment data only (left) and using a moving catchment-area based pool of 31 catchments within each region (right). COLOUR FOR ONLINE ONLY

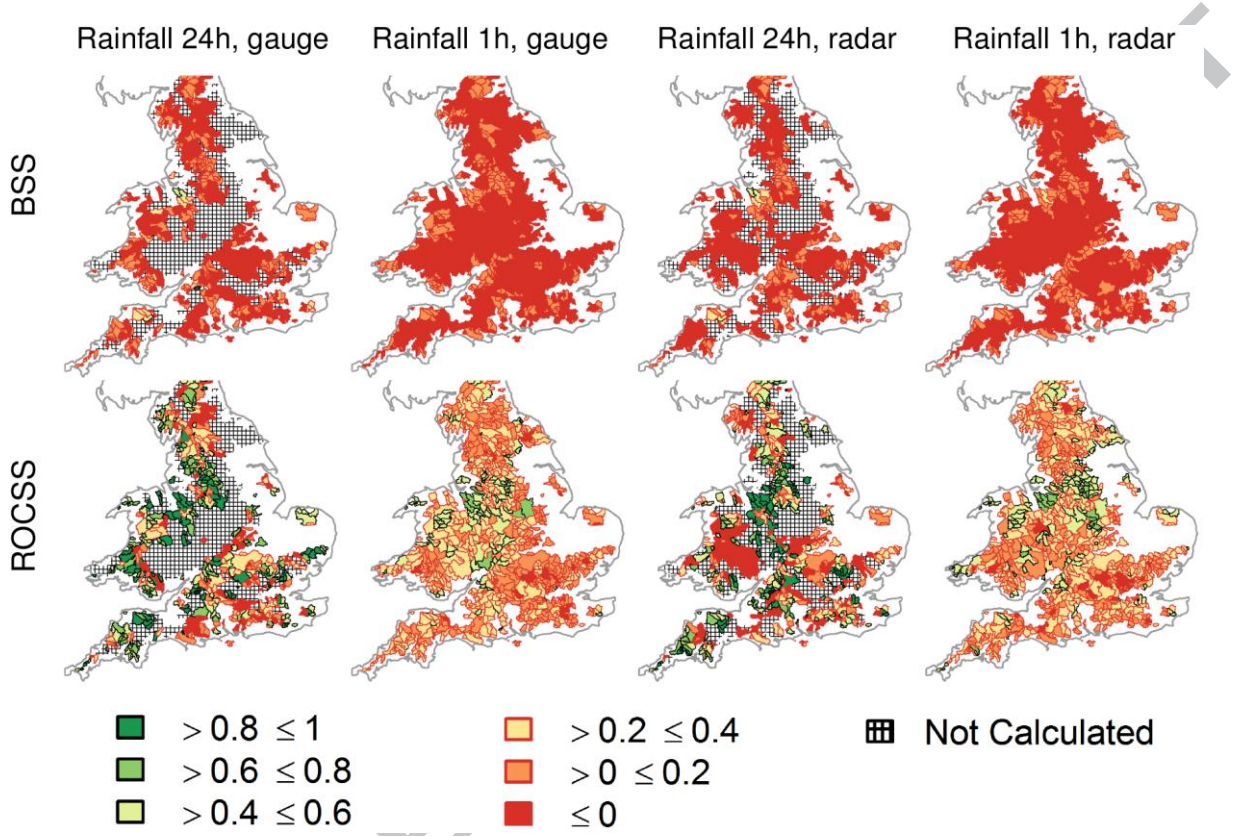

Fig. 11. Maps of the Brier Skill Score (top) and Relative Operating Characteristic Skill Score (bottom) calculated for individual catchment precipitation accumulations in England and Wales using the spatial 95th percentile threshold. From left to right: 24h and 1h precipitation accumulations verified against raingauge data, and then against radar data. COLOUR FOR ONLINE ONLY

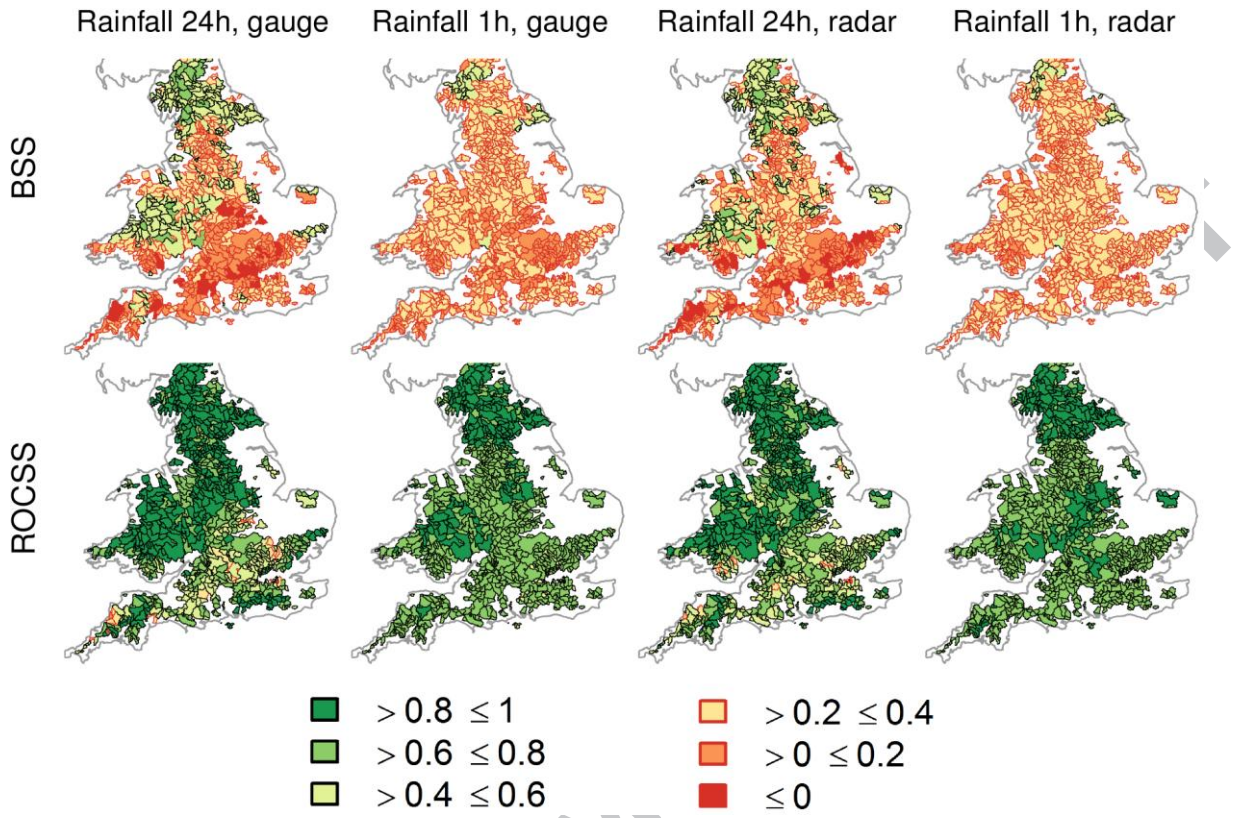

Fig. 12. Maps of the Brier Skill Score (top) and Relative Operating Characteristic Skill Score (bottom) calculated for individual catchment precipitation accumulations in England and Wales, using the temporal 95th percentile threshold for each site. From left to right: 24h and 1h precipitation accumulations verified against raingauge data, and then against radar data. COLOUR FOR ONLINE ONLY

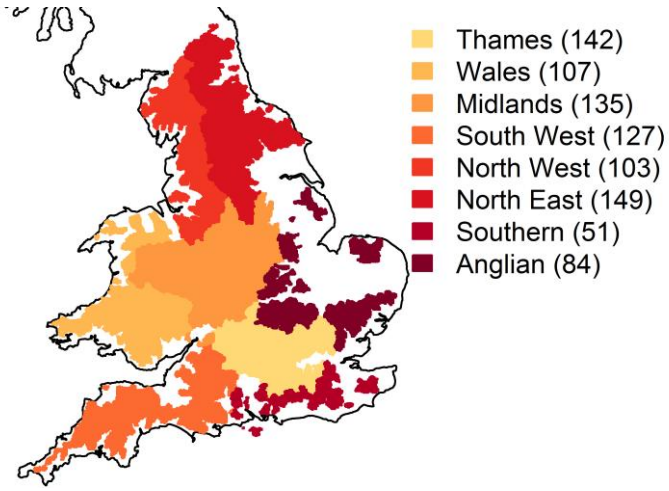

Fig. 1. The eight catchment groups used for regional-scale ensemble verification. The catchment groups are defined based on aggregated river drainage basins aligned to Wales, and Environment

Agency regions. The bracketed numbers give the number of catchments in each region. COLOUR FOR ONLINE ONLY

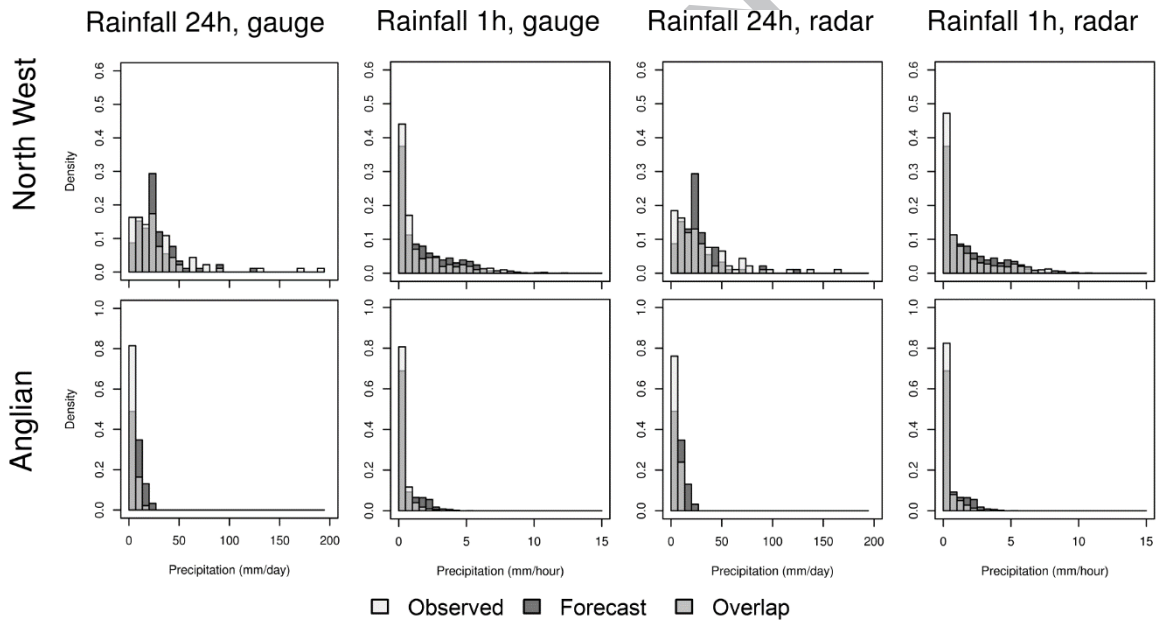

Fig. 2. Frequency histograms (daily and hourly) of observed (raingauge and radar) and forecast precipitation accumulation (daily in mm d^{-1}, hourly in mm h^{-1}) corresponding to the spatial 95th percentile of catchment-average precipitation for all catchments in the North West (top) and Anglian (bottom) regions.

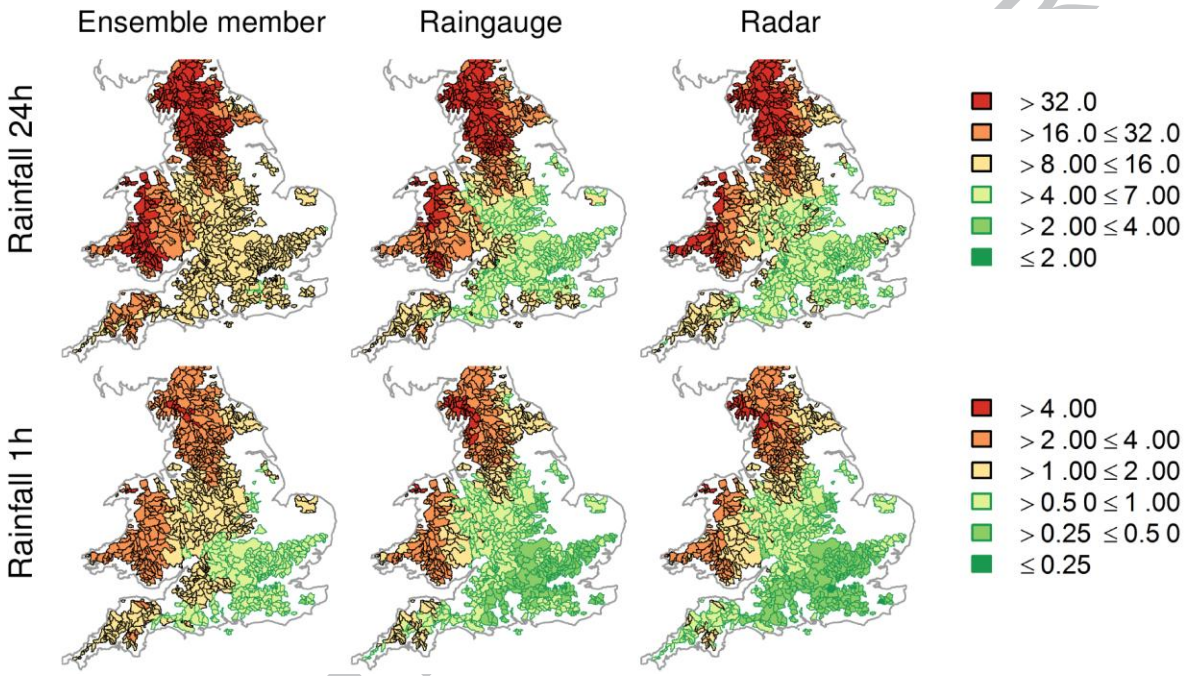

Fig. 3. Temporal 95th percentile precipitation values for daily (top) and hourly (bottom) totals (daily in mm d^{-1}, hourly in mm h^{-1}), calculated from the full 32-day study period 25 November to 26 December 2015. Results are shown for 24h (top) and 1h (bottom) precipitation accumulations from an example ensemble member (left), raingauge data (middle) and radar data (right). COLOUR FOR ONLINE ONLY

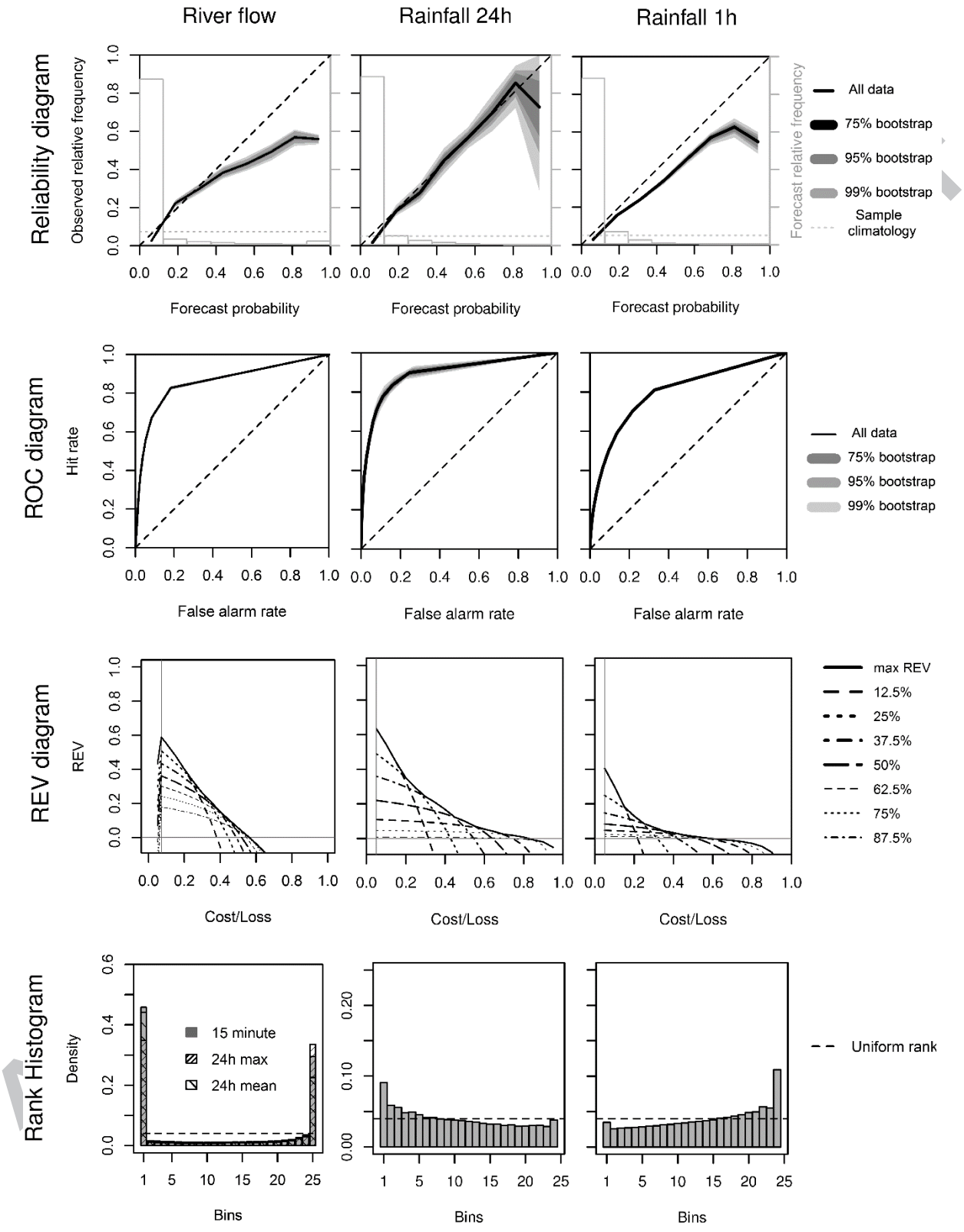

Fig. 4. Verification diagrams calculated using data pooled from all catchments in England and Wales. From top to bottom: Reliability, ROC and REV diagrams, and Rank Histogram. Results are shown for river flow (left) and 24h (middle) and 1h (right) precipitation accumulations (verified against raingauge data). For the Reliability, ROC and REV diagrams the ½Q(2) threshold was used for river flow, and the spatial 95th percentile threshold for precipitation accumulations.

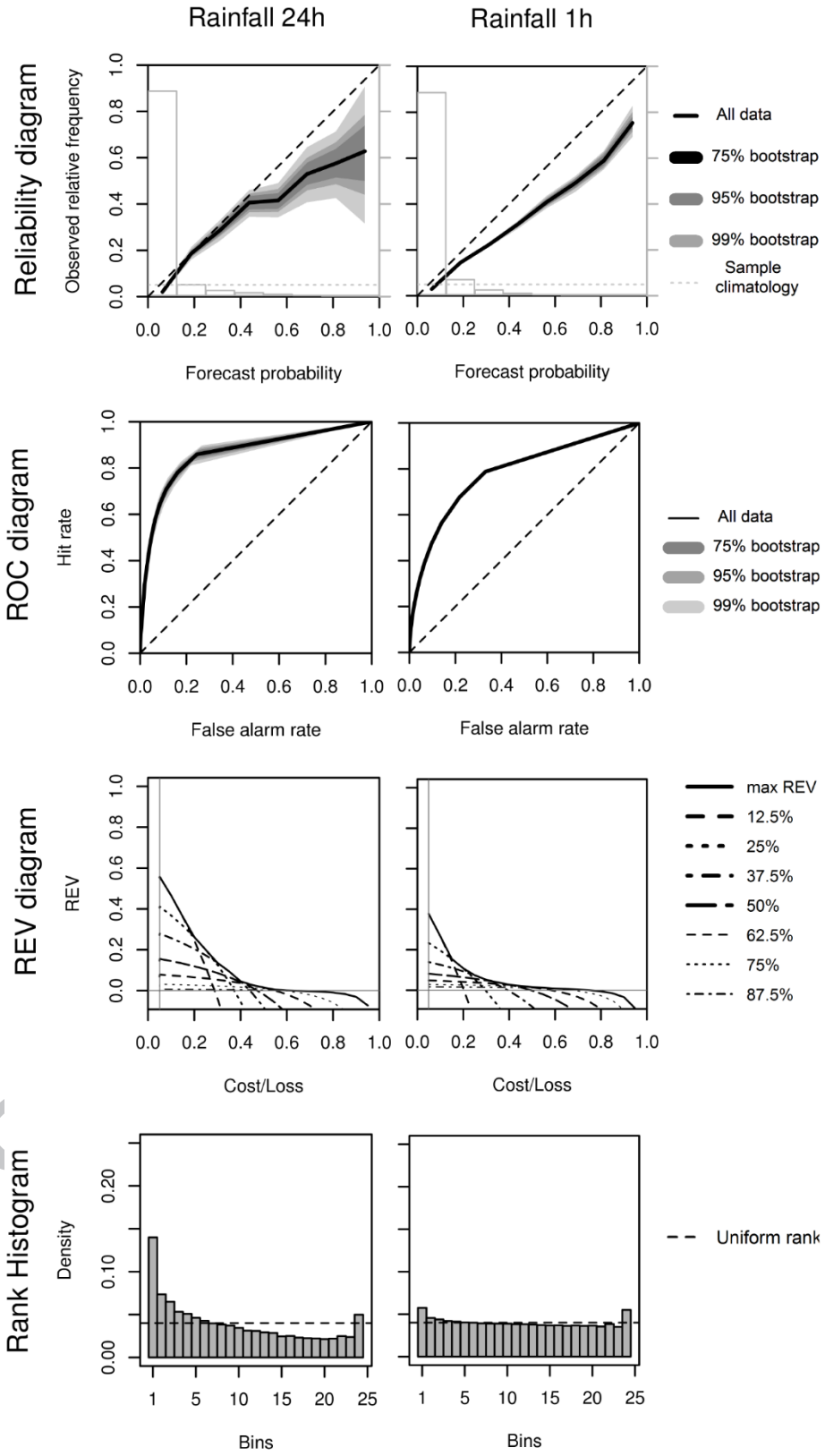

Fig. 5. Verification diagrams calculated using data pooled from all catchments in England and Wales. From top to bottom: Reliability, ROC and REV diagrams, and Rank Histogram. Results are shown for 24h (left) and 1h (right) precipitation accumulations (verified against radar data). For the Reliability, ROC and REV diagrams the spatial 95th percentile threshold was used.

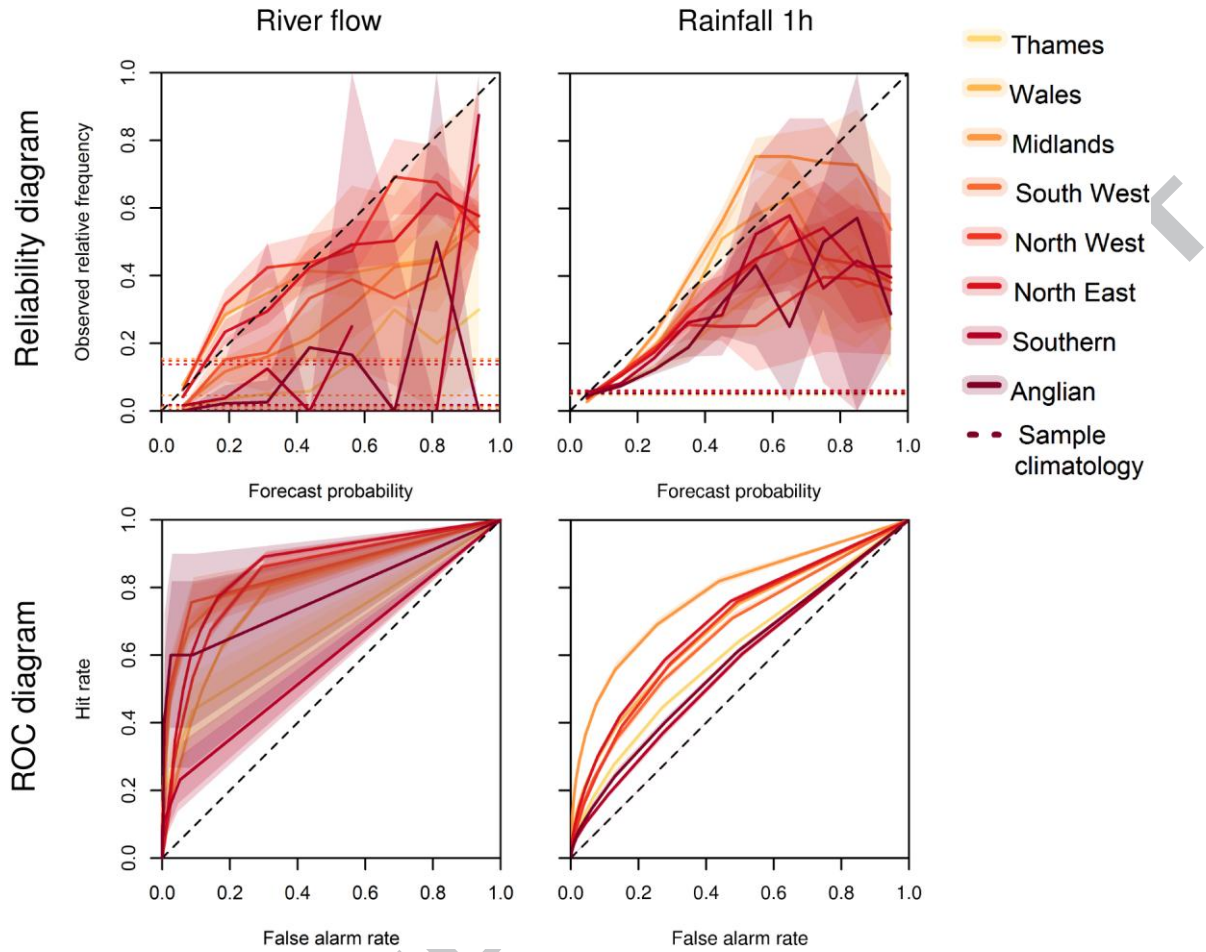

Fig. 6. Verification diagrams for catchments pooled by region. Reliability (top) and ROC (bottom) diagrams are shown for river flow (left) and 1h precipitation accumulations verified against raingauge data (right). The ½Q(2) threshold was used for river flow, and the spatial 95th percentile threshold for precipitation accumulations. Shaded areas around each line show the 99th percentile bootstrap uncertainty, and horizontal dashed lines show the sample climatology. The use of 24h precipitation accumulations, or verification against radar data, lead to similar results.

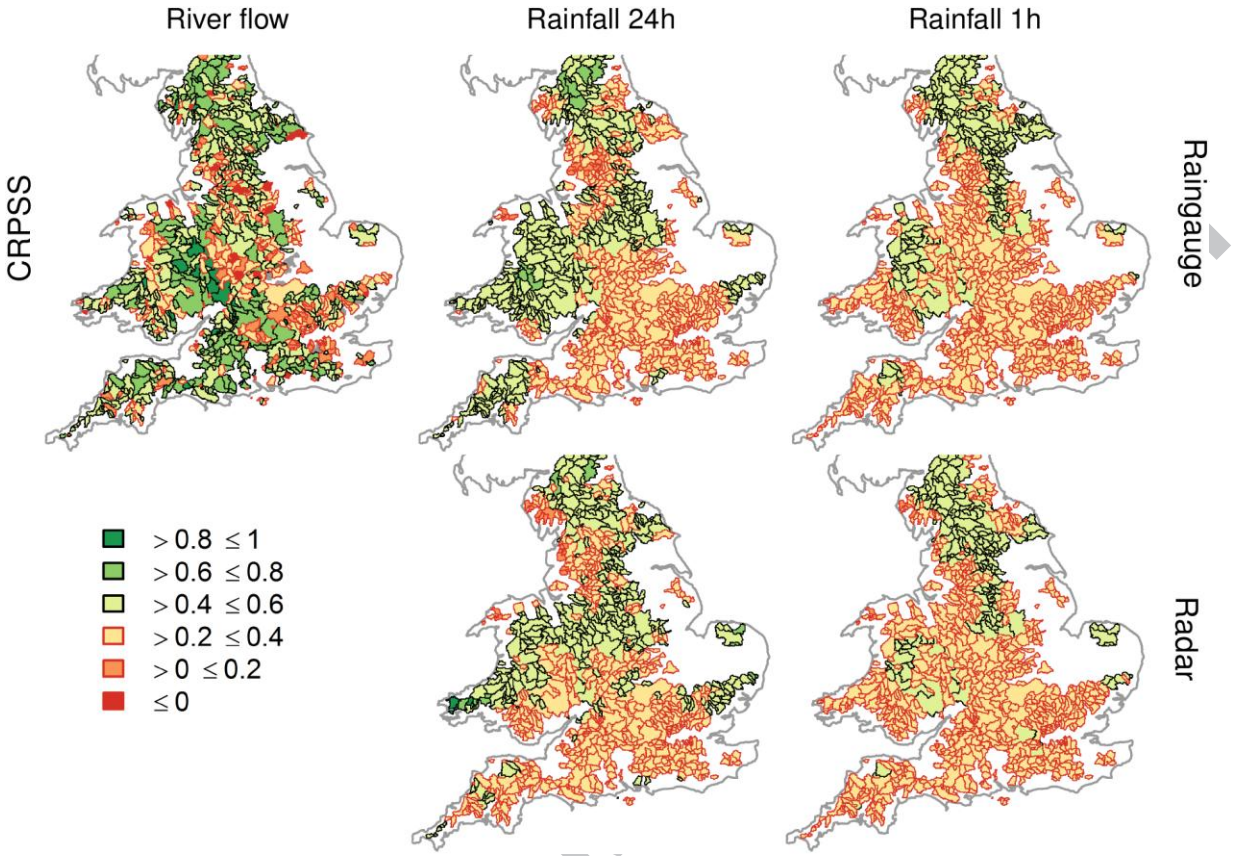

Fig. 7. Maps of the Continuous Rank Probability Skill Score (CRPSS) calculated for individual catchments in England and Wales. Results are shown for river flow (left), and 24h (middle) and 1h (right) precipitation accumulations verified against raingauge (top) and radar (bottom) data. COLOUR FOR ONLINE ONLY

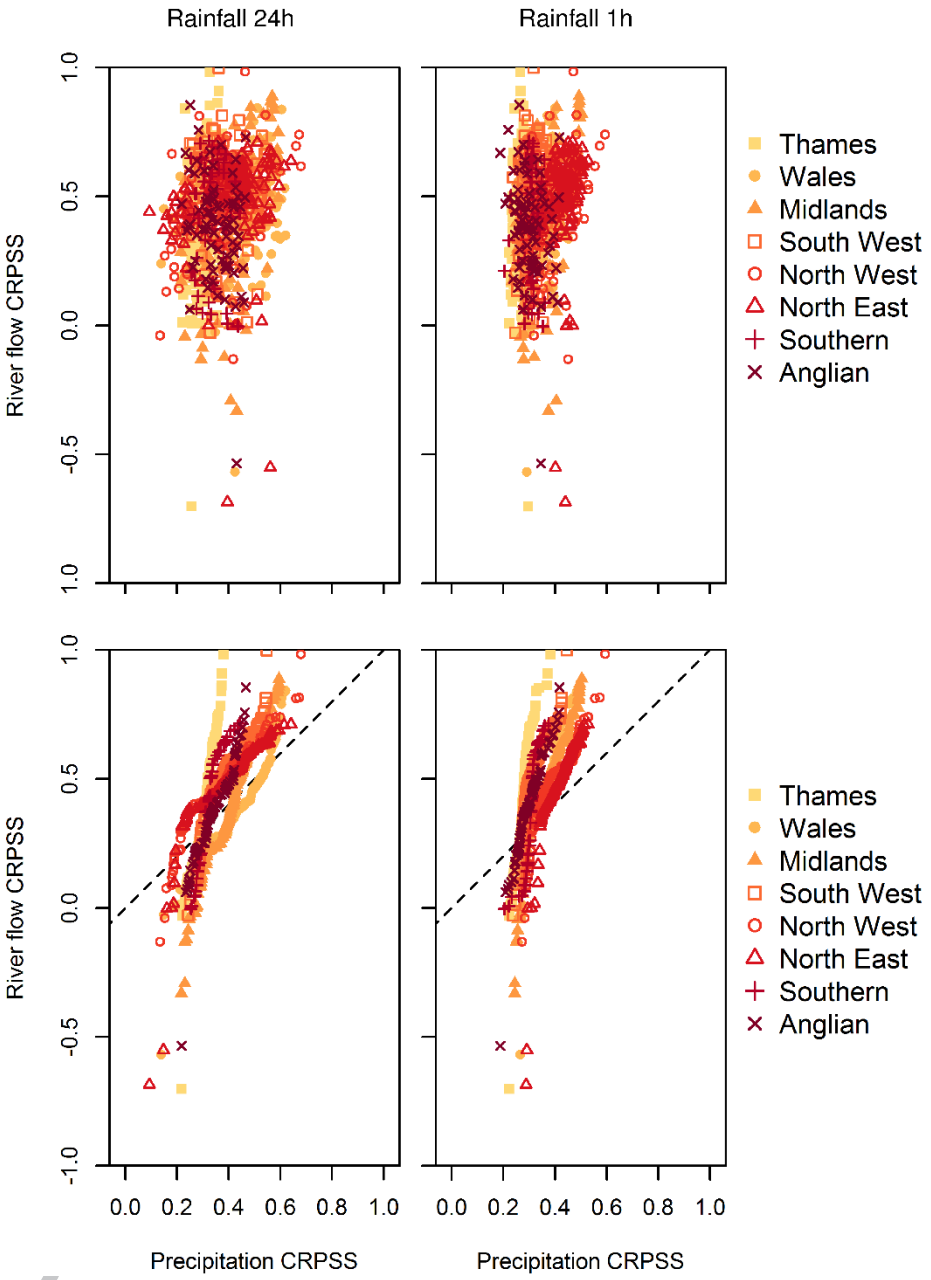

Fig. 8. Scatter (top) and quantile-quantile (bottom) plots of the instantaneous 15-minute river flow CRPSS against the 24h (left) and 1h (right) precipitation accumulation CRPSS, with precipitation verified against raingauge data. For each q-q plot the quantiles of the river flow CRPSS distribution are plotted against the corresponding quantiles of the CRPSS precipitation accumulation distribution. Catchments are pooled by region. COLOUR FOR ONLINE ONLY

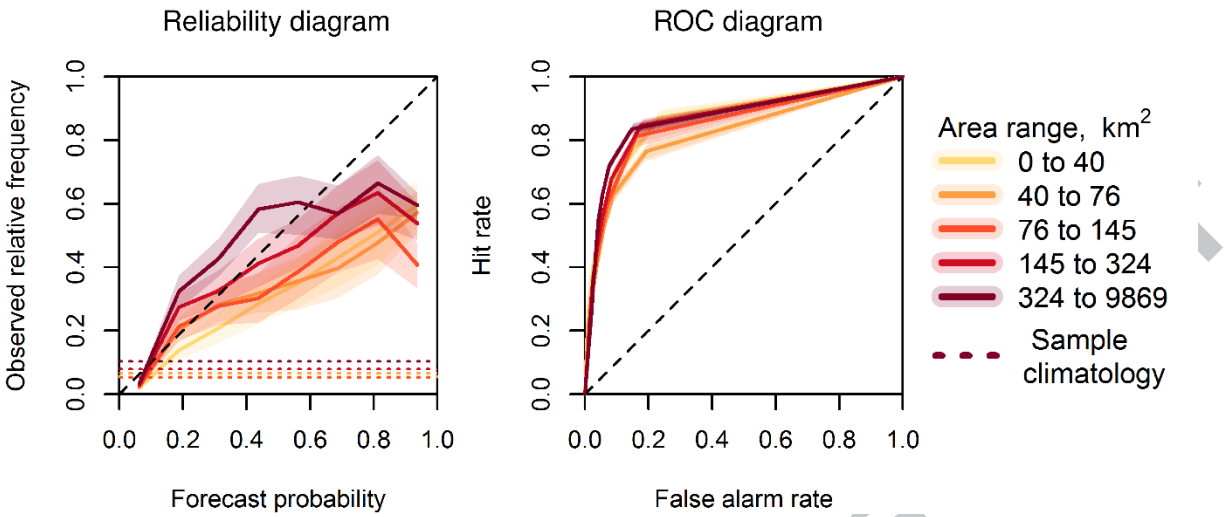

Fig. 9. River flow verification pooled by catchment area using Reliability (left) and ROC (right) diagrams with the ½Q(2) threshold. Five catchment area pooling groups are used, giving around 180 catchments per group. Shaded areas around each line show the 99th percentile bootstrap uncertainty. COLOUR FOR ONLINE ONLY

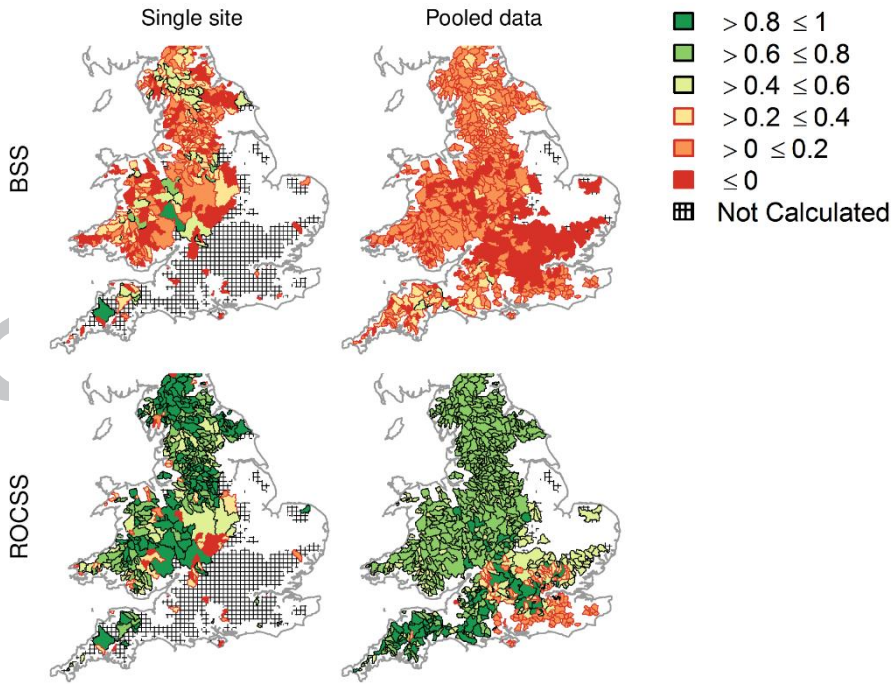

Fig. 10. Maps of the Brier Skill Score (top) and Relative Operating Characteristic Skill Score (bottom) calculated for river flow for all catchments in England and Wales using the ½Q(2) threshold. Results are shown calculated using individual catchment data only (left) and using a moving catchment-area based pool of 31 catchments within each region (right). COLOUR FOR ONLINE ONLY

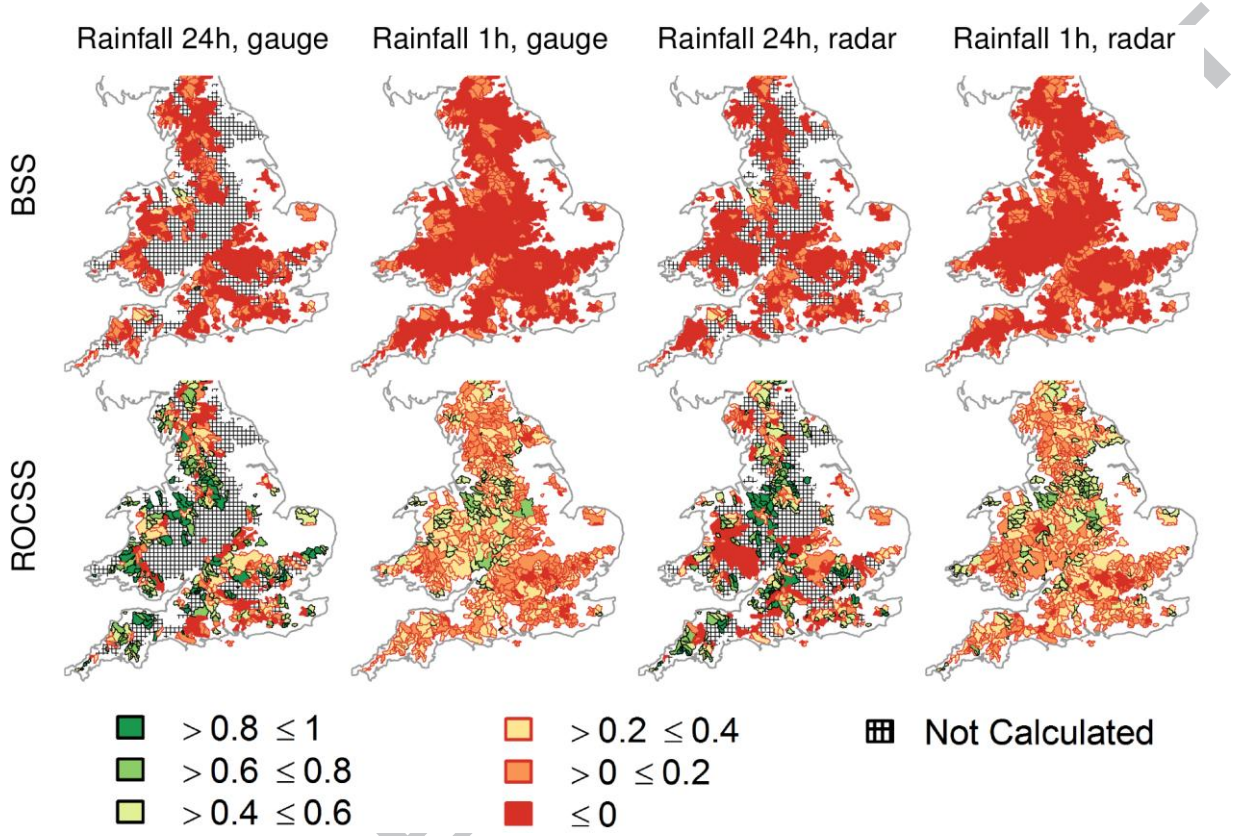

Fig. 11. Maps of the Brier Skill Score (top) and Relative Operating Characteristic Skill Score (bottom) calculated for individual catchment precipitation accumulations in England and Wales using the spatial 95th percentile threshold. From left to right: 24h and 1h precipitation accumulations verified against raingauge data, and then against radar data. COLOUR FOR ONLINE ONLY

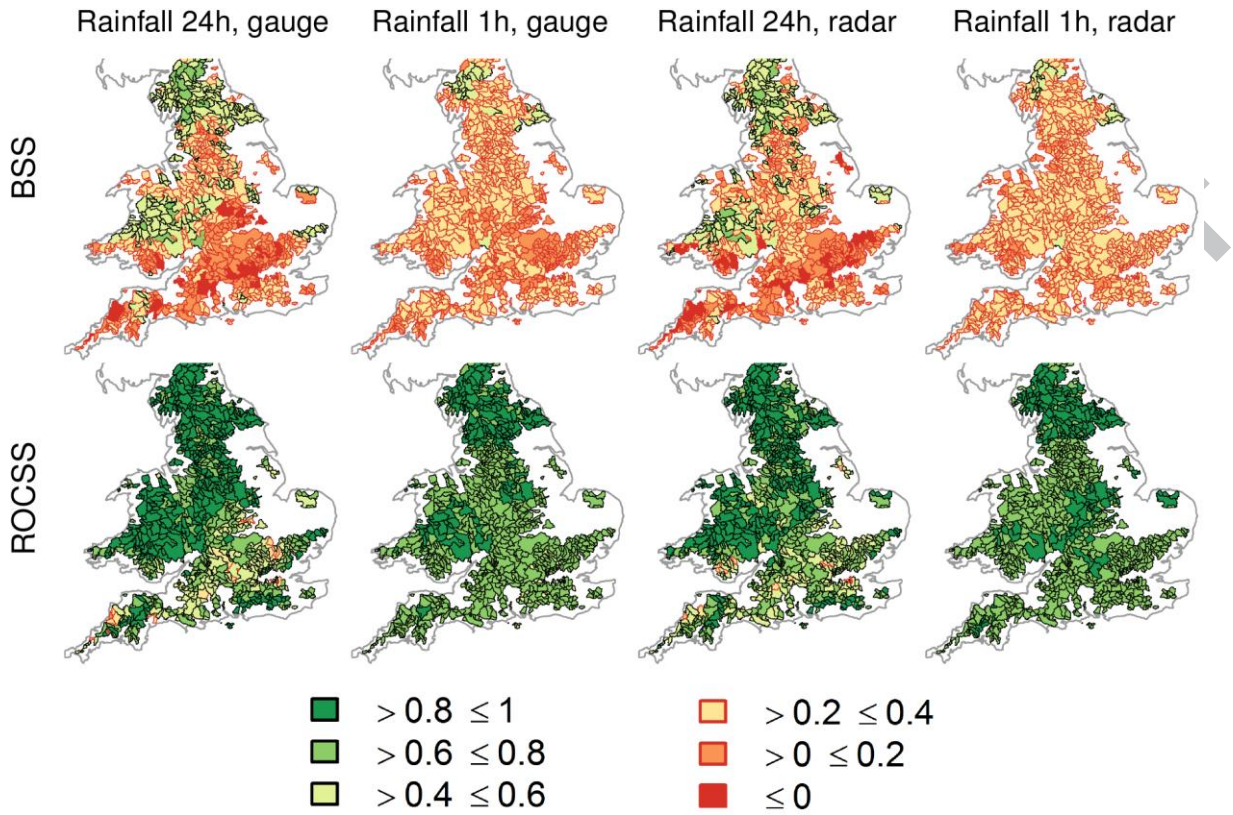

Fig. 12. Maps of the Brier Skill Score (top) and Relative Operating Characteristic Skill Score (bottom) calculated for individual catchment precipitation accumulations in England and Wales, using the temporal 95th percentile threshold for each site. From left to right: 24h and 1h precipitation accumulations verified against raingauge data, and then against radar data. COLOUR FOR ONLINE ONLY

- Precipitation and river flow verification must be joined-up and physically-based
- Multi-scale analysis (national, regional, and catchment) adds relevant information
- Sampling uncertainty reduced by pooling river flow data on catchment area
- Percentile precipitation thresholds allow extremes to be captured when they occur
- Varying effect of observation error on verification metrics influences interpretation

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: