Appendix

Rainfall and River Flow Ensemble Verification: Phase 2
Joint Verification Framework
Final Report Appendix A.1

1. Overview of Framework and metrics for verification

The Prototype Framework for Ensemble Verification is set out in this Appendix as a set of selected metrics (scores and diagrams). The objective of the Framework is to give an overview of performance in general, first individually at each site and then over all sites, possibly split by catchment features (e.g. catchment size). This is where the standard ensemble verification scores are used, both in Numerical Weather Prediction (NWP) and Hydrological Forecasting. It is necessary to compare, contrast and link the different scores and obtain a ''joined up" overview of performance across NWP and Hydrological Forecasting. The aim is to produce information that is ***operationally useful*** for flood forecasting and warning by the FFC and SFFS. Hence the chosen scores must be directly applicable and interpretable in this context.

Different scores can (and arguably should) be used for NWP and Hydrological Forecasting, provided there is a common comparison method. Table 1 summarises the proposed metrics to be used for ensemble verification as part of the Prototype Framework and detailed later in this Appendix.

Table 1 Metrics to be used for NWP and Hydrological Forecast verification.

Common between NWP and Hydrology	NWP only
• Continuous Rank Probability Score (CRPS) with decomposition • Brier Score (BS) with decomposition • Continuous Rank Probability Skill Score (CRPSS) • Brier Skill Score (BSS) with decomposition • Reliability Diagram • Relative Operating Characteristic Diagram and Area Under Curve Skill Score (ROCSS) • Relative Economic Value (REV) • Rank Histogram	• Mean Error (ME, a measure of bias) of areal mean precipitation per member • Root Mean Squared Error (RMSE) of areal mean precipitation per member

Some metrics applied to the precipitation verification focus on the underlying NWP model configuration. As both the FFC and SEPA already have detailed assessments of the G2G deterministic performance - in the form of a Performance Summary for each site (CEH, 2016) - these measures are included for precipitation only. The other metrics focus on the ensemble performance, evaluating the error in probability space. In particular, these assess how well the ensemble captures the spread of possible outcomes (as spatial rainfall or river flow), and how reliable the probabilities are. The ability of the ensemble to discriminate between events (defined as an upward-crossing of a rainfall or flow threshold over a prescribed forecast period) and non-events is also assessed, providing a measure of potential ensemble skill. Skill scores are calculated for a range of different thresholds and lead-times.

A summary of the key features of the verification metrics considered in the Joint Verification Framework, and detailed in this Appendix, is given in Table 2 for ease of reference.

Table 2 Overview of verification metrics used in the Joint Verification Framework

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

2. Deterministic verification metrics applied to individual ensemble members

2.1 Mean error of areal mean precipitation

The Mean Error (ME) is the difference between the forecast mean and the mean of the observations. Hence, the ME is a measure of the forecast bias. Calculating the ME for individual ensemble members allows the distribution of the bias to be assessed. A larger spread of values would indicate a wider range of different precipitation accumulations across the ensemble members.

2.2 RMSE of areal mean precipitation

The Root Mean Squared Error (RMSE) is calculated as

$$\mathrm{RMSE} = \sqrt{n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

where y_i and $\hat{y}_i$ are the i^{th} of n observation and forecast pairs. The RMSE gives the typical magnitude of forecast error, with a higher weighting given to larger errors. Calculating the RMSE for individual ensemble members allows the variation in forecast error across the ensemble to be assessed.

3. Metrics applied to the ensemble as a whole

3.1 Brier Score and Brier Skill Score

The Brier Skill Score provides a relative measure of the skill of a probability forecast, as assessed using the Brier Score (BS) and relative to the BS of a reference forecast, such as climatology. The *Brier Score* - like the Probability of Detection (POD), False Alarm Rate (*F*) and False Alarm Ratio (FAR) - is a categorical form of score based on crossing of a chosen threshold, but appropriate for use when the forecast being assessed is in the form of a probability. It gives the *mean square probability error* over n forecasts.

It is convenient to introduce some notation at this stage to allow the Brier Score to be precisely defined. Let y_i denote the observed value(s) of a quantity of interest (e.g. rainfall, river flow) for forecast i and x denotes a threshold value of interest for this same quantity (e.g. rainfall threshold; Q(2), the flow Q of return period 2 years) and used to define the categories of event-occurrence or non-occurrence. We define Y_i as an indicator variable of the observed event, taking a value 1 if the event does occur and 0 if not. The forecast probability to be assessed, $\hat{Y}_i$, is the forecast probability of the event occurring, taking values in the range 0 to 1.

The Brier Score, giving the *mean square probability error* over *n* forecasts, can be defined through the above notation as:

$$\mathrm{BS} = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 . \tag{2}$$

It is used to express the typical size of error in probability terms on a scale of 0 to 1, with 0 being “best” (a probability forecast associated with no error).

The Brier Score can be decomposed into three terms representing the Reliability (REL), Resolution (RES) and Uncertainty (UNC) of the forecast (Murphy, 1973; Siegert, 2017). Suppose the n forecast probabilities $\hat{Y}_i$ take on only K distinct probability values, that is $\hat{Y}_i \in \{P_1,...,P_K\}$ for all t. Also that

n_k is the number of times the k th forecast probability value is issued, and o_k is the number of events that have occurred when issued. Then the average event frequency for the k th forecast probability value is $\bar{Y}_k = o_k / n_k$. Given these definitions, the Brier Score has the following decomposition:

$$\begin{aligned} \mathrm{BS} &= n^{-1}\sum_{k=1}^{K} n_k (P_k - \bar{Y}_k)^2 - n^{-1}\sum_{k=1}^{K} n_k (\bar{Y} - \bar{Y}_k)^2 + \bar{Y}(1-\bar{Y}) \\ &= REL - RES + UNC \end{aligned} \tag{3}$$

where the climatological event frequency

$$\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i \,. \tag{4}$$

The first term, *REL*, gives a measure of the Reliability of the forecast: that is how close the forecast probabilities are to the true observed probabilities. Lower values of *REL* indicate better reliability, with a perfectly reliable forecast having a *REL* of zero. A system with perfect reliability would have $o_k = n_k$ equal to P_k for all k .

The Resolution term, *RES*, defines the extent to which the conditional probabilities (i.e. the probabilities given the different forecasts) differ from the climatological average. Higher values of *RES* indicate better Resolution, with a *RES* of zero indicating that the model does not give an advantage over climatology.

The third term in the Brier score decomposition, *UNC*, indicates the Uncertainty in the forecasting situation, based on the observations. An instance with a climatological probability of either zero or one will have the minimum Uncertainty (*UNC*=0); a climatological probability of 0.5 will have the highest Uncertainty. A perfect model would have a Resolution term equal to the Uncertainty term, and a Reliability term of zero giving a zero Brier Score.

With $\mathrm{BS}_{\mathrm{ref}}$ denoting the Brier Score for a reference forecast (for example, one based on climatological relative frequencies), then the *Brier Skill Score* (BSS) is given by

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} \,. \tag{5}$$

This provides a relative measure of the skill of a probability forecast, giving the proportion improvement in BS of the forecast relative to the reference forecast (e.g. climatology). At best, BSS takes a value of 1 and values less than 0 indicate the forecast performs worse than the reference over the period of assessment.

3.2 Reliability Diagram

The Reliability Diagram allows the full distributions of probability forecasts generated from the ensemble and observations to be compared for a given threshold. From the Reliability Diagram, both the Reliability and Resolution of the probability forecasts can be assessed. The Reliability Diagram is created by binning forecasts based on the forecast probability (the value for each bin is plotted on the x-axis), and then calculating the conditional probability of the observations given this binning, $P(Y \mid \hat{Y})$, plotting this on the y-axis.

Perfect Reliability (*REL*=0) is found for forecasts which fall on the diagonal line, with *REL* increasing (indicating poorer reliability) further from the diagonal. The value of sample climatology, $\bar{Y}$, is plotted

on the Reliability Diagram as a horizontal line shown in grey in Figure 1. A forecast falling on this line would have zero Resolution, whilst a forecast falling further towards the diagonal signifies it has increased Resolution. Half way between the line of sample climatology and the diagonal lies the no-skill line (not shown): along this line the Brier Skill Score (calculated using the sample climatology as a reference) equals zero.

The Reliability Diagram also indicates whether the forecasts are over- or under-confident. In the over-confident case small probabilities are under-forecast (lie above the diagonal) and large probabilities are over-forecast (lie below the diagonal). In the under-confident case small probabilities are over-forecast (lie below the diagonal) and large probabilities are under-forecast (lie above the diagonal). A systematic bias in forecast probabilities is seen when lines lie fully above the diagonal (under forecasting) or fully below the diagonal (over forecasting). These characteristics are shown schematically in Figure 1.

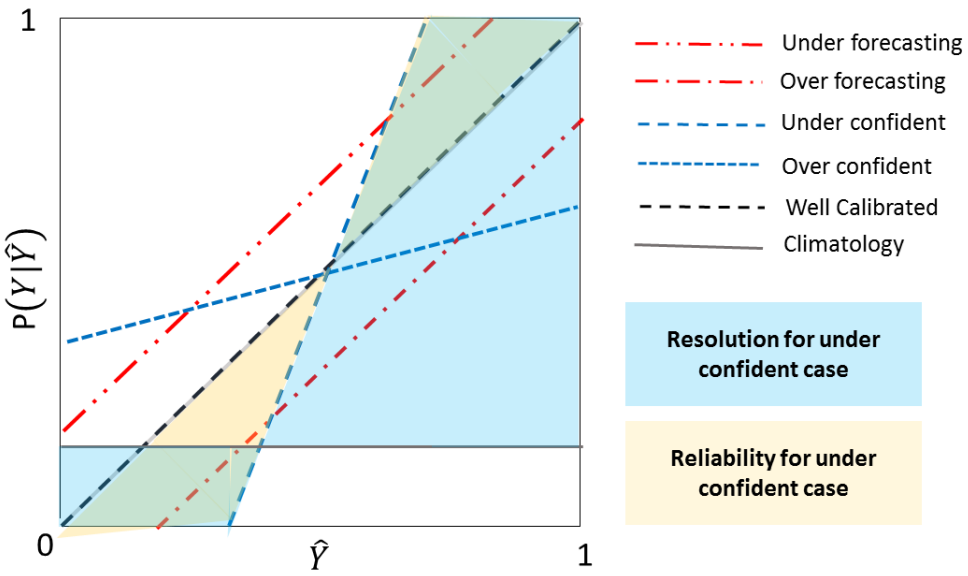


Figure 1 Schematic of a Reliability Diagram showing the key methods for identifying ensemble forecast errors.

3.3 Continuous Ranked Probability Score

A continuous form of the Brier Score follows in a natural way by considering the threshold x to be a continuous variable. We then have an indicator variable $Y_i(x)$ for the event $y_i \geq x$ occurring obtained from observations and $\hat{Y}_i(x)$ the probability of the event as stated in the probability forecast. The Continuous Brier Score (Jones *et al.*, 2003) is then defined as

$$\mathrm{CBS} = n^{-1} \sum \int (Y_i(x) - \hat{Y}_i(x))^2 dx \,. \tag{6}$$

For a single probability forecast, the integral expression is now commonly known as the Continuous Ranked Probability Score (CRPS):

$$\mathrm{CRPS} = \int (Y_i(x) - \hat{Y}_i(x))^2 dx \,, \tag{7}$$

although in practice is averaged over different forecast cases (Hersbach, 2000), and thus equivalent to CBS.

The CRPS measures the difference between the cumulative distribution estimated by the ensemble forecast, and the step-function cumulative density function of the observation. Hence, no thresholds are used to calculate the CRPS, making it a useful overview of the ensemble performance. The CRPS can be thought of as a continuous version of the Brier Score integrated over all possible thresholds. For a deterministic forecast it reduces to the mean absolute error (MAE). As with the Brier Score, it can be decomposed into Reliability (thus related to the Rank Histogram) and Resolution/Uncertainty (thus related to the average spread and outlier behaviour of the ensemble) components (Hersbach, 2000). The CRPS has the same units as the data from which it is calculated. A dimensionless skill score can be formed in the usual way, using the CRPS of a reference forecast such as climatology, giving the Continuous Ranked Probability Skill Score (CRPSS).

3.4 The Rank Histogram

The Rank Histogram (Talagrand *et al*., 1997; Hamill, 2001) assesses the reliability of the ensemble through whether the ensemble and observations have been drawn from the same distribution. If this is the case then each ensemble member forecast is equally likely, and the observation is equally likely to fall between any two ensemble members.

To create the Rank Histogram, the ensemble values are first ranked from smallest to largest for each observation point. This gives N+1 possible *bins* within which the observation may fall, for an ensemble of N members. The bin in which the observation falls is calculated for each observation point: the Rank Histogram is a histogram of the resulting bin populations. The standard interpretation of a Rank Histogram is as follows.

Flat:	The ensemble spread is appropriate to represent the forecast uncertainty.
U-shaped:	The ensemble spread is too small overall (the ensemble is under-spread) with observations frequently falling outside of the ensemble extremes.
Dome-shaped:	The ensemble spread is too large (the ensemble is over-spread) with observations frequently falling towards the centre of the ensemble distribution.
Asymmetric:	The ensemble is biased (high bias, slopes up to left; low bias, slopes up to right).

Thus, the Rank Histogram provides a qualitative, visual measure of the appropriateness of the ensemble spread. Unlike the Reliability Diagram, the Rank Histogram does not provide an absolute measure of forecast Resolution (Sharpness): a random forecast, with observations drawn from the same distribution would give a flat Rank Histogram. In some instances the standard interpretation of the Rank Histogram can be misleading: for example, where conditional biases can lead to a U-shaped histogram (e.g. Hamill, 2001). Hence, it is important to consider the Rank Histogram in conjunction with other complementary verification metrics.

3.5 The ROC Score and ROC Diagram

The Brier Score and Reliability Diagram partition the data in terms of the forecast probability (asking, given a forecast probability, what is the probability that the event was observed?). It is also possible to partition the data according to the observed events. For deterministic forecasts, this leads to the definition of the Probability Of Detection, POD (also known as the Hit Rate, *H*), and the False Alarm Rate, *F*, based on a contingency table of a sequence of binary events shown in Table 3

Table 3 Contingency table for binary deterministic forecasts and observations.

	Event observed	
Event forecast	**Yes**	**No**
Yes	*h* (Hits)	*f* (False alarms)
No	*m* (Misses)	*c* (Correct rejections)

Using the notation of Table 3, then

$$\mathrm{POD} = H = \frac{h}{h+m} \tag{8}$$

$$F = \frac{f}{f+c}\,. \tag{9}$$

The POD and *F* performance measures can be developed to also assess probabilistic forecasts of rainfall or river flow in ensemble form. For a given threshold (such as Q(2)), the POD and *F* value for different probabilities of exceedance ranging from 0 to 1 can be calculated and the paired values plotted on y- and x-axes respectively. This is called the Relative Operating Characteristics (ROC) Diagram (Figure 2).

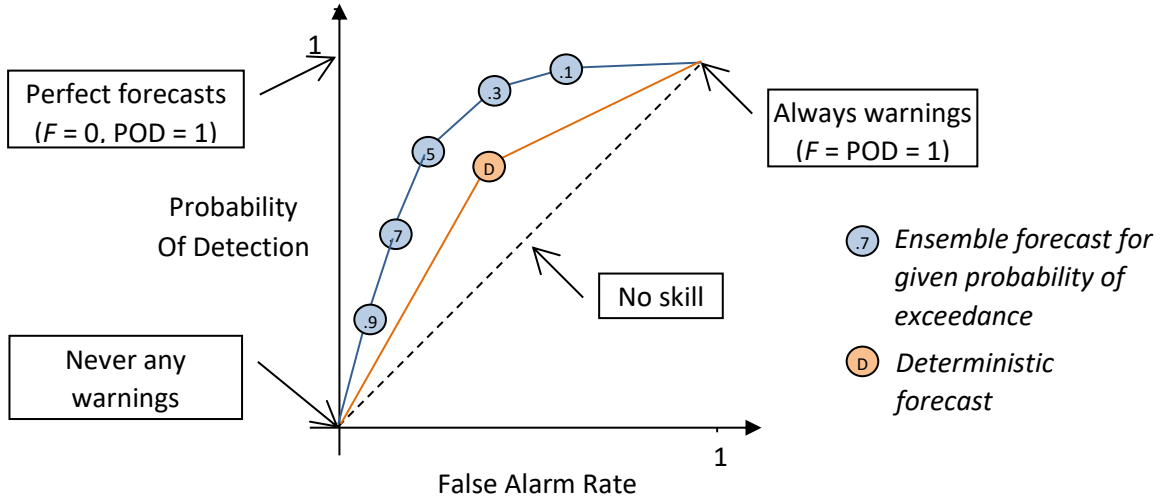


Figure 2 Schematic of the Relative Operating Characteristic (ROC) Diagram.

A perfect forecast (all events detected with no false alarms) corresponds to POD=1 and *F*=0. Thus values in the top left hand corner of the ROC Diagram are associated with highest skill, with a curve showing a significant bend towards this corner indicating forecast skill. The case when POD=*F*=1 (top right corner) corresponds to events always being warned for and when POD=*F*=0 (bottom left corner) no warning of an event is ever given. Above the 1:1 line are outcomes where POD>*F*. Note that the POD and *F* resulting from the deterministic forecast can be plotted on the same diagram as a single value for the chosen threshold.

The area under the ROC curve, AUC, is ideally 1, with a random forecast lying along the 1:1 line with an area of 0.5, and at worst 0 when POD is always 0.

The ROC Score is defined as the AUC normalised with reference to a reference forecast $\mathrm{AUC}_{\mathrm{ref}}$, taken here to be a random forecast with no skill and AUC equal to 0.5, so that

$$\mathrm{ROC\ Score} = \frac{\mathrm{AUC} - \mathrm{AUC}_{\mathrm{ref}}}{1 - \mathrm{AUC}_{\mathrm{ref}}} = \frac{\mathrm{AUC} - 0.5}{1 - 0.5} = 2 \times \mathrm{AUC} - 1.\,. \tag{10}$$

A ROC Score of 1 indicates a perfect forecast and if above 0 the forecast has a skill better than a random forecast.

3.6 Relative Economic Value

The economic benefits of a forecasting system depend on the cost-loss ratio of a particular user. The Relative Economic Value (REV) statistic (e.g. Wilks, 2001; Zhu et *al*., 2002) allows the economic forecast value to be assessed relative to a forecast based on climatological information. The REV is widely used in the verification of both hydrological and meteorological forecasts (e.g. Roulin, 2007; Magnusson *et al.*, 2014).

Like the ROC Diagram and score, the REV is based on a contingency table of a sequence of binary events. The contingency table is extended to include the associated costs *C*, and the total, protected, and unprotected losses (*L*, L_p and L_u respectively). A contingency table of the costs and losses is shown in Table 4.

Table 4 Contingency table for the costs and losses associated with binary deterministic forecasts and observations.

	Event observed	
Event forecast	**Yes**	**No**
Yes	$C+L_u$ (Mitigated loss)	*C* (Cost)
No	$L=L_p+L_u$ (Loss)	*N* (No cost)

From Table 4, the expected expense for a given forecast system can be calculated as

$$E_{forec} = h(C + L_u) + fC + m(L_p + L_u)\,. \tag{11}$$

The expense when only using climatological information is calculated as the minimum expense when protecting or not protecting against potential losses:

$$E_{cl} = \begin{cases} C + (h+m)L_u & C < (h+m)L_p \\ (h+m)L_p + (h+m)L_u & (h+m)L_p \le C \end{cases}. \tag{12}$$

With a perfect forecasting system the user would only take action, and incur costs, when an event occurred. In this situation, the expenses would be

$$E_{perf} = (h+m)(C + L_u)\,. \tag{13}$$

E_{forc}, E_{clim} and E_{perf} are combined to give the REV as

$$REV = \frac{E_{cl} - E_{forc}}{E_{cl} - E_{perf}}\,. \tag{14}$$

Substituting Equations (11) to (13) into Equation (14) and defining the cost-loss ratio $r = C / L_p$ gives

$$REV = \begin{cases} \dfrac{r - (h+f)r - m}{r(1-(h+m))} & r < h+m \\ \dfrac{h+m-(h+f)r-m}{(h+m)(1-r)} & h+m \le r \end{cases}. \tag{15}$$

The REV has a maximum value of one for a perfect forecasting system, a value of zero for a forecast with the same value as climatological information only and is negative for forecasts which have less value than using only climatological information.

4. Quantifying sampling uncertainty using the Bootstrap method

One key question for an operational forecast verification system is "what sample size should be used?". That is, "how many ensemble forecasts should be assessed to obtain meaningful verification?". To help address this question, the sampling uncertainty can be estimated using the bootstrap method. The bootstrap method is based on the principle that sub-samples of the verification data relate to the verification data in the same manner that the verification data itself relates to a much larger sample (the underlying data distribution). By considering the distribution of verification statistics obtained from sub-samples of the verification data, we approximate the distribution of verification statistics which would be obtained by taking different samples of the underlying data distribution. Hence, the spread of verification statistics obtained from sub-samples of the verification data approximates the spread of verification statistics which would be obtained from samples of the underlying data distribution: the sampling uncertainty.

To calculate the bootstrap sampling uncertainty, multiple random samples are selected from the verification data, and the verification statistics calculated for each sample. Samples are selected with replacement, as the occurrence of an event within the verification period does not preclude that event occurring at another time outside of the verification period. Thus, within a bootstrap sample, some of the verification data will be drawn multiple times, and some not at all. The distribution of verification statistics calculated for all bootstrap samples is used to estimate confidence intervals around verification statistics calculated using the full verification data. In this report the 75^{th}, 90^{th} and 99^{th} percentiles are considered, and bootstrapping is used to estimate the uncertainties associated with ROC and Reliability Diagrams. For the verification of river flow forecasts, 500 bootstrap samples are used. It was found that this number of samples were sufficient to approximate the distribution of verification statistics, with similar results obtained for larger sample sizes. For the verification of daily precipitation accumulations 1000 bootstrap samples are used. For hourly precipitation accumulations further investigation suggested that 100 bootstrap samples were sufficient for the Reliability Diagrams and 20 for the ROC Diagrams. The median of the distribution of the bootstrap sample verification statistics was found to approximate well the verification statistics calculated using the full verification period.

5. Calculation of verification metrics

To enable comparison of river flow and rainfall verification metrics, it is essential that the verification metrics detailed in this Appendix are calculated using common code. In this study the verification metrics are calculated using the R package "verification" (NCAR - Research Applications Laboratory, 2015). The functions used to calculate each verification metric, and the output variables used, are summarised in Table 5. Full details of the functions, and function source code, are available from NCAR - Research Applications Laboratory (2015).

Table 5 R verification package functions and output variables used for the calculation of each verification metric.

Verification metric	R verification package function	Package output variables used
BS	brier	BS: bs REL: bs.reliability RES: bs.resol UNCERT: bs.uncert
BSS	brier	BSS: ss
Reliability Diagram	brier	Forecast probability: y.i Observed relative frequency: prob.y Climatological rate: obar.i
CRPS	crpsDecomposition	CRPS: CRPS
ROC Diagram	roc.plot	H: plot.data[,2,] F: plot.data[,3,]
ROC score	roc.area	ROC score: roc.area
REV	value	REV: V r: cl r for maximum V: s

References

Centre for Ecology & Hydrology 2016. Performance assessment of the G2G Model: a guide to the Performance Summary. Report to the Flood Forecasting Centre, Centre for Ecology & Hydrology, Wallingford, UK, 13pp.

Hamill, T.M. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3), 550-560.

Hersbach, H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5), 539-569.

Jones, A.E., Jones, D.A. and Moore, R.J. 2003. Development of Rainfall Forecast Performance Monitoring Criteria. Phase 1: Development of Methodology and Algorithms. Report to the Environment

Murphy, A.H. 1973. A new vector partition of the probability score. *J. Appl. Meteorol.*, 12, 595–600.

NCAR - Research Applications Laboratory. 2015. verification: Weather Forecast Verification Utilities. R package version 1.42. http://CRAN.R-project.org/package=verification

Roulin, E. 2007. Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrol. Earth Syst. Sci.*, 11, 725-737.

Siegert, S. 2017. Simplifying and generalising Murphy's Brier score decomposition. *Q. J. R. Meteorol. Soc.*, 143, 1178-1183.

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

Wilks, D. S. 2001. A skill score based on economic value for probability forecasts. *Meteor. Appl.*, *8*(2), 209-219.

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

Rainfall and River Flow Ensemble Verification: Phase 2
Events for possible case studies
Final Report Appendix A.2

Scotland		
Dates	**Location**	**Comment**
3 & 4 Jan 2016	Tayside and Angus	Flooding in Perth. High levels in North Muirton.
7 & 8 Jan 2016	Aberdeenshire	Unremarkable rainfall on saturated catchments. Record high flows on the rivers Don, Ythan and North Esk. FGS Red for Aberdeenshire. Evacuations took place.
27 Jan 2016	Scottish Borders	Severe Flood Warning issued for Jedburgh. Frontal rain with associated convection in one forecast run, then observed in radar.
6 & 7 June 2017	Findhorn, Lossie and Nairn catchments	Hydrologically significant event. FGS 3x2 (low,sig). Flood defences at Forres and Elgin prevented flooding in these places. River Nairn came close to overtopping defences.
23 to 25 Jan 2018	Scottish Borders	Several linked events. Snowmelt overnight 23-24 Jan.
8 Oct 2018	West and North-West Scotland	FGS amber for Inverclyde & N Ayrshire, yellow for west Highlands.
22 & 23 Oct 2018	Strathoykel (NW Highlands)	High 6-12 hour rainfall totals – highest totals very localised, very localised impacts. 1-in-20 year flow on River Oykel – peaking at 0800 on 23 Oct. Limited impacts.
Wales and England		
Dates	**Location**	**Comment**
8 Aug 2017	Essex, Kent, Surrey, London	EFAS Flash Flood notification. Minor impacts from surface

		water flooding: Essex, Kent, Surrey, Greater London, Suffolk. No fluvial impacts recorded.
23 Aug 2017	Scarborough	Convective event - no river impacts recorded. Flash flooding from surface water, causing travel disruptions. Significant impacts - North Yorkshire (Scarborough), minor for York and West.
30 Sep 2017	Cumbria, Millom	Narrow band of heavy rain over the south of Cumbria. 40-45 mm in 1 hour and 60+ mm in 2-2.5 hrs at Millom and Haverigg between 0800 and 1030 BST. A total of 150-200 properties affected by surface water flooding, largely in the Millom area and with a small number of flooded properties in Windermere and Haverigg.
21 Oct 2017	Storm Brian, Lancashire & W Yorkshire	Missed SIG event. Flood Sirens in Todmorden, Hebden Bridge and Mytholmroyd. 4 properties (river), 1 industrial building (river), 1 pre-school (SW) and a bakery (SW) flooded in Rossendale plus 5 properties in Rawtenstall from river and 8 in Rawtenstall from surface water
3 & 4 Nov 2017	SE England	Rainfall false alarm? No impacts noted. G2G Deterministic supported minimal impacts. Some suggestion of higher flows from G2G ensembles.
22 & 23 Nov 2017	NW England and N Wales	Impacts on 23 recorded as SIG over Cumbria, Lancashire for rivers and also minor for Cumbria, Anglesey, York, and Lancashire. Widespread issues from surface water flooding.

27 Dec 2017	E and SE England	No river flood impacts noted. Minor impacts from surface water flooding: Midlands, SW and SE England.
2 & 3 Jan 2018	Storm Eleanor	EFAS Flash Flood notification. No fluvial impacts recorded. Little response in G2G deterministic but larger response in G2G ensembles.
12-14 Mar 2018	SE and SW England and Derbyshire	Widespread flooding around Burton, South Derbyshire. G2G gave poor advice on the fluvial flood risk in Staffs/High Peaks area. Unusually rapid response given the amount of rain, which lead to quite a few minor impacts from surface water and river flooding. Once the rain was in the gauges, G2G then significantly increased the response within the gridded MRDET through this area.
2-4 Apr 2018	SW, Central and NE England and Wales	Minor river flooding impacts noted on 3 April: N and W Yorkshire and on 4 April in N Yorkshire, Durham and Tyne and Wear. Minor roads around Linton-on-Ouse closed due to flooding from small streams and high flows on River Ouse. Surface water flooding caused closures or partial closure of arterial A roads around Bishops Auckland (Durham) and Tyne & Wear area.
18 Jul 2018	Cornwall, Coverack	Flash Flooding in Coverack with danger to life, helicopter rescues, 50 properties flooded, damaged roads and infrastructure.
20 Sep 2018	Mid Wales, Sheffield, Storms Ali and Bronagh	Primarily, surface water flooding. Transport disruption in Sheffield and Rotherham. Some road flooding in Wales

		(Pontypridd).
12 & 13 Oct 2018	SW Wales, Storm Callum	Main river impacts occurred on 13 Oct 2018 in SW Wales - Powys, Ceredigion and Carmarthenshire. Properties flooded by main rivers. River Towy in Carmarthenshire main focus.
9 Nov 2018	SW England and SW Wales	Frontal orographic + convection rainfall. SIG river impacts recorded in Pembrokeshire, minor in Carmarthenshire. SIG impacts from surface water flooding in Pembrokeshire, with minor in Devon and Cornwall.

Rainfall and River Flow Ensemble Verification: Phase 2
Catchment precipitation processing
Final Report Appendix A.3

1 Overview

Under Phase 2 of the "Rainfall and River Flow Ensemble Verification" project, it was decided that UKCEH would undertake the catchment precipitation processing for the Phase 2 Verification Period (1 June 2017 to 30 September 2018). The gridded precipitation time-series to be processed were the Best Medium Range (BMR) ensemble forecasts and gridded rainfall observations from three sources: raingauge, radar and merged. Specifically, the catchment values were to be extracted from daily and hourly precipitation accumulation grids and saved for use by the Met Office for precipitation ensemble verification.

This document describes the processing methods used, and provides details of the files provided. The primary aim is to serve as a reference guide to the use and interpretation of the files. However, Sections 2 and 3 also provide background information on the available forecast and observed precipitation data, the domains and catchments considered, details of the re-gridding process from 1km to 2km grid-spacing, calculation of hourly and daily precipitation accumulations and threshold-calculations.

2 Precipitation data for processing

2.1 Observed precipitation data

Four types of observed precipitation data were processed for the Phase 2 Verification Period, as detailed in Table 1. To allow comparisons to be made with the Phase 1 (December 2015) analyses, the raingauge and radar precipitation products were also processed for that period. The merged product was not available for use in December 2015.

Table 1 Gridded observed precipitation products processed to obtain catchment values.

Observed precipitation product		Coverage
hkuk_g2g	Hyrad gridded raingauge-rainfall (mm/h)	England & Wales
hkscot_g2g	Hyrad gridded raingauge-rainfall (mm/h)	Scotland
H19	15-min advection accumulation radar-rainfall(mm)	England & Wales, and Scotland to just north of Inverness (up to 879500N)
H23	Merged raingauge-radar rainfall with 1h delay (mm)	England & Wales, and the very south of Scotland (up to 700500N)

These 15-minute gridded time-series data are held at UKCEH in Hyrad SIDB (Spatial Image DataBase) form, on the native 1 km British National Grid.

2.2 BMR precipitation ensemble forecast data

BMR precipitation ensemble forecasts were obtained from MASS for the period 1 June 2017 to 3 September 2018, encompassing the year-long verification period and case-study events of interest. Data from all forecast-origins within this period were processed.

The BMR data are produced on a 2 km British National Grid out to a lead-time of around 6.5 days in Nimrod format, with files containing both 15-minute and 1-hour precipitation accumulations. Beyond a lead-time of 36h, the 15-minute accumulations are created by evenly splitting the 1-hour

accumulations over the 15-minute periods they contain. For the rainfall processing documented here, only 15-minute accumulations are extracted from the Nimrod files.

2.3 Accumulation periods

All data to be processed are available at 15-minute intervals, either as an accumulation over the previous 15 minutes, or as an average rain-rate over that 15-minute period expressed in units of mm/h. To allow precipitation verification of 15-minute, hourly and daily accumulations, hourly and daily quantities were calculated by accumulating the 15-minute values. The accumulations were calculated as follows

Observed data

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

Forecast data

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

For the precipitation catchment processing, all 15-minute and hourly data are converted to units of mm/h, and daily data are converted to mm/d.

3 Methods for precipitation processing

3.1 Re-gridding of precipitation products

It is recognised that the distribution of rainfall intensity will vary as a function of spatial resolution. Thus, to accurately assess the performance of precipitation forecasts, they should ideally be at the same spatial resolution as the observed data.

Upscaling of these 1 km observation precipitation data to the 2 km resolution of the precipitation forecasts to support verification at this resolution has been given careful consideration, noting the points below.

- G2G forecasts are driven by the BMR, which is on a 2 km grid.
- Scientifically, downscaling the precipitation forecast to fit to the observations is not generally recommended, though the Met Office recognise this is precisely what is done to MOGREPS-G forecasts at longer lead-times. The reason for this downscaling is not for verification though, but to create a “seamless” forecast product to drive G2G out to ~7 days. Further downscaling would not be recommended. No detail is being added to the forecast whereas for the observation fields it is generally true that the detail and intensities increase with increasing horizontal resolution.
- The hydrological requirement for precipitation forecasts, in a G2G context, is ideally for a 1 km product. G2G is configured at 1 km resolution through its supporting static spatial datasets on terrain/soil-geology/land-cover and use of dynamic observation sources of precipitation. Perfect foreknowledge of rainfall, assumed in forecast assessments of G2G performance to remove the uncertainty of precipitation forecasts, also employs a 1 km resolution observation source of precipitation as input. Thus, precipitation verification from a hydrological perspective, in this G2G modelling and forecasting context, argues for assessing the 2 km precipitation forecast product against an observation precipitation truth at 1 km scale.

Based on these points, it was decided to process the observed precipitation data both at their native 1 km resolution, and also when up-scaled to a 2 km resolution. The up-scaled version will be used for

the precipitation processing only, with the native 1 km resolution observed precipitation data (raingauge, radar, merged) used as input to G2G for maintaining the initial conditions, as was done in Phase 1 to reflect operational use.

To up-scale the observed precipitation data to the 2 km grid of the BMR ensemble, the average was taken over the four 1 km grid-cells falling within each 2 km grid-cell. As the raingauge-rainfall data only covers 1 km grid-cells falling over land, there are instances where a 2 km grid-cell contains 1 km grid-cells with missing data ("sea-cells"). In this case, the average is taken over land-cells only.

Re-gridding examples are given below in the upper row of Figure 1. To enable the 2 km resolution data to be applied to the 1 km G2G catchment boundaries, the up-scaled data were re-projected back onto the 1 km grid. This is shown in the bottom row of Figure 1, with an example G2G catchment (IWS.Freshw; Western Yar at Freshwater on the Isle of Wight) shown in pink in the right-hand plot.

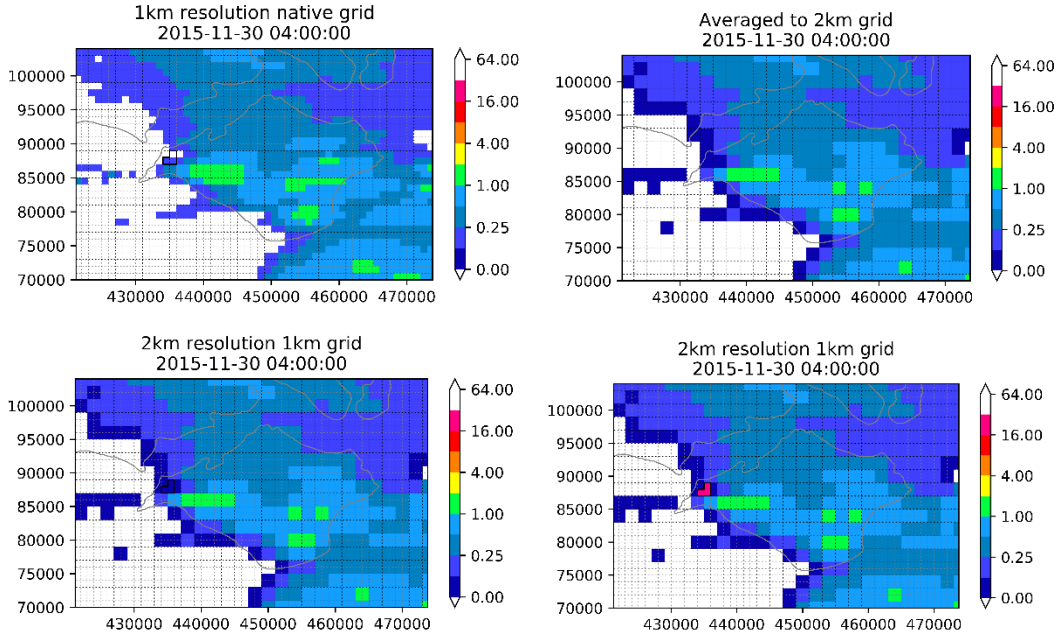


Figure 1 Examples of the different resolution grids used to process catchment observed precipitation. The native 1 km resolution grid is shown first (top left) followed by the same data averaged onto a 2 km resolution grid (top right). These data are then projected back on to the native 1 km grid (bottom left) to allow the 1 km G2G catchment boundaries to be used (e.g. IWS.Freshw shown in pink in bottom right).

3.2 Catchment boundaries

Data were processed for all G2G catchments, and also selected PDM catchments provided by the EA, SEPA and NRW. For each precipitation product processed, all catchments falling within its product domain were used. These catchments for the observed precipitation products are shown in Figure 2.

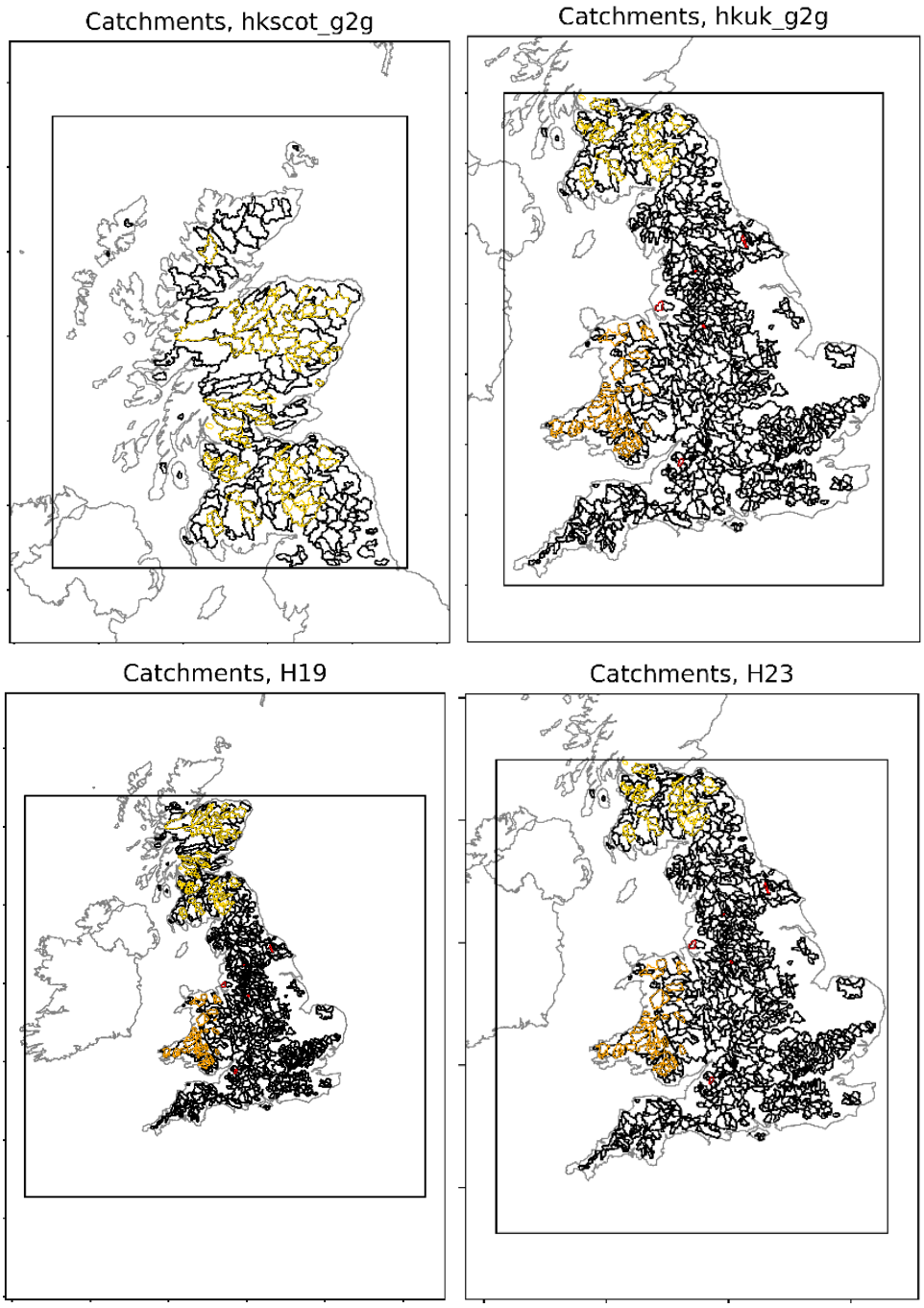


Figure 2 Catchments for which data were processed for each of the observed precipitation products. G2G catchments are shown in black, and PDM catchments in gold for Scotland, in orange for Wales and in red for England.

The G2G catchment boundaries are produced on a 1 km grid, so match the native-resolution observed precipitation data exactly. To apply the 1 km G2G catchment boundaries to the 2 km gridded BMR ensemble forecast data, the BMR data are first projected onto a 1 km grid, with the value of each 2 km grid-cell being assigned to the four 1 km grid-cells covered by it.

The PDM catchment boundaries are true catchment boundaries, and not digitised onto a 1 km grid. When calculating catchment-average rainfall operationally for input to the PDM models (using the Hyrad CatAvg tool) a weighted approach is used, with grid-cells straddling the catchment boundary contributing to the catchment-average with a weighting given by the fraction of that grid-cell falling

within the catchment. For consistency with this operational method, the grid-cell weightings were extracted from CatAvg for all the PDM catchments considered, and used to create a weighted catchment-average rainfall. However, for the other rainfall statistics to be computed (including rainfall percentiles, and the number of cells exceeding a given rainfall threshold, see sections 3.3 and 3.4) a weighted approach is not practical. Thus, for these quantities, a 1 km grid-cell is considered to be "within" a PDM catchment boundary if the mid-point of that grid-cell is within the catchment boundary. For consistency, and to allow the effect of these different methods to be investigated, the catchment-mean was also calculated with the "mid-point" method. Of course, these differences will have most affect for small catchments. Figure 3 shows an example for a large catchment (Alt at Kirkby, 694744, left) and the smallest catchment considered (Earby Beck at Earby Youth Hostel, L1546, right). It can be seen that for Earby, 5 grid-cells contributed to the "mid-point method" mean, whereas 11 grid-cells contribute to the weighted mean (6 with weightings greater than 0.2).

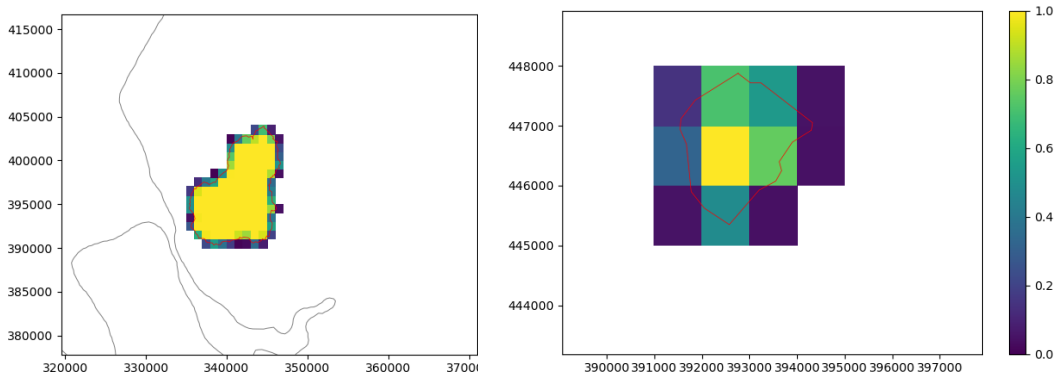


Figure 3 Example PDM catchment grid-cell weightings extracted from CatAvg: Alt at Kirkby (694744) left and Earby Beck at Earby Youth Hostel (L1546) right.

3.3 Within-catchment precipitation distribution

Within-catchment precipitation distribution properties were calculated separately for each catchment from all grid-cells each contain, for each precipitation accumulation grid (15-minute, hourly and daily). The following distribution properties were calculated

- Weighted mean
- Mean
- Percentiles: 50, 75, 90, 95, 99

An example is shown in Figure 4 for the Eden at Sheepmount (G2G ID 765512) during Storm Desmond (5 December 2015). Examples are shown for the raingauge-rainfall and for the BMR ensemble member 00 forecast origin 01:00 4 December 2015 for 15-minute, hourly and daily precipitation accumulations.

The within-catchment distribution data will be used by the Met Office for plotting time-series of catchment precipitation, and calculating the non-threshold-based scores (CRPS, Rank Histogram).

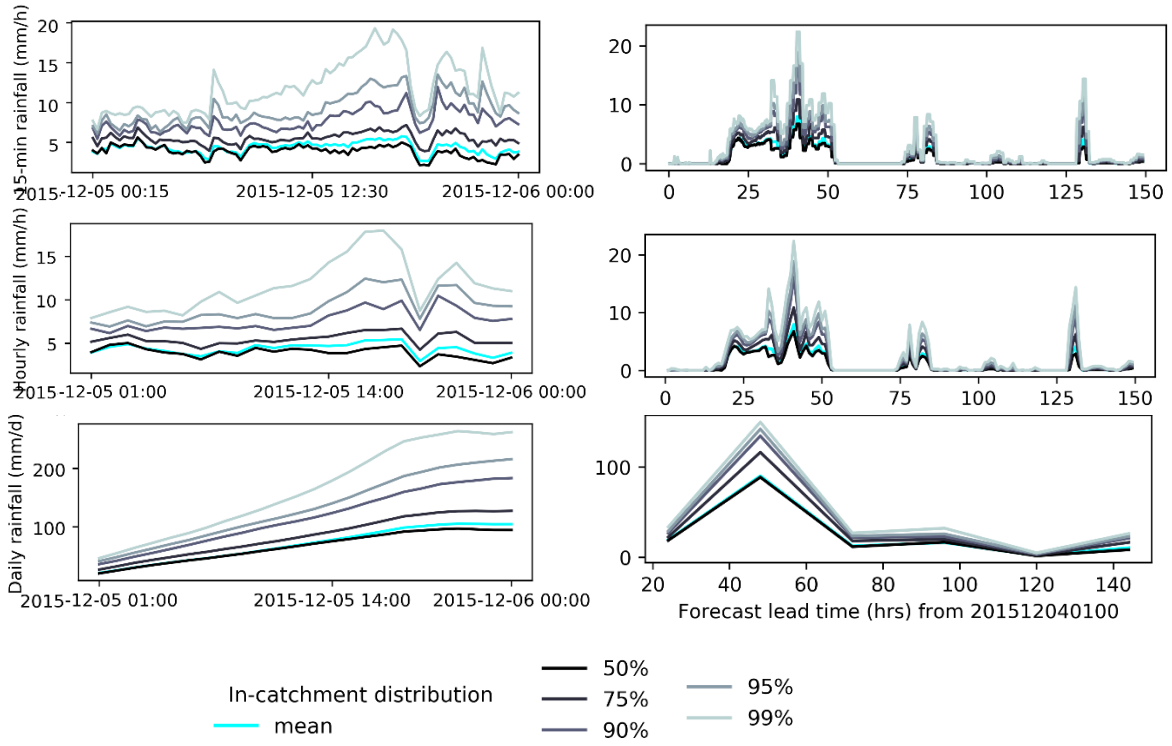


Figure 4 Example of in-catchment distribution properties for the Eden at Sheepmount (G2G ID 765512) for raingauge-rainfall data for 5 December 2015 (left) and for the BMR ensemble member 00 forecast origin 01:00 4 December 2015. Results are shown for 15-minute (top), hourly (middle), and daily (bottom) precipitation accumulations.

3.4 Precipitation threshold calculations

To focus on potentially flood-producing rain, it is necessary to choose appropriate precipitation thresholds. For the catchment-processing, two methods of calculating thresholds are considered: firstly, fixed precipitation values commonly used in precipitation verification and, secondly, catchment-specific thresholds based on the climatological distribution of precipitation for that catchment.

3.4.1 Fixed catchment-precipitation thresholds

Fixed-value thresholds were applied to all catchments considered. For 15-minute and hourly accumulation data, thresholds of 0.1, 1 and 4 mm/h were used, and for daily accumulations thresholds of 0.1, 1, 4 and 8 mm/d were used.

3.4.2 Climatological catchment-precipitation thresholds

Considering the climatological distribution of precipitation for each catchment provides guidance on the variability of precipitation across the UK, and enables a threshold to be calculated that selects the "extreme" precipitation values for that catchment, noting that what is extreme for a low-rainfall catchment may be normal for a wetter catchment.

To create climatological catchment-precipitation thresholds 10 historical years of raingauge-rainfall data from 2007 to 2016 were used. To facilitate this task computationally, each date in the 10-year historical period was separately processed, and the distribution of precipitation values for all grid-cells falling within each catchment saved as a histogram. This was done separately for 15-minute, hourly

(ending on the whole hour) and 24h (ending at 00:00) precipitation accumulations. The histogram bins used were as follows.

> **15-minute and hourly data:** -0.5 to 0.5, 0.5 to 1.5, 1.5 to 2.5,… 49.5 to 50.5 (mm/h)
> **Daily data:** -1 to 1, 1 to 3, 3 to 5 … 199 to 201 (mm/d)

Thus, the sub-daily histograms have a bin-size of 1 mm/h, centred on each whole number of mm/h, whilst the daily histograms have a bin-size of 2mm/d, centred on each even number of mm/d.

To produce the full climatological precipitation distribution for each catchment, the histogram values from each historical year were summed for a given day-in-year. These files are saved in .csv format and are available for the five catchment-sets (G2G catchments in England & Wales and Scotland, and PDM catchments in England, Wales and Scotland) considered (Section 3.2)

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

Overall, a noticeable variation is seen from west to east across the domain, as shown in Figure 5 for an example of the daily precipitation accumulations at the 99th percentile threshold for G2G England and Wales catchments. Precipitation thresholds for spring are generally lower than for other seasons across all parts of the domain, whereas thresholds in autumn and winter are higher for northern England and Wales.

To give an overview of the gradual variation of the seasonal (91 day) precipitation thresholds over the year, Figure 6 shows the rainfall values corresponding to the seasonally-varying 99th and 95th percentile thresholds, with one line drawn per catchment. As there is a larger range of daily precipitation values, covering a larger range of bins in the daily histogram, this plot appears smoother than those for the hourly and 15-minute accumulations. In this example, as is often seen, the hourly and 15-minute precipitation accumulation thresholds are often zero for the majority of sites when the 95th and 90th percentiles are considered. This suggests that these percentiles may be too low to be useful for evaluating the 15-minute and hourly precipitation accumulations.

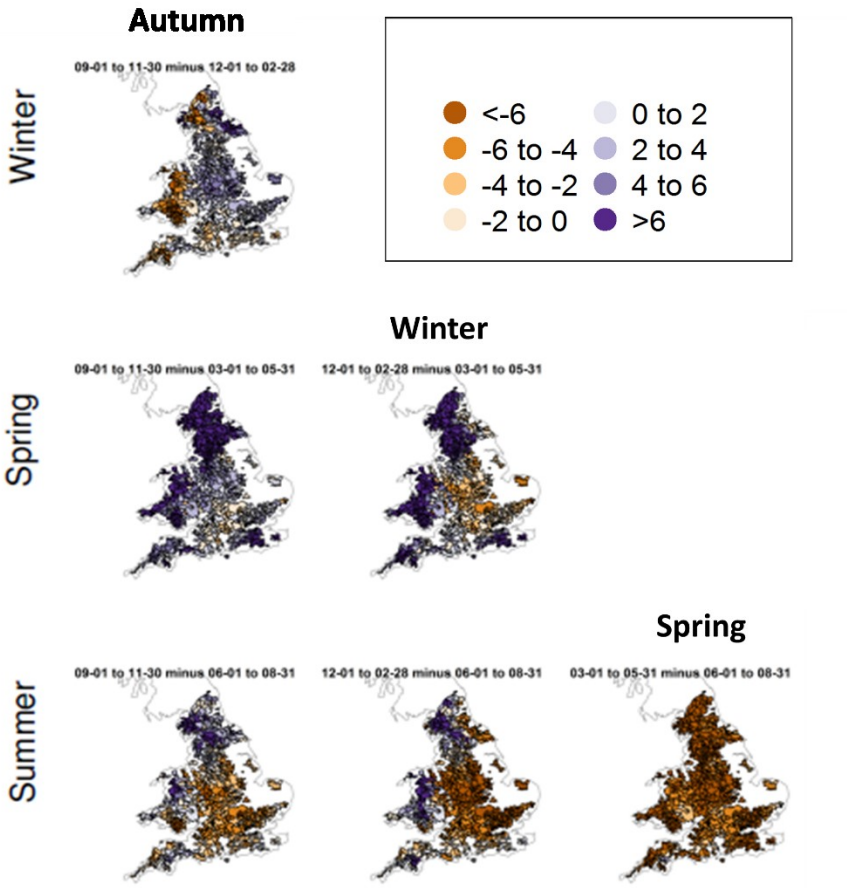


Figure 5 Example maps showing the difference in the 99th percentile daily accumulation precipitation-threshold values (mm/day) between different seasons. The threshold for the season indicated by the row is subtracted from that for the season indicated by the column.

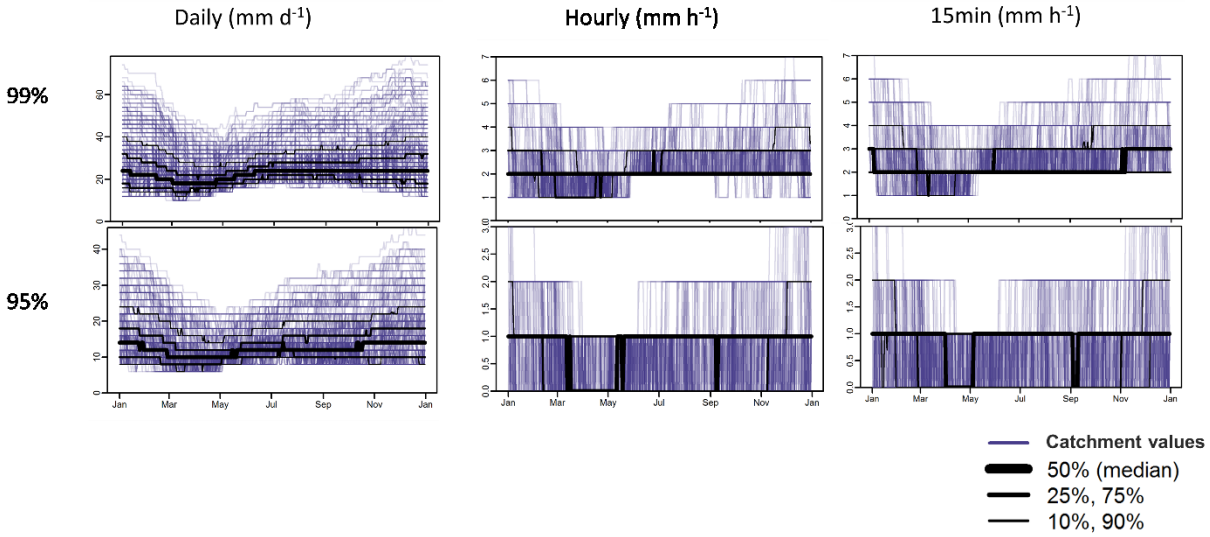


Figure 6 Example for G2G England and Wales catchments showing how seasonal catchment precipitation-thresholds vary with time of year from January to December. Results are shown for the 99th percentile threshold (top) and 95th percentile threshold (bottom) for daily (left), hourly (centre) and 15-minute (right) precipitation accumulations. Thin grey lines are shown for each catchment individually, with black lines indicating the median, quartiles, and 10th and 90th percentiles across all catchment values.

3.5 Precipitation threshold exceedance

In the catchment precipitation processing, the number of grid-cells in each catchment exceeding (≥) the precipitation threshold were calculated separately and saved for each catchment: for each forecast lead-time, accumulation period and ensemble member, and for each observation time. Saving the number of grid-cells gives flexibility for future work to look into the effects of using "time-window threshold exceedance", or requiring a different number of grid-cells in a catchment to exceed a threshold to count as an "event". In particular, these data will be used by the Met Office to calculate threshold-based scores, combining over the forecast assessment periods selected (Day 1, Days 2-3, Days 4-6). This aligns with the treatment of forecast assessment periods for the G2G river flow verification. For each catchment (and ensemble member and lead-time/time-of-day) the total number of grid-cells which exceed the precipitation-accumulation exceedance-threshold will be saved. This is consistent with Met Office code and will provide the flexibility for testing different data-quality cut-offs (especially relevant for radar rainfall data) to define whether an "event" has occurred. Generally, a single grid-cell exceedance is unlikely to be considered as a good enough "event definition".

All of the selected precipitation-thresholds will be applied to each grid-cell in the catchment to create a catchment-grid of binary values: assigning 1 if the threshold is exceeded (≥), 0 otherwise. The output for each catchment and threshold will be the sum of 1-values in that catchment. An illustration of the method for deriving ensemble-based exceedance probabilities of catchment rainfall in a time-window is shown in Figure 7. This considers deriving the time-window probability from daily forecasts for Days 4, 5 and 6 (the rows). For simplicity, the ensemble is assumed to have five members (the columns). At least two grid-cells must exceed (≥) the threshold to avoid spurious exceedances. Beyond that, the time-window exceedance works on the principle that an event for the time-window occurs when there are any exceedances within it.

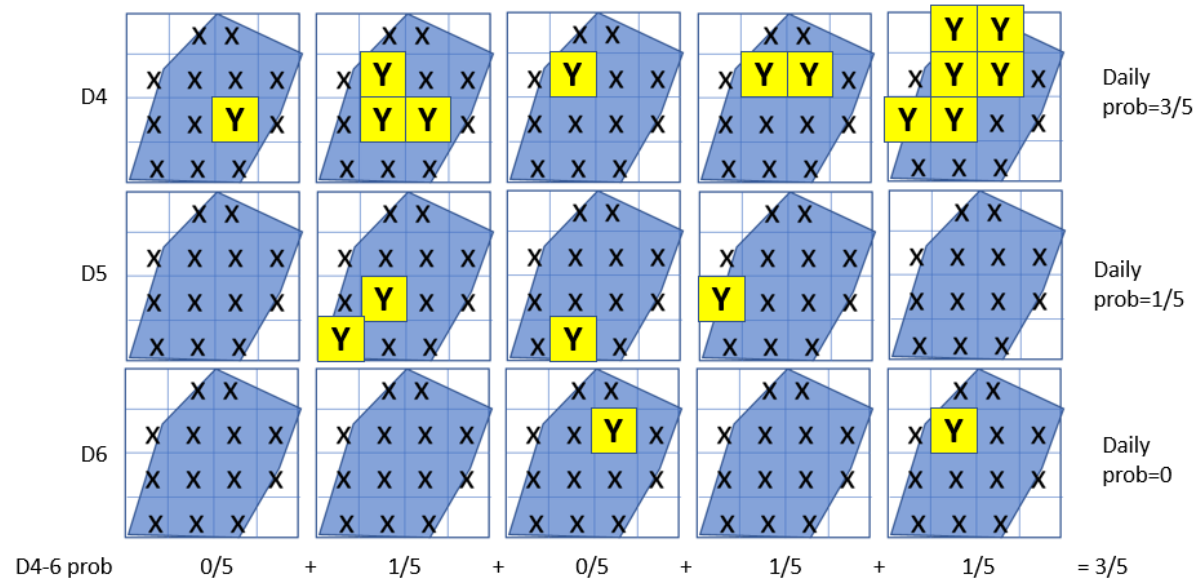


Figure 7 Schematic illustrating the difference between an accumulation-period and time-window probability using daily accumulations for a theoretical catchment (blue) where yellow points marked "Y" exceed a particular precipitation threshold, and other points marked X don't.

4 File structures

This section provides a brief overview of the file-structures used in the precipitation processing, and a guide to the contents of the output files.

4.1 Input files

4.1.1 Catchment shapefiles

There are five catchment shapefiles.

fullcats_g2g_ffc_v1_7_loc.shp	G2G 1km grid catchments for England & Wales
fullcats_g2g_sepa_v1_6_3_loc.shp	G2G 1km grid catchments for Scotland
HyradCatchment.shp	PDM catchments for England
sepa.shp	PDM catchments for Scotland
Wales_PDM_Catchment_Boundaries.shp	PDM catchments for Wales

4.1.2 Catchment look-up tables

There are 20 catchment look-up tables, one per data type per catchment shapefile. The catchment look-up tables provide a table of catchments with associated grid-cell locations (relative to the grid of a specific precipitation product) whose central-point lies within that catchment. This table then provides a fast and computationally efficient method of selecting an individual catchment from the full gridded data array: around 100 times faster than using the original shapefiles to select grid-cells. One catchment-lookup table is provided for each combination of shapefiles and the observed precipitation products. The file labelling is as follows.

***Shapefiles_product**.nc, with*

Shapefiles

fullcats_1km_grid	G2G 1km grid catchments for England & Wales
fullcats_1km_grid_sepa	G2G 1km grid catchments for Scotland
fullcats_pdm_grid_E	PDM catchments for England
fullcats_pdm_grid_S	PDM catchments for Scotland
fullcats_pdm_grid_S	PDM catchments for Wales

Products

hkuk_g2g	Raingauge-rainfall for England & Wales
hkscot_g2g	Raingauge-rainfall for Scotland
H19	15-minute advection accumulation radar rainfall
H23	Merged raingauge-radar rainfall with 1h delay

Within each .nc catchment look-up file, there is an array for each catchment, labelled by the ID from the original shapefile with dimensions [*grid* , *number of grid-cells in catchment*] where *grid* is the y and x coordinates of the grid-cell (relative to the bottom left pixel of the observed precipitation grid).

4.1.3 Catchment look-up tables with weightings

There are 15 catchment look-up tables with weightings, one per data type. Separate files are employed to contain the grid-cells used to obtain the weighted-means for each PDM catchment. The structure is similar to the grid-cell-centre based catchment look-up tables (Section 4.1.2), but with the prefix “weighted”, that is weighted***Shapefiles_product**.nc.* The dimensions of the arrays saved in this file are now [*weighted grid*, *number of non-zero weighted grid-cells in the observed precipitation grid*],

where *weighted grid* is the y-coordinate, x-coordinate, and weighting of each grid-cell. Note that these values are saved as *floats* so the coordinates need to be converted to *integers* before using to access elements in the gridded precipitation arrays.

4.1.4 File containing information for ALL catchments for a given product

To simplify bulk-processing, look-up tables were also produced with one file per gridded precipitation type containing all catchments (i.e. both PDM and G2G) falling within the domain of that data. To ensure unique catchment IDs, the IDs taken from the shapefiles are appended as follows.

_PDM_E	PDM England
_PDM_W	PDM Wales
_PDM_S	PDM Scotland
_G2G_FFC	G2G England & Wales
_G2G_SEPA	G2G Scotland

Note that when combining the *weighted* files, only the PDM catchments are used.

For the BMR ensemble precipitation grids, only catchment look-up files containing *all* the catchments are produced, with filenames fullcats_all_bmr_XX and weightedfullcats_all_bmr_XX. Unlike the all catchment-files for the observed precipitation grids which are a simple combination of the individual catchment-type files, the BMR look-up tables are calculated by converting the catchment look-up tables for the hkuk_g2g and hkscot_g2g products to the BMR grid. This method was needed due to the prohibitive size of the BMR grid for processing the individual catchment shapefiles. It is important to note that, like the BMR data obtained from the Nimrod files, the origin of these BMR grids is in the ***top left-hand corner,*** differing from the ***bottom left-hand corner*** used for the observed precipitation grids.

4.1.5 Files containing the catchment-specific seasonal and annual thresholds

Files containing the catchment-specific thresholds are saved in .csv format. The files are organised into folders identifying the catchment-set used for their creation as follows.

summary_values_E_PDM	PDM England
summary_values_W_PDM	PDM Wales
summary_values_S_PDM	PDM Scotland
Summary_values_E_G2G	G2G England & Wales
Summary_values_S_G2G	G2G Scotland

The files are named as follows.

Seasonally varying 91days_centred_***MMDD***_catchment_thresh_***accum***.csv
Annual Full_year_catchment_thresh_***accum***.csv

Where ***accum*** is the precipitation accumulation (daily, hourly, quart) and ***MMDD*** is the day and month upon which the 91day window for threshold-calculation is centred. Each file contains three columns containing the precipitation threshold values corresponding to the 90^{th}, 95^{th} and 99^{th} percentile of the climatological precipitation distribution for that catchment and time-period. These are in units of mm/h for 15-minute and hourly accumulations, and mm/day for daily accumulations. There is a separate row in the file for each catchment in the catchment-set being considered.

4.2 Output files

The format of the output files is described briefly below. Any missing data values are indicated by the "NaN" value.

Observed precipitation files

Output files for the catchment precipitation processing of the observed precipitation products are saved in directory structures for a given year ***YYYY***, month ***MM*** and day-in-month ***DD***

Native 1 km resolution data: output_data_obs/full_1km_res/***YYYY***/***MM***/***DD***
1 km grid data averaged to 2 km resolution: 2km_res_1km_grid/full_1km_res/***YYYY***/***MM***/***DD***

Within these directories, there is one file per precipitation type per accumulation period
sidb_***YYYYMMDD_Product_accum***_catchment_precip_stats.nc

For ***accum*** *daily*, *hourly* and *quart* and ***Products***

hkuk_g2g	Hyrad raingauge-rainfall	hyradk_nxm_raingauge_surface_15_min_1km_grid_hkukg2g
hkscot_g2g	Hyrad raingauge-rainfall	hyradk_nxm_raingauge_surface_15_min_1km_grid_hkscotg2g
H19	15-minute advection accumulation radar	_u1024_ng_radar_15min_advect_accum_1km_ukpprain_
H23	Merged raingauge-radar with 1h delay	nimrod_ng_radar_merged_accum_composite_24hrdelay_1km _UK_cutout_775X640_eng_Observation

A summary of the file contents is given in Table 2.

Table 2 Summary of the file contents of the output from catchment precipitation processing of the observed precipitation products.

Coordinates	
sites	List of catchments used for the catchment-statistics in this file. Matches those in the all_catchments look-up tables (Section 4.1.4)
times_*accum*	End-time of the accumulations (***accum***) used
precip_thresh	Fixed-value precipitation thresholds used (in **Precip_units**, below)
season_thresh_perc	Percentile thresholds used to calculate the seasonally varying catchment-specific precipitation thresholds
year_thresh_perc	Percentile thresholds used to calculate the annual catchment-specific thresholds
grid_properties	Saved properties of the 1 km BNG grid used

Data variables (per catchment in **sites)**	
mean	Within-catchment distribution mean (Section 3.3)
weighted mean	Within-catchment distribution weighted mean (Section 3.3)
perc_50	Within-catchment distribution 50th percentile (Section 3.3)
perc_75	Within-catchment distribution 75th percentile (Section 3.3)
perc_90	Within-catchment distribution 90th percentile (Section 3.3)
perc_95	Within-catchment distribution 95th percentile (Section 3.3)
perc_99	Within-catchment distribution 99th percentile (Section 3.3)
fixed_th_exceed	Number of grid-cells exceeding **precip_thresh** (Section 3.4.1)

site_th_season_exceed	Number of grid-cells exceeding **season_thresh_perc** (Section 3.4.2)
site_th_year_exceed	Number of grid-cells exceeding **year_thresh_perc** (Section 3.4.2)
grid_info	Grid properties listed in **grid_properties**

Attributes	
Data_grid	The grid used (1km_BMG or1km_BNG_averaged_onto_2km_BNG_ then_projected_back_to_1km_BNG)
Type	The precipitation type for this file (see ***Products*** above)
Accum_label	When **times_accum** refer to (e.g. *end of accumulation period*)
Precip_units	Units of precipitation (*mm/h* or *mm/d*)
thresh_perc_units	Units of the percentile thresholds (*percentile*)
leadtime_units	Units of the lead-time variable (*minutes*)
th_exceed_units	Number of 1 km BNG grid-cells in catchment greater than or equal to thresh
catchment_grid	1km_BNG_G2G
Date_created	YYYY-MM-DD
Creator	UKCEH for Ensemble Verification Project
Accumulation	Accumulation period used (*quart*, *hourly*, *daily*)
percentile_interpolation	Interpolation used when calculating percentiles of in-catchment distribution (*nearest*)
site_thresh_version	Version of the catchment-specific climatological precipitation thresholds, e.g. v1_0_20200619_hkuk_2007_2016

BMR ensemble precipitation forecast files

Output files in NetCDF format were produced for each BMR forecast-origin time in the period 1 June 2017 to 30 September 2018. All files available from the MASS archive were processed. A summary of these is available in the .csv file ***all_processed_BMR_20170601_20180930.csv***. For all but one BMR forecast-origin time in the period considered, data from all ensemble members (e.g. members 00 to 23) were available. The exception is the BMR forecast for 1400 7 July 2017, where members 01, 10, 13 and 22 are missing from the MASS archive.

BMR output files for the catchment precipitation processing of the observed precipitation products are saved in directory structures for a given year ***YYYY***, month ***MM*** and day-in-month ***DD***

output_data/***YYYY***/***MM***/***DD***

Within these directories there is one file per ensemble member ***XX*** per accumulation period and per forecast-origin at hour ***hh*** and minute ***mm***

YYYYMMDDhhmm_u1096_ng_bmr***XX***_precip_2km_***accum***_catchment_precip_stats.nc

A summary of the file contents is given in Table 3.

Table 3 Summary of the file contents of the output for catchment precipitation processing of the BMR ensemble precipitation forecasts.

Coordinates	
sites	List of catchments used for the catchment-statistics in this file. Matches those in the all_catchments look-up tables (Section 4.1.4)
leadtimes	**Intended to be**: all forecast lead-times in the forecast (minutes) **Actually**: the whole hour of all forecast lead-times (minutes) ***Correct if re-run. In meantime use VT-DT (see below)***
leadtimes_*accum*	Forecast lead-times at the end of each accumulation period considered (for *quart* these are again the whole hour-part of the lead-time only)
precip_thresh	Fixed-value precipitation thresholds used (in **Precip_units**, below)
season_thresh_perc	Percentile thresholds used to calculate the seasonally varying catchment-specific precipitation thresholds
year_thresh_perc	Percentile thresholds used to calculate the annual catchment-specific precipitation thresholds
gen_ints_entries	Positions of general integer header from the Nimrod file
gen_real_entries	Positions of general real header from the Nimrod file
spec_real_entries	Positions of specific integer header from the Nimrod file
char_entries	Positions of character header from the Nimrod file
spec_int_entries	Positions of specific integer header from the Nimrod file

Data variables from nimrod file header	
DT	Data time (origin) of the forecast
VT	Validity time of the forecast
nimrod_gen_int	general integer header from the Nimrod file
nimrod_gen_real	general real header from the Nimrod file
nimrod_spec_real	specific integer header from the Nimrod file
nimrod_char	character header from the Nimrod file
nimrod_spec_int	specific integer header from the Nimrod file

Data variables (per catchment in **sites)**	
mean	Within-catchment distribution mean (Section 3.3)
weighted mean	Within-catchment distribution weighted mean (Section 3.3)
perc_50	Within-catchment distribution 50th percentile (Section 3.3)
perc_75	Within-catchment distribution 75th percentile (Section 3.3)
perc_90	Within-catchment distribution 90th percentile (Section 3.3)
perc_95	Within-catchment distribution 95th percentile (Section 3.3)
perc_99	Within-catchment distribution 99th percentile (Section 3.3)
fixed_th_exceed	Number of grid-cells exceeding **precip_thresh** (Section 3.4.1)
site_th_season_exceed	Number of grid-cells exceeding **season_thresh_perc** (Section 3.4.2)
site_th_year_exceed	Number of grid-cells exceeding **year_thresh_perc** (Section 3.4.2)
grid_info	Grid properties listed in **grid_properties**

Attributes	
Type	Identifying part of Nimrod file used e.g. ***YYYYMMDDhhmm***_u1096_ng_bmr***XX***_precip_2km
Accum_label	When **leadtimes_accum** refer to (e.g. *end of accumulation period*)
Precip_units	Units of precipitation (*mm/h* or *mm/d*)
thresh_perc_units	Units of the percentile thresholds (*percentile*)

leadtime_units	Units of the lead-time variable (*minutes*)
th_exceed_units	Number of 1 km BNG grid-cells in catchment greater than or equal to thresh
catchment_grid	1km_BNG_G2G
Date_created	YYYY-MM-DD
Creator	UKCEH for Ensemble Verification Project
Accumulation	Accumulation period used (*quart*, *hourly*, *daily*)
percentile_interpolation	Interpolation used when calculating percentiles of in-catchment precipitation distribution (*nearest*)
site_thresh_version	Version of the catchment-specific climatological precipitation thresholds, e.g. v1_0_20200619_hkuk_2007_2016
member	Ensemble member is ***XX*** in 201712250100_u1096_ng_bmrXX_precip_2km (**extract only XX for future versions**)

Rainfall and River Flow Ensemble Verification: Phase 2

Definition of Time-Window Probabilities (TWPs)

Final Report Appendix A.4

In Phase 1 of the project a model-oriented verification approach was followed in which the precipitation ensemble forecast is evaluated sequentially from start to finish, with each forecast accumulation period (15-min, hourly, and daily) precisely matched to the observed period in space and time. Moreover, the catchment-mean precipitation value was derived and used to reflect ensemble performance. This is a perfectly valid thing to do and demonstrated the skill of the precipitation ensemble as used as input to G2G.

As summarised in Appendix B.4 it was found that the 15-min precipitation verification analyses are very similar to the hourly, with the daily giving the best overall steer for heavy rainfall given the relatively modest extremes over the UK at the hourly time-scale, where many places have 99th percentiles less than 4 mm/h (see Appendix A.5). This sequential way of assessing the weather model ensemble precipitation means timing errors can dominate the verification analyses, even fairly early on in the forecast, leading to poor scores even when percentile thresholds were used to capture the most extreme rain at any given time. Unfortunately, these percentile thresholds were not very extreme on many occasions, though probably more so given the wet nature of the Phase 1 study period. Fixed precipitation thresholds applied everywhere can be highly unsatisfactory because the thresholds would need to be kept very low to ensure all catchments are sampled. Pushing the thresholds too high everywhere reduces the number of contributing catchments to the extent that the sample size is unsuitable for computing robust verification statistics.

When it comes to assessing the risk of potential flood-producing rainfall, from the user perspective it is more common to “scan” a particular-time window for possible events. This heuristic process can be replicated by computing using Time-Window Probabilities (TWPs), and is closer to the way a hydrometeorologist would look at a precipitation forecast, especially at longer lead-times where the objective is to look for exceedance events that can pose a flood risk or threat. Therefore, under Phase 2, the conventional “model-based” precipitation verification has been computed alongside the much more user-focused TWP verification.

Another enhancement under Phase 2 concerns the precipitation thresholds that have been used. Long-term climatological precipitation distributions have been computed at the catchment-scale (see Appendix A.3). From these, three specific percentiles were examined in more detail: 90th, 95th and 99th. These are plotted as precipitation percentile maps for England & Wales in Appendix Y. These maps are a useful resource in their own right, giving a comprehensive view of what can be considered extreme precipitation for the UK climate. In many instances the 90th percentile is close to zero, especially for the east and south-east parts of England and even the 99th percentile values derived from hourly precipitation accumulations are less than 5 mm/h: hardly flood-producing rain but nevertheless rare. From a catchment-based verification perspective, with interest in capturing the performance of the forecast for potential flood-producing rainfall, there is a clear mismatch between the magnitude of the exceedance events of interest and the sample size needed to compile robust verification statistics. Under Phase 2, the focus has been on the 95th percentile as the 90th is too dry and the 99th provides inadequate samples, even for a 12-month period.

In the first instance the following schematic (Figure 1) depicts the concept of the Time-Window Probability in terms of sequential “scanning”.

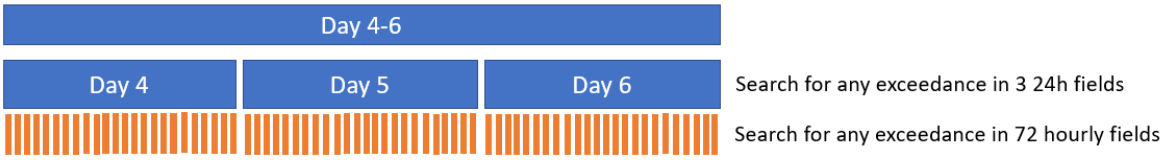


Figure 1 Schematic of the concept of Time-Window Probability

Days 4-6 consists of either three non-overlapping 24h accumulations or 72 hourly accumulations of precipitation. In a sequential "Phase 1" sense, each of these precipitation accumulations is verified separately and analyses are aggregated (averaged) to compute daily or hourly *average* analyses for the Days 4-6 range. Another point to note is that TWPs are probabilities, i.e. they can only be computed with a precipitation threshold and the ensemble precipitation forecast has to exceed that threshold in order to produce a probability. So, for constructing a TWP a precipitation threshold is applied to each of the 72 hour or 3-day accumulations and if *any* of the 72 hour or 3-day accumulations exceed the threshold, they would contribute to a probability which represents the *entire* Days 4-6 period, with a few additional caveats.

The spatial view of the TWP derivation is illustrated below. One might well want to ask: "when would I be sure that an event (threshold exceedance) has or will occur within a catchment?" The answer matters because it is related to the ability to verify and our belief that something happened, i.e. "when do I think I know I have detected an event?" In the schematic below (Figure 2) the precipitation ensemble forecast consists of 5 members.

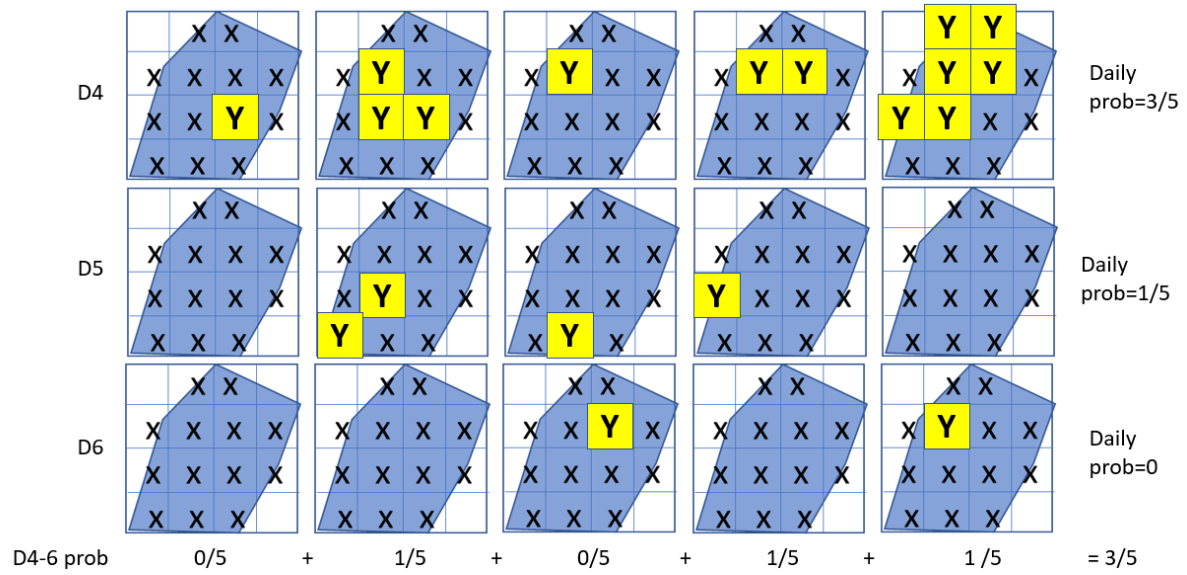


Figure 2 Schematic illustrating the difference between a conventional (accumulation-period) and Time-Window Probability using daily accumulations for a theoretical catchment (blue) where yellow points marked "Y" exceed a particular precipitation threshold, and other points marked X don't.

The X's indicate the weather model grid-cells that are identified as falling within the catchment in blue. The yellow squares indicate where the forecast values for the grid-cells exceed a specified rainfall threshold. But is it enough if one grid-cell in the catchment exceeds the rainfall threshold to be counted? Put differently, if in a gridded raingauge or radar rainfall field only one grid-cell in the catchment exceeded the threshold, would you consider that to indicate an event has occurred? Given observation uncertainty, the answer should be "no", especially for

radar rainfall data, where spurious single grid-cell values might still be fairly common, despite rigorous quality control. As it is advisable to apply the same rules to both forecast and observation fields, a criterion has been adopted that at least two grid-cells must exceed the precipitation threshold within a catchment for the ensemble member to be included in a probability calculation.

By working *across a row*, a *conventional probability* is derived where each of the ensemble members is considered giving a probability for a given day. In this instance a D4 (Day 4) daily probability of exceeding the threshold results in a probability of 3/5 or 60%. Whereas D5 yields a probability of 1/5 or 20% etc. For deriving TWPs, the first step in the process is to *scan across the time-slices denoted by the columns*. i.e. for ensemble member 1 (in the first column), do *any* of the grid-cells in the catchment daily totals exceed the threshold? In this instance the answer is no. The process is repeated, e.g. for the second column, two of the three members have grid-cells in the catchment that exceed the threshold and *so the ensemble member count that exceeds the threshold for the time-window is incremented*. In the end the TWP probability for the Days 4-6 window is the sum of those shown along the bottom of the schematic (Figure 2), 3/5 or 60%.

TWPs will:

- tend to be higher than ordinary or conventional probabilities, which is beneficial when searching for higher threshold events
- allow for the use of somewhat higher thresholds because they are based on individual grid-cell values
- use all the grid points in the catchment to check for exceedance and are not derived based on the catchment mean or even the median. Computing the mean or median removes the peaks from the within-catchment distribution and, especially for situations where the catchment is not covered by extensive heavy rain, anything more localised (e.g. isolated showers or thunderstorms) may be misrepresented or not be detected at all when the catchment mean or median is used.
- provide a truer reflection of localised rainfall impact in probability space.

Furthermore, TWPs:

- are a form of post-processing in which the time-dimension is collapsed
- cannot account for inherent forecast intensity biases, as is the case for conventional probabilities
- may not be reliable: calibrating the probabilities may still be necessary.

It is worth bearing in mind that the conventional probabilities computed in Phase 1 (and in Phase 2) were/are calculated based on the catchment mean. These are also available in Phase 2 for comparison, but note they are very different by construction.

This is clearly manifested in Figures 3 and 4 which show the distribution of probabilities for the 0.5 and 8 mm/d thresholds using catchment means to compute conventional probabilities of exceedance and TWPs which are based on considering all the individual grid-cells in a catchment. Note the x-axis represents the sequence of probability bins between 0 (left) and 1 (right) used to create Reliability Diagrams. The probability distribution is shifted to the right, i.e. TWPs are generally going to be larger than the conventional kind. This effect is more pronounced for higher precipitation thresholds where the catchment means are likely to be much lower than individual grid-cell values, skewing the distribution further.

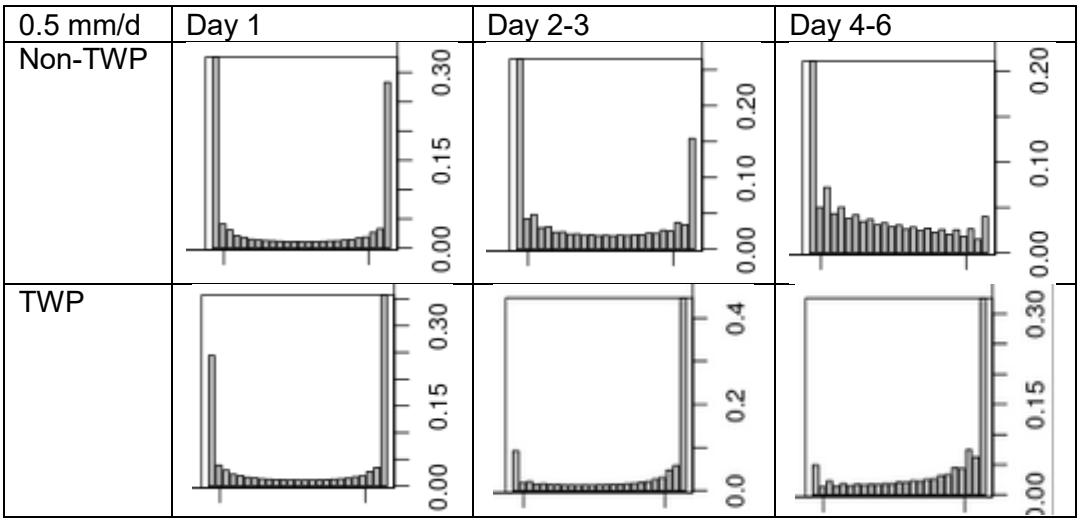


Figure 3 Comparison of forecast precipitation ensemble probability distributions for the 0.5 mm/d threshold using conventional (non-TWP) exceedance probabilities based on the catchment mean (upper row) and TWPs (lower row).

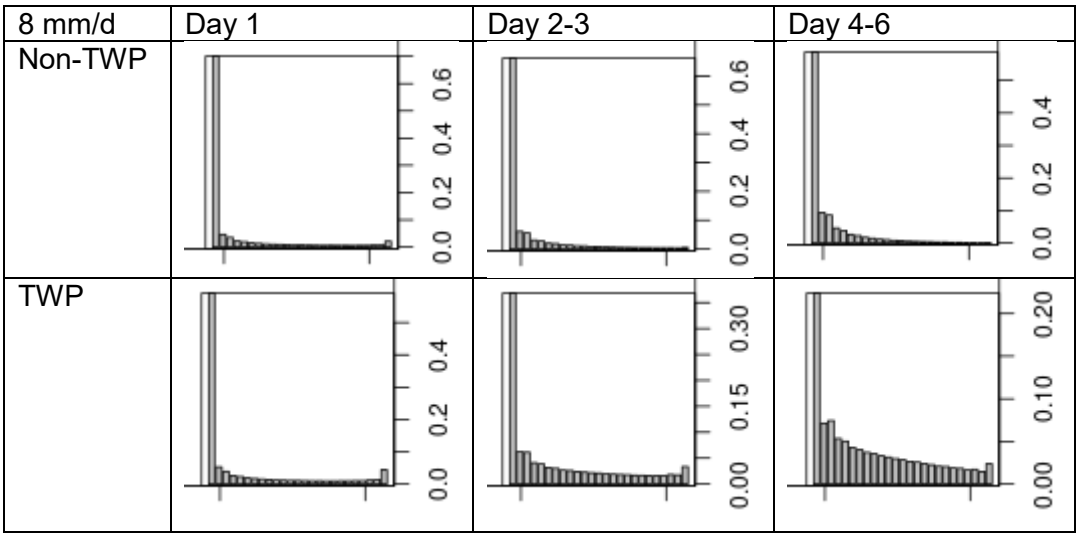
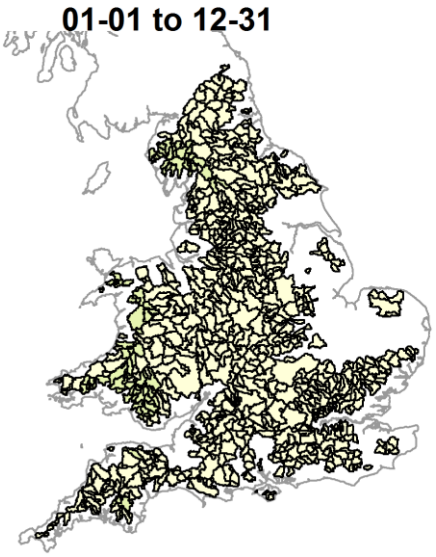
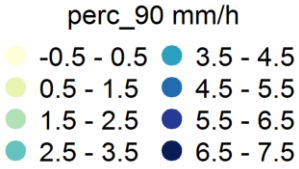


Figure 4 As Figure 3 but for the 8 mm/d precipitation threshold.

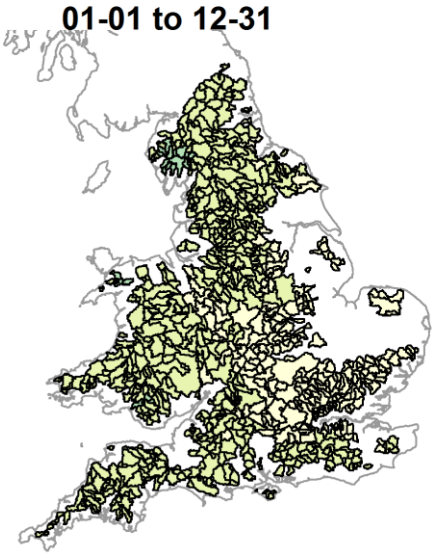
Rainfall and River Flow Ensemble Verification: Phase 2
Climatological threshold maps for England & Wales and Scotland
Final Report Appendix A.5

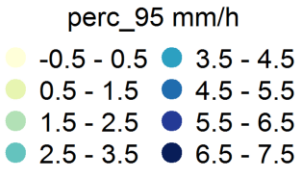
Annual hourly thresholds:



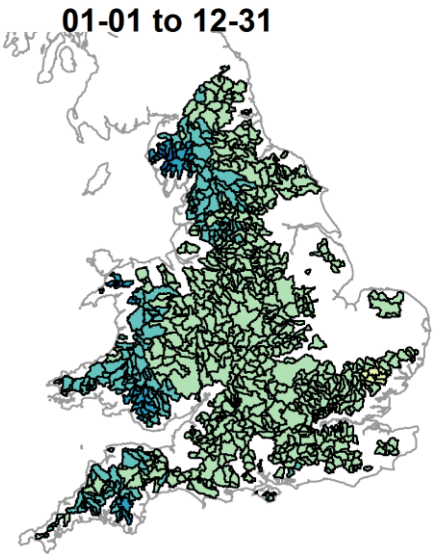


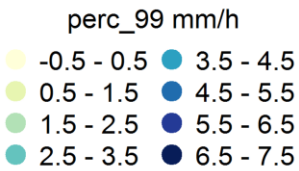
90th percentile is close to 0 except in the western upslopes of Wales and NW England





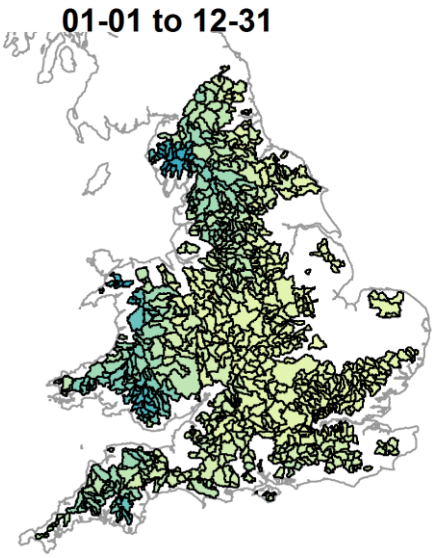
95th percentile is generally less than 1.5 mm/h.





99th percentile is generally less than 2.5 mm/h except on western fringes where it is ~5 mm/h.

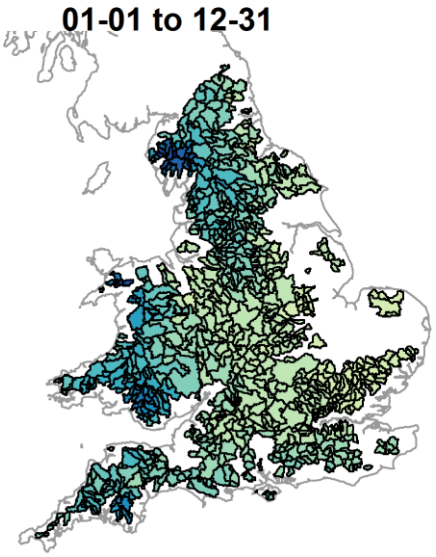
Annual daily thresholds:



perc_90 mm/day

-1 - 1	19 - 21
1 - 3	21 - 23
3 - 5	23 - 25
5 - 7	25 - 27
7 - 9	27 - 29
9 - 11	29 - 31
11 - 13	31 - 33
13 - 15	33 - 35
15 - 17	35 - 37
17 - 19	37 - 39

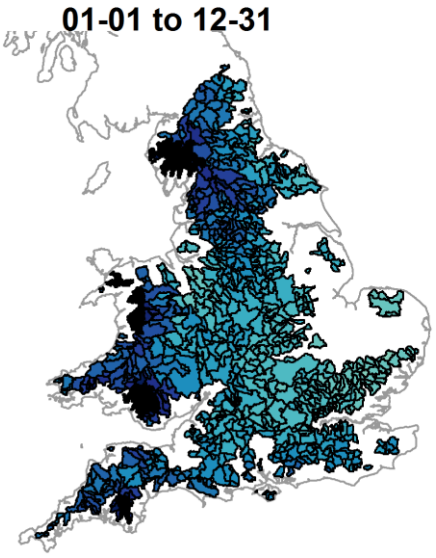
90th percentile is less than 10 mm except for western fringes/upslopes where values 15-20 mm



perc_95 mm/day

-1 - 1	19 - 21
1 - 3	21 - 23
3 - 5	23 - 25
5 - 7	25 - 27
7 - 9	27 - 29
9 - 11	29 - 31
11 - 13	31 - 33
13 - 15	33 - 35
15 - 17	35 - 37
17 - 19	37 - 39

95th percentile is less than 15 mm except for western fringes/upslopes where > 20 mm

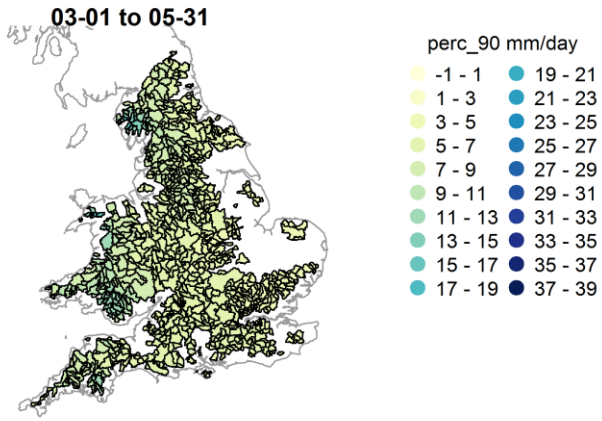


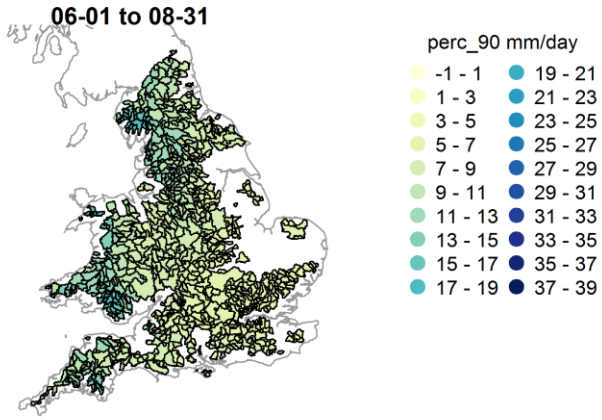
perc_99 mm/day

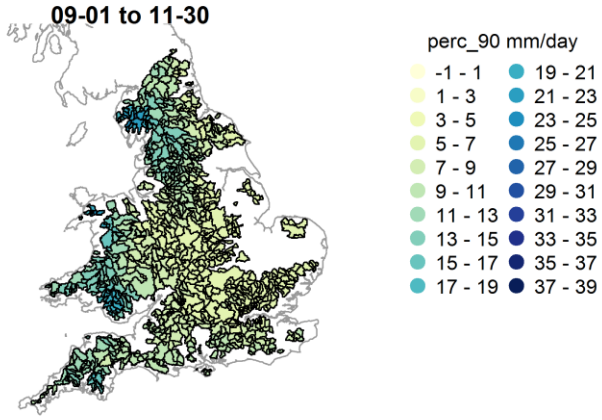
-1 - 1	19 - 21
1 - 3	21 - 23
3 - 5	23 - 25
5 - 7	25 - 27
7 - 9	27 - 29
9 - 11	29 - 31
11 - 13	31 - 33
13 - 15	33 - 35
15 - 17	35 - 37
17 - 19	37 - 39

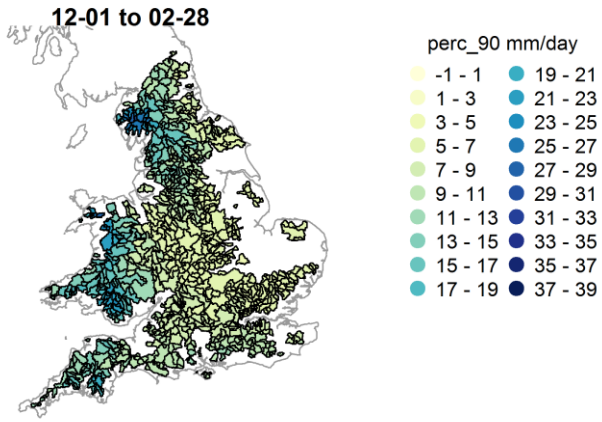
99th percentile is ~20 mm except for western fringes/upslopes where > 30 mm

Seasonal daily thresholds:

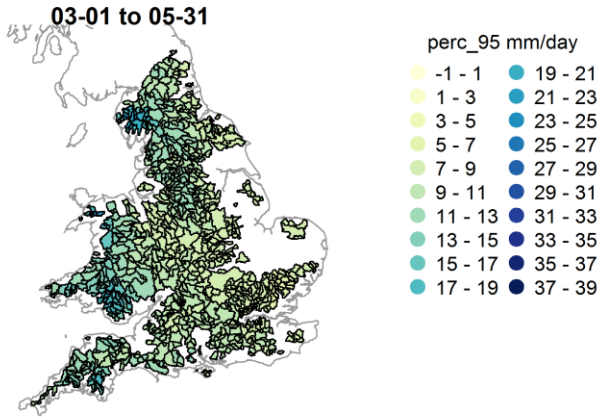


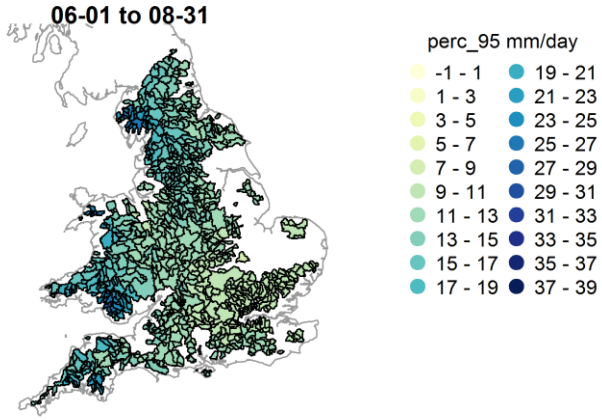


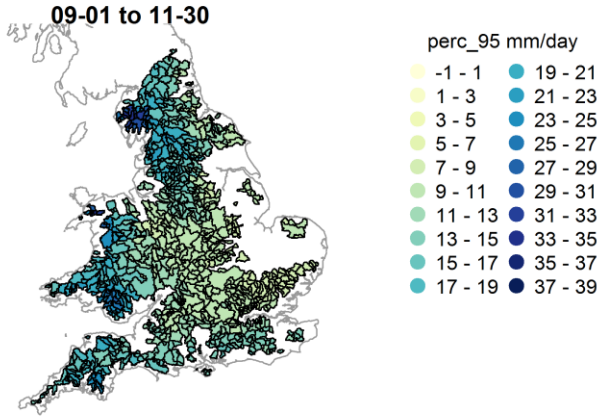


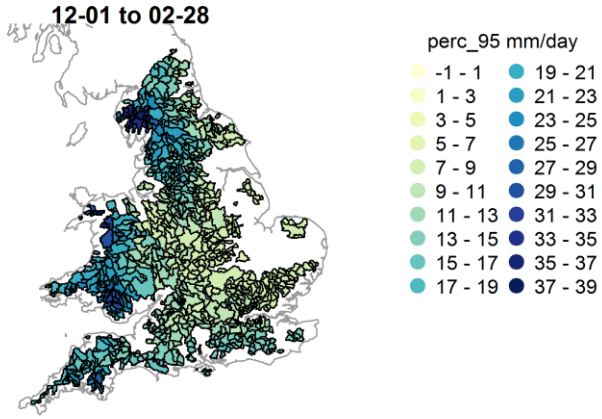


Broadly speaking 90th percentile thresholds below 10 mm for all seasons in the E. Upslopes and W/S coastal regions have slightly higher values, especially in the colder seasons, with values in excess of 15 mm.

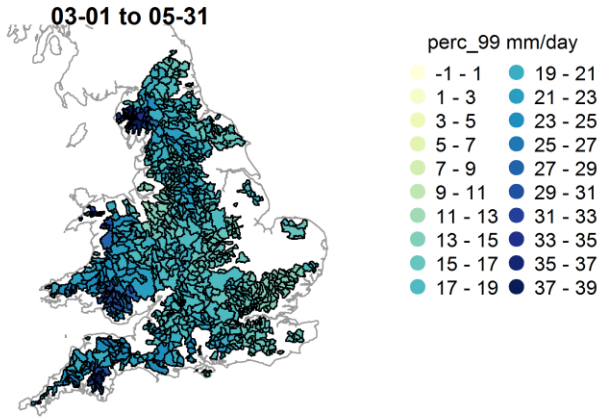


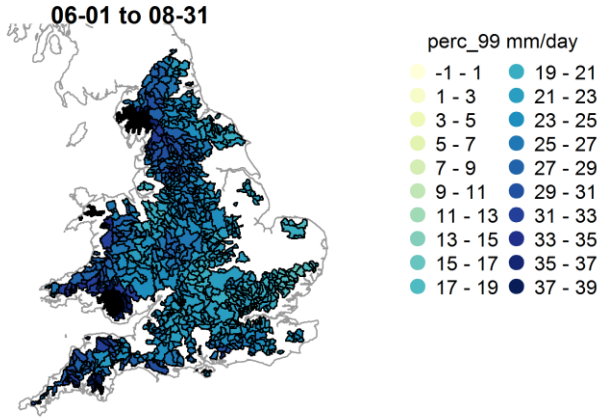


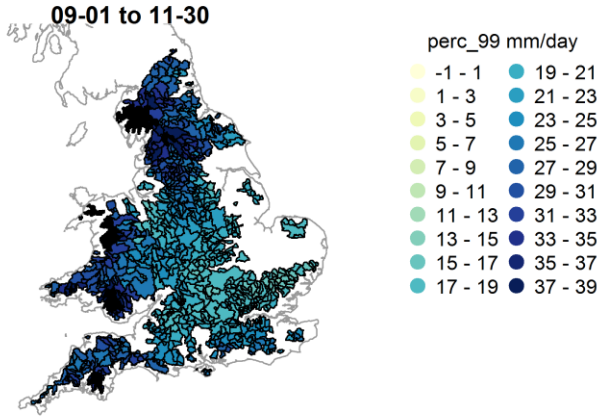


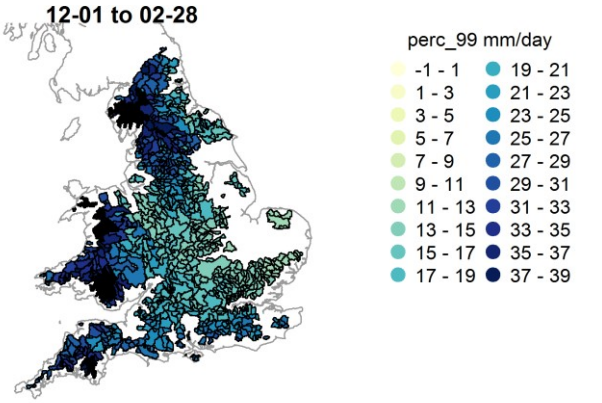


Seasonal 95th daily percentiles show the biggest contrast in winter western fringes and uplands in excess of 25 mm in winter, otherwise ~20 mm or more.



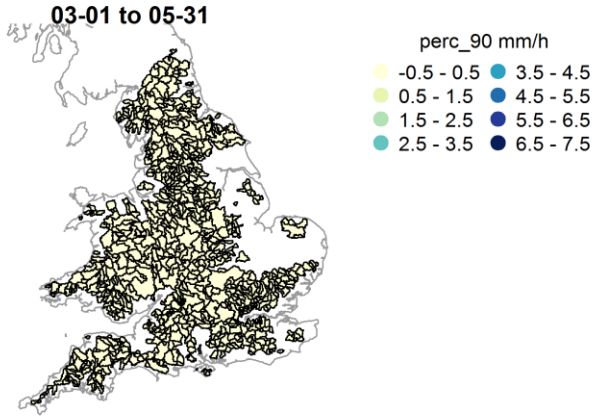


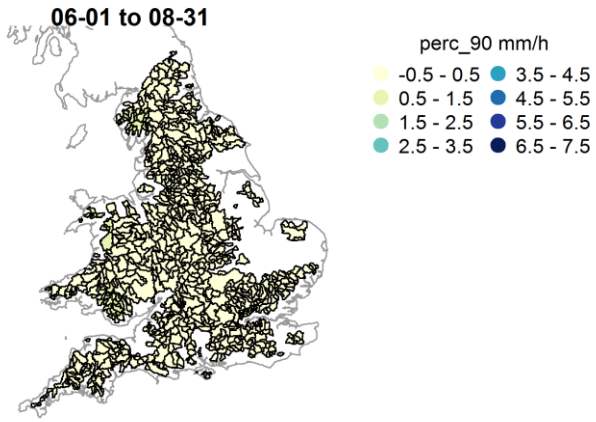


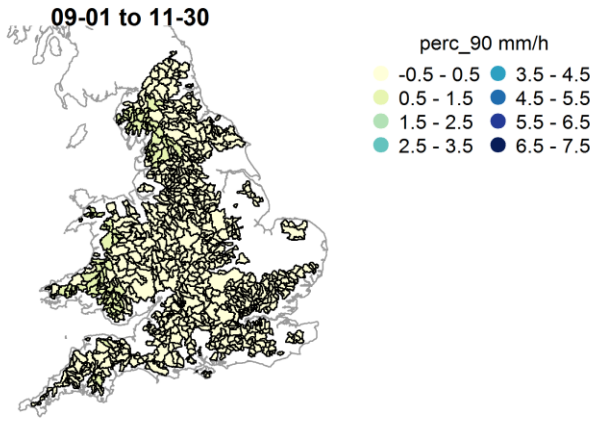


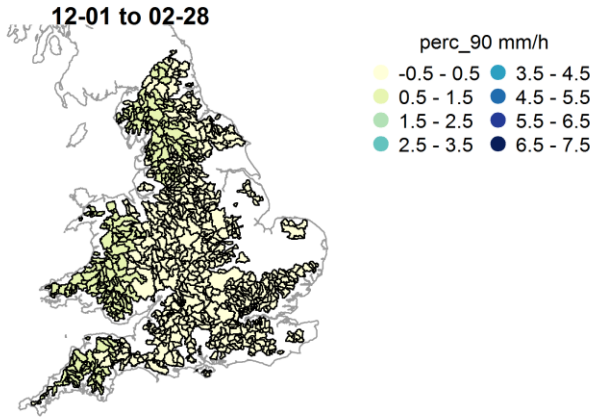
Seasonal 99th daily percentile is above 30 mm for western fringes/upslopes and comfortably above 15 mm everywhere.

Seasonal hourly thresholds:

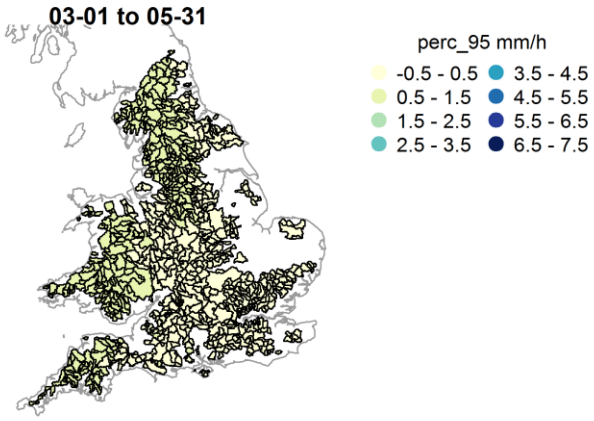


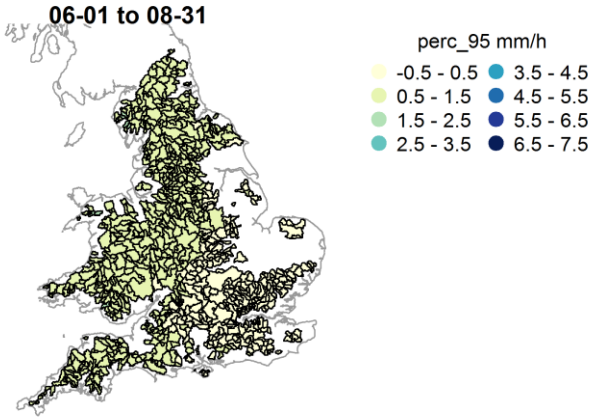


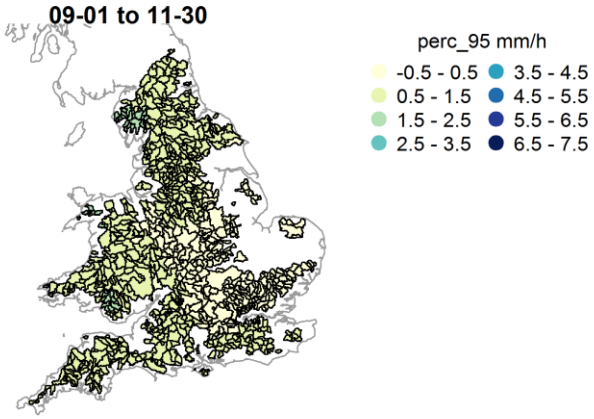


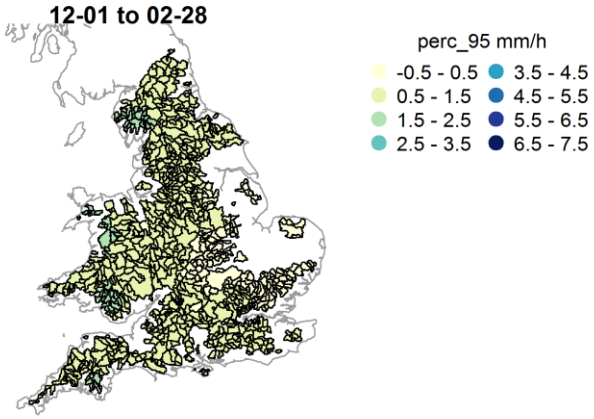


Seasonal 90th hourly percentile is generally < 0.5 mm except for western fringes/upslopes in the colder seasons.

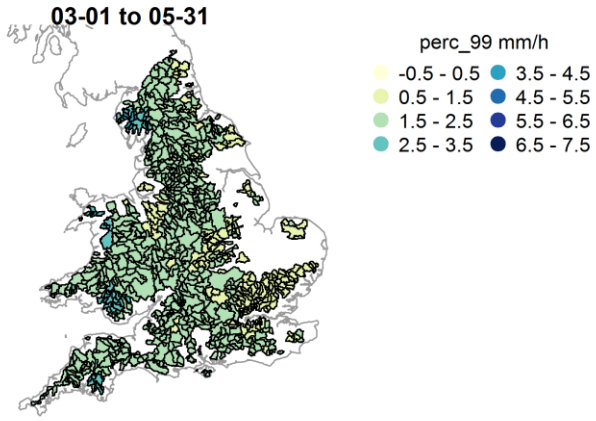


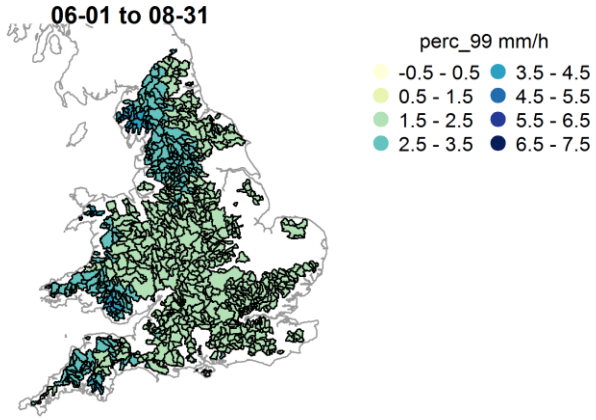


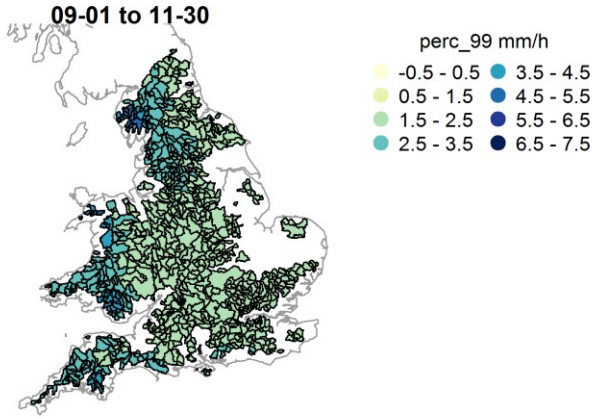


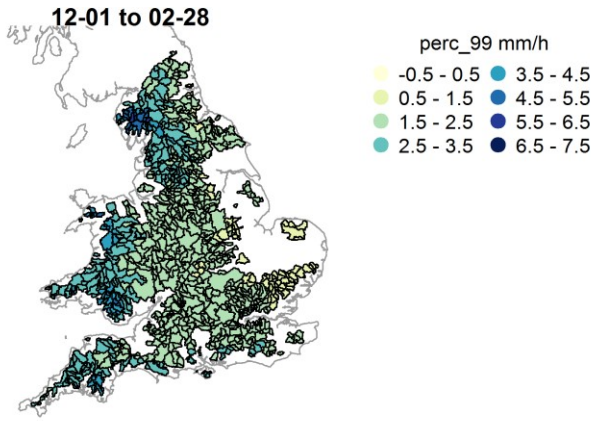


Seasonal 95th hourly percentile is still < 0.5 mm for the E and SE with slightly higher values W, but even in the colder season values are generally < 4mm.



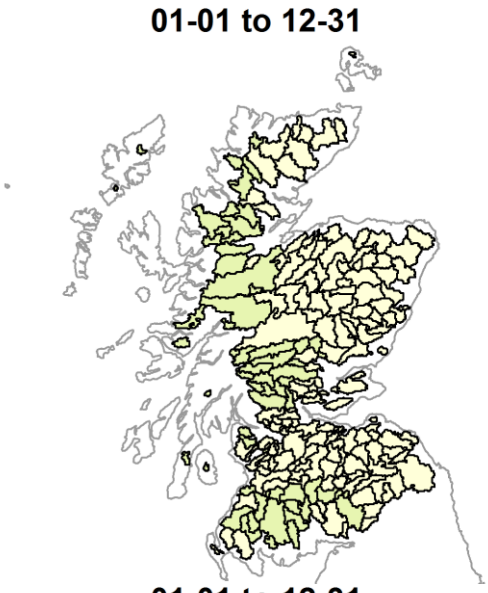




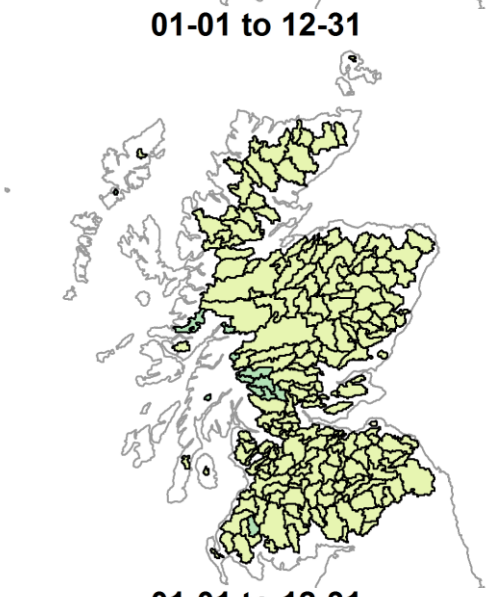


Seasonal 99th hourly percentiles are comfortably in the < 3 mm for large parts of the UK irrespective of season; western fringes and upslopes have values in excess of 4mm.

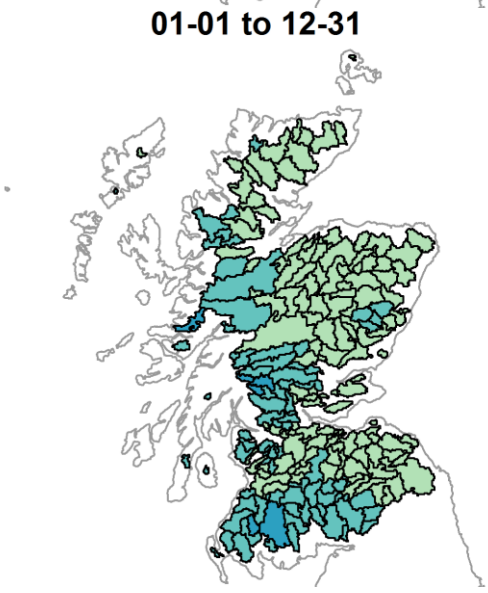
Annual hourly thresholds:



As in E&W hourly 90th percentile values are widely less than 1 mm/h

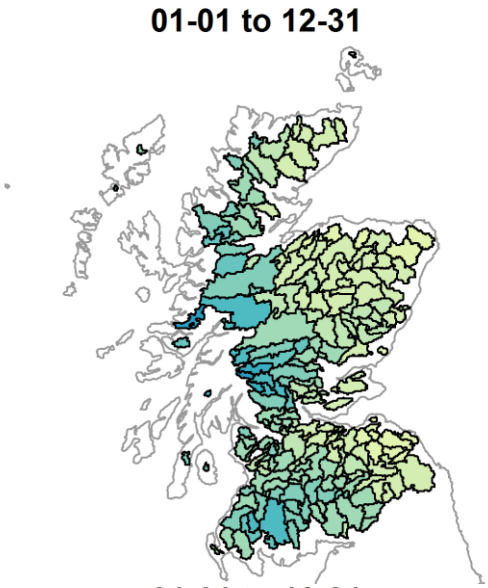


Even hourly 95th percentile values are less than 3 mm/h



Hourly 99th percentile values approach 5 mm/h in western upslopes but remain lower in the E.

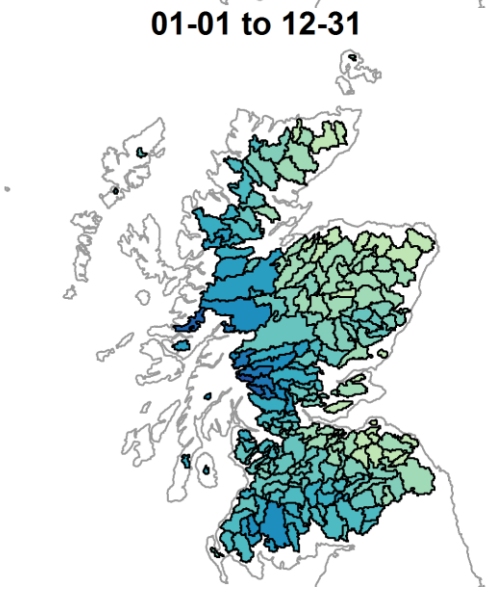
Annual daily thresholds:



perc_90 mm/day

-1 - 1	19 - 21
1 - 3	21 - 23
3 - 5	23 - 25
5 - 7	25 - 27
7 - 9	27 - 29
9 - 11	29 - 31
11 - 13	31 - 33
13 - 15	33 - 35
15 - 17	35 - 37
17 - 19	37 - 39

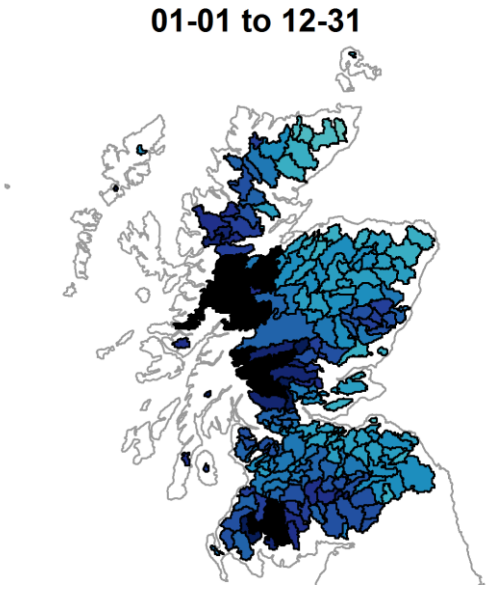
Daily 90th percentile values are higher for W upslopes, typically of the order of 15-20 mm/d



perc_95 mm/day

-1 - 1	19 - 21
1 - 3	21 - 23
3 - 5	23 - 25
5 - 7	25 - 27
7 - 9	27 - 29
9 - 11	29 - 31
11 - 13	31 - 33
13 - 15	33 - 35
15 - 17	35 - 37
17 - 19	37 - 39

Daily 95th percentile values exceed 25 mm/d in some W catchments.

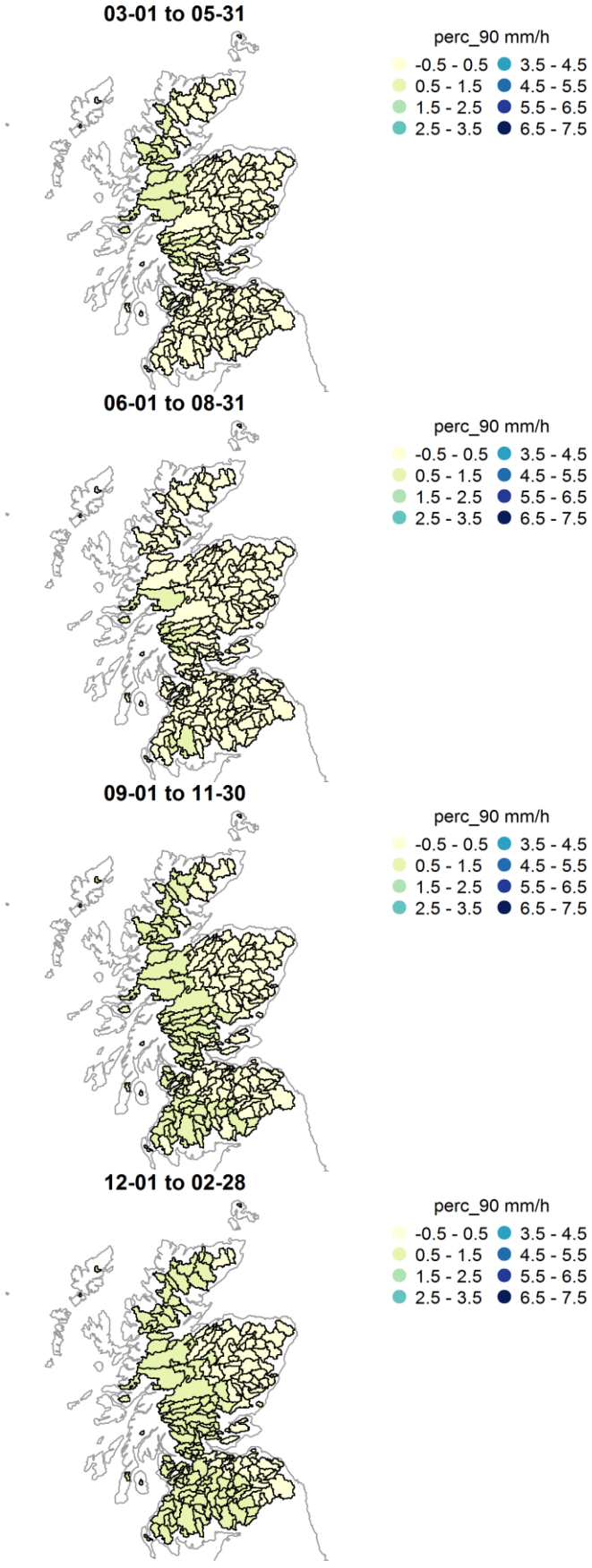


perc_99 mm/day

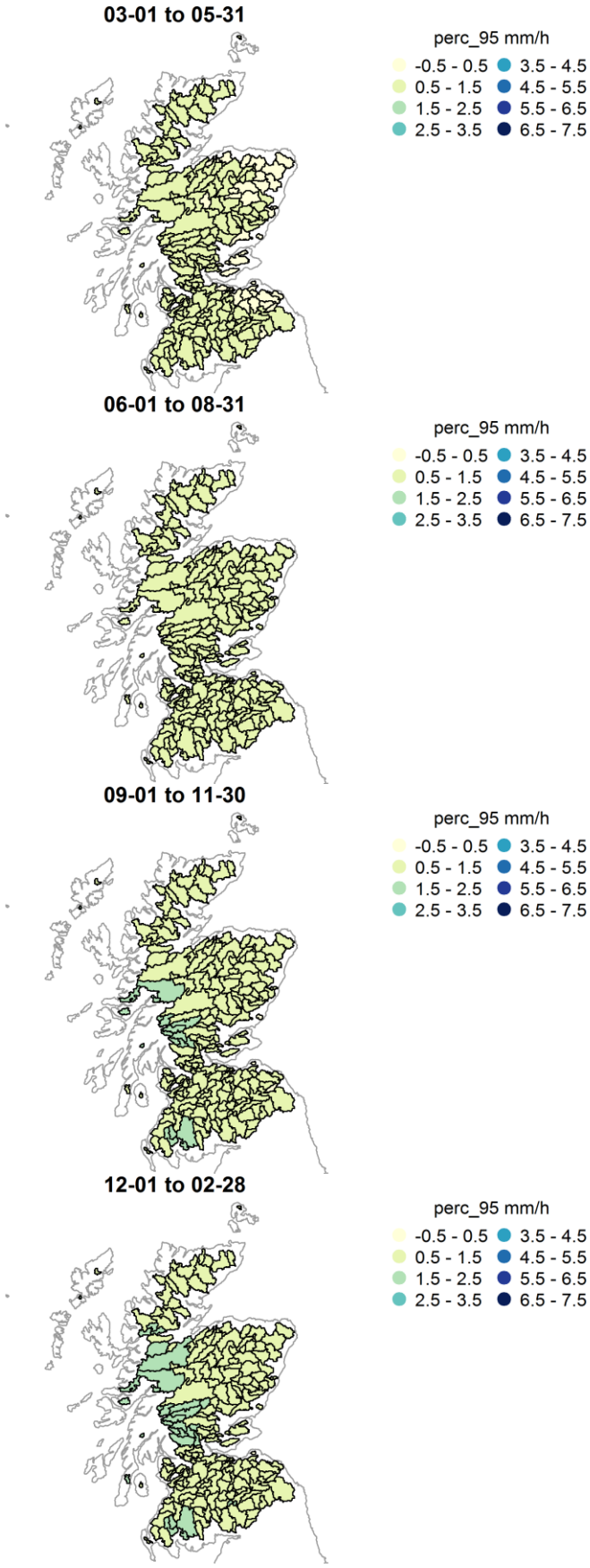
-1 - 1	19 - 21
1 - 3	21 - 23
3 - 5	23 - 25
5 - 7	25 - 27
7 - 9	27 - 29
9 - 11	29 - 31
11 - 13	31 - 33
13 - 15	33 - 35
15 - 17	35 - 37
17 - 19	37 - 39

Daily 99th percentile values exceed 30-35 mm/d with W or SW exposure.

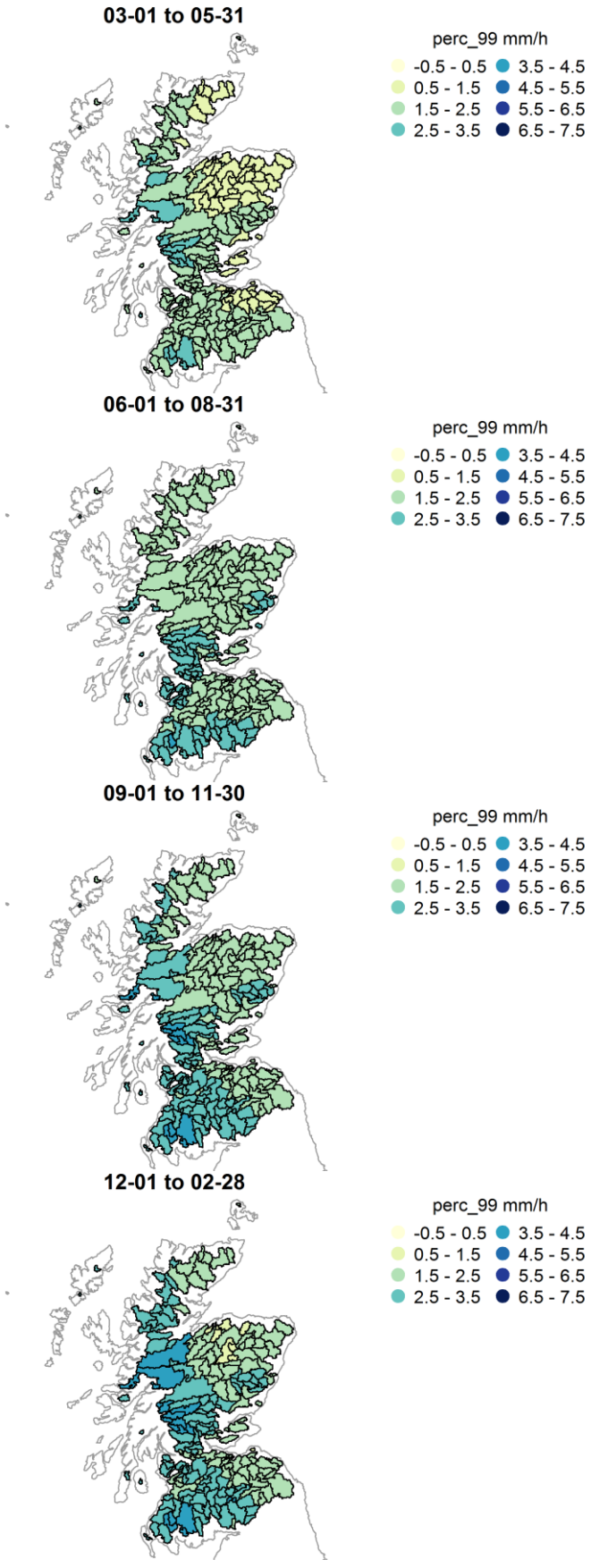
Seasonal hourly thresholds:



Hourly 90th percentile values split by season remain low but are higher in the autumn and winter.



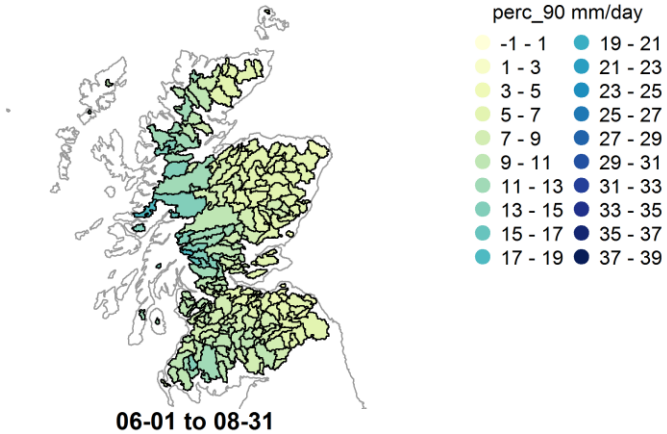
Hourly 95th percentile values show the same trends with values still modest overall.



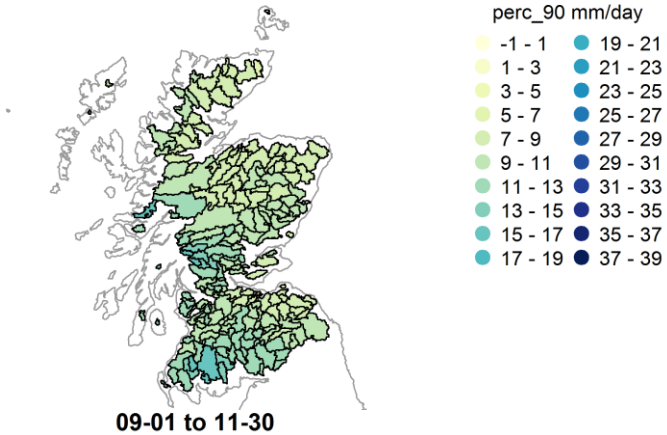
Hourly 99th percentile values show that Spring is the driest, and winter the wettest, though few catchments have values ~5 mm/h

Seasonal daily thresholds:

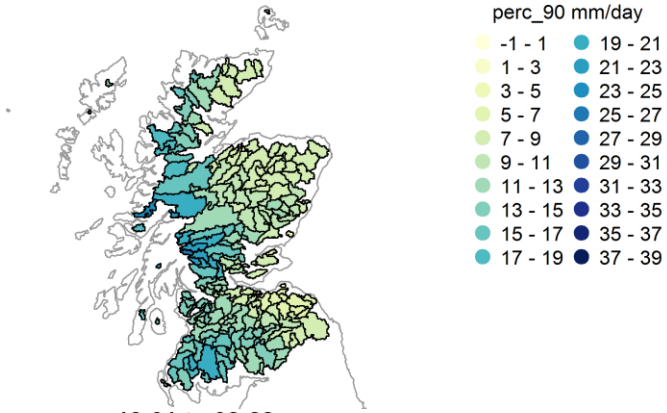
03-01 to 05-31



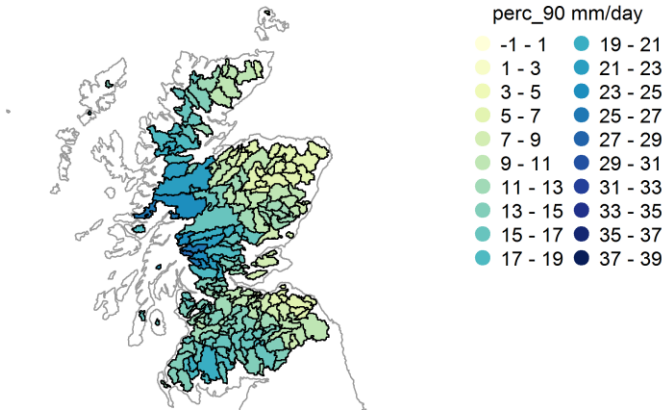
06-01 to 08-31



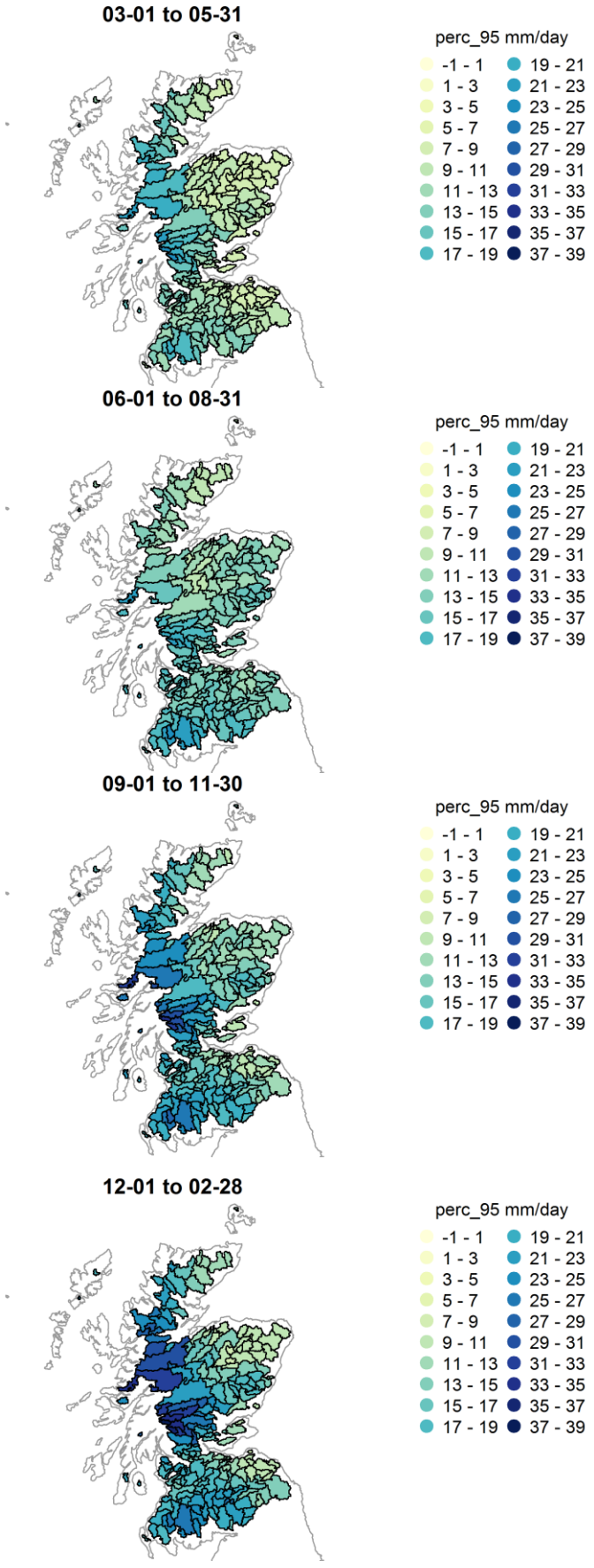
09-01 to 11-30



12-01 to 02-28

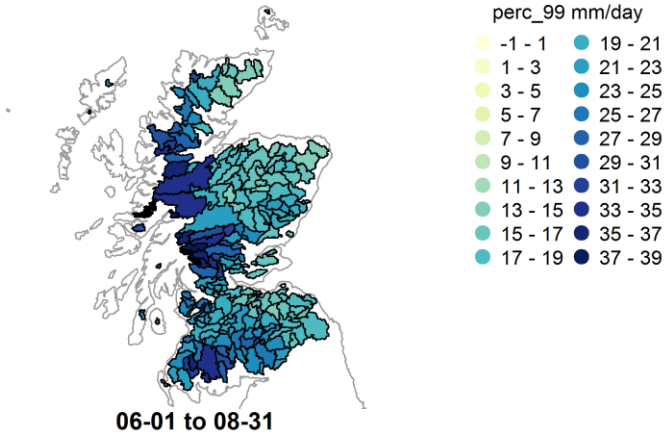


Daily 90th percentile values show the W-E gradient which is more pronounced in the autumn and winter, where values can exceed 20 mm/d in western upslope regions.

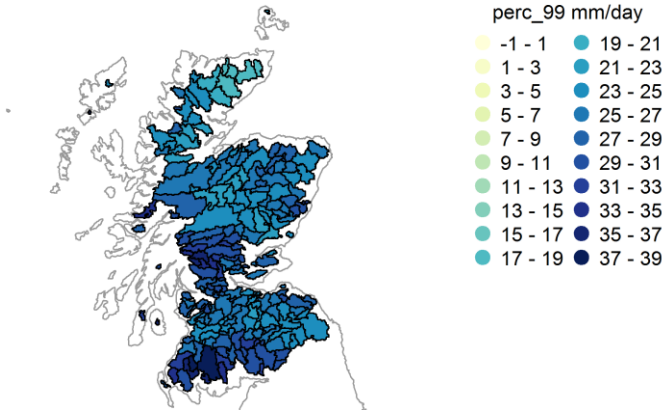


Daily 95th percentile values show how the W-E gradient is enhanced with western upslope values in excess of 25 and even 30 mm/d in autumn and winter.

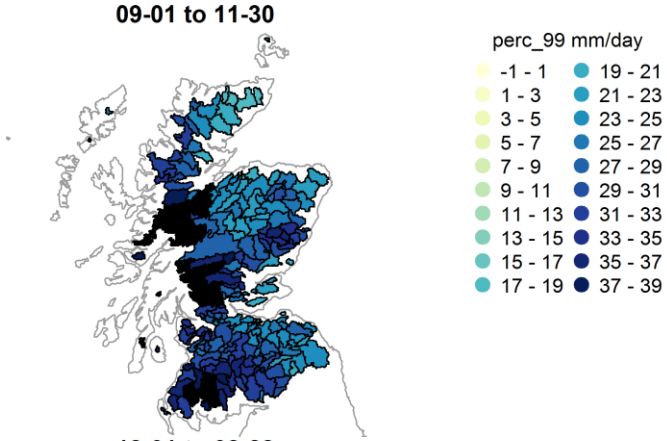
03-01 to 05-31



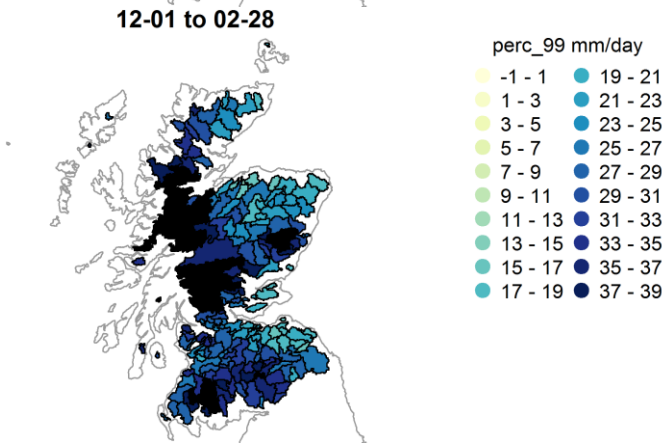
06-01 to 08-31



09-01 to 11-30



12-01 to 02-28



Daily 99th percentiles are markedly higher everywhere, but the differences in the seasons remains the same with the cooler seasons showing values in excess of 30 mm/d.

Rainfall and River Flow Ensemble Verification: Phase 2
Forecast triggering investigations
Final Report Appendix A.6

0	7	
100	465	
200	5	
300	5	
400	0	
500	1	
600	5	
700	470	
800	4	
900	3	
1000	0	
1100	1	
1200	3	
1300	476	
1400	3	
1500	2	
1600	0	
1700	3	
1800	11	
1900	461	
2000	8	
2100	4	
2200	1	
2300	2	
	Number	Percent
Total	1940	100
Ontime	1872	96.49485
Early	33	1.701031
Late	34	1.752577
Other	1	0.051546

Figure 1 Forecasts as a function of trigger times. Main times are 01, 07, 13 and 19 UTC.

In Phase 1 of the project there was considerable variability in the availability of new ensemble precipitation forecasts. There was concern that if there were many forecasts that deviated from the expected times of 01, 07, 13 and 19 UTC it could affect results and would need to be considered when it came to the combination of results. As a consequence, an activity was included in the work plan for Phase 2.

Once the study period was finalised, an examination of the forecast trigger times shown in Figure 1 indicates that only ~1.7% of forecasts (34 out of 1940) fall into the late category (forecasts badged as 1 to 2 hours after expected) and similar for early (badged 1 o 2 hours before expected).

Monthly scores, which would be the absolute minimum, are produced from a sample of ~120 forecasts. Samples of ~30 forecasts are insufficient to analyse separately and can be safely removed.

Operationally, trigger times should remain fairly stable and delayed (late) forecasts should only happen by exception. If this is not the case, then it should be investigated from an operational perspective rather than be accounted for in the verification.

Rainfall and River Flow Ensemble Verification: Phase 2
Commentary on precipitation verification maps and plots
Final Report Appendix B.1.1

Note: score definitions and references to the Ensemble Verification Framework are provided in Appendix A.1. This commentary refers to the plots provided in Appendix B.1.2.

1. Weather model focused precipitation verification

Weather model focused verification is important because it aims to measure the underlying performance of the precipitation forecasting system, ensemble rainfall forecasts from which are used as input to G2G to produce ensemble river flow forecasts. The forecast is verified by precisely matching the ensemble precipitation forecasts in time and computing some aggregates, either in time or both in time and in space over catchments.

In this verification framework the ensemble characteristics are captured through the Rank Histogram and the Continuous Ranked Probability Score (CRPS) and its skill score (CRPSS). The CRPS is sensitive to the intensity bias, recalling that it collapses to the Mean Absolute Error (MAE) for a single (deterministic) forecast. This bias is translated into ensemble space and will affect the scores such as the CRPS directly and the probabilities too (indirectly). To this end the weather model characteristics (i.e. the Met Office Unified Model (UM)), which determine aspects such as precipitation intensity are examined using the mean error (ME) of the ensemble control member.

Understanding the bias in the model accumulations is always important, but especially so given the construction of the Best-Medium Range (BMR) ensemble. The 2.2 km MOGREPS-UK ensemble is used for ranges up to 36h whereas the 20 km (from July 2017, ~34 km before) MOGREPS-G is used for all ranges beyond this. Given the differences in resolution alone, these two model configurations produce very different precipitation fields, both in terms of texture and intensity, and these combined ensemble precipitation forecasts are used as input to the G2G model to produce ensemble river flow forecasts, on a 1km grid at 15 minute intervals at national scale.

The weather model ensemble error can depend on location (e.g. hills and valleys, coasts), which may be associated simultaneously with greater predictability (e.g. orographic enhancement of precipitation is predictable) but also be subject to larger errors in magnitude (e.g. due to the way that the orography is resolved). These larger *perceived* model errors are, however, also subject to uncertainty given the difficulties of measuring precipitation in complex terrain for example, where all observation sources have problems of one kind or another.

It is important to remember that the atmosphere is a continuum and, as far as the modelling of the atmosphere is concerned, the same model physics is applied everywhere in the domain. As a result, one would not expect the *long-term* behaviour of model precipitation to show large variations in *skill* (i.e. compared to a reference forecast such as climatology), whilst short-term variations can be large and volatile. Skill can, and does, vary as a function of time-of-year, which is linked to the type (and often amount) of rainfall. There is a distinct weather dependence on the month-to-month performance of the forecast.

Long-term trends in skill evolve slowly and are more often associated with changes in resolution than model physics changes: see, for example, Mittermaier *et al.* (2013) which

shows the step-change in skill between the 4 and 12 km versions of the UM. From a weather model development perspective, the desire is to see an upward trend in skill scores over time. The only way to see this is to form aggregated scores over longer time periods and/or computing moving averages over time. Here monthly, seasonal and annual scores are compared to gain an understanding of what is sensible for monitoring catchment-scale precipitation. Using the catchment as the unit of comparison implies either a form of upscaling (computing a catchment mean or median from all the weather model grid-cells in a catchment) or abstraction (checking whether a threshold is exceeded by a subset of the grid-cells within a catchment). Secondly it implies that the unit of interest, the catchment, is in a fixed location. Therefore, spatial accuracy *is* important for hydrometeorological applications.

The weather model precipitation intensity biases are examined first. Monthly ME in the daily catchment mean precipitation for the control member of the ensemble is shown in Figure 1. The biases are shown for all three observation sources along with the number of catchments used to calculate the bias. Recall that error is defined as forecast *minus* observation. Therefore positive (negative) ME means the weather model has too much (too little) precipitation. Whilst there is month-to-month variability, there is a general under-estimation trend beyond Day 2 onwards. The weather model is generally assessed as least biased when compared to the raingauge rainfall, though it often has the largest positive bias for Day 1. The behaviour of the weather model bias against the merged radar-gauge rainfall product is mixed, though generally it produces biases that are between the raingauge and the radar rainfall observations. The overall trend in the ME over the 16 months is shown in Figure 2, showing that over ~5 seasons the forecasts are generally under-estimating catchment mean precipitation against all observation sources except raingauges on Day 1, with a step-change in bias from Day 3 onwards where the ensemble contains only MOGREPS-G. Days 1 and 2 are similar, being a blend of radar nowcast and MOGREPS-UK on Day 1 and a blend of MOGREPS-UK and MOGREPS-G for Day 2. The precipitation ensemble forecasts have the largest under-estimation bias against radar rainfall. This is perhaps not surprising given that MOGREPS-G at 20 km does not resolve showery precipitation elements very well. These may also not be well represented in gridded raingauge analysis, which can explain why MOGREPS-G seems less biased compared to the raingauge analysis. This simply highlights the representativeness differences between the different observation sources and also between the observation sources and the weather model. From a G2G perspective, it is therefore important to note that the rain volumes ingested have a trend with lead-time and this is likely to have an impact on modelled peak flows in different lead-time windows. Although the bias of G2G and PDM ensemble forecasts was not assessed directly, the case-study analysis does not suggest an overall pattern of under-estimation with some case-studies showing the river flow peaks increasing with decreasing lead-time, whilst others showing the opposite. Of course, the case-study events here were selected based on having river flow impacts so are not generally representative. Thus it is possible that the precipitation underestimation may link to an overall river flow underestimation if all data are considered.

Figures 3 and 4 show the Rank Histograms for daily precipitation given the entire 16-month period to give a sense of the ensemble spread and how this is affected by observation source. In Phase 1, only radar and raingauge rainfall data were available, whilst under Phase 2 the merged radar-gauge rainfall product has been added. For England & Wales, the catchments are covered by all three observation sources (see Figure 3). The BMR precipitation ensemble is under-spread with the distinctive U-shape for Day 1 and Days 2-3. The middle part of the Rank Histogram is relatively flat, but the Days 4-6 time-horizon shows progressively more observations falling in the higher bins, suggesting that often the observation is above the spread of model rainfall values. This is consistent with the bias findings in Figures 1 and 2. All three lead-time horizons show that the merged product has more instances where the

observation falls outside the ensemble lower bound and fewer instances than the other observation sources of being above the ensemble upper bound. Both are a little curious. The lower bound case is probably more so, and one wonders whether this is a somewhat artificial outcome of the merging process. At longer lead-times the radar-based lowest bin tends to be more populated than for the raingauge, owing to the greater discrimination of 0 and the interpolation that is used to create the gridded raingauge rainfall product. The highest bin does show the largest number for the radar-rainfall observations lying outside the ensemble, especially for Days 2-3 and Days 4-6. The merging does reduce this, which may mean that there is a subtle downward shuffling in all the bins, and certainly, the middle bins support this notion. For Scotland in Figure 4 the signal is similar, with a few larger differences for Day 1 where the radar-rainfall shows a much more distinct pattern of being outside the ensemble range of values. The results change little depending on whether 12-months are used. Monthly and seasonal Rank Histograms for daily precipitation accumulations show similar, but noisier results. For brevity these are not shown.

For completeness the hourly Rank Histograms are shown in Figures 5 and 6, again for the entire period from June 2017 to September 2018. The results for England & Wales show very little evidence of a conventional U-shape but do show that the ensemble is under-spread given the number of observations in the highest bin. Surprisingly, it shows the precipitation ensemble is the most under-spread against raingauge rainfall for all lead-time horizons. The merged rainfall product does not sit between the radar and the raingauge values, which again is curious. There is a definite trend in the raingauge-rainfall histogram, whilst the number of observations is more evenly spread between the bins for the merged and radar observations. One possibility is this behaviour is linked to the raingauge-rainfall gridding process and the spreading of rainfall accumulations spatially. Again, the results for Scotland in Figure 6 are broadly similar though the difference in the highest bins between observation sources is less pronounced.

The CRPS and CRPSS (shown here) provide a good summary of the forecast ensemble error in the magnitude of precipitation. When viewed as a skill score this puts all the values into context and enables a better way of aggregating over climatologically different catchments. Here, the whole period Mean Absolute Error (MAE) for each observation source is used as the score reference. For example, all errors are capped or limited by the total amount of rainfall in the observed or forecast totals. If the totals are small the errors are small, if the totals are large, the errors can be larger. Some parts of the country receive more rain than others, and therefore have a greater capacity to have large CRPS. It is only through creating a skill score that these discrepancies can be accounted for in a fair way, especially when aggregating over regions with different climatologies. Figures 7 and 8 show the monthly fluctuations in the CRPSS for the daily and hourly precipitation accumulations for different time-horizons, and against the three observation sources. One rule of thumb (that may not hold very often!) is that scores are lower in the warmer months (spring/summer) and higher in the colder months (autumn/winter). The situation is often a lot more complicated over the UK than that. On the graph, May-Aug 2018 can be seen to perform really well everywhere, and better than Jun-Aug 2017. There appears to be no clear winner in terms of the observation sources, though the precipitation ensemble performs fairly consistently against the radar rainfall accumulations. Differences between the CRPSS computed against the different observation sources over England & Wales are unlikely to be statistically significant on a consistent basis though there are occasions where it may be the case. The picture is somewhat different over Scotland. Here, the weather model performance against raingauges over Scotland in Figure 7(b) appears more variable with large differences, and trends in scores going one way for Day 1 and Days 2-3, with an opposite trend for Days 4-6! As shown in Figure 1, for February 2018 the weather model bias with respect to raingauge rainfall appears better by some margin

compared to that against radar and the merged rainfall product. The CRPS is sensitive to the bias and will reward (or penalise) unbiased (biased) ensemble forecasts. The hourly CRPSS values in Figure 8 show less variation between the observation sources for England & Wales, though again there is more variability over Scotland. There is possibly a clearer upward trend in the 16-month period overall but there appears very little evidence of a seasonal cycle per se.

The spatial uniformity in skill is illustrated using the Year 1 and Year 2 verification periods. Figure 9 shows the CRPSS for the daily precipitation accumulations for the three forecast lead-time horizons against the raingauge rainfall. (Note, most of the maps in this appendix will be shown against raingauge rainfall because of its complete coverage for England & Wales and Scotland.) There is little tangible difference between Year 1 and Year 2 and a slow decrease in skill with forecast lead-time horizon. Scores are generally above 0.4, even for Days 4-6. The exception is over the high ground in Scotland, even on Day 1, where skill is close to, or even less than 0, suggesting the climatological reference is more skilful. This is an emerging theme, evident in all observation sources, and may be related to snow in the colder months. Increasing horizontal resolution does not appear to entirely resolve this issue as the Day 1 results are also poor. tThis apparent deficiency seen for all observation sources is potentially due to deficiencies in the weather model.

In Figure 10 the previously mentioned characteristics of the CRPS and intra- and inter-seasonal variability are shown as monthly maps for the three observation sources of precipitation. As an example, the Days 2-3 lead-time is shown. Firstly, the west-east rainfall gradient across the UK is often apparent, irrespective of observation source. West-facing slopes and uplands are associated with the largest errors (recalling that CRPS for precipitation has units of mm). April 2018 shows a pattern more common under showery conditions, with an increased prevalence of larger errors over South, Central and East England. From these maps the coverage/extent of the different observation sources is also clear. These months illustrate the principle that errors are small where it is driest and sometimes this can have some seasonal dependencies, but this does *not* have to be the case. Overall, the CRPS as a function of lead-time is remarkably consistent between the different observation sources. This could be as much due to the changes in weather model characteristics where the absolute value of the precipitation accumulation error is approximately the same, though Figures 1 and 2 indicate that the sign of the bias changes with lead-time.

Hourly CRPSS also shows considerable uniformity across the UK (not shown). The seasonal CRPSS on the other hand does show some inter-seasonal variability as indicated in Figure 11. Here, the scores are computed using gauge precipitation and all three lead-time horizons are provided to show the decrease in skill with increasing lead-time. One region that stands out again is the Scottish Highlands into the lowlands to the east, which seems to have greater variations in skill and some of the lowest skill. The region is particularly poor in autumn and winter with negative seasonal scores even for Day 1. This may be snow related but could also be linked to the placement of precipitation with respect to higher ground and lack of a rain-shadow effect. Day 1 seems to show some higher variability in scores over Wales too, which is less evident for longer lead-times.

2. User-focused precipitation verification

When the ensemble precipitation forecasts are turned into probabilities the output becomes more user-focused. These probabilities are not used as input to the G2G model for river flow but should be seen as an important component of the flood forecasting and warning decision-making process.

The section on verification analyses employing probabilities begins with Figure 12 which compares the Brier Skill Score (BSS) for two methods of deriving probabilities (as outlined in Appendix A.4) using daily precipitation accumulations for Days 2-3 forecasts. Here the scores obtained using raingauge rainfall are shown. To highlight month-to-month and inter-seasonal variability, four months have been chosen to illustrate the differences between the use of a fixed traditional (non-TWP) and Time-Window (TWP) probabilities for the 8 mm/d precipitation threshold as well as the seasonal and annual 95th climatological percentile thresholds (see Appendix A.3 and A.5). The climatological thresholds were only used with TWPs since the daily 95th percentile annual thresholds range between 15 and 20+ mm across the UK. The daily seasonal 95th percentile thresholds show more contrast between the western fringes and further east, especially in the winter, where thresholds are in excess of 25 mm. Otherwise a general rule of thumb is ~20 mm. Thus, the TWPs derived from the climatological thresholds are for much larger accumulations than what would be possible otherwise. For clarity, the reader is reminded that the results for the traditional non-TWPs are based on the catchment mean; the time-window score is the average Brier Skill Score (BSS) over the time increments in the time-window. For example, for the score representing the Days 2-3 time-window the BSS is calculated from the combined Day 2 and Day 3 probabilities.

In Figure 12 catchments coloured dark red indicate negative skill scores, where the forecasts (on average) are worse than the sample climatology. At first glance results are much noisier and more variable than what has been seen thus far. **For summer, the TWPs for the fixed 8 mm/d precipitation threshold are considerably more skilful with many regions that have negative BSS for the traditional probabilities showing good skill with BSS exceeding 0.4.** The seasonal and annual precipitation thresholds show a lot more negative scores but, given that in most instances the threshold used is above 20 mm, there is still a surprising amount of skill on offer, even at the monthly scale. This is most noticeable in the south, which is perhaps unexpected as this is the driest part of the UK in the summer. The pattern of improvement in the skill of the precipitation forecast through the use of TWPs (by comparing columns 1 and 2) is evident for all months. Skill in the south appears to be worst in April (spring). Whether this is purely down to weather dependence or not is unclear, but springtime is often dominated by small-scale showers which may not be captured accurately by either weather model configurations in terms of location. There seems to be little to choose between the seasonal and annual results, and this is probably because in many instances the precipitation thresholds are extremely similar. Seasonal precipitation thresholds may be more peaked with slightly higher values. On balance, (rolling) seasonal precipitation thresholds may be preferable but come with a bigger maintenance overhead. For this reason, annual precipitation thresholds may be the more pragmatic choice.

The variations in skill across the UK reduce with increasing lead-time and with the verification window used. Figure 13 shows the seasonal BSS for daily totals within Days 4-6 against raingauge rainfall for the predefined seasons. **Again, the 8 mm/d TWPs show considerable improvement over the traditional probabilities in the summer.** This could be ascribed to the fact that timing errors are very damaging to precipitation forecast skill, especially for more convective rain. **TWPs mitigate or remove this timing error, leading to higher skill**. A lot of skill is also gained in the winter, with more subtle gains in the spring and autumn. For the climatological precipitation thresholds the TWPs have lower skill than the 8 mm/d TWPs but in some instances are not that different to the 8 mm/d traditional TWPs, again highlighting how the use of TWPs enables the consideration of much higher thresholds (where they occur) to be considered and to have some positive skill. From a product-generation and user-perspective the 90th percentile daily precipitation thresholds would also be available and could be more skilful but are not shown here. The purpose here was to show the limits of useable precipitation forecast skill. [Results for the 99th percentile are not shown, because the skill for

this precipitation threshold, even when aggregated over seasons or 12 months, remains worse than climatology.]

It is well known that the rainfall intensity-duration relationship is non-linear and capturing short-duration localised events remains a challenging precipitation forecasting problem. For the most part this project has shown that the weather model has better skill at capturing events at the daily time-scale, and this has been further enhanced by the use of TWPs so that precipitation thresholds could be pushed higher. Extreme hourly rainfall is very rare, something which was mentioned in Phase 1 but is well illustrated in Appendix A.5, with 4 mm/h sitting near the maximum value for many catchments, or outside the climatology for some. These are not flood-inducing rainfall amounts. Still, when considering daily rainfall, it is impossible to know whether the precipitation accumulation fell at a steady pace over many hours or whether it all fell in the space of 30 minutes. Therefore, the daily verification results provide the best steer in terms of rain volume to indicate flooding potential because the frequency of truly short-duration sub-hourly downpours is so rare, i.e. difficult to forecast and impossible to verify reliably.

To illustrate this further, Figure 14 shows quite clearly that even on Day 1 the use of TWPs cannot achieve much in terms of extracting skill when applied to a fixed precipitation threshold of 4 mm/h, irrespective of the season. The seasonal and annual climatological precipitation threshold TWPs on the other hand appear to show a much more positive view of forecast skill but the thresholds are all lower than 4 mm/h, and really not of hydrological interest. Most of the thresholds are in the region of 1-2 mm/h of rain. The exception again, is much of the Scottish Highlands and adjacent lowlands, where even precipitation accumulations in excess of 1 or 2 mm/h are struggling to show any skill. Therefore, from a flood forecasting perspective these probabilities are not very useful and certainly not skilful. The use of hourly precipitation forecasting products needs to be considered very carefully. They are valuable in understanding the evolution of rainfall, providing the context for the intensity-duration relationship: that is, is the rainfall short-and-sharp or steady-continuous (building up over time). There is possibly more value in deriving other precipitation products such as the number of hours with more than *x* mm of rain, to capture events that build up over time. Whilst hourly intensity biases were not explored, the intensity biases seen at the daily time-scale are likely to be translated to the hourly one too.

To round off this section, a comparison of the observation sources is provided in Figure 15 which shows the Day 1 BSS for daily TWPs exceeding the 0.5 mm/d precipitation threshold (defining the rain-no-rain boundary) for Year 1 and Year 2. The 12-month scores are very similar between Year 1 and Year 2, and the radar and merged results are more similar to each other than the raingauge results. Overall, the raingauge scores are locally higher than the radar or merged ones, and the lower scores over the eastern Scottish Highlands and adjacent lowlands are seen in the maps produced using radar and raingauge rainfall sources, though the scores using raingauge rainfall are worse. Even the BSS for a 12-month verification window achieves a considerable degree of uniformity across the UK. The anomaly is the very poor performance over south Scotland against the radar and merged observations. This is thought to be signalling a radar artefact or lack of coverage, rather than a weather model issue. The BSS for the hourly Day 1 results for the 0.5 mm/h TWPs are presented for Year 1 and Year 2 in Figure 16. The largest difference is the lower magnitude in the scores (compared to those in Figure 15) with the same spatial patterns. The poor raingauge scores are probably still due to the gridding process in complex terrain. It is unfortunate the merged precipitation product does not extend further north. Radar, for all its failings, does provide better information about the structure and texture of precipitation, even in complex terrain (unless the signal is

completely blocked, which is not the case here). Creating a merged product that covers the whole of Scotland should be a priority.

3. Reliability of probability forecasts of precipitation

A key question for any ensemble forecasting system of precipitation is answering the question: “are the probabilities reliable?”. This is true for the raw precipitation (viewed in the traditional sense) as well as for derived probabilities like the TWPs used here. The first thing that is needed though is resolution or discrimination. Can the ensemble system discriminate between events and non-events? In Reliability Diagram terms, this means the relationship between the forecast probabilities and the observed frequency of occurrence must follow the diagonal or have a decent slope. If that is the case, then the ensemble probabilities could be calibrated. Some aspects of over- and under-confidence can also be related to physical bias. If the underlying weather model has a low bias, then the probabilities are likely to be under-confident, and vice versa. Sometimes a Reliability Calibration can account for these deficiencies in the underlying model, but often it is worth investigating whether the underlying model can be bias-corrected before probabilities are derived. Precipitation is notoriously difficult to post-process in physical space and in this study the raw ensemble is used. Calibrating the probabilities is often the (slightly) easier option but does not solve the issue of translating model biases to downstream flood forecasting models such as G2G.

- The method of deriving the *conventional (non-TWP) precipitation exceedance probabilities is based on the catchment mean* whilst the *TWPs are derived from individual weather model forecasts for grid-cells within the catchment*. They are fundamentally different in construction.
- Days 2-3 is the cross-over between weather model configurations and Figures 1 and 2 suggest that the sign and/or the magnitude of the biases changes substantially, i.e. the probability of exceeding progressively higher precipitation thresholds will decrease with lead-time, and this will affect the conventional probabilities more than the TWPs because there will always be individual weather model grid-cell precipitations exceeding the catchment mean value, so that there will be (many) instances where the catchment mean precipitation does not exceed a given threshold but the TWP for the same catchment will.
- The difference in the catchment precipitation means between the weather models is also determined by the model distribution of rainfall, which has not been discussed so far. This is another feature of the underlying weather model configuration (and horizontal resolution) which has not been investigated here. Other studies have found that the km-scale UM (up to 36h) is somewhat deficient in light precipitation and has been sparser in terms of spatial coverage, with the distribution skewed towards a few large values. The presence of any large weather model grid-cell precipitation values will however lead to a large catchment mean and could lead to threshold exceedance (for Day 1, maybe Day 2), even for the catchment mean (non-TWP) or individual grid-cells (TWP). MOGREPS-G (Day 2 onwards) on the other hand has a large 20 km footprint. Given the size of the grid-cell there tends to be much less (average) rainfall per grid-cell and a more muted distribution overall, which when downscaled to 2 km provides a very homogeneous (possibly bland) rainfall pattern. For the Days 2-3 window these two weather model configurations are combined and provide a very different view, with potentially very different characteristics.
- In Appendix A.4 it is illustrated how the use of TWPs skews the precipitation forecast probability distribution from being positively skewed (small probabilities dominate) to being negatively skewed (larger probabilities dominate), i.e. the TWP distribution has a fatter tail.
- Finally, there are the observation sources. In the plots below the weather model probabilities (non-TWP and TWP) are known to be based on the same model configuration

for all the periods analysed: the differences in the results are related to the observation characteristics and how these compare to the weather model characteristics. Overall,

- catchment precipitation means (for the weather model or any of the observation sources) are going to provide fewer threshold exceedances than individual grid-cells.
- MOGREPS-UK is more likely to produce larger catchment precipitation means and also more large individual grid-cell values than MOGREPS-G.
- Equally, the radar rainfall product is more likely to produce larger catchment precipitation means than the raingauge and merged rainfall products, with more large grid-cell totals in the radar rainfall than the other observation sources.

Figure 17 shows the Reliability Diagrams for Year 2 England & Wales Days 2-3 daily precipitation accumulations against all the different observation sources and the different methods of deriving probabilities and thresholds. All the different thresholds are shown in the same plot together.

The precipitation threshold sequence is described in the left-hand column of Figure 17, with the first row showing the fixed conventional exceedance probabilities (non-TWP) results. The lowest three thresholds are generally under-confident for the probabilities less than 60%. The behaviour for higher probabilities switches to being over-confident, where the precipitation forecast is too keen with the greatest degree of over-confidence evident against the radar rainfall observations and somewhere in-between for the merged product. What is interesting is how good the 8 mm/d non-TWP reliability is against raingauges. This does not seem intuitive. Recall that for this lead-time horizon more than half the precipitation forecasts are from MOGREPS-G which has more muted rainfall. This also implies the catchment means are smooth and on the lower end, producing lower probabilities of exceedance which is a good match for the gridded raingauge analyses, where within-catchment variability is likely to be much lower than for the radar rainfall analysis, for example. It therefore does make sense that the reliability could be quite good against raingauges. For large probabilities the precipitation ensemble forecast remains over-confident against all observation sources, which does suggest that the weather model configuration and ensemble generation is not providing sufficient spread in outcomes.

Considering now the TWPs for the fixed precipitation thresholds in the second row, the TWPs show excellent reliability against raingauges for all thresholds. This simply suggests that the frequencies of occurrence in the raingauge observations and the TWPs is well matched. What is perhaps surprising is the exaggeration of the under-confidence against radar and merged rainfall observations, for a larger range of probabilities. How can this be explained? Recalling that these are the same precipitation forecasts, the answer lies in the identified events in the observations and given that for TWPs the event definition is related to identifying at least two grid-cells in the catchment that exceed the threshold. There are many more identified events in the radar and merged rainfall products, so that the ensemble TWPs are now deficient, except for the highest probabilities. This is reassuring in that when presented with a large probability the precipitation forecast is likely to be fairly reliable and a good steer. Using the seasonal and annual thresholds, the TWPs - which are generally at least 8 mm/d - show over-confidence against raingauges, and a measure of under-confidence against an observation source which is likely to detect more events. Again, the largest TWPs over 80% are exceptionally reliable.

The under/over confidence could be remedied through calibration with the exception of the 99th percentile where in reality even a 12-month sample of precipitation forecasts is insufficient to sample the entire spectrum of probabilities, falling short in detecting enough events with

probabilities exceeding 60-70%. One could try to fit something through a truncated distribution where the sampling is sufficient and assume that the shape of the distribution would follow subject to sufficient sampling.

The results for Scotland are a little different, as shown in Figure 18, which provides the results for Year 2 against raingauges only for the three different lead-time horizons. This reflects the weather model chain fairly well. For the traditional probabilities, for Day 1 a general over-confidence is seen in the probabilities, which exists for all probabilities for the higher two thresholds. This is when MOGREPS-UK dominates. For Days 2-3 there is a shift to more under-confidence and improved reliability for the lower probabilities exceeding 4 and 8 mm/d. This lead-time is a mixture of MOGREPS-UK and MOGREPS-G. For Days 4-6 the shift continues to the left with increased under-confidence of the precipitation forecast probabilities, which is most evident for the lower thresholds. The 4 mm/d precipitation threshold becomes quite reliable whilst the 8 mm/d threshold also has fairly good reliability compared to raingauges, achieving about the same frequency of occurrence based on the catchment means compared to that observed in the raingauge catchment means. It is the only precipitation threshold that continues to provide slightly over-confident probabilities. These lead-times are all based on MOGREPS-G.

The TWPs for the fixed precipitation thresholds in the second row largely mirror the behaviour for Day 1, though the over-confidence for the lowest thresholds increases, which is to be expected. TWPs shift the distribution to the right. For Day 2, the TWPs show considerable improvements in reliability compared to the conventional probabilities, suggesting that the weather model differences are somehow mitigated against, though the TWPs appear to be unable to fully mitigate MOGREPS-G characteristics for Days 4-6, where results look fairly similar to the traditional probabilities, though the 8 mm/d threshold looks to be the most reliable, and especially good for the higher probabilities. The biggest difference seems to be between the seasonal and annual precipitation threshold TWPs against raingauges for England & Wales (in Figure 17) and Scotland (in Figure 18). TWPs appear to be substantially over-confident for Day 1 and Days 2-3. One has to speculate as to how much this is down to the raingauge analysis characteristics in complex terrain and its ability to account for local extremes. Snow could be another observation-related factor. The other possibility is MOGREPS-UK characteristics and the tendency to over-estimate orographic precipitation/enhancement. It is after all the same forecast in England & Wales and Scotland, and whilst the England & Wales results suggest that the probabilities can be improved by calibration, they are not suggesting this level of correction. The 99th percentile results are also comparatively noisy, though again appear to follow a somewhat different path to those over England & Wales.

Figure 19 shows the results for Day 1 over England & Wales using the merged radar-raingauge product. For brevity only the fixed traditional probabilities, fixed TWPs and seasonal TWPs are shown, given how similar the annual and seasonal results are. Recall that there are only three thresholds considered for the hourly totals. Broadly speaking the results vary little with the seasons. This is reassuring, in the sense that any calibration could be independent of seasons. The more distinct differences are related to the probability derivation. Traditional probabilities using the fixed precipitation thresholds tend to be over-confident with the 4 mm/h thresholds suffering from insufficient samples to fully map the full range of probabilities. Fixed TWPs show good reliability on the whole: for the 0.5 and 1 mm/h precipitation thresholds they are somewhat under-confident for low probabilities and over-confident for larger probabilities. The 4 mm/h probabilities tend to be either fairly reliable or over-confident but also vary more by season. In the summer there is a tendency to be under-confident as well. The seasonal TWPs show a somewhat different picture here. The 90th percentile curves are quite noisy

because the precipitation threshold is often very close to 0 (or 0!) for some catchments and therefore has a sampling problem in the warmer seasons. Otherwise the 95th and 99th percentile values are similar to the fixed precipitation thresholds but with somewhat better reliability overall.

The results for Scotland are shown in Figure 20, though here the gridded raingauge product is used. Once again it is best to ignore the 4 mm/h precipitation threshold for the traditional probabilities as having insufficient sample size, along with the 90th percentile TWPs, for the same reason. The results appear to be fairly consistent between seasons. The hourly results are mirroring the daily results for Scotland. The TWPs improve the reliability, especially for the larger probabilities, irrespective of season. Using the seasonal thresholds to compute the TWPs appears to improve the reliability further: the most marked improvement is for the 99th percentile where some of the sampling deficiencies appear to have been removed. All probabilities are generally over-confident, except when very small, so some form of calibration would seem to be necessary to improve reliability further.

This analysis is completed by showing some Reliability Diagrams for monthly samples for hourly precipitation accumulations over England & Wales using the seasonal percentile thresholds against all the observation sources. Figure 21 attempts to show some of the issues that arise with reducing the sample size when using less than ~90 days. Again, the forecasts are the same so that the differences seen are due to the observation source alone. The 90th percentile precipitation thresholds are noisy because they often relate to thresholds which are near 0. It's worth remembering that for the hourly precipitation totals even the 99th percentile is only between 2 and 3 mm/h in most cases. The TWPs are fairly reliable for this precipitation threshold against raingauges and show similar tendencies towards under-confidence for Days 2-3 (see Figure 17). The reliability against the merged rainfall product is somewhere in-between. Against the raingauges the over-confidence is primarily in the cooler months, but overall, the reliability for the 95th and 99th percentile shows little sign of sampling problems. The 90th percentile has greater sampling problems against radar and merged rainfall products, with reliability extremely good against these products in the cooler months. In the spring there is good reliability against radar rainfall and the merged rainfall product for the larger probabilities but shows signs of being under-confident at the lower end. The weather model may not be capturing showery precipitation with enough spatial density that is seen in the radar rainfall observations.

4. Summary and recommendations

In this study 16-months of ensemble precipitation forecasts have been evaluated from which the following thematic commentary can be drawn.

On time-window probabilities

- TWPs extract useful information content by removing or mitigating against timing errors.
- TWPs allow the use and evaluation of higher precipitation thresholds, which better reflect the user needs for identifying and assessing potential flood risk from heavy precipitation.
- TWPs change the reliability and can switch an under-confident traditional probability to being reliable or over-confident. The distribution of precipitation forecast probabilities shifts from being positively skewed to negatively skewed, i.e. a so-called "fat tail" with more large probabilities which increases user confidence.
- TWPs improve the sample size for larger precipitation thresholds because of considering individual grid-cell values, which helps to improve the stability and robustness of skill scores.

- TWPs computed using the seasonal and annual percentile precipitation thresholds show much better reliability and for higher thresholds. This is good news from the user perspective.
- There is little to choose between seasonal and annual precipitation thresholds. If a simple solution is sought then annual precipitation thresholds would do, though seasonal thresholds preserve more of the intra-annual variability which exists across the UK.
- How would these be computed in real-time? Catchment-scale is possible, but one could choose to compute TWPs over clusters of catchments, e.g. the EA regions. (Though at this point it is worth pointing out that using TWPs - even with higher precipitation thresholds - clearly shows the issue at the catchment-scale is not a lack of detection!) Nevertheless, the case studies would suggest that regional probabilities of precipitation alongside more specific catchment information may be beneficial. This need could be satisfied through the generic neighbourhood-processed IMPROVER output.

On observation sources and biases

- Characteristics of the observations can have a strong influence on results, which can lead to the drawing of opposing conclusions about the performance of the same weather model forecast, either over the same area (England & Wales) or over England & Wales and Scotland. The latter would seem to be unrealistic, especially if this is against the same observation source!
- Though an assessment of the precipitation forecast bias would suggest that the forecasts are the least biased against the raingauge analysis (based on catchment means), this is considered somewhat misleading and gratuitous as at longer ranges the forecasts are strong under-estimates and the raingauge analysis may provide a poorer reflection of localised maxima relative to radar and merged rainfall products.
- The merged rainfall product would appear to be a good compromise for providing the texture that an interpolated raingauge analysis does not have, whilst improving an inherent radar rainfall bias. This should be available right across the UK.
- Good precipitation observation source QC is essential. The results over the Scottish Borders point to some kind of observation issue in the radar rainfall data which is translated to the merged rainfall product.

On biases

- The fact that the weather model bias against the merged rainfall product follows the raingauge bias quite closely shows how the raingauge data are used to correct the observed radar rainfall amounts.
- MOGREPS-G is unable to resolve the detail in the radar rainfall fields, so it is unsurprising that the catchment means compare well to the gridded raingauge product given that they are the most similar in nature: smooth with many features similar at sub-grid-scale for both.
- There should be some concern (or at least acknowledgement) that the raingauge mean error here is gratuitous rather than a true reflection of the precipitation forecasts being the least biased against the gridded raingauge analysis. Any mismatches in more localised maxima of precipitation (which act to shift the catchment mean), tend to be reflected more in the weather model biases computed against the merged and radar rainfall products, and make them larger. This should be considered the more realistic view.
- Physical biases can feed into the biases of the probabilities that are described here. The precipitation ensemble forecast is not seamless in time. Where the weather models are joined together is evident in the results. From a flood forecasting perspective, the volume of water is of interest, highlighting the need and benefit of adjusting the forecast rainfall values. However, precipitation is difficult to adjust, especially in the extremes. This is

definitely a long-term research problem, especially what impact any adjustments may have to the downstream tuning of a river flow model.

On verification periods

- Precipitation ensemble forecast skill does vary from month-to-month and season-to-season, but regional differences are likely to be bigger and more persistent. Annual results appear to be fairly robust though 99th percentile results are still somewhat sparse.
- Seasonal results appear to be stable too. If recent performance is of particular interest, a rolling 3-month window may well be very useful alongside something that tracks performance for 12 months or more. This ensures that the weather dependencies are better accounted for.
- Monthly results can be useful but, in terms of inferring continual performance, some form of rolling performance information would be more useful.
- Some form of reliability calibration *could* be beneficial to make the precipitation ensemble forecast probabilities more reliable. This could be done without addressing the underlying physical biases. Given the differences in the precipitation observation sources, some careful thought would need to be given to which observation data to use for this purpose (both physical and probability).

References

Mittermaier, M.P., Roberts, N. and Thompson, S.A. 2013. A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Meteorol. Apps*, **20**, 176-186.

Rainfall and River Flow Ensemble Verification: Phase 2
Precipitation verification maps and plots
Final Report Appendix B.1.2

This appendix contains a subset of the total precipitation verification maps and plots produced by the project, relevant to explaining its aims and outcomes. The commentary on these is provided in Appendix B.1.1. Whilst most elements are explained in the commentary it is worth pointing out a few here to make the figures easier to browse.

The selection of plots attempts to provide an overview of all of the following in as succinct a way as possible:

- *three* precipitation observation sources: raingauge-only, radar-only and merged radar-raingauge
- *daily* versus *hourly* precipitation accumulations
- *monthly*, *seasonal* and *12-monthly* statistics
- *fixed* precipitation thresholds (as used in Phase 1) compared to the use of *long-term annual and seasonal climatological percentile thresholds* values shown in Appendix A.5
- *conventional* probabilities (for fixed thresholds using the catchment mean rainfall) compared to *Time-Window Probabilities* (TWPs). The derivation of these is explained in Appendix A.4.

Several specific periods were investigated to understand the inter-annual and intra-annual variability in scores and metrics.

- Y1 Year 1 June 2017 to May 2018
- Y2 Year 2 September 2017 to August 2018
- S1 Spring Year 1 JJA 2017
- A1 Autumn Year 1 SON 2017
- W1 Winter Year 1 DJF 2017/18
- Sp2 Spring Year 2 MAM 2018
- S2 Summer Year 2 JJA 2018

The following scores and metrics are shown.

- Brier Score and/or Brier Skill Score (against sample climatology) to consider the skill in the precipitation ensemble probabilities
- Continuous Ranked Probability Score and/or Skill Score to assess the distribution of the precipitation ensemble values
- Rank Histograms to assess the spread of precipitation ensemble values
- Reliability Diagrams to assess the precipitation ensemble probability bias

Finally, the mean error (bias) in the daily catchment mean precipitation, as provided by the control member of the precipitation ensemble was computed to quantify the underlying rainfall intensity bias. This is important because the Best Medium-Range (BMR) ensemble evaluated here is a combination of three forecasts: STEPS (0-6h), MOGREPS-UK (0-36h) and MOGREPS-G (36-149h). The first 36 hours are strongly modulated by MOGREPS-UK, but Day 2 is a mixture of MOGREPS-UK and MOGREPS-G, whereas Days 3 to 6 are all MOGREPS-G. These weather models are very different in resolution and underlying behaviour. Results are provided for the different observation sources available over England & Wales and Scotland.

Figure 1 Monthly mean error (bias) in the daily catchment-mean precipitation from the BMR control member.

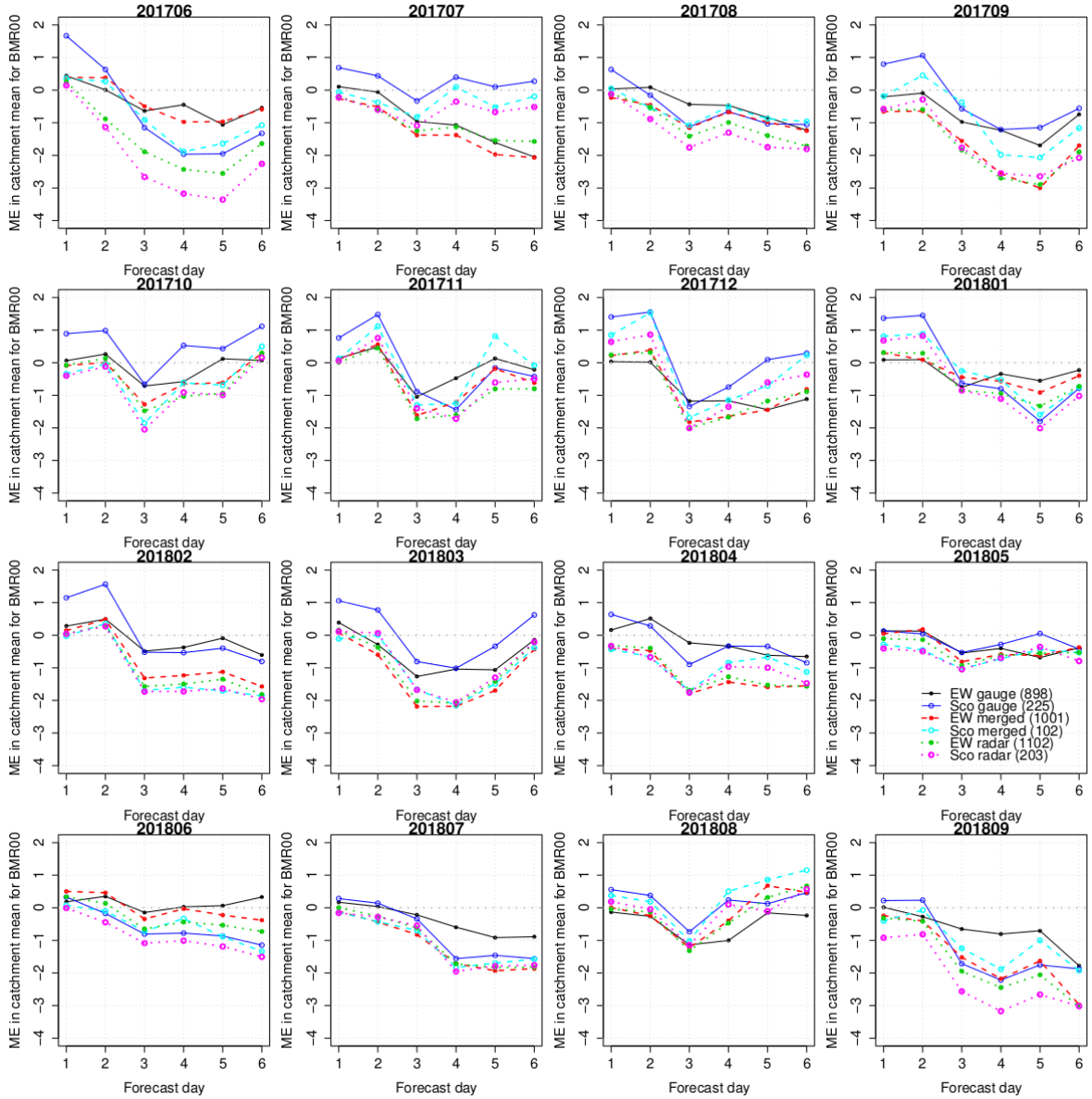


Figure 2 Average mean error (bias) in the daily catchment-mean precipitation from the BMR control member.

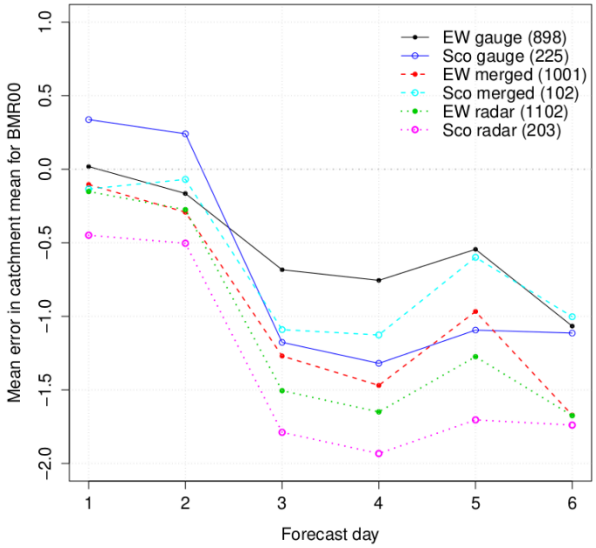


Figure 3 Rank Histogram (equalised across observation sources) for daily England & Wales for the three different observation sources. Whole Period (June 2017 to September 2018).

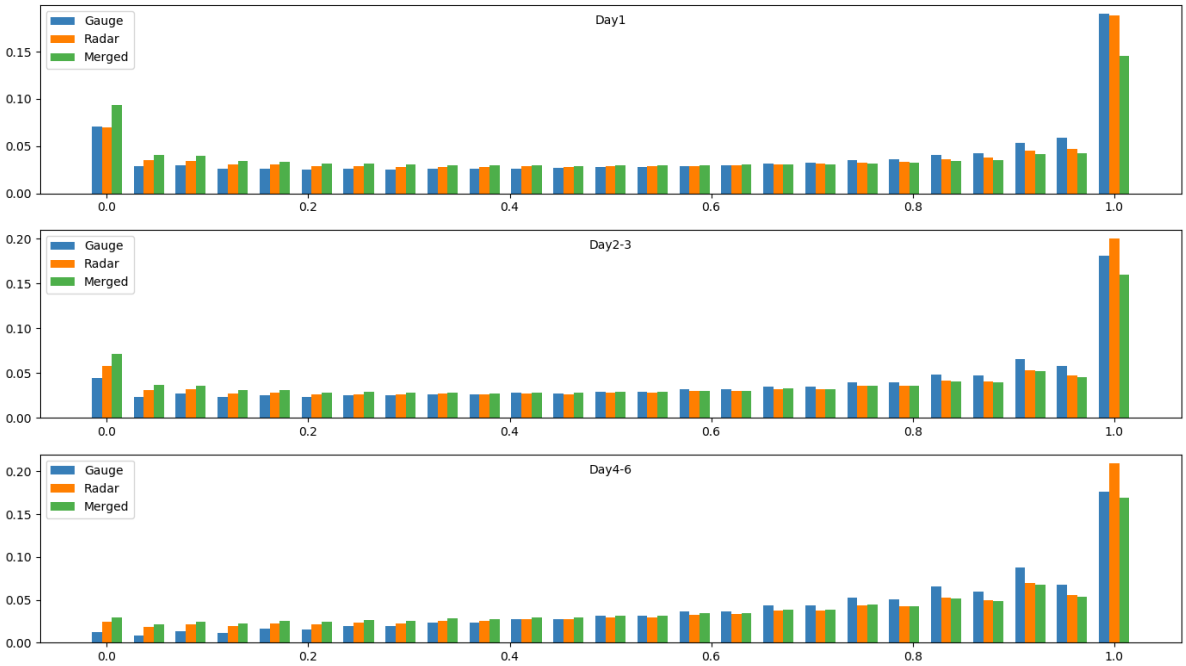


Figure 4 Same as Figure 3 but for Scotland.

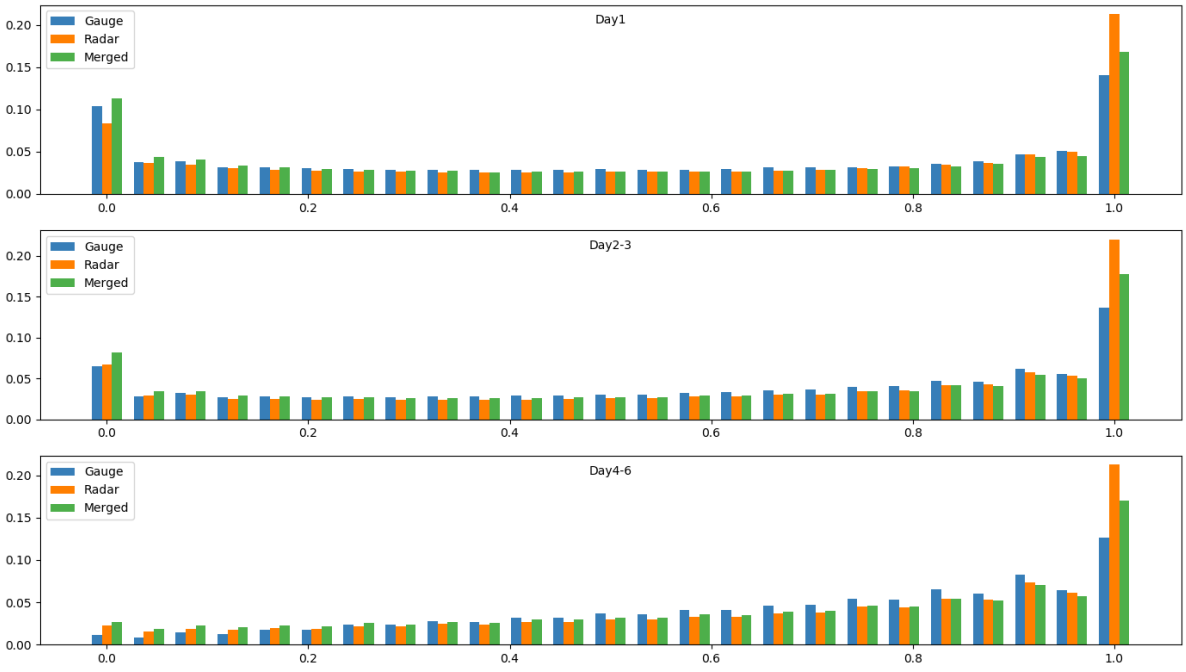


Figure 5 Rank Histogram (equalised across observation sources) for hourly precipitation accumulations in England & Wales for the three observation sources. Whole Period (June 2017 to September 2018).

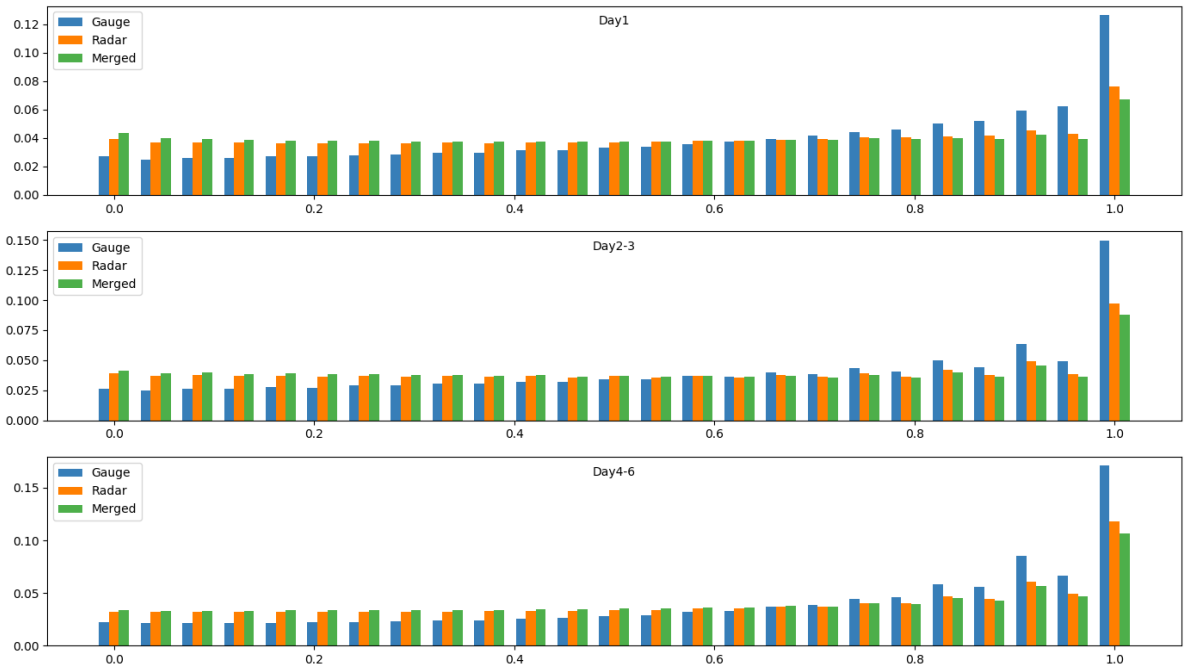


Figure 6 Same as Figure 5 but for Scotland hourly precipitation accumulations.

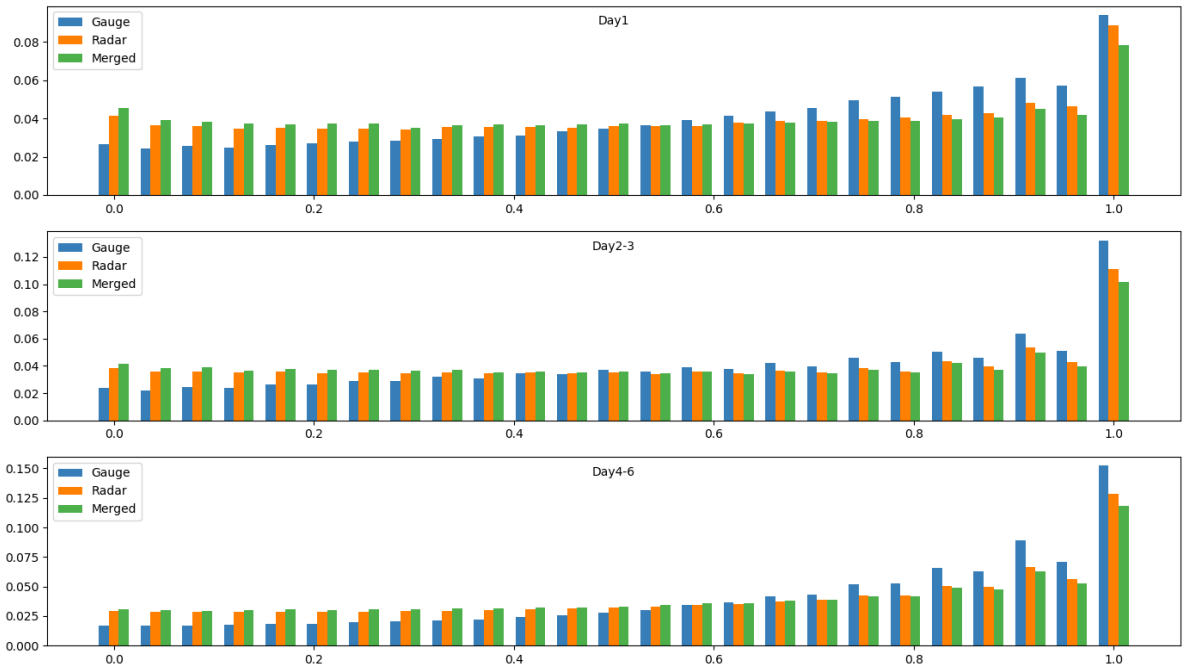
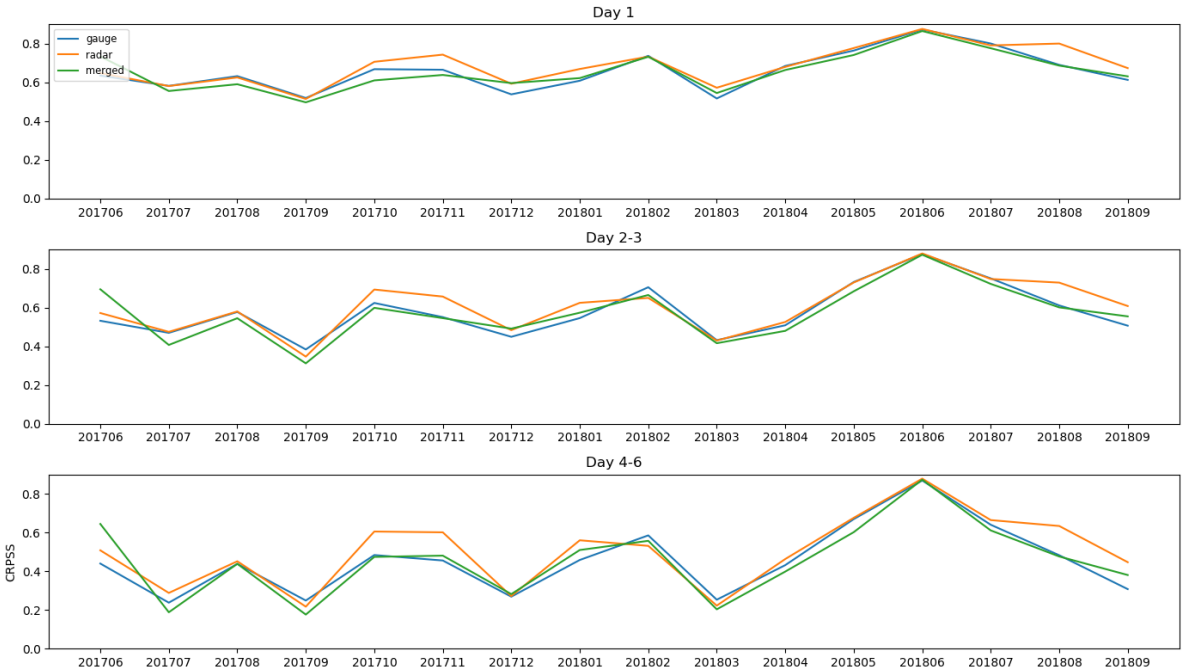


Figure 7 Monthly CRPSS for daily precipitation accumulations (referenced with respect to the whole-period MAE for each observation source), equalised across observation sources and using an agreed site list (148 sites over Scotland and 731 sites over England & Wales).

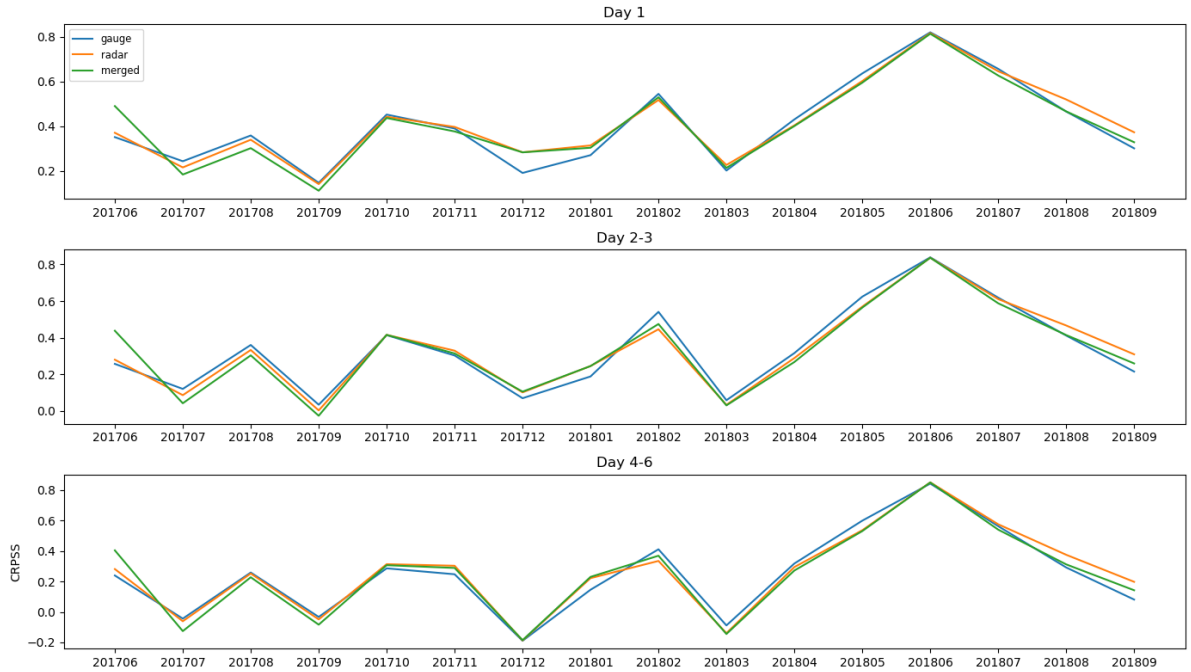
(a) England & Wales



(b) Scotland

Figure 8 Same as Figure 7 but for monthly CRPSS values based on hourly precipitation accumulations.

(a) England & Wales



(b) Scotland

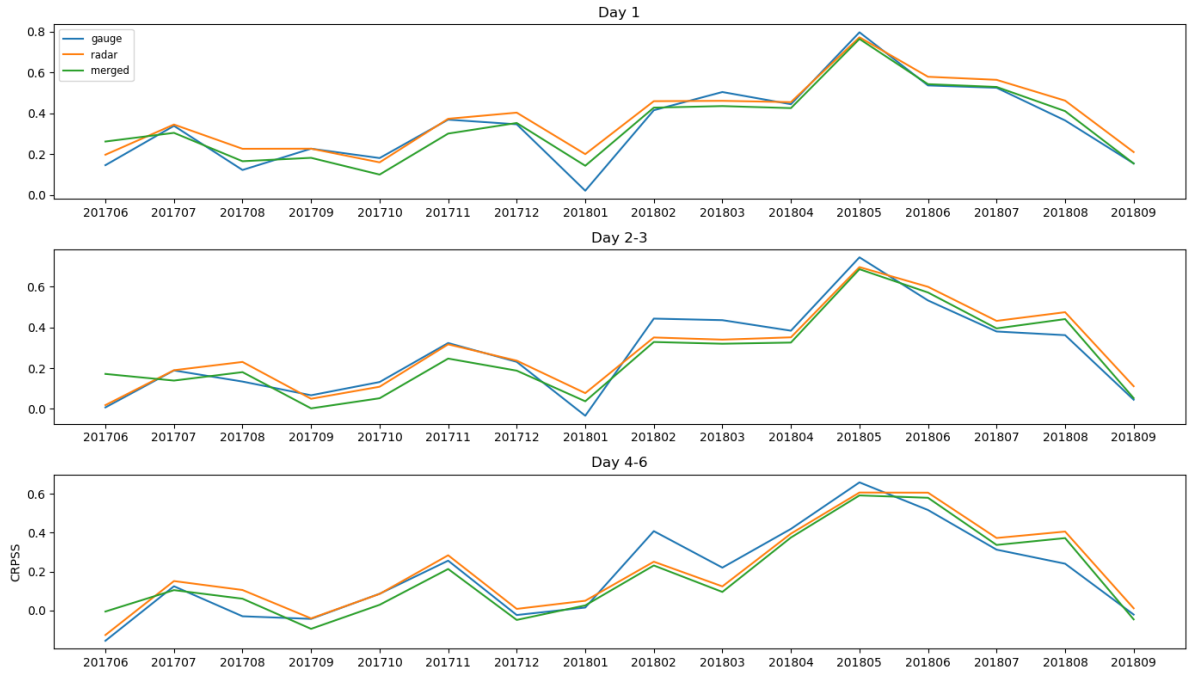


Figure 9 CRPSS for Year 1 and Year 2 for the daily precipitation accumulations compared to raingauge rainfall.

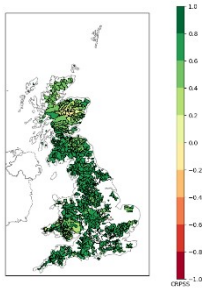
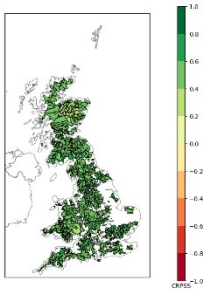
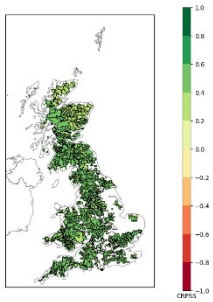
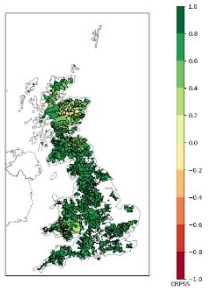
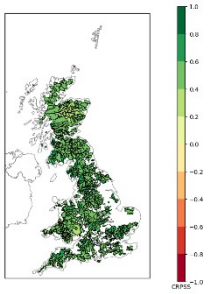
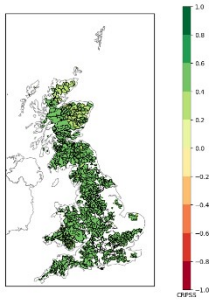
	Day 1	Days 2-3	Days 4-6
Y1			
Y2			

Figure 10 Monthly CRPS in units of mm for the daily precipitation accumulations for Days 2-3 comparing the different observation sources. It illustrates the CRPS association with the MAE and shows that the errors scale with the rainfall amounts, i.e. when the CRPS was low the rainfall amounts were relatively low too.

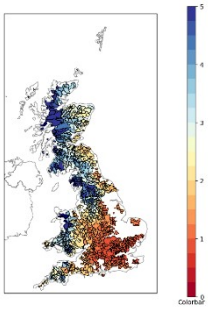
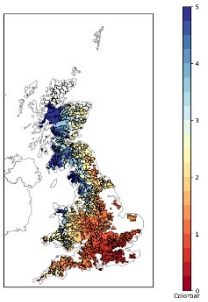
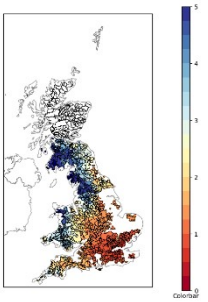
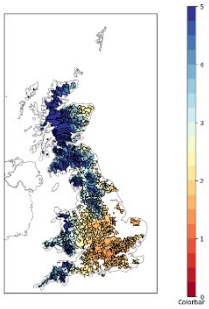
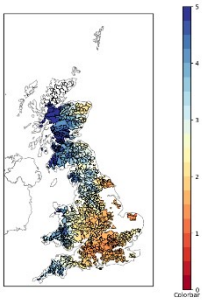
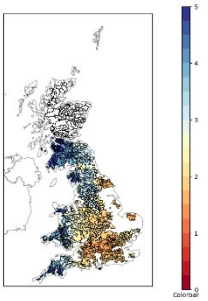
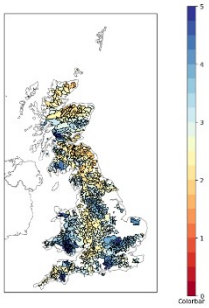
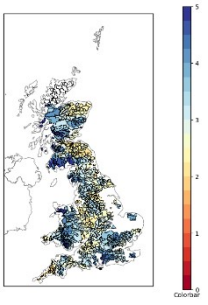
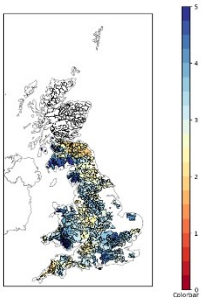
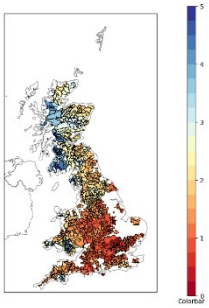
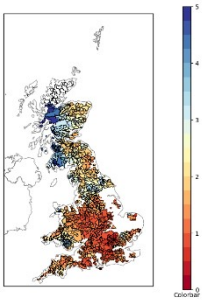
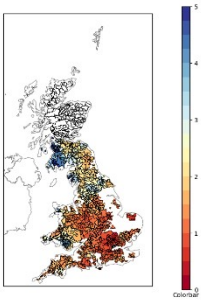
	Gauge	Radar	Merged
Oct 2017			
Jan 2018			
Apr 2018			
Jul 2018			

Figure 11 CRPSS for hourly precipitation accumulations over different against gridded gauge rainfall as providing the most comprehensive coverage.

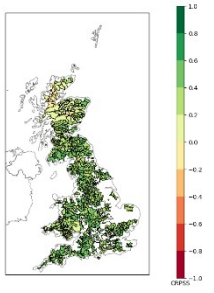
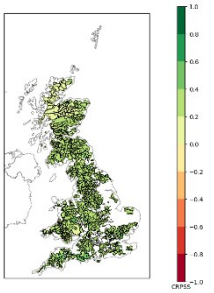
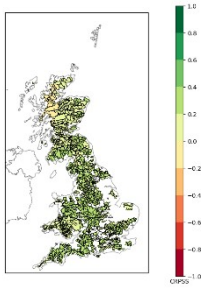
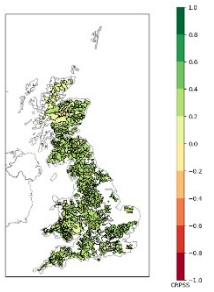
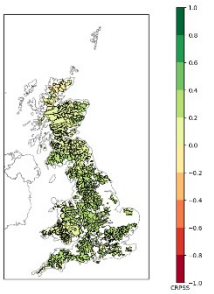
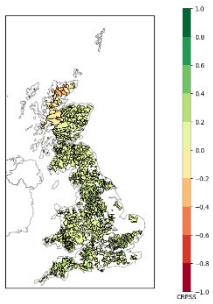
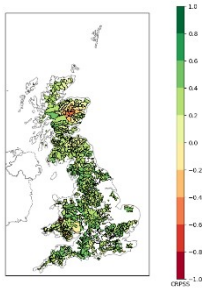
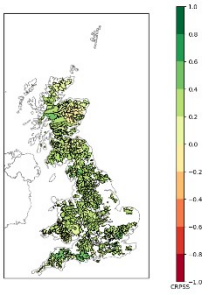
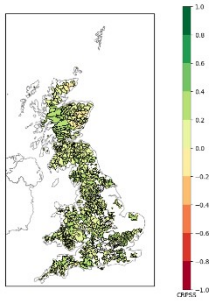
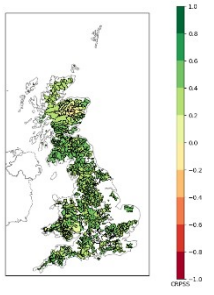
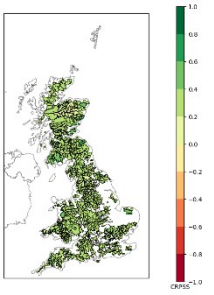
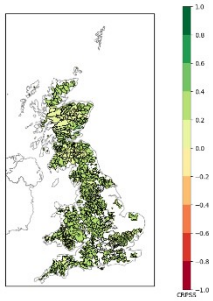
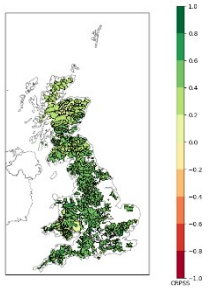
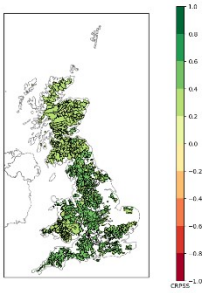
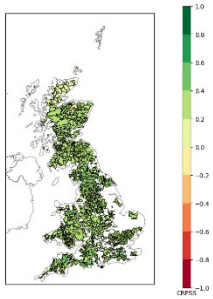
	Day 1	Days 2-3	Days 4-6
S1			
A1			
W1			
Sp 2			
S2			

Figure 12 Monthly BSS for daily precipitation accumulations within Days 2-3 against gridded gauge rainfall comparing the different methods for deriving probabilities and thresholds.

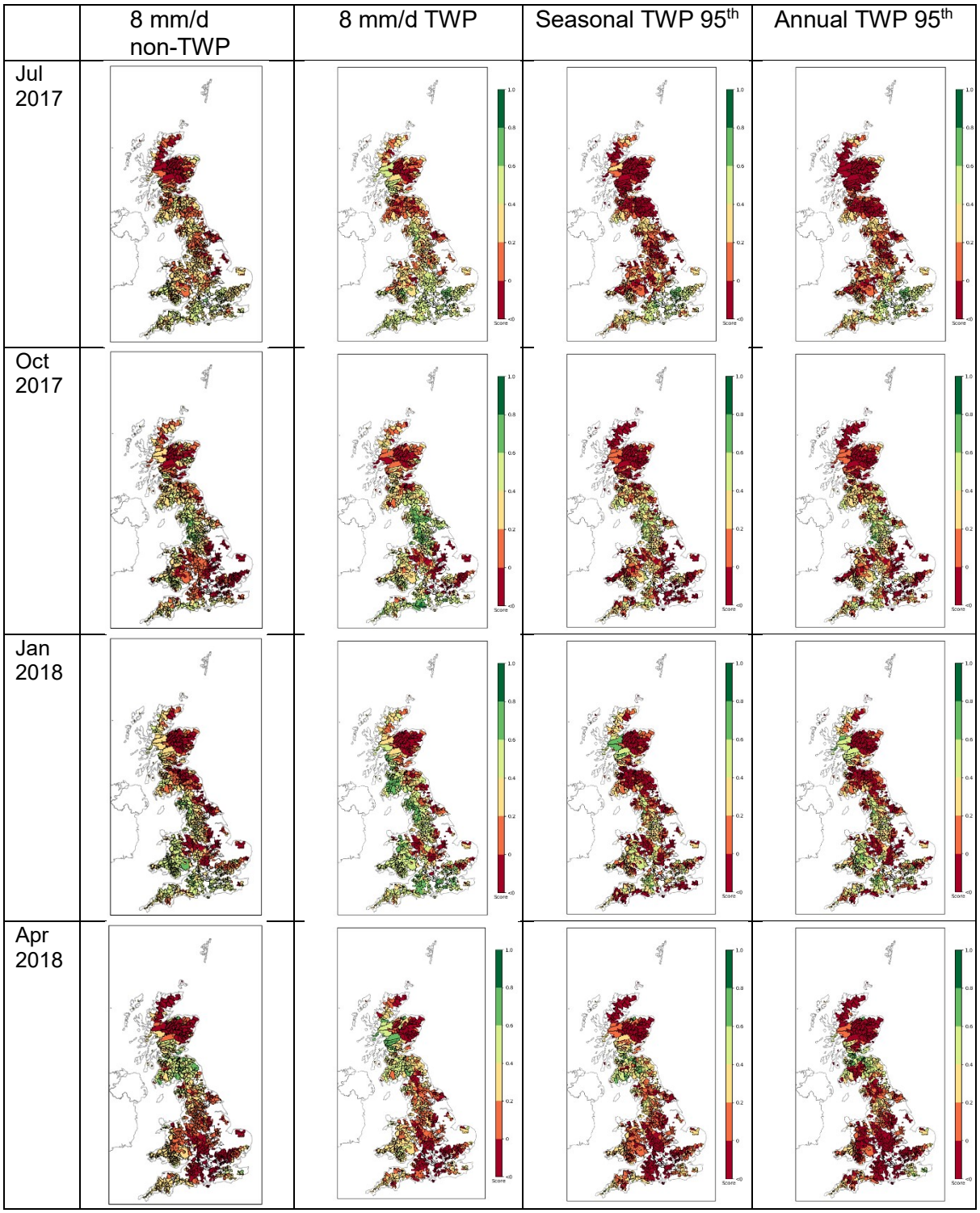


Figure 13 BSS for daily precipitation accumulations over different considering Days 4-6 forecasts against gridded gauge rainfall, comparing different methods for deriving probabilities and thresholds.

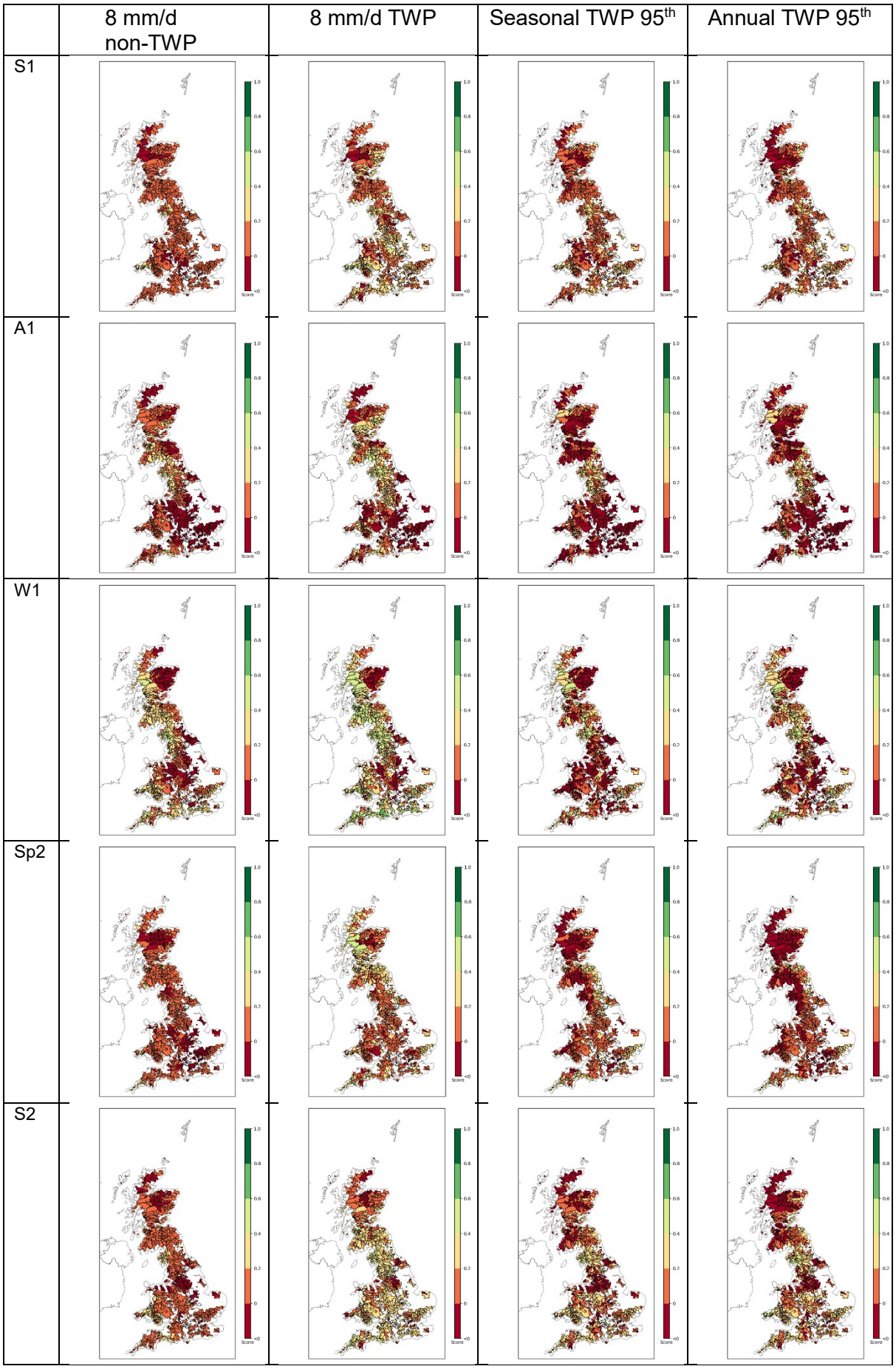


Figure 14 BSS for Day 1 hourly precipitation accumulations over different seasons considering Day 1 forecasts against gridded gauge rainfall, comparing different methods of deriving probabilities and thresholds.

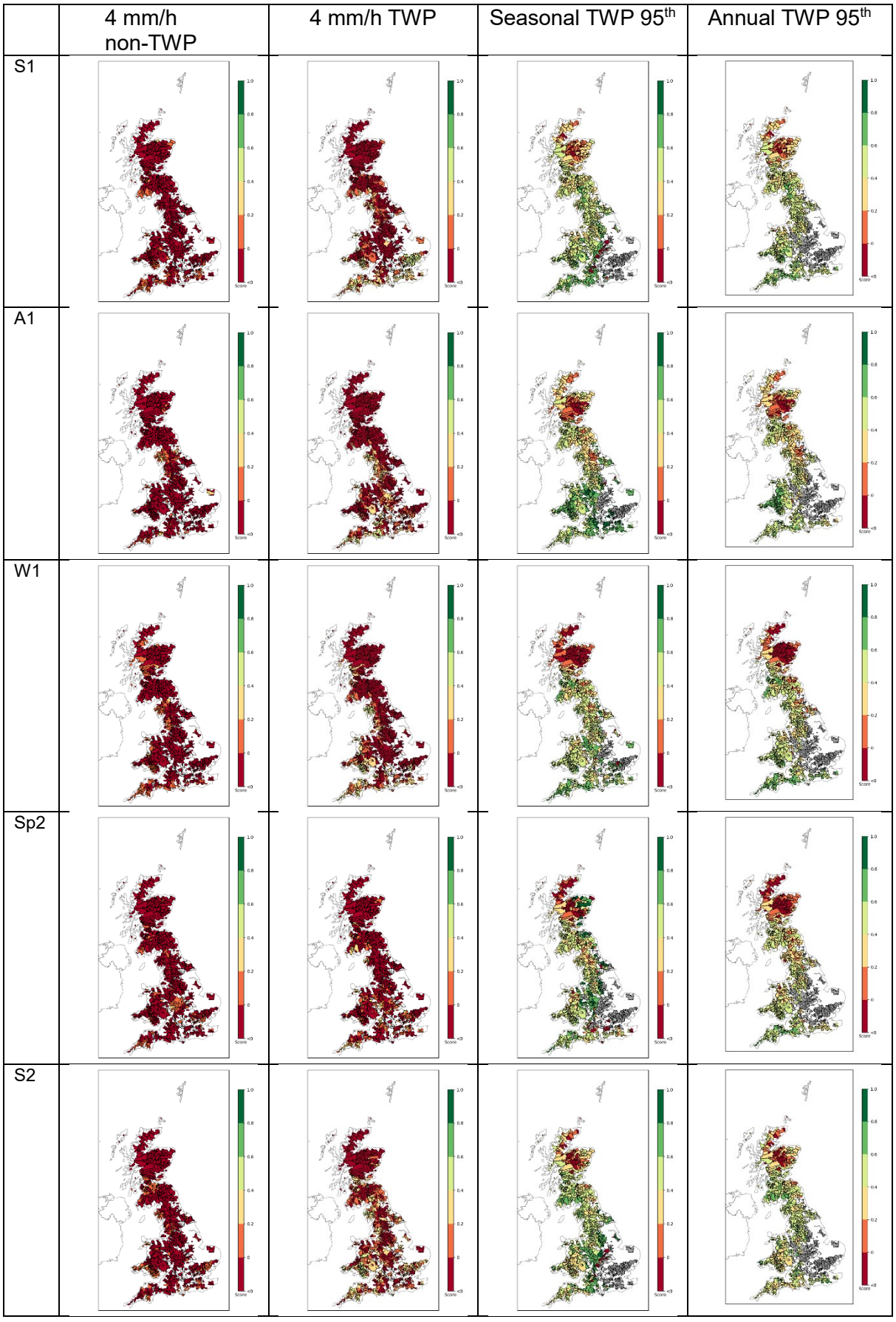


Figure 15 BSS for Day 1 daily precipitation accumulations over Year 1 and Year 2 for the 0.5 mm/d TWPs comparing the different observation sources.

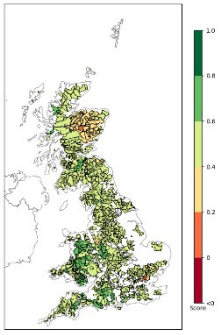
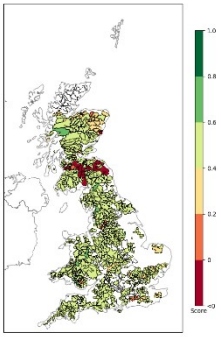
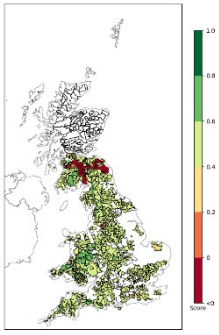
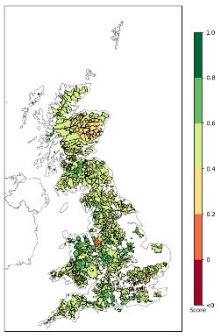
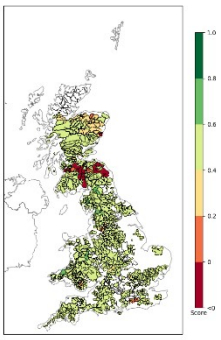
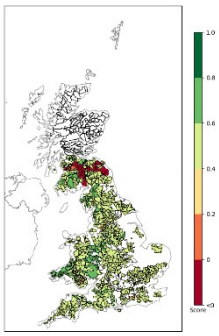
	Gauge	Radar	Merged
Y1			
Y2			

Figure 16 BSS for Day 1 hourly precipitation accumulations over Year 1 and Year 2 for the 0.5 mm/h TWPs comparing the different observation sources

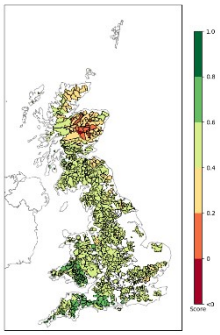
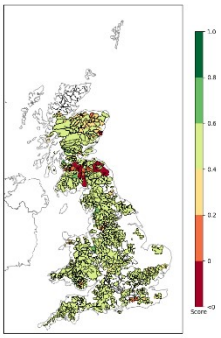
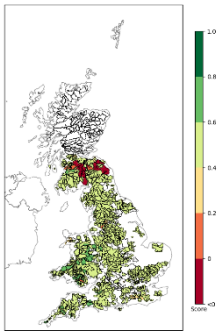
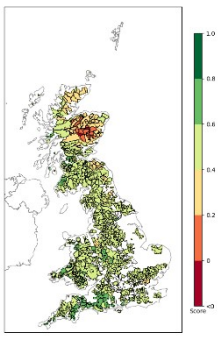
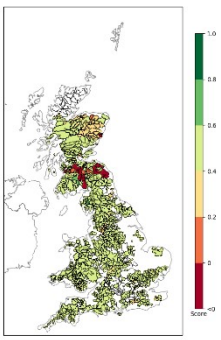
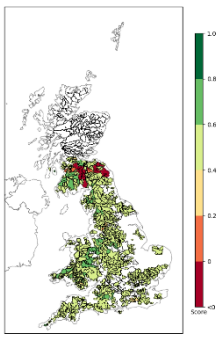
	Gauge	Radar	Merged
Y1			
Y2			

Figure 17 Reliability Diagrams for Days 2-3 daily precipitation accumulations for Year 2 over England & Wales showing the differences by observation source and threshold-probability derivation.

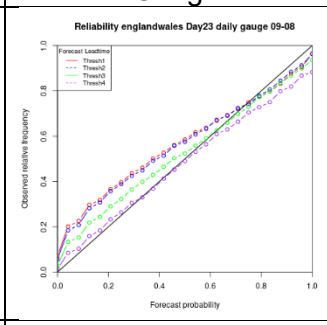
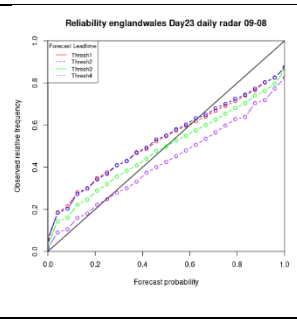
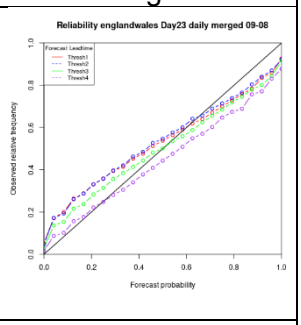
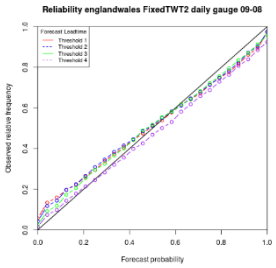
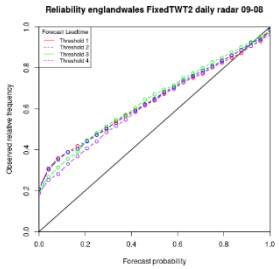
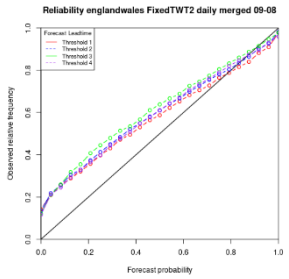
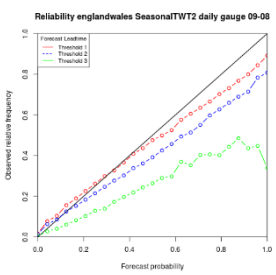
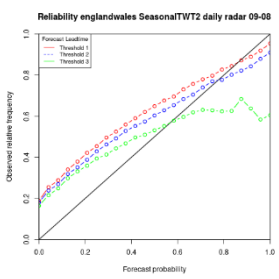
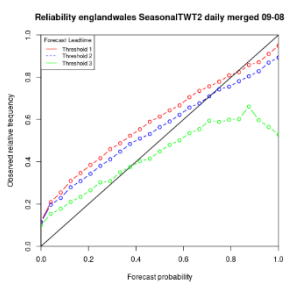
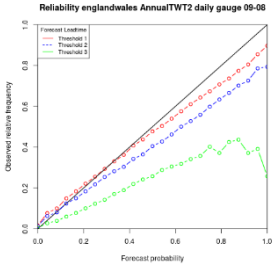
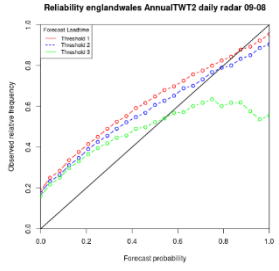
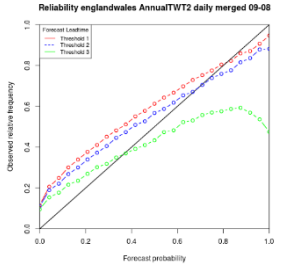
	Gauge	Radar	Merged
Fixed non-TWP (0.5, 1, 4, 8)			
Fixed TWP (0.5, 1, 4, 8)			
Seasonal TWP (90, 95, 99)			
Annual TWP (90, 95, 99)			

Figure 18 Reliability Diagrams for daily precipitation accumulations over Year 2 for Scotland based on gridded gauge rainfall.

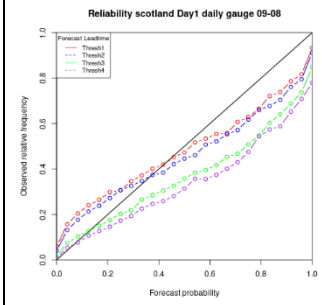
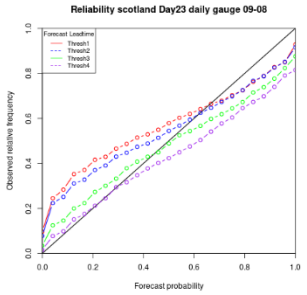
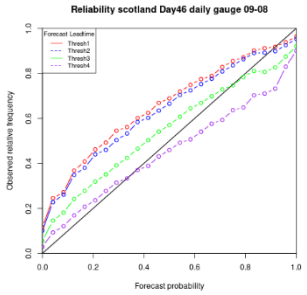
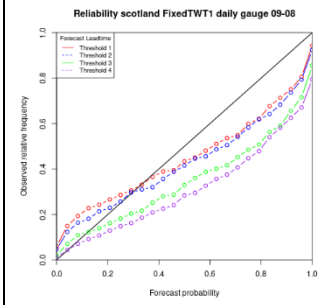
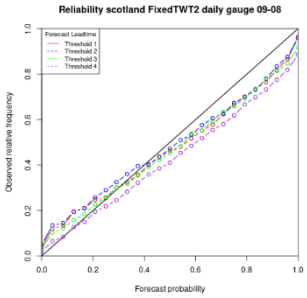
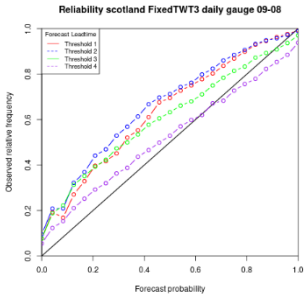
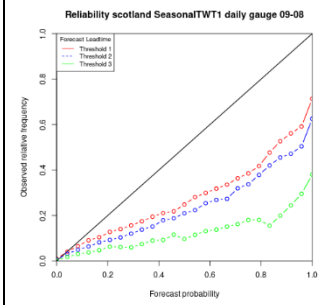
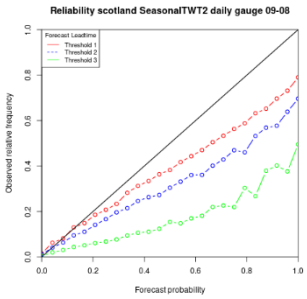
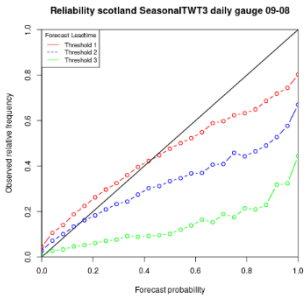
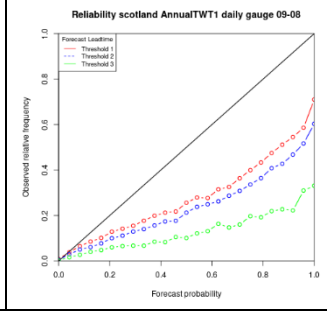
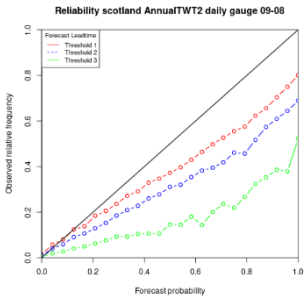
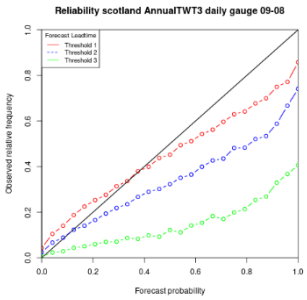
	Day 1	Days 2-3	Days 4-6
Fixed non-TWP (0.5, 1, 4, 8)			
Fixed TWP (0.5, 1, 4, 8)			
Seasonal TWP (90, 95, 99)			
Annual TWP (90, 95, 99)			

Figure 19 Reliability Diagrams for Day 1 hourly precipitation accumulations over England & Wales for different seasons against the merged radar-gauge rainfall product for different threshold-probability derivation methods.

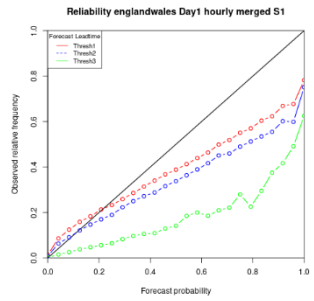
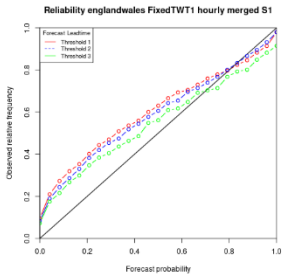
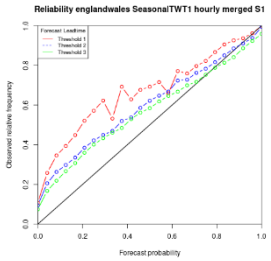
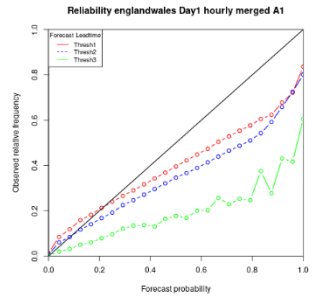
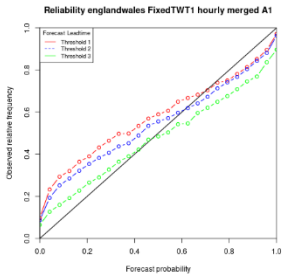
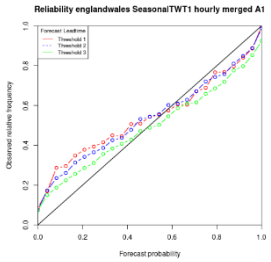
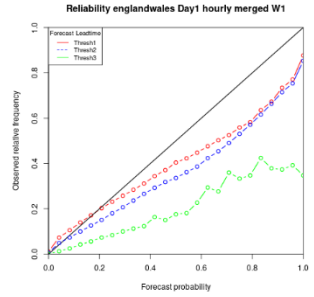
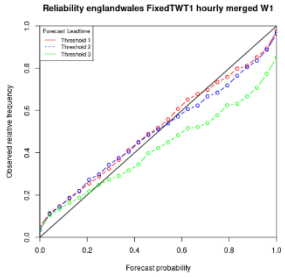
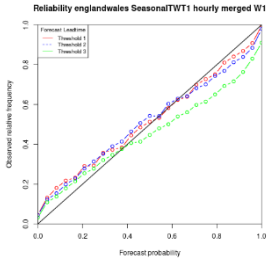
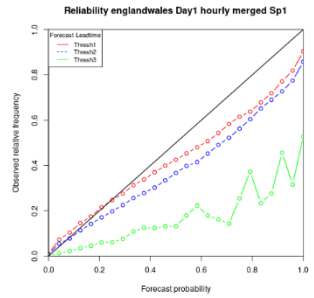
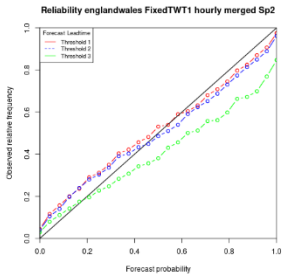
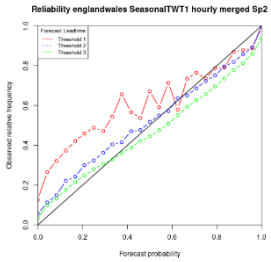
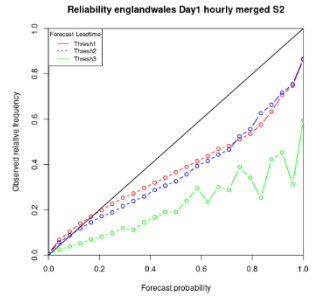
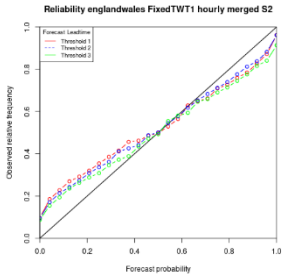
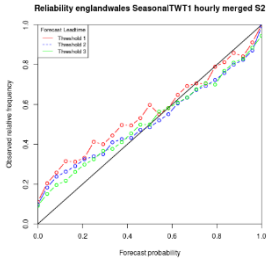
	Fixed non-TWP (0.5, 1, 4)	Fixed TWP (0.5,1,4)	Seasonal TWP (90,95,99)
S1			
A1			
W1			
Sp2			
S2			

Figure 20 Reliability Diagrams for Day 1 hourly precipitation accumulations over Scotland for different seasons against the gridded gauge rainfall for different threshold-probability derivation methods.

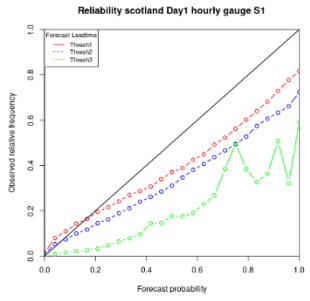
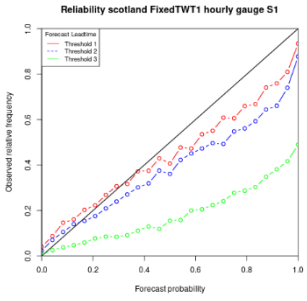
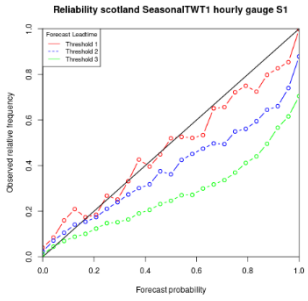
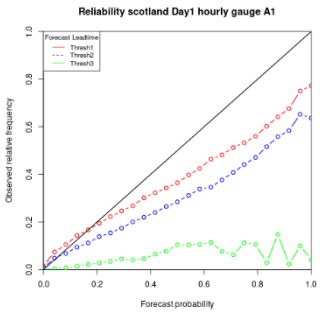
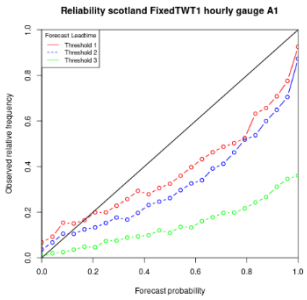
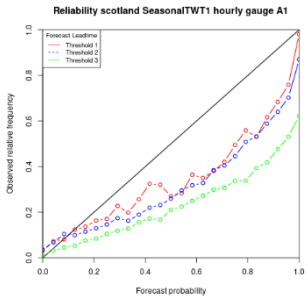
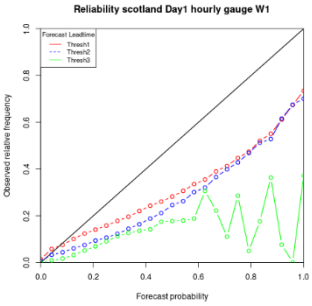
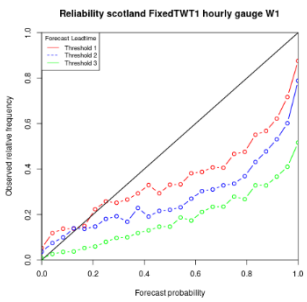
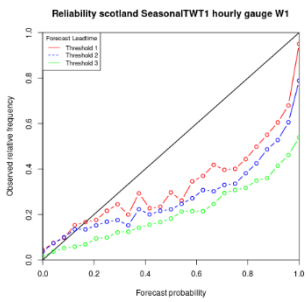
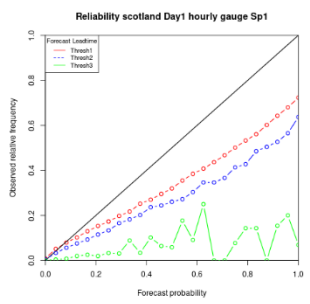
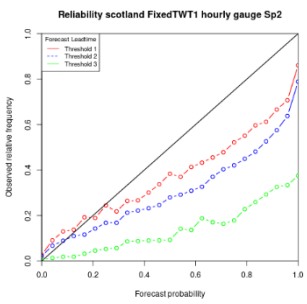
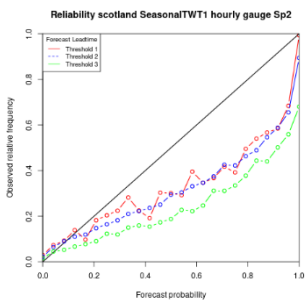
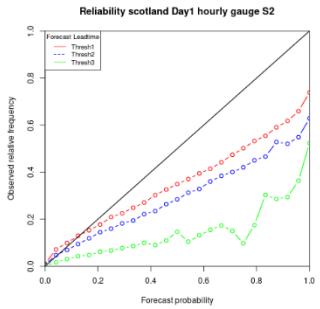
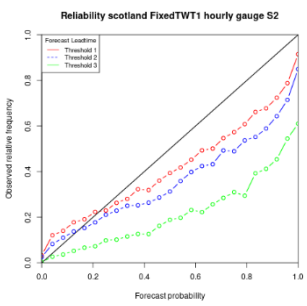
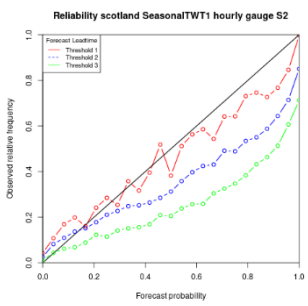
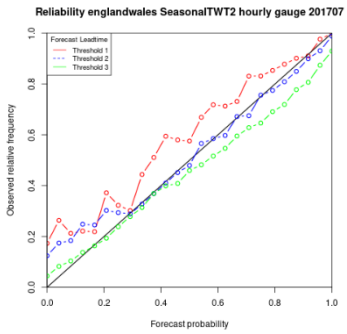
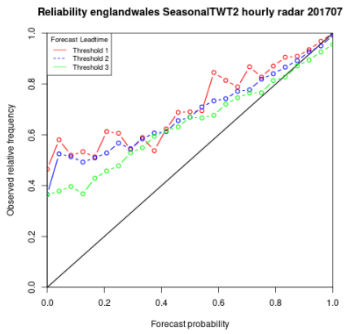
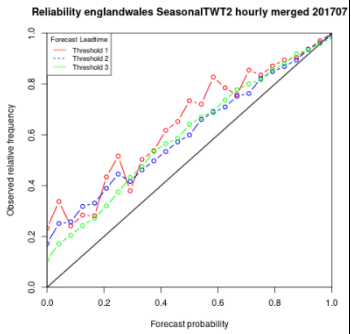
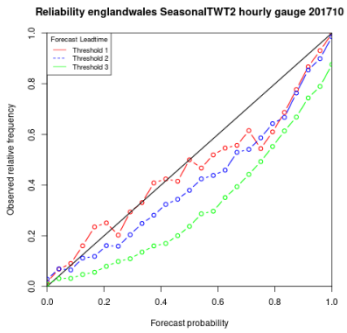
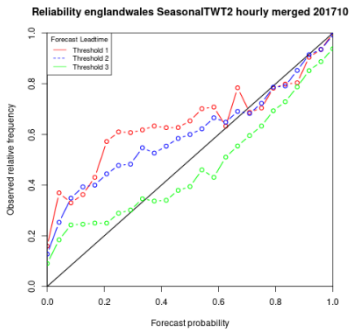
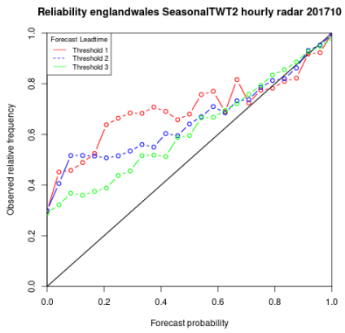
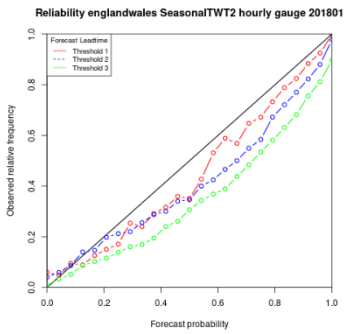
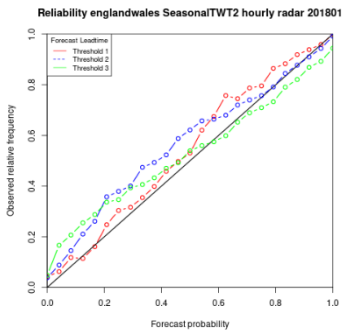
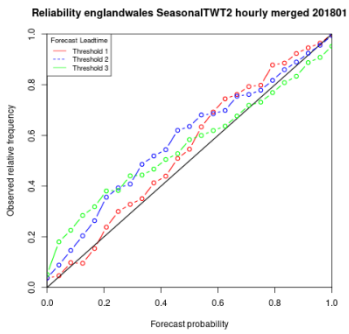
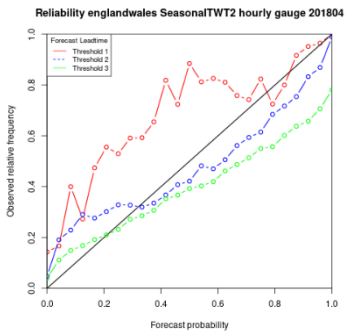
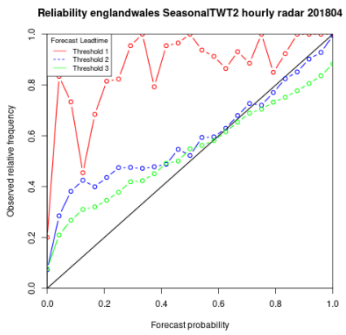
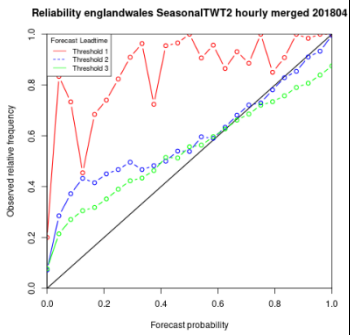
	Fixed non-TWP (0.5, 1, 4)	Fixed TWP (0.5,1,4)	Seasonal TWP (90,95,99)
S1			
A1			
W1			
Sp2			
S2			

Figure 21 Monthly Reliability Diagrams for Days 2-3 hourly precipitation accumulations over England & Wales, computed for the seasonal percentile thresholds showing the impact of observation source. Thresholds: 90th, 95th and 99th percentiles.

	Gauge	Radar	Merged
Jul 2017			
Oct 2017			
Jan 2018			
Apr 2018			

Ensemble rainfall and river flow verification: Phase 2
Overall verification summary: river flow analyses
Final Report Appendix B.2.1

This document first summarises the findings from the G2G river flow analysis for the Phase 2 verification periods (falling between 1 June 2017 and 31 August 2018) in Section 1. The aim is to use the 15-month period to assess the sensitivity of verification diagrams and scores to the length of assessment period, and provide guidance on the appropriate verification period length for an operational River Flow Ensemble Verification System. The effect of seasonality on the verification output is considered, along with the appropriate spatial scales needed to obtain meaningful verification analyses. A comparison is made with the Phase 1 analyses from the much-shorter, but abnormally wet, Phase 1 analysis period of December 2015. The focus here is on providing a written overview, discussing output verification diagrams and maps, and providing key conclusions. For clarity of presentation, only key summary figures are presented. For completeness, all other river flow verification plots are provided in appendix

Appendix_B_2_2_Overall_Verification_Summary_River_Flow_plots.zip,

with plots grouped into separate PDF files by G2G domain (the filename prefix "SEPA_" indicating the Scotland domain) and verification period

Table 1 Verification periods considered within the Phase 2 period 1 June 2017 to 31 August 2018. The period starts from the first day of the "Start" month listed and ends on the last day of the "End" month listed.

	Summer2017	Autumn	Winter	Spring	Summer2018	Year 1	Year2
Start	Jun 2017	Sep 2017	Dec 2017	Mar 2018	Jun 2018	Jun 2017	Sep 2017
End	Aug 2017	Nov 2017	Feb 2018	May 2018	Aug 2018	May 2018	Aug 2018

Based on the overall findings from these analyses, the full Year 2 period (Sep 2017 to Aug 2018) was selected to verify the PDM local models. In contrast to the different spatial-scale verification analyses used for G2G, a number of example single-site PDMs have been verified. These results are presented in Section 2.

The Precipitation verification analyses are summarised separately in Appendix

Appendix_B_1_1_precipitation_verification_commentary.pdf

with plots provided in

Appendix_B_1_2_precipitation_verification_plots.pdf.

Details of the verification metrics considered are given in

Appendix_A_1_Joint_Verification_Framework.pdf

1 River flow verification analyses for G2G

1.1 Threshold-based score analyses

1.1.1 Seasonal variation

The number of threshold-crossings (Figure 1 of River Flow Verification Summary) varies with season, with the smallest number seen in the summer when baseflows are lower and flood events tend to be linked to intense convective precipitation. In autumn (at least for the 2017 period considered), river flow threshold-crossings are more widespread, but limited to catchments in Wales, Scotland, and the north and west of England. In winter and spring, river flow threshold-crossings are seen throughout England & Wales, although the number of catchments experiencing threshold-exceedances of above Q(2)/2 is still relatively small. An example for the Q(2)/2 threshold is shown in Figure 1. For Scotland, the spring 2018 season has very few threshold-crossings, even for the Q(2)/2 threshold, and predominantly affecting catchments in the south and east. The winter 2017-18 season has more threshold-crossings for Scotland but, as for England & Wales, the number of catchments experiencing threshold-exceedances of above Q(2)/2 is still relatively low. Of course, these results are to be expected as the return-period thresholds relate to the expected return-period of an event. For example, for the Q(2) threshold with a return period of 2 years, this would not be expected to be exceeded for the majority of catchments in any one year. This again highlights the exceptional nature of the December 2015 period used for the Phase 1 analysis where many catchments had crossings of the higher Q(T) thresholds, even though only one month of data were considered.

Comparing the two summer periods considered, 2017 and 2018, there are far fewer threshold-crossings in the abnormally dry summer 2018 period than for 2017 when more normal meteorological conditions prevailed (e.g. Figure 1, left-hand side). This highlights the effects of inter-annual variability on the verification sample sizes, and suggests that caution is needed when drawing verification conclusions from one season of data. This effect of *sampling uncertainty* when looking at individual seasons is also seen when comparing the verification diagrams (Figures 2 to 7 and 9 to 14 of the River Flow Verification Summary). Even for the lowest threshold considered (Q(2)/2), the summer-season Reliability and ROC diagrams are dominated by sampling uncertainty, even for low forecast probability values. This is also true for the spring-season Reliability and ROC diagrams for Scotland. For the other seasons, the national-scale verification diagrams (e.g. Figure 2 of the River Flow Verification Summary) are consistent, at least for the Q(2)/2 threshold, although the effects of sampling uncertainty are still evident for the highest forecast probability bins. Overall, these diagrams lead to similar conclusions to those obtained from the Phase 1 analyses (e.g. Phase 1 Report Figures 4.1, 4.6, 4.11) with performance decreasing with forecast lead-time, and the river flow ensemble tending to over-forecast (the probability values tend to be too high) and also be over-confident (larger probabilities are more over-forecast). Example Reliability Diagrams, calculated over different verification periods for England & Wales, are shown in Figure 2. At the regional-scale (River Flow Verification Summary Figures 6 and 7), sampling uncertainties dominate for the majority of regions in south and east England suggesting that the seasonal sample-size is not sufficient to obtain meaningful inferences at this scale. An example for England & Wales is shown in Figure 3. Similar conclusions are drawn from all catchments at the catchment-scale (River Flow Verification Summary Figures 9 to 11), even when catchments are pooled by catchment size (River Flow Verification Summary Figures 12 to 14).

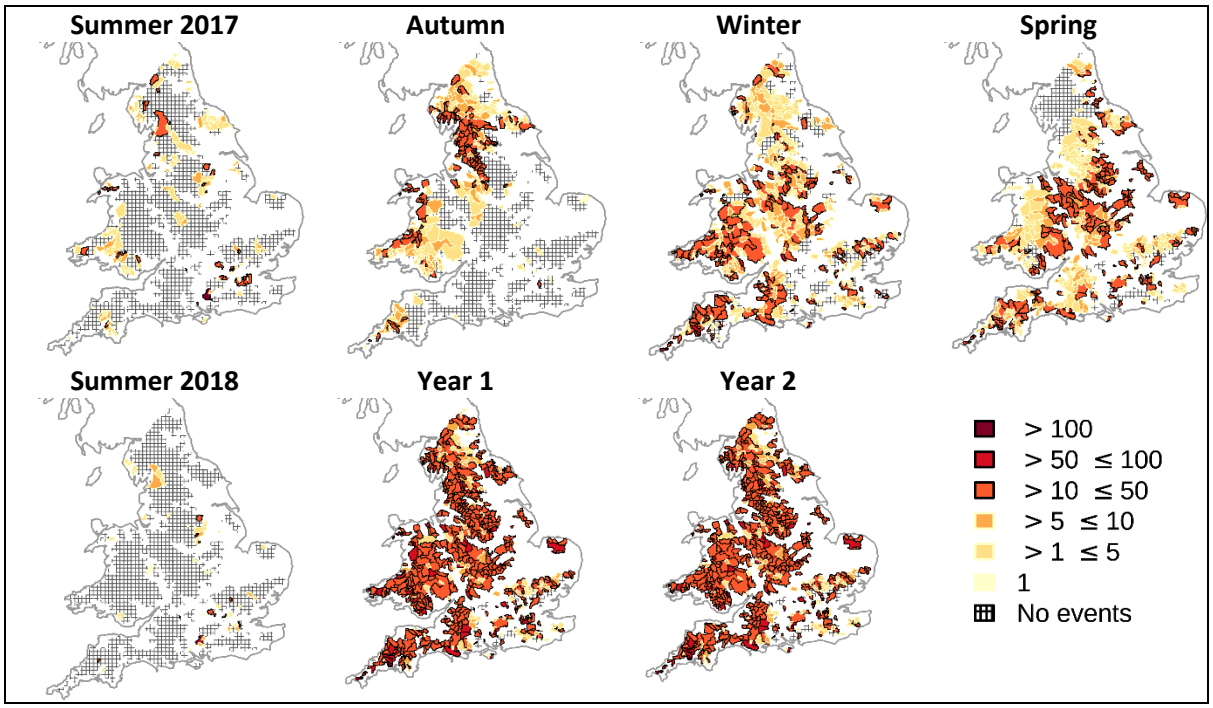


Figure 1 Number of river flow forecasts having observed threshold-crossings over different verification periods. For river flow threshold Q(2)/2 and time-periods corresponding to Day 1 forecasts.

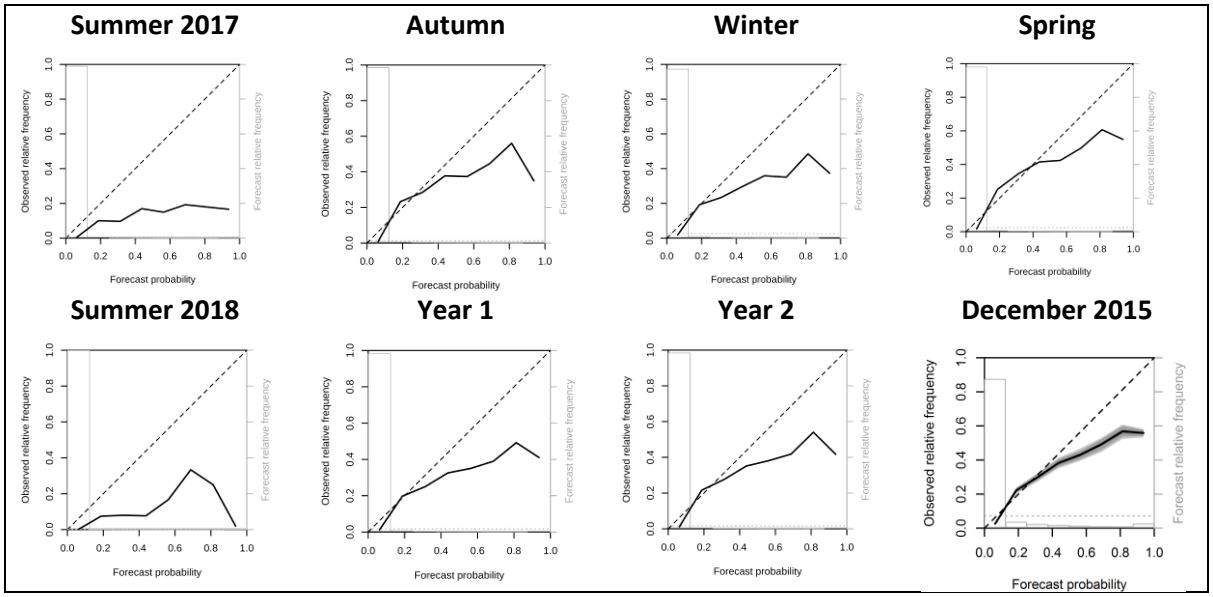


Figure 2 G2G river flow Reliability Diagrams over different verification periods, pooled over all sites in England & Wales. For river flow threshold Q(2)/2 and Day 1 forecasts.

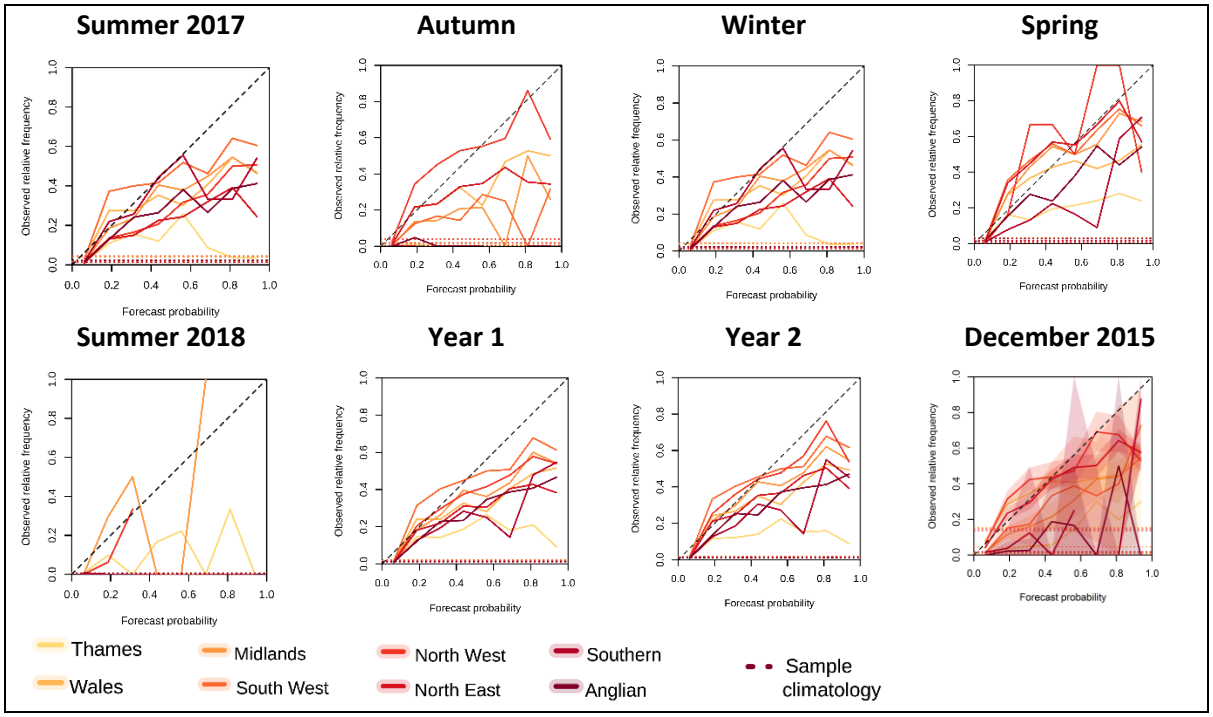


Figure 3 G2G river flow Reliability Diagrams over different verification periods, for all sites in England & Wales split by region. For river flow threshold Q(2)/2 and Day 1 forecasts.

1.1.2 Annual variation

To assess the effect of considering slightly different 12-month verification periods, verification analyses were compared for the periods 1 June 2017 to 31 March 3018 (i.e. including summer 2017) and 1 September 2017 to 31 August 2018 (i.e. including summer 2018). These periods are denoted as Year 1 and Year 2 respectively. Although the two summer periods were noticeably different, with 2018 being much drier with fewer threshold-crossings, their 12-month period analyses are expected to be similar as they will be dominated by a majority of threshold-crossings which occur in the months of September to May (e.g. Figure 1). This is indeed seen when comparing verification analyses at the national-, regional- and catchment-scales for the Q(2)/2 and Q(2) thresholds. Similar to the analyses for autumn and winter seasons, those for the national-scale 12-month verification period generally agree with those from Phase 1. Figure 4 shows an example for England & Wales.

For Scotland, some differences are seen in the national- and regional-scale Reliability Diagrams, with the full 12-month verification period analyses showing the ensemble to over-forecast more (the probability values tend to be higher) and also be more over-confident (larger probabilities are more over-forecast). This brings the Scotland analyses closer to those seen for England & Wales, and suggests that the abnormal December 2015 period was influencing the interpretation of the Reliability Diagrams for Scotland. For thresholds above Q(2), high sampling uncertainties are seen for the 12-month analyses suggesting that there are insufficient threshold-crossings to support this type of verification analysis. Figure 5 shows a comparison with the Phase 1 analyses (provided in full in the Phase 1 Report Figures 4.1 and 4.2). It suggests that the sampling issue for high thresholds is worse for the more-normal 2017 to 2018 12-month periods of Phase 2 than was the case for the extremely wet December 2015 period of Phase 1. This is an important consideration for an operational verification system: although a long and recent verification period is desired to capture up-to-date

weather model behaviour, to capture extreme events it may be necessary to include a verification period longer in the past.

For catchment-scale analyses (e.g. maps of Brier Skill Score and ROC Skill Score; River Flow Verification Summary Figures 9 to 14, summarised below in Figure 6), the spatial pattern of ensemble skill is different from that seen for the Phase 1 analyses (e.g. Phase 1 Report Figures 5.1 and 5.4). In particular, the trend of poorer skill measured by the Brier Score for the southeast of England seen in the Phase 1 analyses is not seen for the 12-month verification period. As discussed in the Phase 1 Report, this was thought to be associated with only a small number of threshold-crossings occurring in southeast England in December 2015, which is not the case for the 12-month verification period which has much more spatially-uniform coverage of threshold-crossings. Where the Phase 1 analyses showed a clear trend of skill decreasing with forecast lead-time, this is not clear from the 12-month analyses. This may again be related to the sample size, with larger sample sizes associated with the longer-duration 48h Days 2-3 and greater than 48h Days 4-6 (England & Wales) verification time-windows. This conjecture is supported by consideration of the verification analyses for Scotland, where a shorter post Day 3 time-window is used and individual site performance at these lead-times is poorer than that seen for Days 2-3.

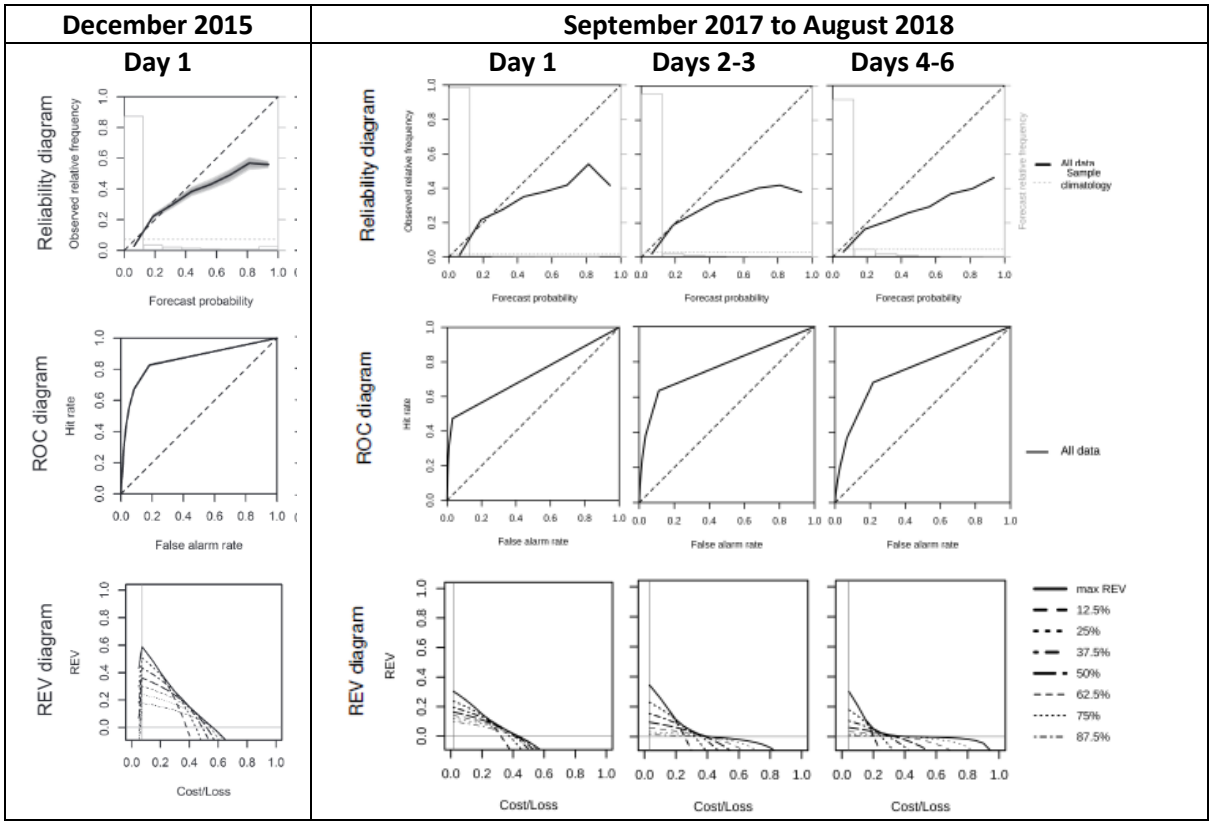


Figure 4 River flow verification diagrams calculated using data pooled from all catchments for England & Wales. Reliability, ROC, and REV diagrams are shown using the Q(2)/2 threshold over the Phase 1 December 2015 period for Day 1 forecasts (left) and over the Phase 2 12-month period September 2017 to August 2018 for Day 1, Days 2-3 and Days 4-6 forecasts (right).

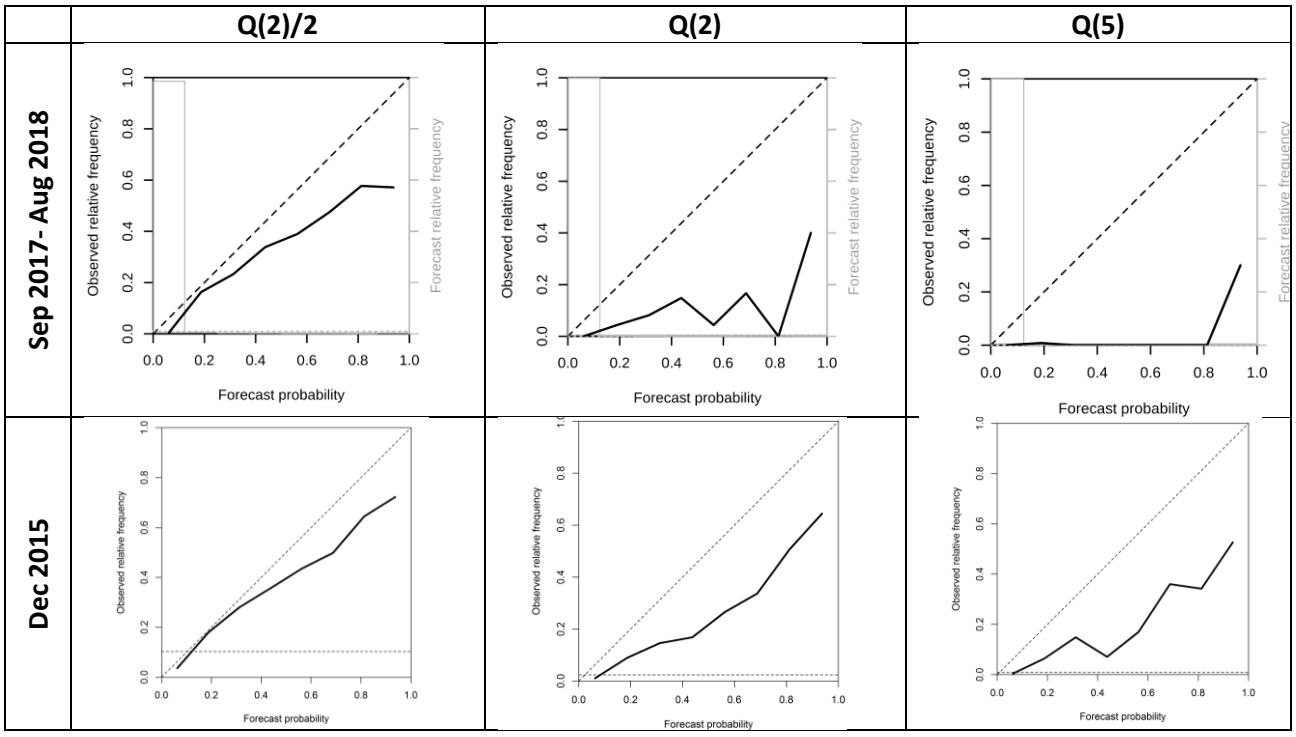


Figure 5 River flow Reliability Diagrams calculated using data pooled from all catchments for Scotland. For Day 1 forecasts and Q(2)/2, Q(2) and Q(5) thresholds over the Phase 1 period December 2015 (bottom) and Phase 2 12-month period September 2017 to August 2018 (top).

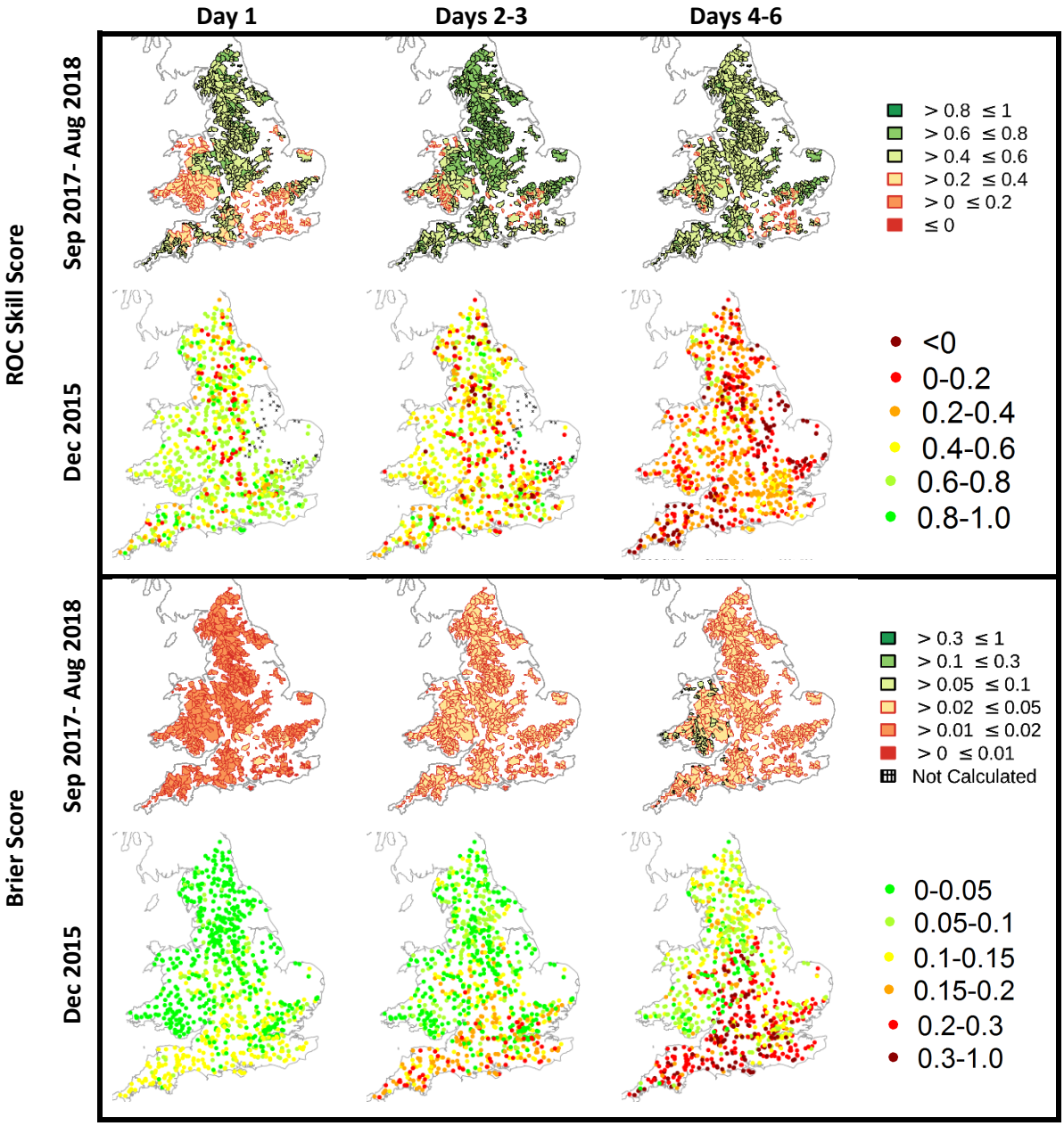


Figure 6 ROC Skill Score (upper panel) and Brier Score (lower panel) calculated for individual catchment river flows using a region-based moving catchment-area based pool of 31 catchments across England & Wales. Scores are calculated for both the Phase 2 12-month period September 2017 to August 2018 (inset top) and for the Phase 1 period December 2015 (inset bottom).

1.2 Non-threshold based score analyses

Analyses using the non-threshold based Rank Histogram over the Phase 2 12-month period (River Flow Verification Summary Figure 8) agree overall with those seen for the Phase 1 December 2015 period (Phase 1 Report Figures 4.16 to 4.18). The ensemble appears under-spread overall due to individual forecasts being under-spread, or to conditional biases in the ensemble, or a combination of these. Slight differences are seen between the relative populations of the upper and lower histogram bins

(i.e. the relative proportion of observations that fall above or below the ensemble members), but these are not thought to be significant. The single-season Rank Histograms show a more coherent pattern of differences between the relative populations of the highest and lowest histogram bin. In particular, for autumn and winter the lowest histogram bin has the highest population, whereas for the spring and summer the largest population falls in the highest histogram bin. This suggests that observed river flow is likely to be lower than all ensemble members in the winter (high-bias), but higher than ensemble members in the summer (low-bias). This interesting behaviour warrants further investigation.

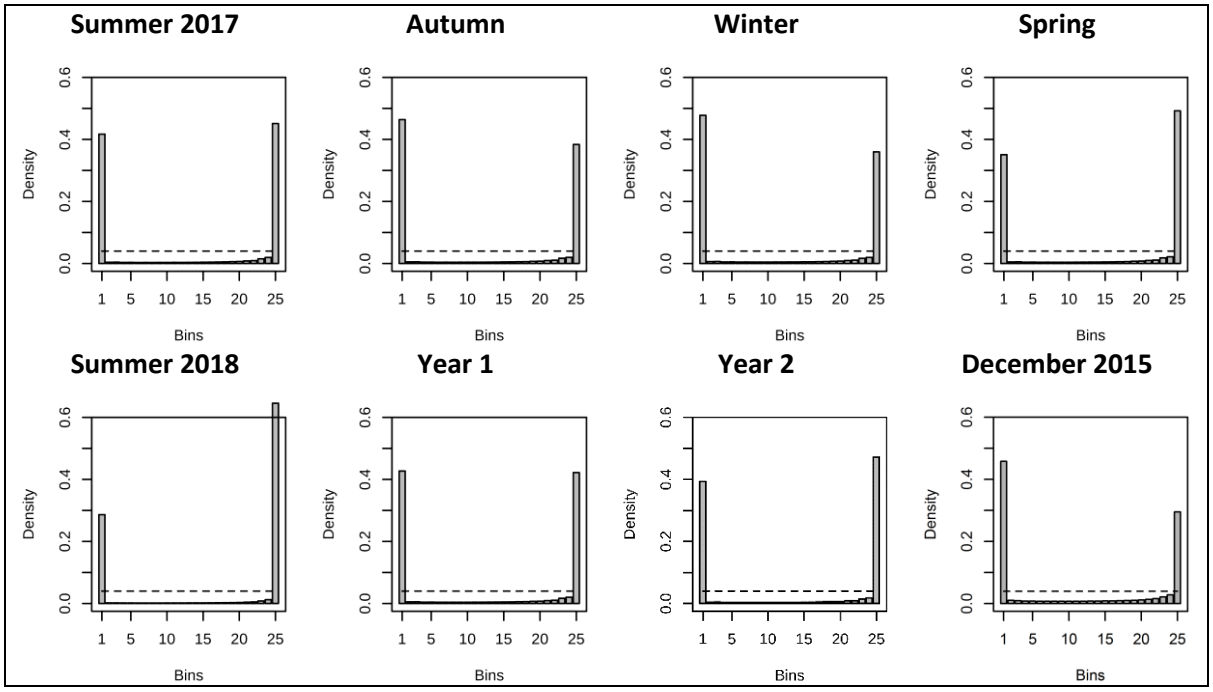


Figure 7 G2G river flow Rank Histograms for different verification periods. Calculated for all sites in England & Wales using Day 1 forecasts.

For the non-threshold based CRPSS, the Phase 1 analyses showed no clear national pattern of ensemble performance. This is also seen for the Phase 2 12-month period analyses. Note that, as a different type of reference is used in the CRPSS calculations for Phase 2 (based on the Ensemble MSE over the full 12-month period instead of the CRPS of the ensemble mean), the magnitude of the CRPSS values cannot be compared between the Phase 1 and Phase 2 analyses.

2 River flow verification analyses for PDM local models

Overall, the river flow verification analyses for PDM local models show high sampling uncertainties as expected due to the calculation of verification statistics at the catchment-scale over only a 12-month period. As scores for the PDMs are calculated for individual sites only, there are too few threshold-crossings to calculate verification statistics above the Q(2)/2 threshold (the exception being Beddgelert, although the Q(2) verification analyses for this site are also suspect). Example verification diagrams are shown in Figure 8 for the catchments Riccal at Nunnington (England), Glaslyn at

Beddgelert (Wales) and Findhorn at Shenachie (Scotland); diagrams for the other sites considered are provided in

Appendix_C_3_2_Case_Study_Analysis_Hydrological_Impacts_plots.zip

To reduce the effect of sampling uncertainty it would be necessary to either

1. pool the calculations across multiple catchments having PDMs, as done for catchments within the G2G model domain,
2. use a much-longer verification period, or
3. choose a specific verification period known to contain a number of threshold-crossings.

Of course, each of these options has both advantages and disadvantages. Pooling over multiple catchments may not be appropriate for different local-models which, unlike G2G, may lack spatial-consistency. Although Option 2 would give the most representative analyses, it would require the BMR rainfall ensemble to be run in hindcast-mode over a much-longer period, a practise not currently supported at the Met Office. There would also be additional overheads in making the long-duration PDM ensemble runs and analysing the longer time-series of forecast river flows. Although Option 3 is based on the use of a “non-representative” period of data with above-average rainfall and river flow events, it is a possible workable solution for analysing ensemble forecast performance in the unusual, high-impact situations of particular interest.

In general and noting the differences in score-calculation discussed above, better performance is seen for the PDMs than G2G for corresponding catchments: this is as expected when comparing site-calibrated local models with a national-scale distributed model at gauged catchment locations.

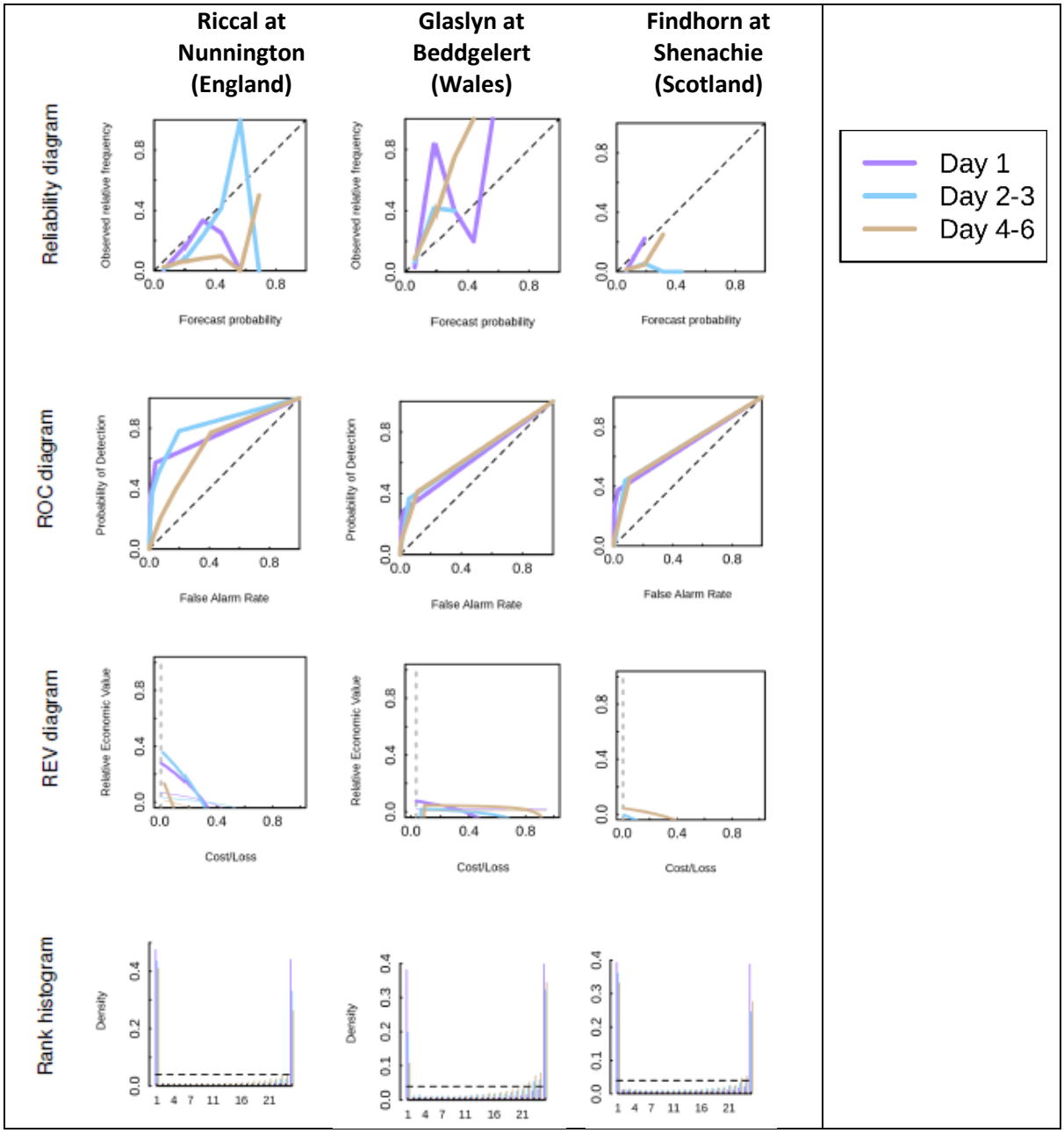


Figure 8 River flow verification diagrams calculated for single site PDMs. Reliability, ROC, and REV diagrams are shown using the Q(2)/2 threshold for the Day 1, Days 2-3 and Days 4-6 forecasts. The final row shows the Rank Histograms (no threshold dependency).

3 Key conclusions from the river flow verification analyses

- Apart from anomalously wet periods (e.g. December 2015 as used in Phase 1), the number of river flow threshold-crossings for Q(T) thresholds appropriate for flood forecasting (e.g. a minimum of Q(2)/2) is ***not sufficient*** for ensemble verification at ***sub-national scales*** when forecasts from only one season are considered.
- For autumn, winter and spring seasons, the number of river flow threshold-crossings can be sufficient to give meaningful verification analyses at the ***national scale*** for the ***Q(2)/2 threshold.*** This depends on there being a **large-enough pool of sites nationally** and is, for example, not true for the spring season over Scotland with only 225 sites (compared to the 731 sites for England & Wales).
- For the lower river flow thresholds (Q(2)/2 and Q(2)) a ***12-month*** verification period ***can be sufficient*** to give meaningful verification analyses. If a rolling 12-month verification period were to be used, the analyses would be expected to be ***more sensitive*** to changes in the ***winter months*** used as these contain the majority of threshold-crossing events.
- For ***sub-regional scale*** analyses, sampling size and forecast skill are influenced by the ***time-window increasing*** in length with increasing lead-time.
- If threshold-crossings are ***unevenly distributed*** across the domain, then the interpretation of individual-catchment maps may be influenced by ***spatially-varying sampling uncertainties***. Threshold-based score maps should be ***viewed alongside*** maps of ***the number of threshold-crossings*** occurring in the verification period.
- Verification scores ***can be calculated for local models*** (e.g. PDM) in the same manner as that shown for G2G. Local model verification analyses ***show high sampling uncertainties*** when calculated for ***single catchments*** using 12-months of ensemble forecasts, even for the lowest threshold considered, Q(2)/2.
- The local model verification analyses can be used to inform prototype real-time displays. However, given the ***high sampling uncertainties***, these should be considered as ***demonstrative*** rather than generally representative.
- In an operational system, the ***local model sample size*** would need to be ***increased*** through either ***multi-catchment pooling*** of analyses, consideration of a ***longer verification period***, or through using a ***fixed historical period known*** to have ***sufficient threshold-crossings***.

Rainfall and River Flow Ensemble Verification: Phase 2
Impact of observation uncertainty on verification metrics
Final Report Appendix B.3

1. Background

Accounting for observation uncertainty is essential for the assessment of ensembles. Forecast ensembles are used to provide spread (uncertainty) information about what is going to happen.

Observations are imperfect and provide an observation of the true state with an error, which is generally unknown or not fully known. Ferro (2017) observes: *"In these circumstances, proper scoring rules favour good forecasts of observations rather than of truth and yield scores that vary with the quality of the observations. Proper scoring rules thus can favour forecasters who issue worse forecasts of the truth and can mask real changes in forecast performance if observation quality varies over time.".* Observations therefore also come with error bounds, which are made up of many sources. For example, the uncertainty can be related to whether the quantity is measured directly (instrument error) and/or whether it is inferred or derived from something that is measured (estimation). Without having an estimate of the observation error, it is impossible to differentiate between what is true ensemble spread (uncertainty) and what is due to the observation uncertainty.

Several authors have explored the idea of including observation error or attributing an error to the observations (Saetra et al. 2004, Candille and Tallegrand, 2005, 2008, Bowler, 2006, 2008, Santos and Ghelli, 2012, Koh et al. 2012, Rőpnack et al., 2013).

Given this background, within the Ensemble Verification project there exists the opportunity to estimate the observation error of two of the observation sources, using well-established data assimilation techniques. These estimates can be applied, using an approach similar to Bowler (2008) wherein the forecast ensembles are perturbed with the observation error before the verification scores are computed.

2. Methodology

Two estimates of precipitation are available: a gridded raingauge analysis and a radar-rainfall accumulation (which has had some mean-field raingauge-based bias adjustment). Both have specific error characteristics. The first is based on point-based observations which have been interpolated onto a grid, making assumptions about what happens in between the point observations. The second is an estimate on a grid but it is derived or inferred from the observed quantity, radar reflectivity. Both have observation errors, but originating from very different sources. Also there is a weather model forecast for this quantity. In this instance the forecast can be used to constrain the differences in the observations**.**

Using a method first proposed by Hollingsworth & Lönnberg (1986) for deriving observation error statistics in data assimilation, let ***a*** be a vector of forecast values and ***b*** and ***c*** the vectors of two different observation types of the same quantity: here, *hourly* precipitation accumulations. Both of these are estimates of the true observation ***t*** which is unknown.

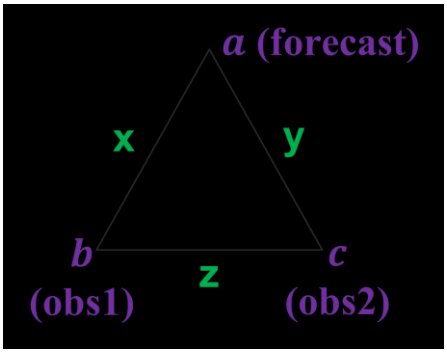


Figure 1 Schematic illustrating the three-way pairwise comparisons, where the forecast constrains the difference between the observation types.

First proceed by computing the pairwise mean squared errors (MSE) between all possible pairs as shown in Figure 1. If both observations were perfect estimates of the true state then $\boldsymbol{a} = \boldsymbol{t}$ and $\boldsymbol{b} = \boldsymbol{t}$ and $E[(\boldsymbol{a} - \boldsymbol{b})^2] = 0$. However this is known to be untrue, so by introducing $\boldsymbol{t}$ the MSE of the differences (or deltas) represent an estimate of the observation error. These are referred to here as *x, y* and *z*.

For example, the MSE between $\boldsymbol{a}$ and $\boldsymbol{b}$ is:

$$x = E[(\boldsymbol{a} - \boldsymbol{b})^2] = E[((\boldsymbol{a} - \boldsymbol{t}) - (\boldsymbol{b} - \boldsymbol{t}))^2] = E[\boldsymbol{\delta}_a^2] + E[\boldsymbol{\delta}_b^2] - 2\,E[\boldsymbol{\delta}_a\,\boldsymbol{\delta}_b] \quad (1)$$

The method makes the assumption that the errors are not correlated so that the product term in Eq. 1 dissapears. This is a reasonable assumption to make given that the observations are measured and derived in very different ways, and are assumed to be much smaller than the deviations from the true value. The presence of zeros can break the assumption. For precipitation this issue is addressed by binning the values.

Based on the pairwise MSE equations, a set of three equations with three unknowns can be constructed to solve for *x*, *y*, and *z*:

$$E[\delta_a^2] = 0.5\,(x + y - z)$$
$$E[\delta_b^2] = 0.5\,(x - y + z) \quad (2)$$
$$E[\delta_c^2] = 0.5\,(-x + y - z)$$

As alluded to, variables such as precipitation which are dominated by zeros need to be binned before the method can be applied. In this initial study 5 bins were used with a bin size progression which reflects the shape of the underlying distribution. For precipitation the lognormal distribution is often used. The bin sizes are: BIN1: [0,0.1), BIN2: [0.1,1), BIN3: [1, 2), BIN4: [2, 4), BIN5: [4, ∞) mm/h.

The values of *x*, *y* and *z* are then calculated for each of the bins to obtain an error estimate for a given range of precipitation values, as determined by the bin size.

3. Initial results

First, there is a check to consider whether the underlying distributions between the three data series are similar. For the forecast ensemble there are two choices. Either a single member (usually the control) or the ensemble mean. Table 1 shows that the bin-based distributions of the ensemble mean and the observations are more different from each other, whereas the control member seems to be closer to the observations in most of the bins. Based on this, the control member is used as the basis for binning as the ensemble mean is very smooth, with potentially larger values removed. What does the basis for binning mean? In this instance, to compute the pairwise MSE, the bin of the control member determines the comparison, i.e. if the control forecast value is in BIN1, it doesn't matter in which bin the other value in the pairwise comparison is located. If it is in BIN2 or BIN3, it simply means that the MSE value will be higher the further away it is and contribute more to the magnitude of the error estimate for that bin.

Table 1 Summary of the proportion of hourly precipitation values per bin for two different forecast options and the two observation types. BIN1: [0,0.1), BIN2: [0.1,1), BIN3: [1, 2), BIN4: [2, 4), BIN5: [4, ∞).

%	BIN 1	BIN 2	BIN 3	BIN 4	BIN 5
Ensemble mean	59.4	31.4	6.3	2.7	0.2
Radar	74.5	17.8	4.3	2.6	0.8
Gauge	71.1	21.4	4.3	2.5	0.7
Control forecast	72.1	17.7	5.5	3.7	1.0

Figure 2 shows how the MSE converges as a function of sample size, based on the control member at t+12h. The errors are generally bigger for the forecast (here the ensemble mean is shown) than for the two observation types. By increasing the sample size, the error values converge towards a near constant value in each bin. It is clear that the sample is too small for the bins 3, 4 and 5, but especially so for bins 4 and 5, though it is clear that (even with the small sample size) convergence towards a robust estimate is beginning to happen. It is these robust error estimates that are being sought.

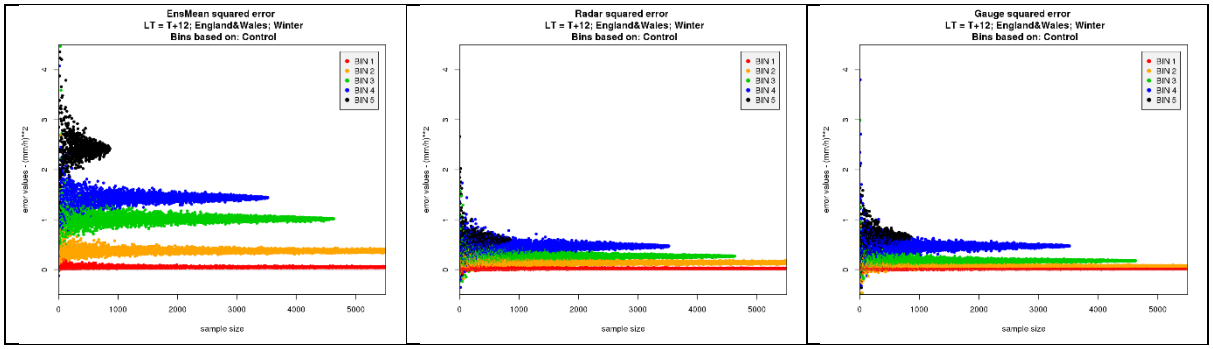


Figure 2 Convergence of MSE as a function of sample size based on the t+12h control member bins and applied to the ensemble mean and observation types.

In Figure 3 the error estimates are plotted as a function of lead-time (Figure 2 was only for t+12h). for the three middle bins. Recall that BIN 1 contains values which are very close to zero and the mathematical assumption is not valid. BIN 5 was excluded as the sample size was just deemed to be too small at this stage. (Arguably, for BIN 4 the sample size is not enough.)

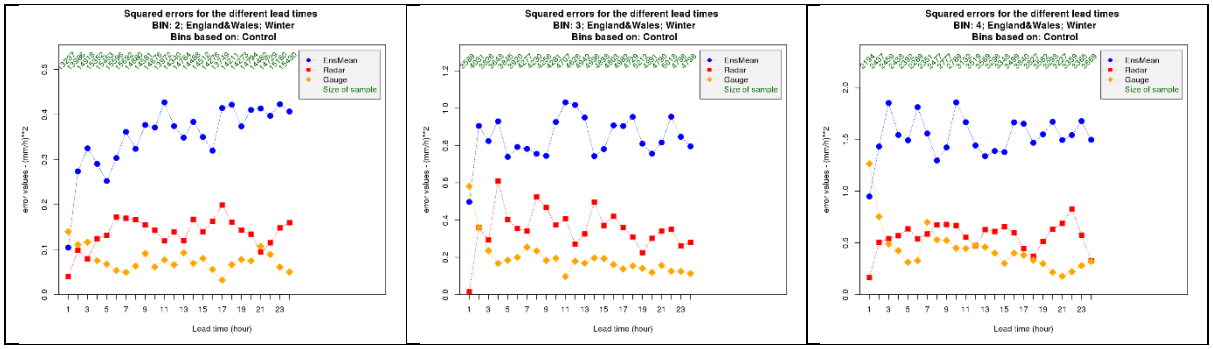


Figure 3 Error estimates as a function of lead-time for BINS 2, 3 and 4, showing that the model error is about 2 to 3 times the size of the observations. For the low intensity bins the radar error appears to be slightly larger, whereas for BIN 4 the differences between the observation types seems less. Numbers in green show the sample size per hour for the first day (up to t+24h).

Figure 3 broadly reflects expectations: namely that the observation errors should not change with lead-time, whereas, the forecast error should. This behaviour is only really seen for BIN 2. There is a curious discrepancy at t+1h for the two observation types, which is possibly due to the characteristics of the forecast which, at this lead-time, is almost entirely based on the STEPS nowcast, which is largely an extrapolation of radar rainfall (~80%). This could explain why the error estimates for the radar data are so low for t+1h. For any subsequent analysis of data under Phase 2 of the project and any aggregated error estimates, t+1h ought to be removed (this was not done here). The behaviour as a function of lead-time is still a little noisy due to the small sample size of just one month, but a larger sample should resolve this.

Given the overall behaviour is as expected it allows all the hourly samples to be combined to increase the sample size for computing error estimates for Day 1 (for example). This is shown in Figure 4. On this graph, the error estimates (MSE) obtained from a combined Day 1 sample (made up of 24 hourly samples) are plotted against the bin mid-points. The relationship is approximately log-linear, so a log-linear model can be fitted to these data points to provide a function which can be applied to compute the effect of observation error on verification statistics for all possible precipitation values.

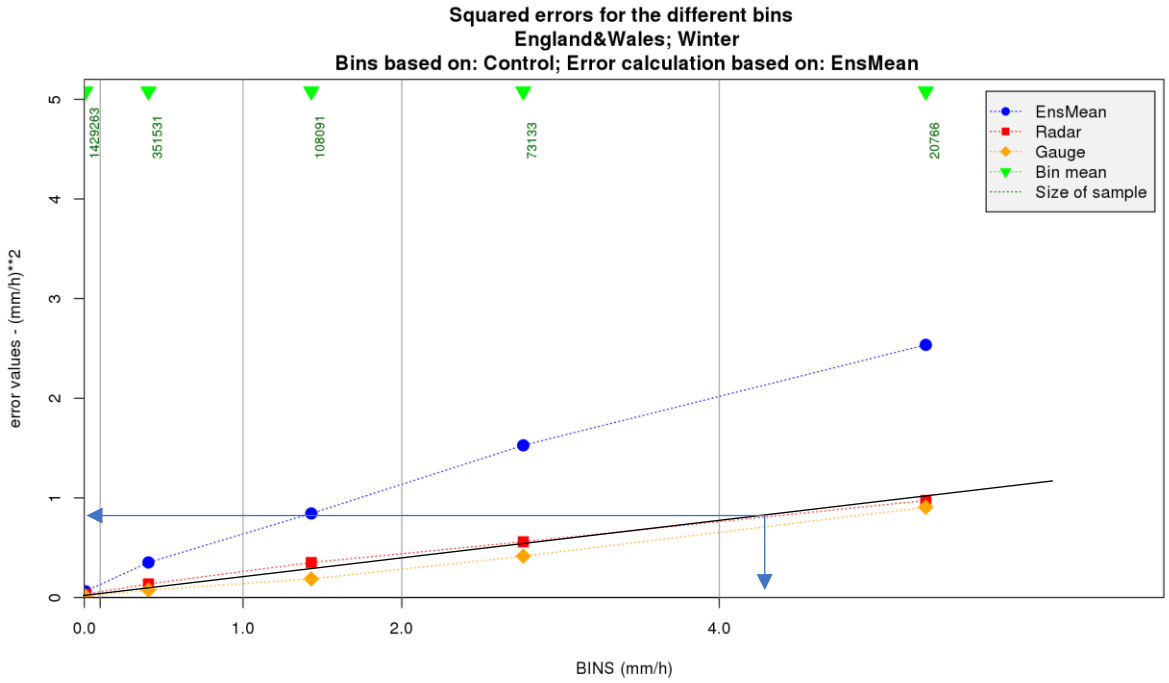


Figure 4 Error estimates as a function of bins (intensity) over the sample of all hourly values in Day 1. Green triangles show the bin mid-point and green numbers indicate the sample size. A log-linear fit (illustrative only) has been added to the forecast points, to illustrate how the relationship is applied, e.g. a radar rainfall value of 4.2 mm is associated with a 0.9 mm^2 MSE.

For each hourly time-step and for each catchment, the log-linear relationships are used to obtain the observation uncertainty associated with that hour's forecast catchment-mean value (here based on the control). This is done separately for the radar and raingauge observations using their respective relationships.

Similar to Bowler (2008), the forecast ensemble hourly catchment-mean values are perturbed using the observation error derived for a given observation type. This is done by first creating a random sample (equal to the ensemble size) from a suitable error distribution (e.g. Gaussian). This random sample of perturbations (of 24 values in this case) is scaled to have a mean of zero and a standard deviation equal to the observation error estimate. These 24 perturbations are added to the corresponding ensemble member catchment-mean forecast value, capping at zero precipitation if necessary (to avoid negative values). At this stage, the process has yielded a modified set of 24 ensemble member catchment-mean forecast values which have been adjusted to account for observation uncertainty, and these can now be verified using the standard methods of the Ensemble Verification Framework.

Figure 5 shows the impact of including observation uncertainty on reliability. Both radar and gauge results are shown for 1, 2 and 4 mm of rainfall. The blue line represents the original unperturbed forecasts and the red line the forecast perturbed with the observation error.

The perturbed results show that the ensemble remains over-confident. In fact, the results look to be very close together, especially for low thresholds where the impact is likely to be less. Nevertheless, even for the lower thresholds there is a slight improvement in reliability for the higher probabilities, which appear to be outside the confidence intervals (which are so narrow they are invisible).

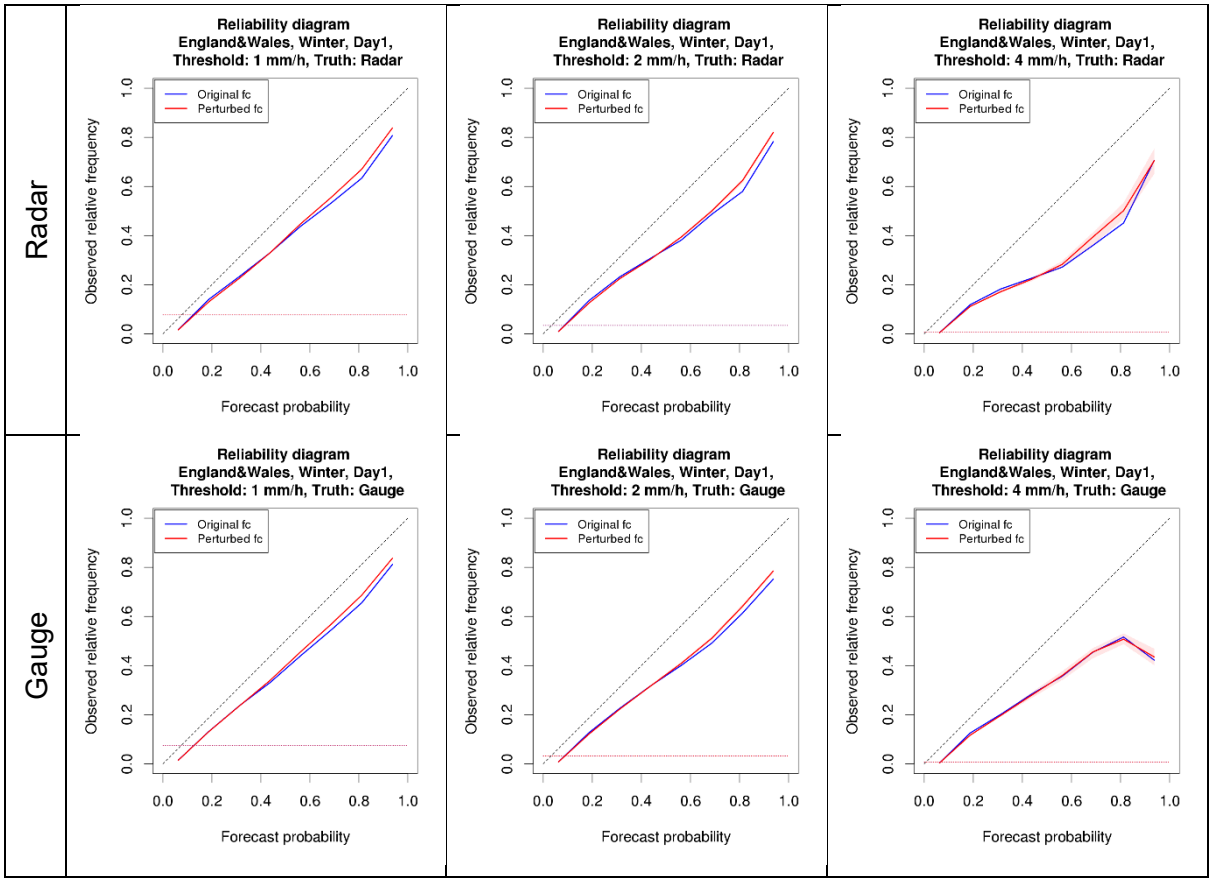


Figure 5 Illustration of impact of observation uncertainty on reliability. Day 1 results, for both raingauges and radar and using three thresholds, are shown for winter over England & Wales.

For the 4 mm/h threshold, the sample size is too small, as indicated by the drop-off in the larger probabilities. This is especially evident in the raingauge results. Differences in reliability between the perturbed and unperturbed forecasts appear larger but this could be due to the small sample size and subsequent random sampling. Confidence intervals are so narrow that they are invisible for the lower thresholds but for the 4 mm/h threshold there is more tangible evidence that the original forecast reliability based on the radar data lies outside the spread of the perturbed forecast for the larger probabilities. However, it is unwise to read too much into this given that, overall, the sample is too small to be robust.

4. Future work

The Phase 1 dataset sample, which consisted of only one month of forecasts, is too small to draw any definitive conclusions other than to say that there is an impact and the error estimates are definitely non-negligible. So, it can be concluded that observation uncertainty will have an impact and some estimate of observation uncertainty should be included in any future ensemble verification framework. This will be especially important for larger thresholds, which are of greater interest for hydrological applications. It is planned to repeat what has been described in this report using the larger Phase 2 dataset, although it remains to be seen whether a sample of 12 months will be sufficient to extend the range of thresholds beyond 4 mm/h.

References

Bowler, N. (2006). Explicitly accounting for observation error in categorical verification of forecasts. Mon. Wea. Rev., **134**, 1600–1606.

Bowler, N. (2008). Accounting for the effect of observation errors on verification of MOGREPS. Meteorol. Apps., **15**, 199–205.

Candille, G. and Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. Q.J.R. Meteorol. Soc., **131**, 2131–2150.

Candille, G. and Talagrand, O. (2008). Impact of observational error on the validation of ensemble prediction systems. Q.J.R. Meteorol. Soc., **134**, 959–971.

Ferro C.A.T. (2017). Measuring forecast performance in the presence of observation error. *Q.J.R. Meteorol. Soc*, **143:** 2665-2676.

Hollingsworth, A., and P. Lönnberg. (1986). The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus*, **38A,** 111–136.

Koh, T.-Y., Bhatt, B., Cheung, K., Teo, C., Lee, Y., and Roth, M. (2012). Using the spectral scaling exponent for validation of quantitative precipitation forecasts. Meteorol. Atmos. Phys., **115**, 35–45.

Rőpnack A., Hense, A., Gebhardt, C. and Majewski, D. (2013). Bayesian model verification of NWP ensemble forecasts. Mon. Wea. Rev., **141**, 375--387.

Saetra, O., Hersbach, H., Bidlot, J.-R., and Richardson, D. (2004). Effects of observations error on the statistics for ensemble spread and reliability. Mon. Wea. Rev., **132**, 1487–1501.

Santos, C. and Ghelli, A. (2012). Observational uncertainty method to assess ensemble precipitation forecasts. Q.J.R. Meteorol. Soc., **138**, 209–221.

Rainfall and River Flow Ensemble Verification: Phase 2
Fifteen-minute precipitation verification results and future plans
Final Report Appendix B.4

1. Summary

Whilst it is true that G2G is driven by 15-min precipitation accumulations and it is at some level important to check that the 15-min accumulations have some accuracy and skill, the true utility of the precipitation forecasts in a flood forecasting context is not in the 15-min accumulations (because they tend to be small). *River-based flooding is the result of consecutive 15-min accumulations which lead to a flood response.* In this instance, the previously analysed hourly and daily accumulations are far more useful to gain an understanding of shorter- and longer-duration event totals and compare to the river response. There is some utility of the precisely matched 15-min results for Day 1, where forecast errors are not as dominated by timing errors.

2. Results in brief

Figure 1 shows "Figure 5.29" of the Phase 1 Report augmented with the 15-min results for Day 1. It shows a slight broadening of the scatter of the precipitation scores for the 15-min accumulations. All precipitation scores remain positive (i.e. skilful though a score of near-zero is marginal). It is worth pointing out that there is a fundamental mismatch between how the scores were derived. *Apart from the CRPSS shown here,* for river flow the scores have always been derived as "events", searching for exceedances in all lead times for Day 1, for example, rather than a precise matching, as is the case for the precipitation.

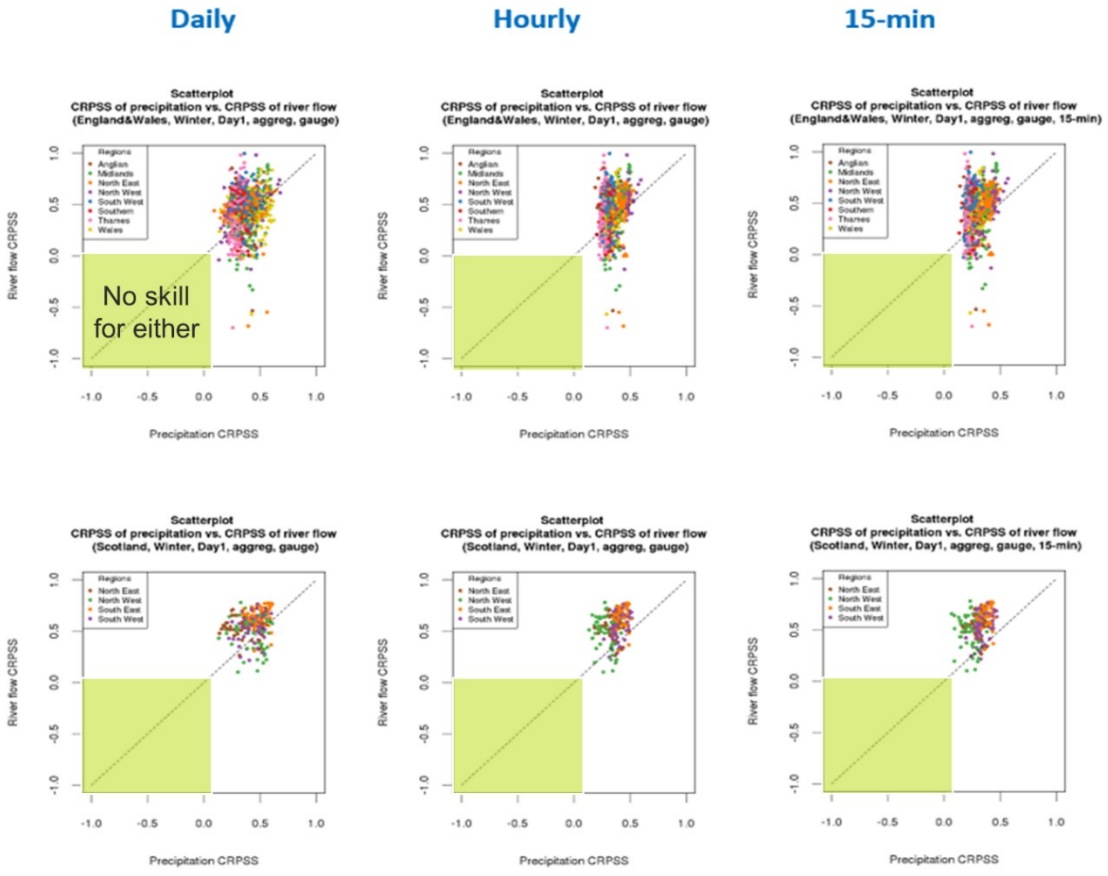


Figure 1 Comparison of CRPSS between precipitation and river flow for Day 1. Note that the river-flow scores between "Daily", "Hourly" and "15-min" are all 15-min scores.

Figure 2 is an extension from what was done in the Phase 1 Report, showing the results for all lead-time horizons. The decrease in both river flow and precipitation scores with lead-time is clear, though the reduction in scores for precipitation is much stronger, especially for England. This is probably linked to the study period used. By Day 4-6 very few catchments show any positive precipitation skill whilst a fair proportion of river flow forecasts still have positive scores.

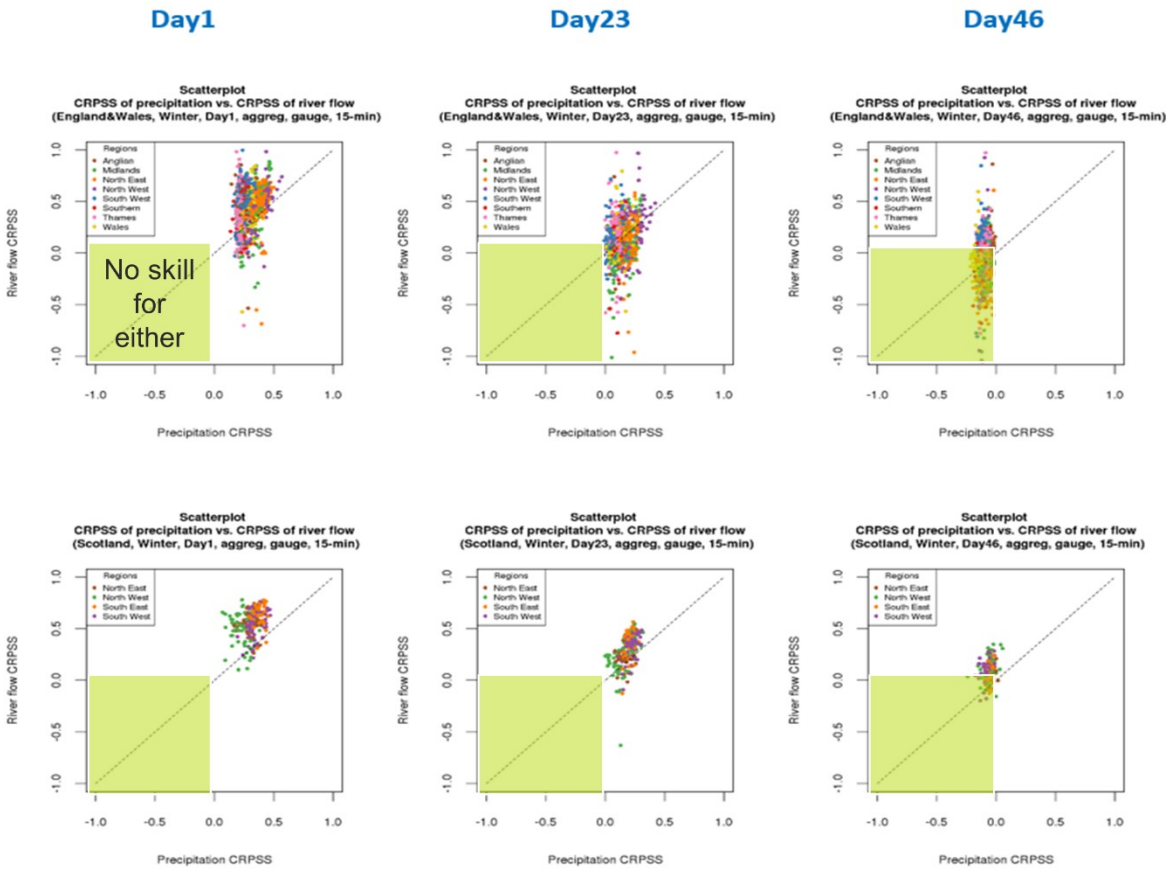


Figure 2 As Figure 1, but showing 15-min precipitation and river flow scores only, for three lead-time horizons Day 1, Day 2-3 and Day 4-6.

So, whilst Phase 2 planned to undertake more work on the precisely matched 15-min accumulations for the dataset used in Phase I, an inspection of the results obtained thus far corroborates the view above. Figure 3 compares the verification measures - ROC Diagram, Reliability Diagram, Rank Histogram, and map of Brier Skill Score (BSS) - for Day 1 with daily, hourly and 15-min precipitation accumulations against raingauge data for England & Wales. Even for Day 1, the measures are slightly worse for 15-min than hourly precipitation accumulations.

Figure 3 shows that Area under the ROC Curve is systematically reduced as the accumulation window decreases. The Reliability between daily and hourly is broadly similar but for the 15-min results there is a sampling problem for the largest probabilities. In terms of the Rank Histogram the hourly and 15-min ones are very similar: therefore the hourly version provides the information needed with the 15-min version not adding further to this. In terms of the BSS, the proportion of poor scores (red) increases with decreasing accumulation window.

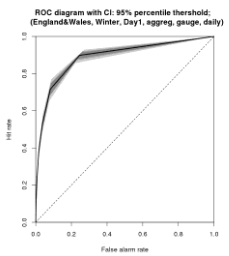
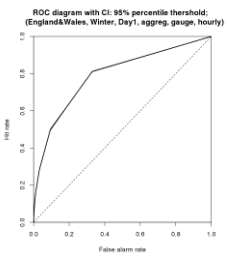
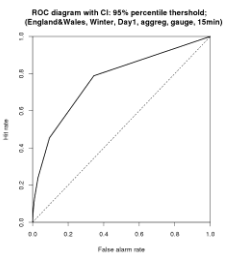
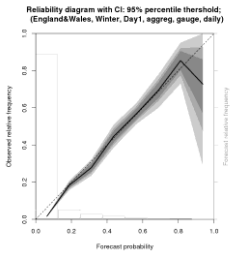
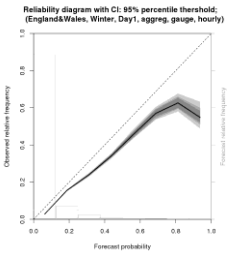
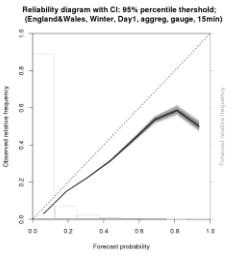
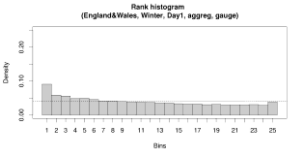
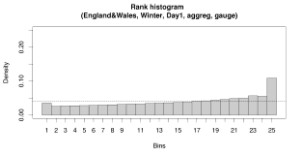
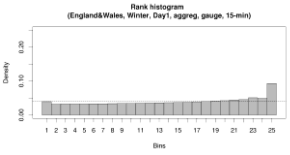
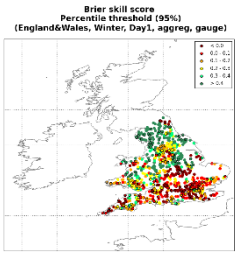
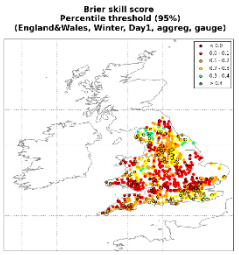
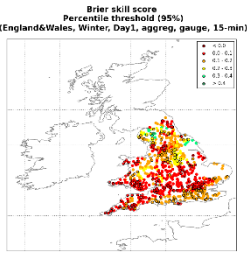
	Daily	Hourly	15-min
ROC			
Reliability			
Rank Histogram			
Map (BSS)			

Figure 3 Selection of plots (ROC Diagram, Reliability Diagram, Rank Histogram, and map of BSS) of Day 1 aggregated results for England & Wales in the winter period calculated using the 95th percentile threshold with verification against raingauge data.

Figure 4 shows only the 15-min results for Scotland in winter as a function of lead-time. Potential skill, as defined by the Area under the ROC Curve, is not high to begin with, but degrades to very low levels when using precisely matched time-windows. The Reliability also is also not high, even on Day 1, with further degradation for later days. There is also a systematic degradation of *the Brier Skill Scores (BSS)* beyond Day 1, with scores predominantly poor (and many catchments having negative scores, i.e. worse than the sample climatology). All this can, at least in part, be attributed to the shortness of the time-window, and the increasing likelihood of timing errors and mismatches in the precisely matched forecast-observed pairs. None of these results represent the practical utility (other than in driving G2G) of the precipitation forecast. Additionally, it is worth bearing in mind that at longer lead-times the hourly totals are split into four equal parts to represent 15-min accumulations.

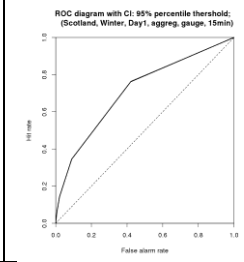
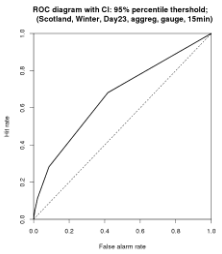
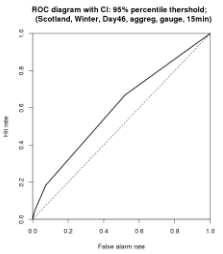
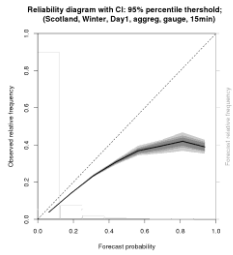
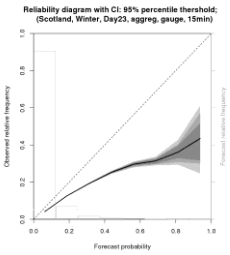
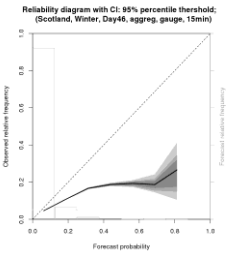
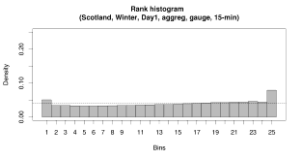
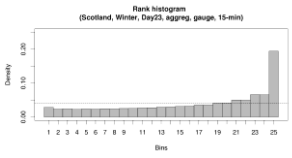
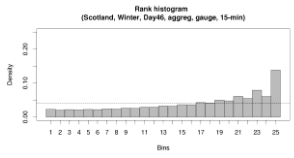
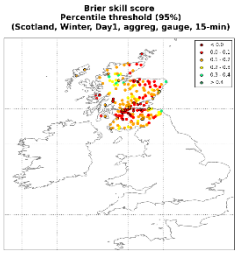
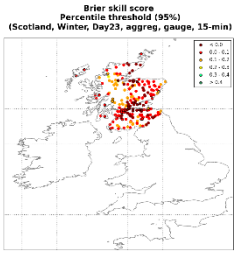
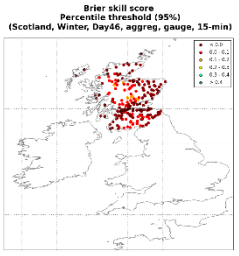
	Day 1	Day 2-3	Day 4-6
ROC			
Reliability			
Rank Histogram			
Map (BSS)			

Figure 4 Selection of plots (ROC Diagram, Reliability Diagram, Rank Histogram, and map of BSS) of aggregated results for Scotland in the winter period calculated using the 95th percentile threshold, 15-min accumulations, with verification against raingauge data.

The results would suggest that the G2G forecasts seem to be relatively insensitive to the lack of accuracy and skill of the 15-min precipitation accumulations. As in Figures 1 and 2, CRPSS values are predominantly positive for most catchments despite the apparent lack of skill in the 15-min precipitation. This is potentially useful information and encouraging. The outcomes of this verification project should not be judged on only a precise-matching approach for 15-min precipitation accumulations. Therefore, the results produced under Phase 1 - focusing on daily and hourly totals - are emerging as being far more appropriate and sufficient for evaluating precipitation forecast utility (having seen and analysed the 15-min results). Monitoring 15-min results for Day 1 (before other elements of the forecast error dominate) is perhaps still useful to ensure that these behave similarly to the hourly accumulation results.

3. Recommendation

Going forward, any future analysis work involving identifying events within specified time-windows needs to be aligned to match what is being done for river flow. This will be expected to show more utility of the precipitation forecast. The analyses might consider using physical

rather than percentile thresholds (especially given the additional data under Phase 2) for daily and hourly totals.

It is *still* unclear whether the analysis of 15-min precipitation accumulations in a case study event-based framework will yield anything useful, given that very few 15-min accumulations are that extreme. Time permitting, some of the cases in the Phase 1 dataset might be revisited, but generally the focus (going forward) will be on the Phase 2 study period. For longer verification periods, the focus will be on daily and hourly precipitation accumulations, and taking into account the precipitation timing uncertainty within the forecast horizon as is done for river flow.

Rainfall and River Flow Ensemble Verification: Phase 2
Comparison of G2G river flows using different rainfall sources as input
Final Report Appendix B.5.1

1 Background

The joint Met Office and Environment Agency (EA) project on merging of radar and raingauge data (Project SC100004: Radar Rainfall Merging)) was carried out over the period 2011 to 2016 with the main results published in Jewell and Gaussiat (2015). This reported on a comparison of different merging methods, using raingauge data as a reference "ground truth", and identified a Kriging with External Drift (KED) method as the preferred approach for operational implementation. Subsequently, two products were made available (from 29 February 2016) to the EA for trial - Merged 1h and Merged 24 h - with receipt delays of 1 h and 24h, the Merged 24 h utilising more raingauges and quality control. Both products provide 1 km 15 minute precipitation accumulations over an England & Wales domain.

The EA recognised that the assessments of the new merged rainfall products to date had been referenced to raingauge data. Further trials were needed when using the merged rainfall as input to the hydrological models supporting flood warning and guidance. It was important to gain experience of using the merged rainfalls as input to hydrological models, such as the PDM catchment rainfall-runoff model used in local river network models and the G2G distributed model applied nationally. The opportunity to assess use of the merged rainfalls as input to G2G could be achieved efficiently through including the work in Phase 2 of the "Rainfall and River Flow Ensemble Verification" project. It also aligned to interest within the project on how the rainfall source (radar, raingauge, merged) impacted on the verification of ensemble rainfalls. This assessment is reported on here with the aim of providing the EA (along with NRW, SEPA and FFC) with evidence of benefits and/or issues with using the new merged rainfalls as input to flood forecasting models.

2 Purpose and approach

The aim here is to compare the effect on G2G modelled river flows of using different observation sources of rainfall. Four different rainfall sources are compared.

1. **Gauge**. Raingauge data (from the EA and NRW raingauge networks over England and Wales) gridded by multiquadric interpolation with zero offset. This rainfall source is currently used for maintaining G2G states.

2. **Radar**. Radar rainfall data from the Met Office RadarNet system.

3. **Merged 1h**. Radar rainfall data from the Met Office RadarNet data merged with raingauge data from a Met Office network using a Kriging with External Drift (KED) method (Jewell and Gaussiat, 2015). Available with a 1 hour delay in real-time.

4. **Merged 24h**. As 3. above, but available with a 24 hour delay in real time, allowing more raingauge data to be included and with better quality-control procedures applied.

All four rainfall sources are available as 15 minute accumulations on a 1km grid covering England & Wales. The first two rainfall sources were compared from a rainfall perspective in Phase 1 of the "Rainfall and River Flow Ensemble Verification" project. To focus on the effect of the different rainfall sources, G2G was run in simulation-mode (that is, no state updating or flow insertion was used). Comparisons were limited to England & Wales where the merged product has coverage.

Initially, G2G simulations were made over the period 1 March 2016 to 31 March 2017. This period was selected to allow analysis of a full year of river flows when rainfall data from all four sources were available. To allow spin-up from the raingauge-data maintained G2G initial conditions, the first month of the simulations were not analysed, so the results reported on here are for the year 1 April 2016 to 31 March 2017. These results are presented in Section 3, with Figures included at the end of this document.

As more-recent data up to September 2018 became available later in the project, the above analysis was completed for the additional periods:

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

The results from these periods are available in separate .pdf documents of the form

Compare_rainfall_source_**YYYYMM**to***YYYYMM***_731sites.pdf

where **YYYYMM** is the start month of the verification period, and ***YYYYMM*** the end month of the verification period. These documents are contained in Final report Appendix B.5.2..

Results are computed for 731 of the 898 operational G2G gauged catchments in England & Wales where observed river flows are suitable for verification of the current G2G configuration. To test the effect of this reduction in catchments, results are also included for the 1 April 2016 to 31 March 2017 period using the originally-available observed river flow and raingauge data, but using the reduced number of catchments. Section 4 provides a discussion of these additional results and compares and contrasts with those presented in Section 3.

3 Initial G2G simulations for the period 1 March 2016 to 31 March 2017

3.1 Effect on the river flow hydrograph: goodness-of-fit measures

The G2G river flow simulations are compared first using the three goodness-of-fit measures: absolute relative Bias, Correlation, and R^2 Efficiency. These measures were chosen to analyse the effect of the different rainfall sources on the full flow hydrograph. They give an overview of the overall differences between the four hydrograph simulations, without focussing on the crossing of particular river flow thresholds. These statistics were calculated by comparing, at 15-minute intervals, the instantaneous river flows from G2G with observed river flows.

Box Plots of the Bias, Correlation, and R^2 Efficiency are shown in Figure 1, with bars grouped by region of England and Wales (NE: North East, NW: North West, MI: Midlands, AN: Anglia, TH: Thames, SO: Southern,

SW: South West, WA: Wales), and lastly for all sites in England & Wales. Overall, the G2G flow simulations with Raingauge data as input perform best, with nearer-zero bias values, and higher values of Correlation and R^2. Analysis of the hydrographs for individual catchments (e.g. Blackwater (SO) at Ower (042014) and Roch at Albert Royds Bridge (690207)) showed that this is related to spurious peaks in the Radar data that are not removed in the merging process. Overall, the merged products show slight improvement over using only the Radar data, with the Merged 24h product outperforming the Merged 1h. However, this is not the case for all regions: for example, Midlands and Southern have higher values of Bias for the Merged 24h product. Of course, these differences at the regional scale are small and based on a smaller sample of sites: it will be interesting to see if similar results are obtained when analysing simulations for the 2017-18 period (Section 4).

3.2 Effect on upward crossing of flow thresholds: Categorical Skill Scores

To analyse how the different rainfall sources affect the ability of G2G to capture upward crossings of a selected flow threshold (aligned to how flood warnings might be triggered, or severity assessed for flood guidance), three Categorical Skill Scores based on the Contingency Table are considered. The Probability of Detection, POD, measures the proportion of observed events that were correctly forecast. The False Alarm Rate, *F*, measures the proportion of non-events that were incorrectly forecast as occurring (False Alarms). The Critical Success Index, CSI, measures the proportion of all events (forecast or observed) that were correctly forecast (Hits). CSI can be seen as combining the POD and *F* scores, and is included to give an overall measure of success with regard to forecasting "threshold events".

To focus on daily and hourly periods, an "event" is defined as an upward crossing of a river flow threshold occurring anywhere within a time window of either 1 hour (4 time-steps) or 1 day (96 time-steps). These windows were applied at each 15-minute time-step in the G2G simulations and the observed river flows. The windows were symmetric about the time-step in question. Given the year-long period analysed, edge effects were not considered to have an impact on the overall conclusions: shorter windows were used in the first and last days of the simulations.

Overall, the number of Hits was found to be much smaller when a 1-hour time window was used, resulting in POD and CSI values being calculated for only a small number of sites. Specifically 64 sites over England & Wales for the Q(2)/2 threshold (Q(T) signifies the flow Q having a return period of T years) . This sample was not considered to be sufficiently large to give meaningful results. Although *F* values were calculated at more sites, these values are all very close to zero due to the lack of observed events, and are not useful to analyse further. Thus, POD, *F*, and CSI results are only presented here for the 1-day window. These measures were calculated for the river flow thresholds Q(2)/2, Q(2) and Q(5). However, due to the small sample size found at Q(5) (less than 50 sites with non-zero POD over England & Wales), differences between the simulations are not considered robust at this threshold and are not presented here.

The number of sites where scores can be calculated depends on the number of observed and forecast events and thus varies between the different G2G simulations. Two possible approaches are considered for combining the results of multiple sites when looking at the performance regionally or nationally.

> **Method 1**. Only including sites where the score is non-zero for *all four* simulations. This is fair in the sense that the same number of sites are compared for a given region.

Method 2. Considering sites where the scores were non-zero calculated separately for *each* of the four simulations. Although this results in different numbers of sites being compared, it ensures that all False Alarms or Hits are accounted for. For example, if one rainfall source gives a non-zero number of Hits (performs better), or gives a non-zero number of False Alarms (performs worse) these will be captured in the statistics.

All results combining multiple sites were calculated using both these methods.

Results including only sites with non-zero scores for all sites (Method 1) are shown in Figures 2 and 3 for the Q(2)/2 and Q(2) thresholds respectively. At the Q(2)/2 threshold the Gauge simulation has lower (worse) POD scores than the other simulations which include the use of radar data. This makes sense as the multiquadric raingauge interpolation will not capture small localised peaks in the rainfall pattern for spatial scales smaller than the distance between raingauges. These peaks are expected to be better-detected by the radar data. The opposite performance is seen for the *F* scores, with lower (better) *F* scores for the Gauge simulations. Agreeing with the Bias, Correlation and R^2 statistics (Figure 1) there are an increased number of spurious peaks when rainfall sources that include radar data are used. The number of these False Alarms is not noticeably reduced by merging the radar and raingauge data. Overall, the Gauge simulation does better as shown by higher CSI scores, although these differences are generally small, balancing the effects of fewer Hits and fewer False Alarms. Overall, little difference is seen between the Merged 1h and Merged 24h products, although regional differences exist (with the 24h performing both better or worse depending on the region). For this analysis period, regions to the north and west of England, and Wales tend to benefit from the 24h delay, with regions to the south and east performing worse. However, these differences are small and based on a limited sample. (It will be interesting to see whether similar results are found for the 2017-18 period.) .Of course, in an operational setting, the 1h delay product would be used. The analysis here suggests that this will not significantly affect the overall quality of G2G performance compared to the 24h delay product.

For the Q(2) (median flood) threshold (Figure 3) the *F* and CSI results are similar to those obtained for the Q(2)/2 threshold. However, the poorer performance for the Gauge simulation at the Q(2)/2 threshold is not seen here: at the Q(2) threshold, *F* values are similar overall across the four rainfall sources.

This is somewhat unexpected, but suggests that the predominantly longer-duration, larger-area precipitation leading to the higher threshold exceedances is being better captured by the raingauge interpolation than the lower precipitation peaks. However, it is also possible that these differences are related to the small sample size used at the Q(2) threshold with POD scores only calculated for 93 catchments over England & Wales. This will be tested when analysis has been completed for the 2017-18 period.

Similar conclusions can be drawn from Box Plots when including sites with non-zero scores for each rainfall source separately (Method 2, Figures 4 and 5), although some subtle differences are seen, particularly in the south and east of England. It is interesting that the number of catchments included in this Method 2 approach is much larger for all simulations than used for Method 1 when all simulations are required to have non-zero scores (e.g. a minimum of 482 sites for the POD in Figure 4, but only 382 sites for the POD in Figure 2). Thus, zero-scores (resulting from no Hits, or no False Alarms at a given site) are often found at *different catchments* for the each rainfall source.

Although the Box Plots give a useful overview of the performance across multiple catchments, it is also informative to look at maps of the individual catchment scores. Maps of the POD and *F* scores for the Q(2)/2 threshold are shown in Figures 6 and 7. For ease of comparison, Figure 8 shows the difference in POD and *F* scores between Gauge and those from Radar and Merged simulations of G2G. Overall, the maps are consistent with the Box Plot analysis presented above. In particular, it can be seen that the overall spatial pattern of performance is similar between the different rainfall sources. Looking at the detail, it can be seen that the improvement in POD seen between the Gauge and Radar simulations is predominantly for sites with poor-performance. In contrast, when moving from the Merged 1h to Merged 24h, it tends to be the good sites which are improved. From the maps of *F* and their associated differences (Figure 7, lower panel 8) it can be seen that the Radar, Merged 1h and Merged 24h products have an increased number of poor (high) *F* scores in the north-west and south of England. This is particularly seen for small catchments which will be more sensitive to spurious peaks in the radar data.

Maps of the CSI (Figure 9) show the Gauge source performing better in well performing (shown in green) small catchments to the north of England and along the Pennines. Interestingly, the catchments in these areas with poorer performance are better captured using the radar and merged rainfall as input to G2G. This is consistent with the improvement in POD between Gauge and Radar simulations being seen for the catchments with poorer performance, and the better *F* scores for simulations using the Gauge source as input being seen for the smaller catchments. Future work on verification for the 2017-18 period, including case-study analysis will investigate these catchment-specific differences further.

Figures 10, 11 and 12 show the equivalent maps to Figures 6, 7 and 8, but for the Q(2) threshold. These maps visually highlight the small sample size used for the Q(2) analysis, and reinforce why differences in the Box Plots at this threshold (Figure 3) should be interpreted with this in mind.

3.3 Conclusions for initial analysis of period 1 April 2016 to 31 March 2017

G2G flows in simulation-mode, using four rainfall sources as alternative inputs - Gauge, Radar, Merged 1h, Merged 24h - have been compared for the one-year period 1 April 2016 to 31 March 2017. The key conclusions resulting from the comparison are summarised below.

- Better performance was seen overall when G2G employed Gauge rainfall as input. This was seen when analysing the Bias, Correlation and R^2 Efficiency statistics, and also the CSI categorical skill score.
- Better Probability Of Detection (POD) scores were often seen for the Radar and Merged simulations, particularly at the Q(2)/2 threshold.
- Better False Alarm Rate (*F*) scores were often seen for the Gauge simulations. This was particularly seen over smaller catchments to the northwest and south of England, and was linked to spurious high-precipitation values in the radar data.
- Sample size remains an important consideration, even when a full year of G2G river flows are evaluated in simulation-mode. This is particularly true for thresholds of Q(2) and above. To obtain a meaningful number of scores, even at the Q(2)/2 threshold, a moving time-window of 1 day (96 time-steps) was applied to both the observed and simulated river flows, with any upward crossing of a flow threshold within this window counting as an observed or forecast event.

- Spurious peaks in the radar data can make the merged rainfalls less robust, and a problem for hydrological use. The cause(s) need to be diagnosed and rectified. One possible source is in the mean field bias adjustment of the radar data: this can introduce transient errors that are not sufficiently suppressed by the radar-raingauge merging process.

4 G2G simulations for additional verification periods

4.1 Updated observed river flows, updated raingauge data and a reduced number of sites

Overall, reducing the number of catchments considered, or changing the version of the observed river flow and raingauge data, has little impact on the goodness-of-fit measures (Figure 1). Although slight differences can be seen in the position of outlier catchments, or the exact positioning of the across-catchment median, the relative performance of the different rainfall types remains unchanged and the overall conclusions from the initial analysis period still hold. Differences in the categorical skill scores (Figures 2 to 13) are slightly more noticeable, but again the overall conclusions from the original analysis still hold. When higher thresholds are used (e.g. QMED, Figures 3 and 5) the sample size is smaller so there is more sensitivity to individual catchments and threshold crossings. Thus, as the threshold increases, the exact details of the categorical skill score values vary more with small changes to the underlying data and catchments considered. When considering individual site performance maps (e.g. Figures 6 to 13), the effect of reducing the number of catchments considered from 898 to 730 is noticeable, particularly for the higher thresholds where sample sizes are small. Although the two sets of maps are not contradictory, it is difficult to draw conclusions about the spatial distribution of performance when a large number of catchments, particularly in the south-eastern regions of England, are not included. Thus, the discussion below comparing the performance over different verification periods does not include the single-site maps. For completeness, all maps are still included in the plot documents.

4.2 More recent 12-month verification period 1 September 2017 to 31 August 2018

Comparing the two sets of 12-month results (2016-2017 and 2017-2018), a similar pattern overall is seen for the goodness-of-fit measures, particularly for the Correlation. However, there are also some noticeable differences. In particular, both the merged products show differences in the Bias, with results for the most recent period tending to have Bias values closer to those seen for the Raingauge data, whereas those for the initial period tended to be closer to the Radar values. This suggests that improvements to the merging algorithm are giving more weight to the Raingauge data in terms of absolute Bias (noting that the Raingauge data generally has lower-magnitude biases than the Radar data). Differences between the R^2 values for the different rainfall-types are negligible, although all R^2 values are generally slightly higher for the more recent period, particularly over the Midlands and Anglian regions.

4.3 Longer verification periods 1 October 2016 to 30 September 2018 and 1 April 2016 to 30 September 2018

As expected, the results from the two longer periods – for two water years and the full period of comparison - fall somewhere in between those for the separate two-year periods. For the Q(2) threshold, the improvement in sampling uncertainty is clear for the categorical skill scores, with less (noisy) variation seen between different regions, and more consistent differences seen between the rainfall-types. In general, differences between the rainfall-types were more noticeable for the higher Q(2) threshold, with the overall CSI measure showing that G2G performs best with Raingauge data as input, followed by the 24h delay merged data, 1h delay merged data, and then the Radar data.

5 Conclusions

Overall, the conclusions previously set down - in Section 3.3 for the 1 April 2016 to 31 March 2017 verification period - are found to still hold for other more recent verification periods, and when a longer 2-year verification period is used.

Three additional conclusions can be made.

- Reducing the number of catchments considered (e.g. from 898 to 730) makes it harder to discern spatial patterns at the catchment-scale, especially if the selected catchments are unevenly distributed.

- The merged products perform better for more-recent periods (e.g. 1 September 2017 to 31 August 2018). In particular, the Bias in G2G simulated river flows is improved for the more recent verification period, and lies closer to that seen when Raingauge data are used as input, in contrast with earlier periods where the Bias is more-similar to that when Radar data are used as input.

- Considering a longer 2-year verification period gives clearer, less-noisy results for higher thresholds, with more consistency both across regions and with the Q(2)/2 threshold results. This is due to a reduction in the sampling uncertainty.

REFERENCES

Jewell, S.A. and Gaussiat, N. 2015. An assessment of kriging-based rain-gauge–radar merging techniques. *Q. J. R. Meteorol. Soc.*, 141, 2300–2313.

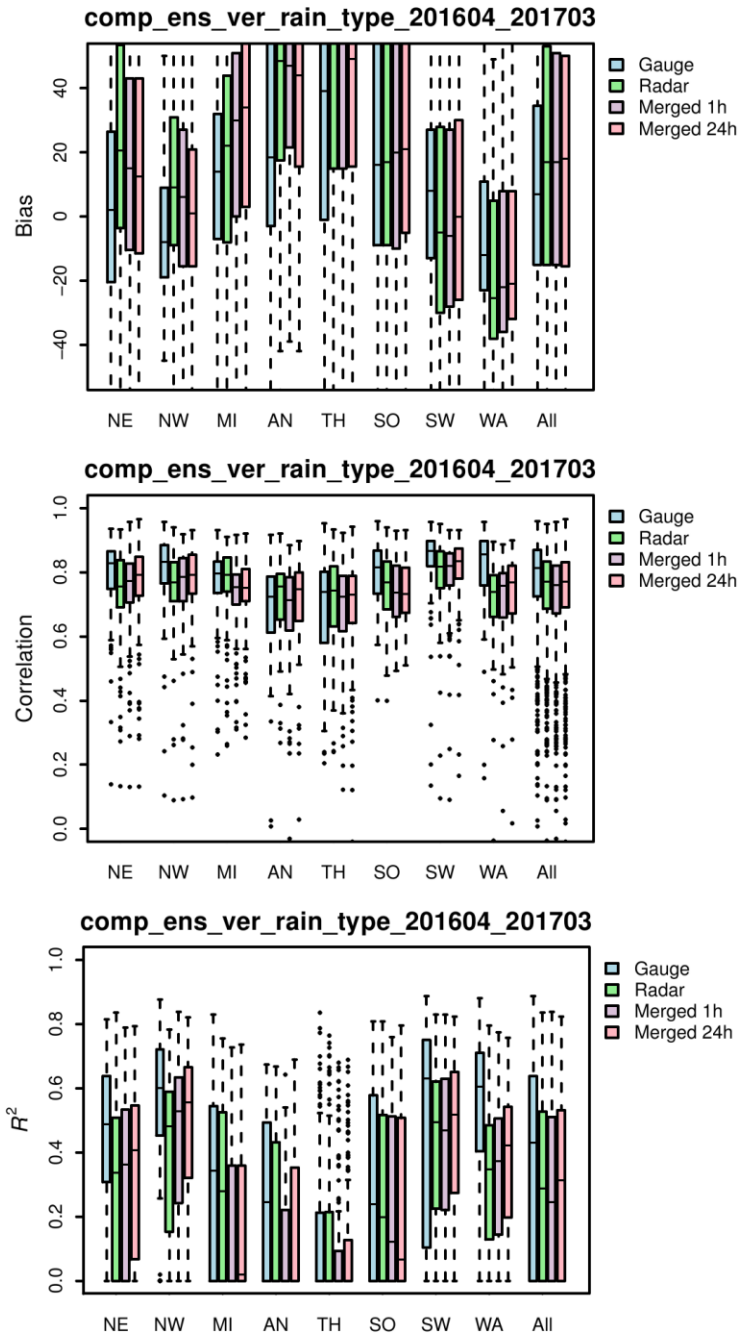


FIgure 1 Box Plots comparing the performance of G2G river flow simulations using different observed precipitation sources as input. The Bias (top), Correlation (middle), and R^2 Efficiency (bottom) goodness-of-fit statistics are shown. Bars are for each grouping of catchments considered: each region in England, for Wales, and for all catchments in England & Wales. Each set of bars contains (from left to right) results for G2G simulations using Gauge, Radar, Merged 1h and Merged 24h observed precipitation data as input. Each bar shows the median (solid line) and interquartile range (coloured box) of the distribution of statistics over the set of catchments. Dashed lines extend to 1.5 times the interquartile range from the box, and indicate the typical range of the data. Outlying points are shown by black dots.

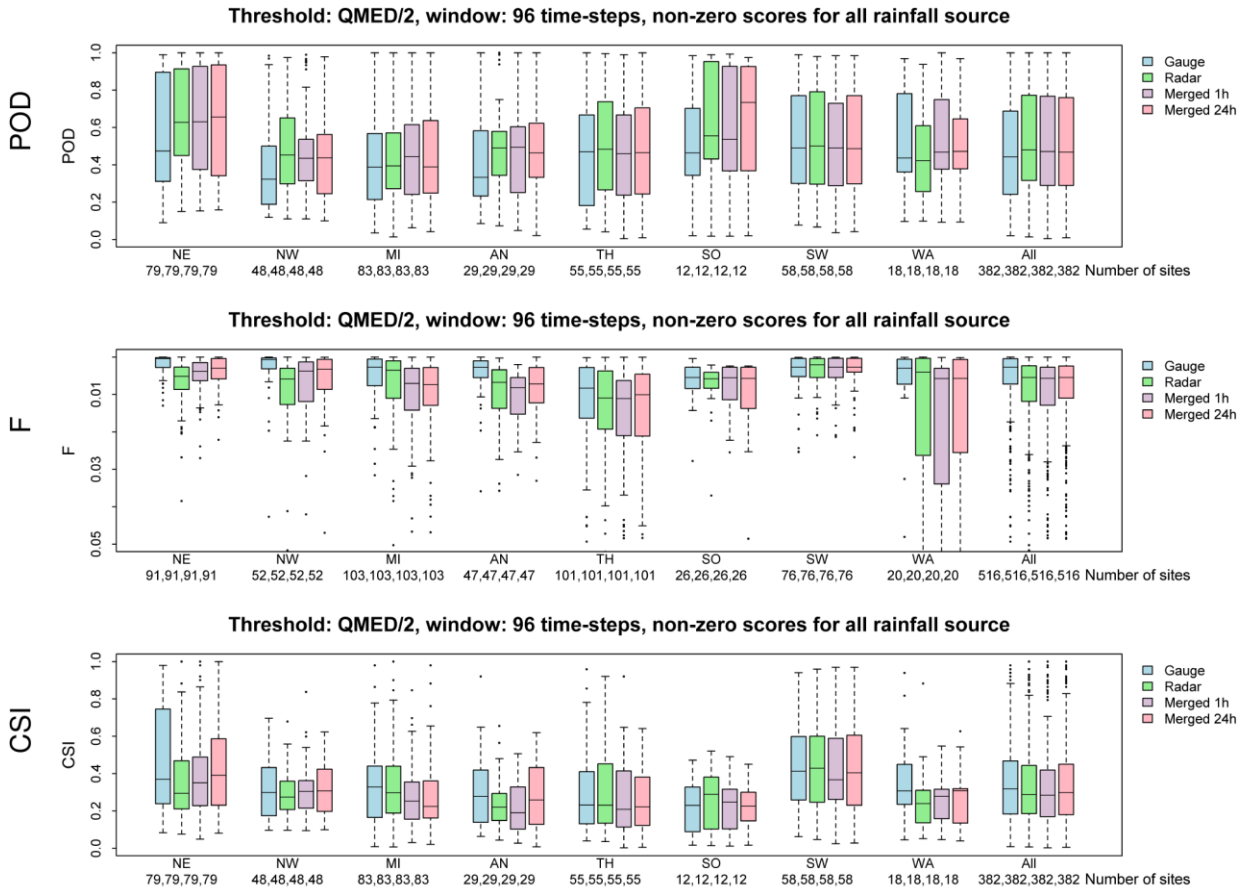


Figure 2 Box Plots comparing the performance scores (POD, *F*, CSI) of G2G river flow simulations - using different observed precipitation sources as input - for the Q(2)/2 threshold and 24h moving window. Bars are for each grouping of catchments considered: each region in England, for Wales, and for all catchments in England & Wales. Each set of bars contains (from left to right) results for G2G simulations using Gauge, Radar, Merged 1h and Merged 24h observed precipitation data as input. Each bar shows the median (solid line) and interquartile range (coloured box) of the distribution of scores over the group of catchments. Only catchments with non-zero scores for all precipitation sources (Method 1) are included (the number of catchments is indicated beneath the bars). Dashed lines extend to 1.5 times the interquartile range from the box, and indicate the typical range of the data. Outlying points are shown by black dots.

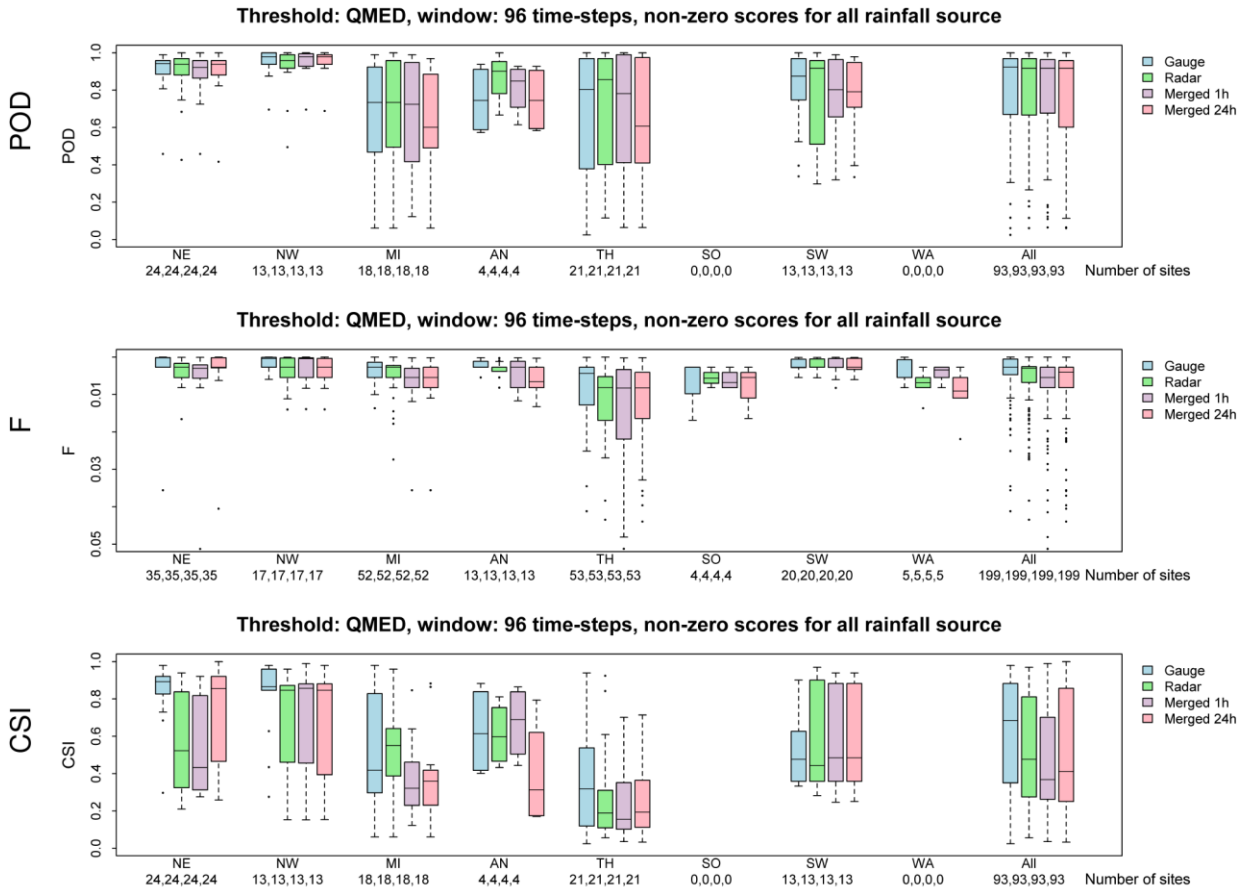


Figure 3 Box Plots comparing the performance scores (POD, *F*, CSI) of G2G river flow simulations - using different observed precipitation sources as input - for the Q(2) threshold and a 24h moving window. Bars are for each grouping of catchments considered: each region in England, for Wales, and for all catchments in England & Wales. Each set of bars contains (from left to right) results for G2G flow simulations using Gauge, Radar, Merged 1h and Merged 24h observed precipitation data as input. Each bar shows the median (solid line) and interquartile range (coloured box) of the distribution of scores over the group of catchments. Only catchments with non-zero scores for all precipitation sources (Method 1) are included (the number of catchments is indicated beneath the bars). Dashed lines extend to 1.5 times the interquartile range from the box, and indicate the typical range of the data. Outlying points are shown by black dots.

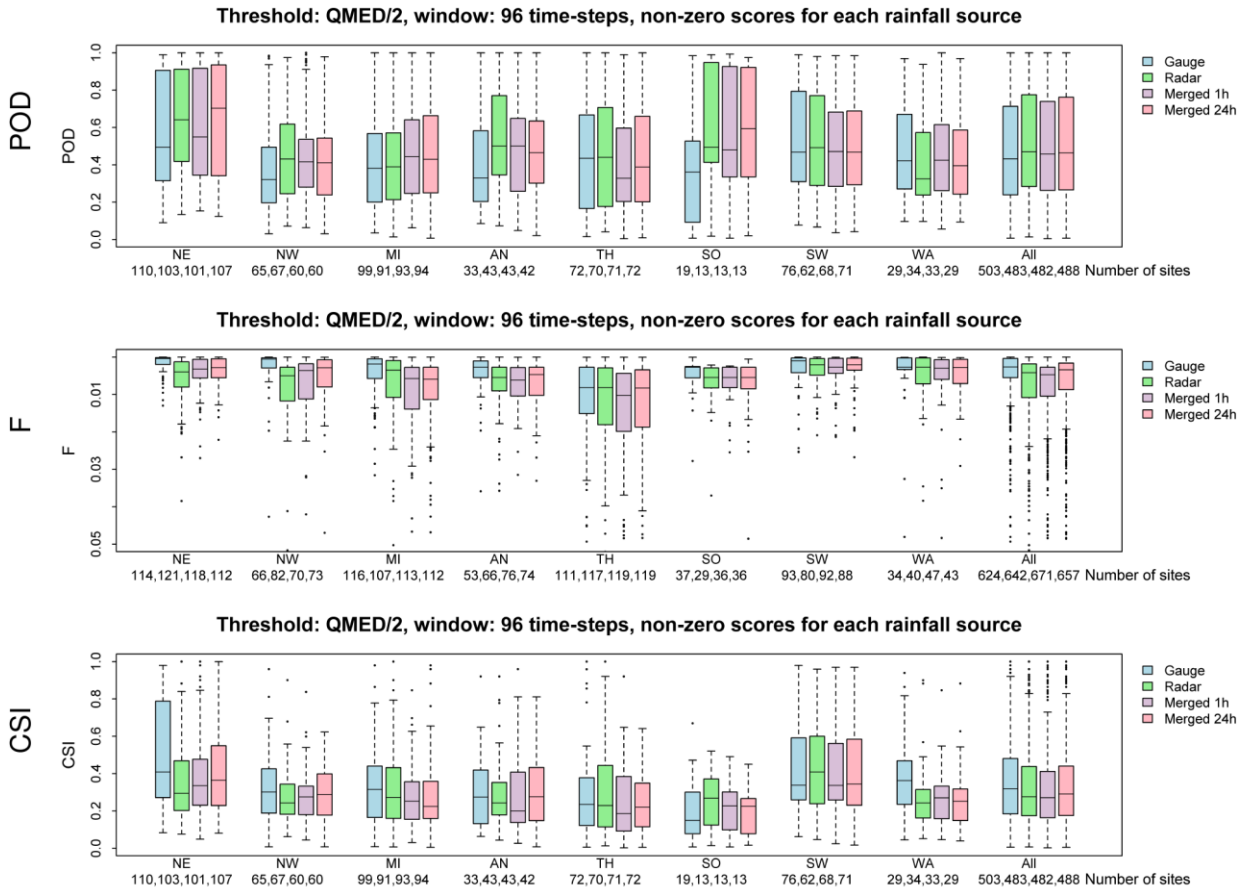


Figure 4 Box Plots comparing the performance scores (POD, *F*, CSI) of G2G river flow simulations - using different observed precipitation sources as input - for the Q(2)/2 threshold and a 24h moving window. Bars are for each grouping of catchments considered: each region in England, for Wales, and for all catchments in England & Wales. Each set of bars contains (from left to right) results for G2G simulations using Gauge, Radar, Merged 1h and Merged 24h observed precipitation data as input. Each bar shows the median (solid line) and interquartile range (coloured box) of the distribution of scores over the set of catchments. Catchments with non-zero scores for each precipitation source (Method 2) are included (the number of catchments included is indicated beneath the bars). Dashed lines extend to 1.5 times the interquartile range from the box, and indicate the typical range of the data. Outlying points are shown by black dots.

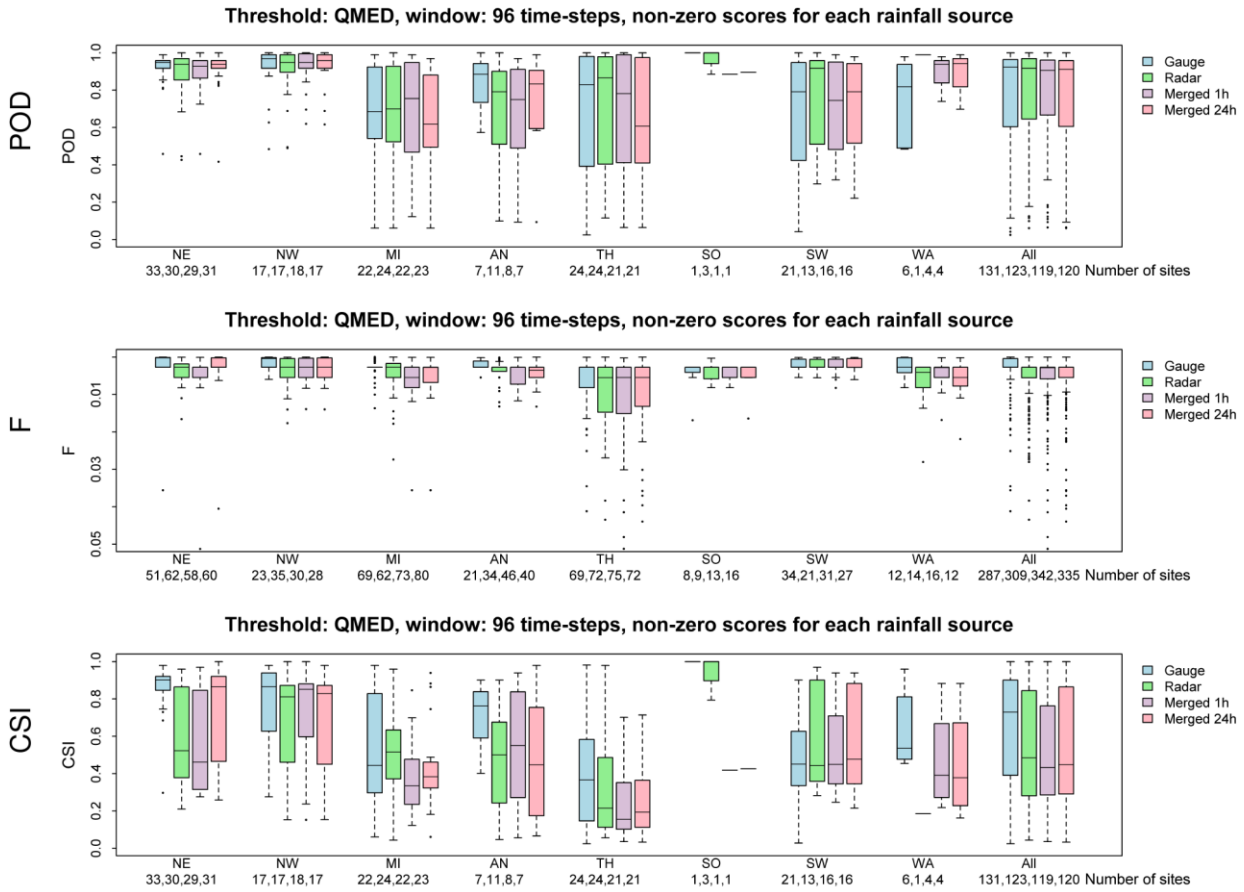


Figure 5 Box Plots comparing the performance of G2G river flow simulations - using different observed precipitation sources as input - for the Q(2) threshold and a 24h moving window. Bars are for each grouping of catchments considered: each region in England, for Wales, and for all catchments in England & Wales. Each set of bars contains (from left to right) results for G2G simulations using Gauge, Radar, Merged 1h and Merged 24h observed precipitation data as input. Each bar shows the median (solid line) and interquartile range (coloured box) of the distribution of scores over a grouping of catchments. Catchments with non-zero scores for each precipitation source are included - the number of catchments included is indicated beneath the bars. Dashed lines extend to 1.5 times the interquartile range from the box, and indicate the typical range of the data. Outlying points are shown by black dots.

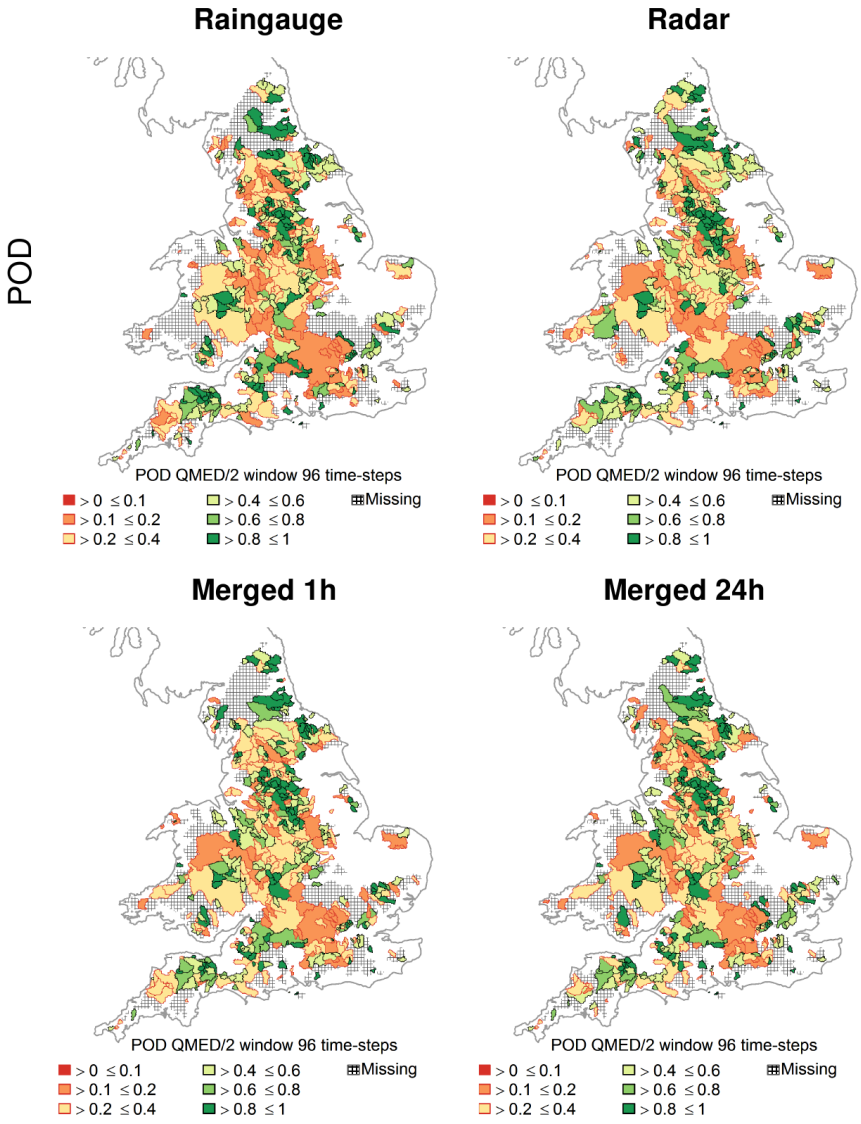


Figure 6 Maps of POD scores calculated for the Q(2)/2 threshold and a 24h moving window. POD scores are shown from red with no outline (poor) to green with outline (good) for the G2G river flow simulations using Gauge, Radar, Merged 1h and Merged 24h precipitation data as input.

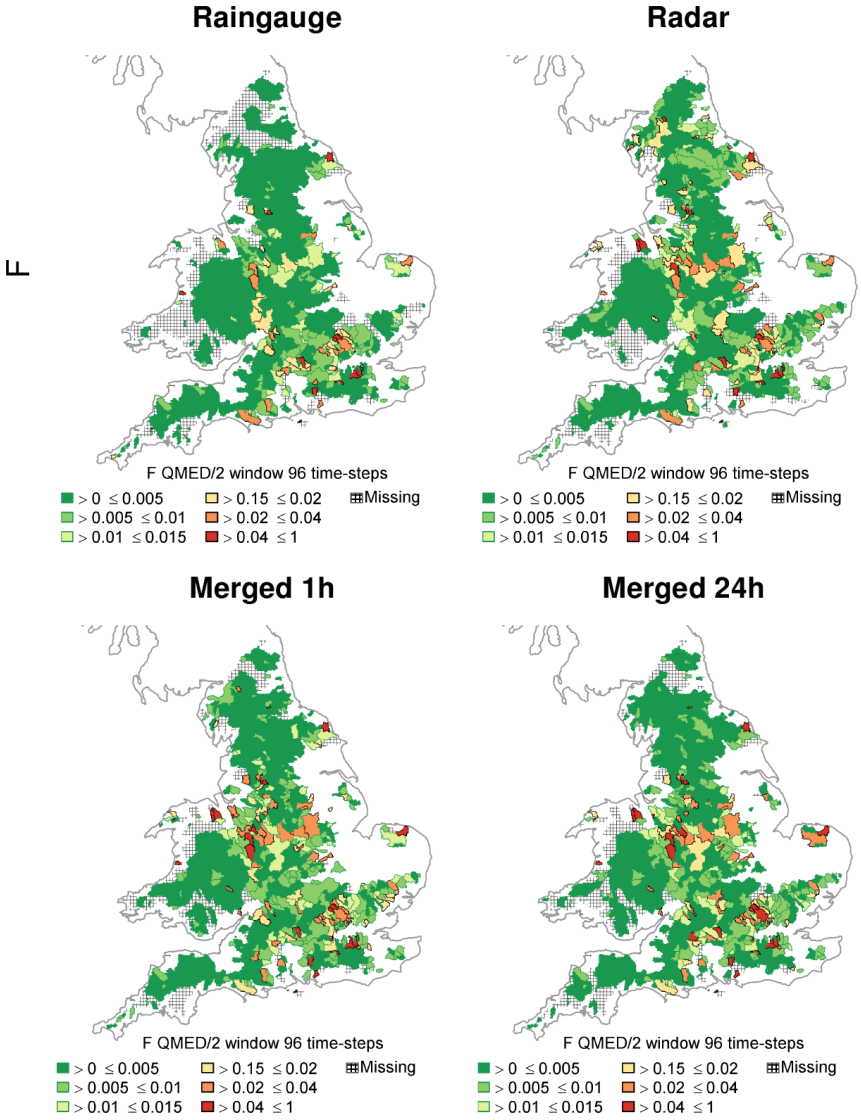


Figure 7 Maps of *F* scores calculated for the Q(2)/2 threshold and a 24h moving window. *F* scores are shown from red with no outline (poor) to green with outline (good) for the G2G river flow simulations using Gauge, Radar, Merged 1h and Merged 24h precipitation data as input.

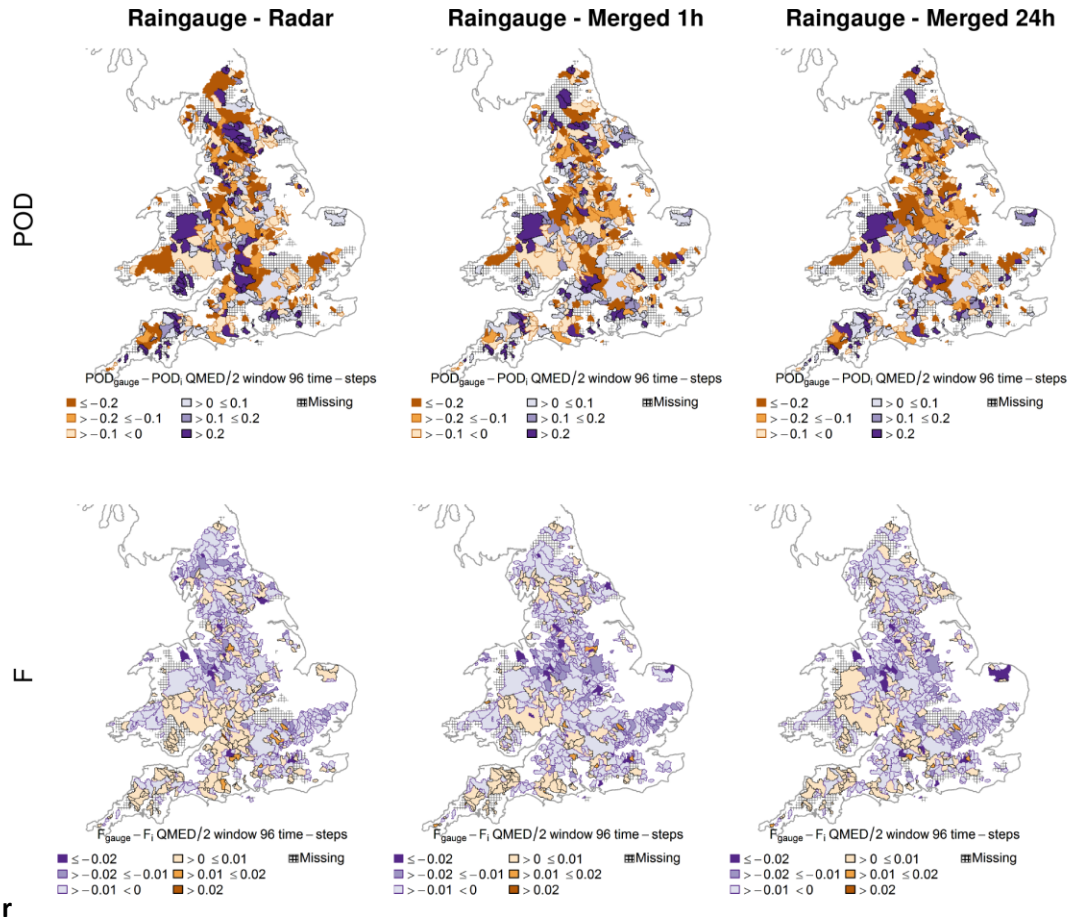


Figure 8 Maps of the difference in POD and *F* scores calculated for the Q(2)/2 threshold and a 24h moving window. Gauge-Radar (left), Gauge-Merged 1h (middle) and Gauge-Merged 24h (right). Purple colours show Gauge performing better, orange colours show Gauge performing worse.

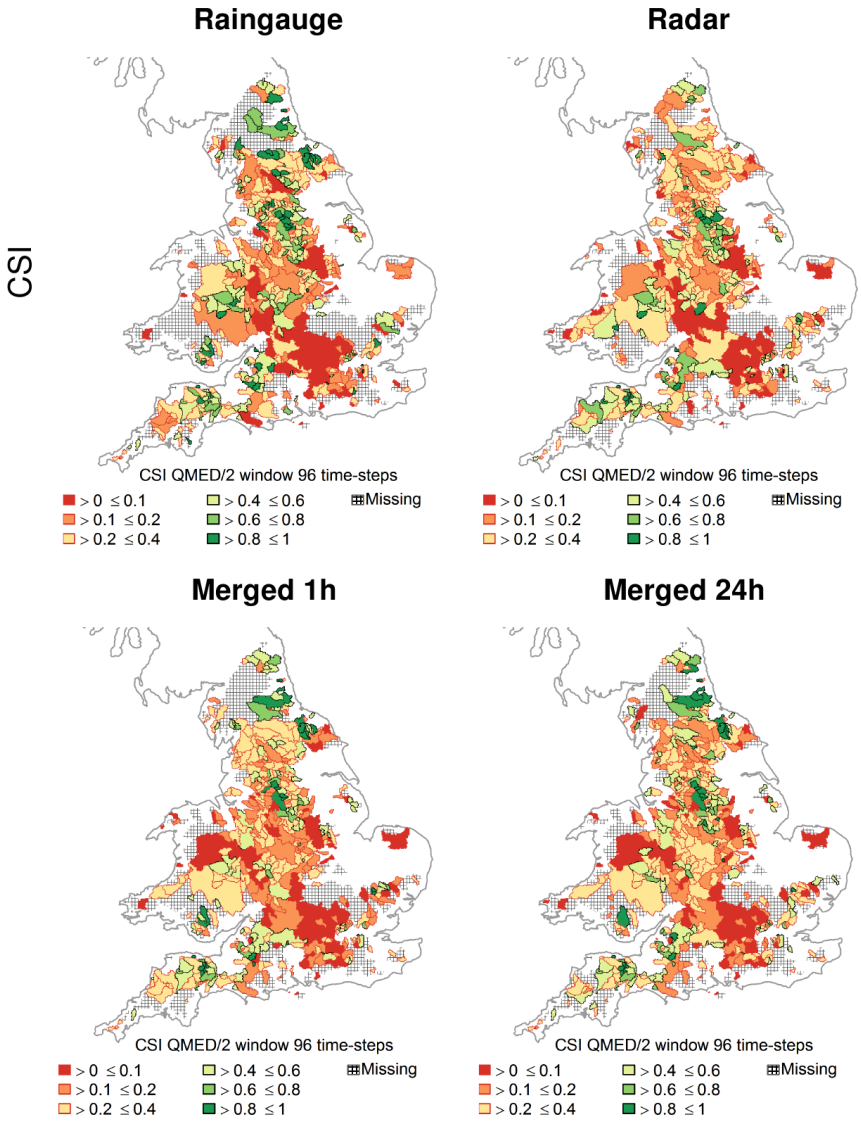


Figure 9 Maps of CSI scores calculated for the Q(2)/2 threshold and a 24h moving window. CSI scores are shown from red with no outline (poor) to green with outline (good) for the G2G river flow simulations using Gauge, Radar, Merged 1h and Merged 24h precipitation data as input.

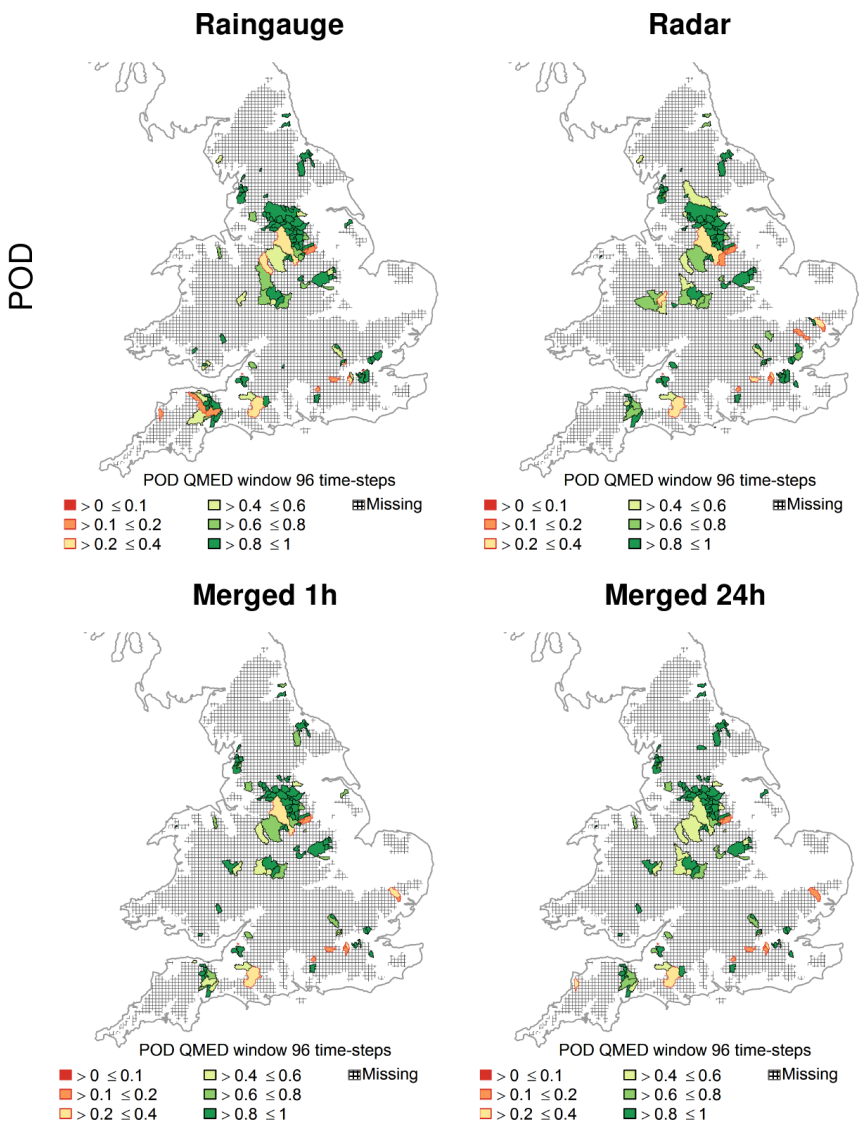


Figure 10 Maps of POD scores calculated for the Q(2) threshold and a 24h moving window. POD scores are shown from red with no outline (poor) to green with outline (good) for the G2G river flow simulations using Gauge, Radar, Merged 1h and Merged 24h precipitation data as input.

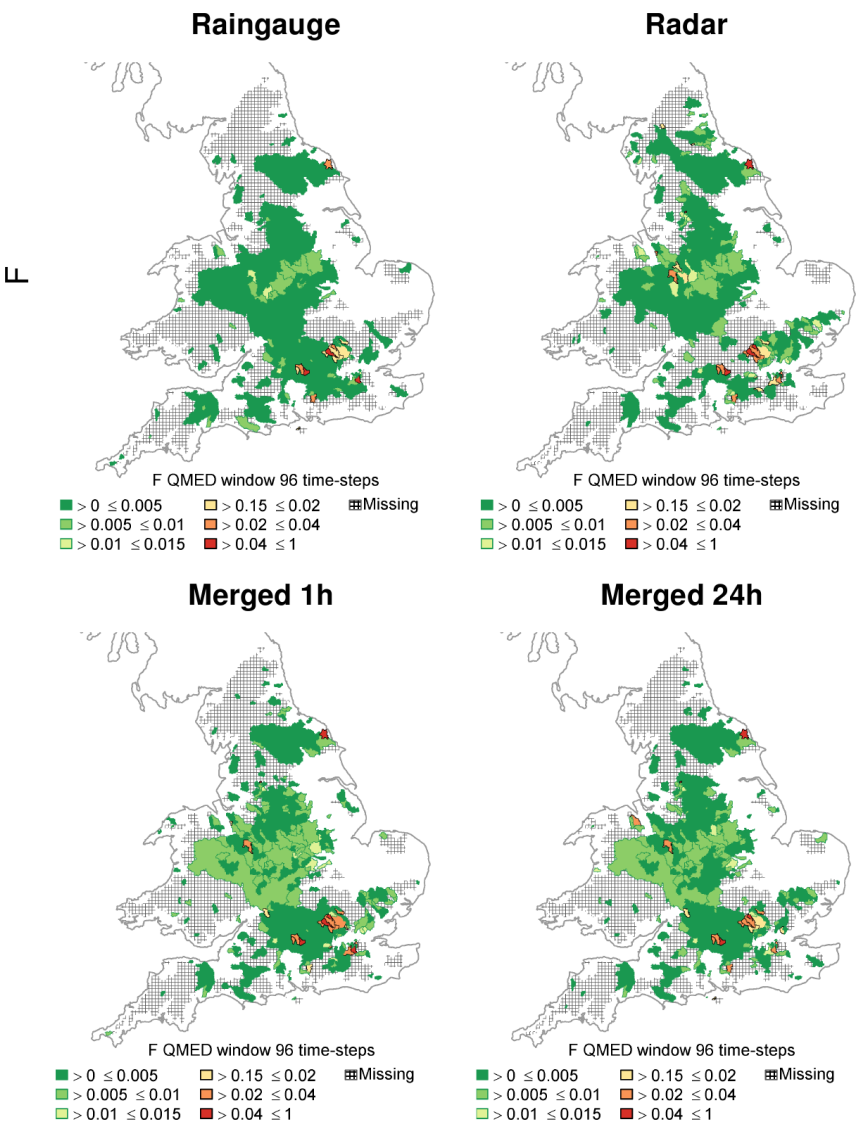


Figure 11 Maps of *F* scores calculated for the Q(2) threshold and a 24h moving window. *F* scores are shown from red with no outline (poor) to green with outline (good) for the G2G river flow simulations using Gauge, Radar, Merged 1h and Merged 24h precipitation data as input.

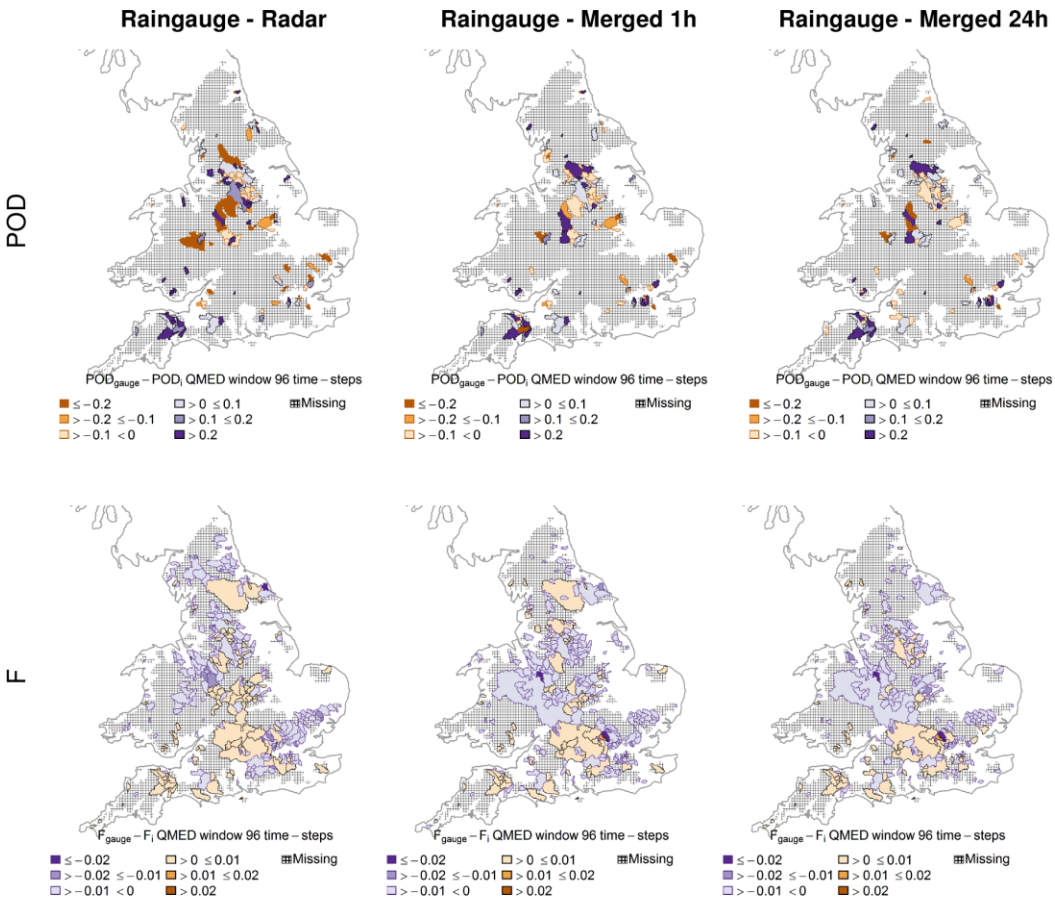


Figure 12 Maps of the difference in POD and F scores calculated for the Q(2) threshold and a 24h moving window. Gauge-Radar (left), Gauge-Merged 1h (middle) and Gauge-Merged 24h (right). Purple colours show Gauge performing better, orange colours show Gauge performing worse.

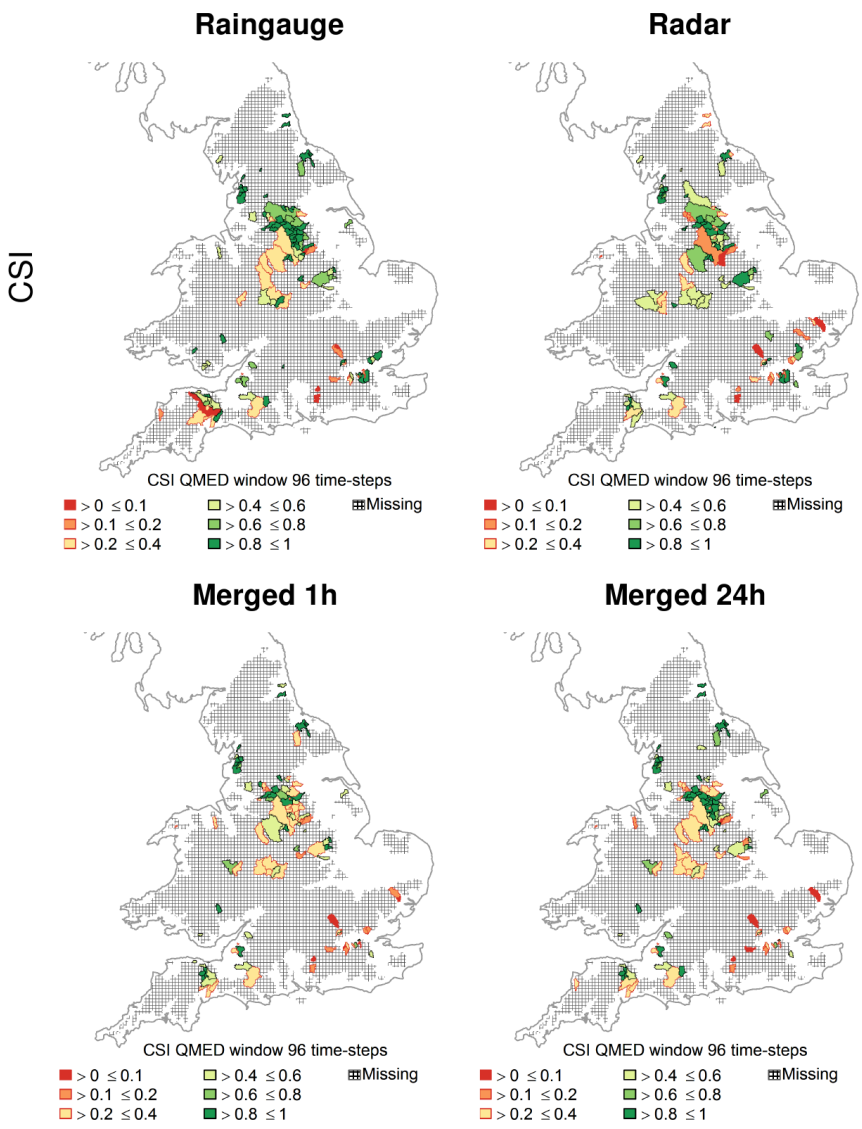


Figure 13 Maps of CSI scores calculated for the Q(2) threshold and a 24h moving window. CSI scores are shown from red with no outline (poor) to green with outline (good) for the G2G river flow simulations using Gauge, Radar, Merged 1h and Merged 24h precipitation data as input.

Rainfall and River Flow Ensemble Verification: Phase 2
Evaluation and comparison of December 2015 case study storms
Final Report Appendix C.1

1. Introduction

For each of the case study storms - Desmond, Eva and Frank - an informal assessment of the evolution of the precipitation ensemble forecasts is first carried out by comparing the catchment-mean precipitation accumulations from the ensemble members with those from raingauges. This is consistent with the approach followed in Phase 1. Catchment-mean precipitation accumulations from radar rainfall are also considered. Under Phase 2, the assessment is extended to include comparison of the 99th percentile in-catchment values for each ensemble member with those from radar and raingauges, to consider the efficacy of using catchment-mean precipitation for ensemble verification.

For all plots featuring in this assessment, the y-axes are capped at 20 mm for the catchment-mean precipitation and at 40 mm for the 99th percentile in-catchment values to help with the comparison. Throughout, the observed precipitation time-series are in blue and the ensemble members in grey. Each panel represents a 7-day forecast, with forecasts issued at six-hourly intervals at 01:00, 07:00, 13:00 and 19:00, covering the period of interest (and subject to forecast availability).

In the following sections all precipitation "events" or episodes are discussed with respect to the *first* time-series plot (depicting the first forecast in the series), plotted in the top left panel.

For each case study, the precipitation time-series are visually compared with the river flow hydrographs for each forecast issued. These hydrographs were presented in Section 7 of the Phase 1 Report but are included here to provide a self-contained comparison.

2. Case study catchments

The case study catchments are those considered In the Phase 1 Report (Section 6.1.2) and summarised in Table 1 and mapped in Figure 1.

Table 1 Catchments used for individual site analysis

Catchment name	G2G ID	G2G catchment area (km^2)	Region
Eden at Sheepmount	765512	2274	North West England
Lune at Caton	724629	984	North West England
Dee at Park	234291	1834	North East Scotland
Dee at Polhollick	234294	691	North East Scotland
Dee at Mar Lodge	234274	289	North East Scotland

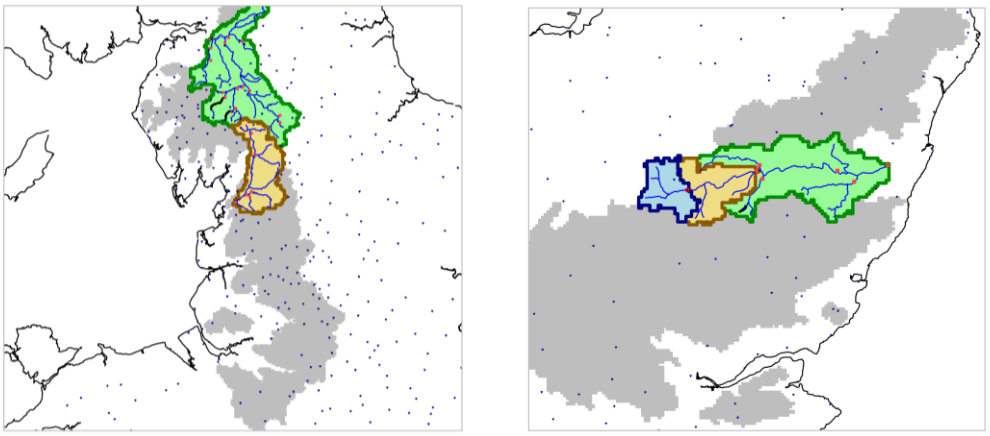


Figure 1 Location maps for catchments used for individual site analysis. Top left: Eden at Sheepmount (green) and Lune at Caton (brown) from the North West England region; top right: Dee at Park (green, brown and blue), Polhollick (brown and blue) and Mar Lodge (blue) from the North East Scotland region. Grey shading shows all G2G catchments within a given region. Raingauges are shown by blue triangles, river gauges by red squares.

Two catchments from different river basins in the North West region of England are selected to represent catchments with many threshold crossings during the winter period. Considering catchments from different river basins, but with results from a similar pool of sites, allows the effects of pooling to be demonstrated. Salient points for the two catchments selected are noted below.

Eden at Sheepmount (G2GID 765512). High flows, 2274 km^2 catchment. Greater than 200-year return period event with record peak flow on 6 December 2015 (Storm Desmond) and with mean flow for period Nov 2015 to Jan 2016 being 232% of long-term average flow (Barker *et al.*, 2016). Moderate to high solid geology permeability in the valley, with low permeability in the western headwater areas (the Lake District), and low superficial deposit permeability. This catchment is mainly natural, but contains the towns of Carlisle and Penrith. Reservoirs control around 2% of the catchment (NRFA).

Lune at Caton (G2GID 724629). High flows, 984 km^2 catchment. 100 to 200 year return period event with record peak flow on 5 December 2015 (Storm Desmond). Moderate to high solid geology permeability with headwaters in the Pennines and Shap Fell, and low superficial deposit permeability (where present). Note that this site contains some artificial influences (notably reservoirs for public water supply).

Three nested catchments from the Dee (Grampian) basin in the North East Scotland region are selected which experience Q(5) threshold exceedances for all three storms Desmond, post Storm Eva and Frank. All three catchments are natural to within 10% at the 95th percentile flow (NRFA). Overall, G2G performs reasonably with better performance seen for the larger, downstream, catchments. All three catchments have low solid geology permeability, and where there are superficial deposits their permeability is also low. Notes on the three catchments follow.

Dee at Park (G2GID 234291). 1834 km^2 (3rd largest) catchment in the North East Scotland region. Greater than 200-year return period event with record peak flow on 30 December 2015 (Storm Frank), mean flows Nov 2015 - Jan 2016 are 236% of long-term average flow (Barker *et al.*, 2016). Low solid and superficial deposit permeability.

Dee at Polhollick (G2GID 234294). 691 km^2 catchment, upstream of Park in the middle reaches of the River Dee (Grampian). Upland catchment with mountainous headwaters, snow-covered in winter. This river gauging station is just upstream of Ballater, which was badly flooded around the end of December 2015 (Storm Frank).

Dee at Mar Lodge (G2GID 234274). 289 km^2 headwater catchment**,** upstream of Polhollick. Catchment rainfall may be significantly underestimated (NRFA).

3. Case Study 1: Storm Desmond

Figure 2 sets the scene for this period of unusually wet weather over large parts of the UK, showing the daily radar-rainfall maps for each day (00:00 to 00:00) from 31 October to 6 December 2015.

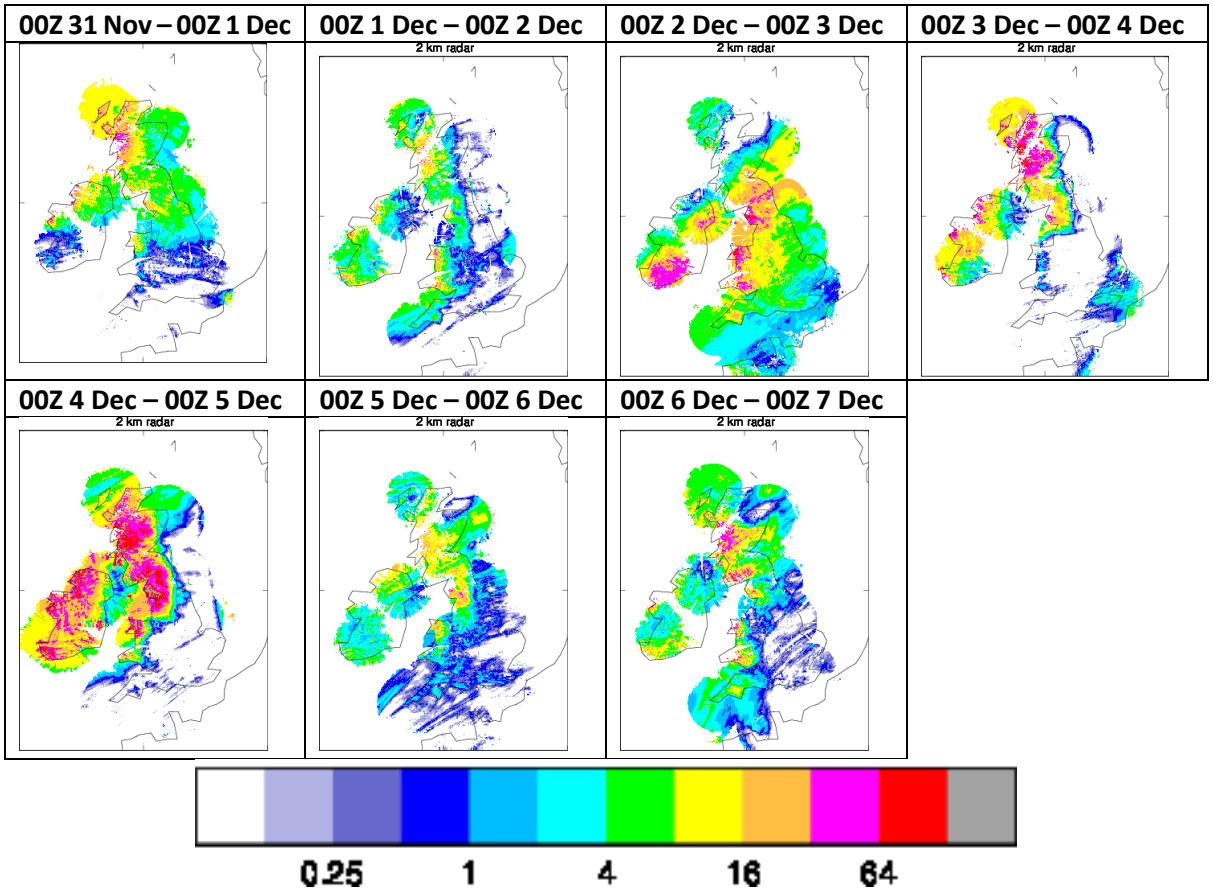


Figure 2 Sequence of daily radar-rainfall accumulations (mm) over the period 31 Nov to 6 Dec 2015

Figure 3 covers all the forecasts issued between 07:00 30 November and 01:00 6 December 2015 and compares the catchment-mean precipitation of the ensemble and the raingauges for the River Eden at Sheepmount (765512). The first panel shows that the raingauge time-series has five distinct precipitation "events" or episodes. Events 1 and 2 are well captured in the first forecast (07:00 30 November), whilst events 3 and 4 are merged into one in the ensemble which spans the period between the observed precipitation events. Event 5 is again well captured. Catchment-mean precipitation totals for the ensemble are comparable to those from raingauges for events 1 and 2 but for the other events the ensemble gives around twice those observed. The timing mismatch for the third and fourth observed peaks persists for several forecast cycles. Forecasts of the fifth peak are generally good from the first forecast onwards, especially in terms of timing.

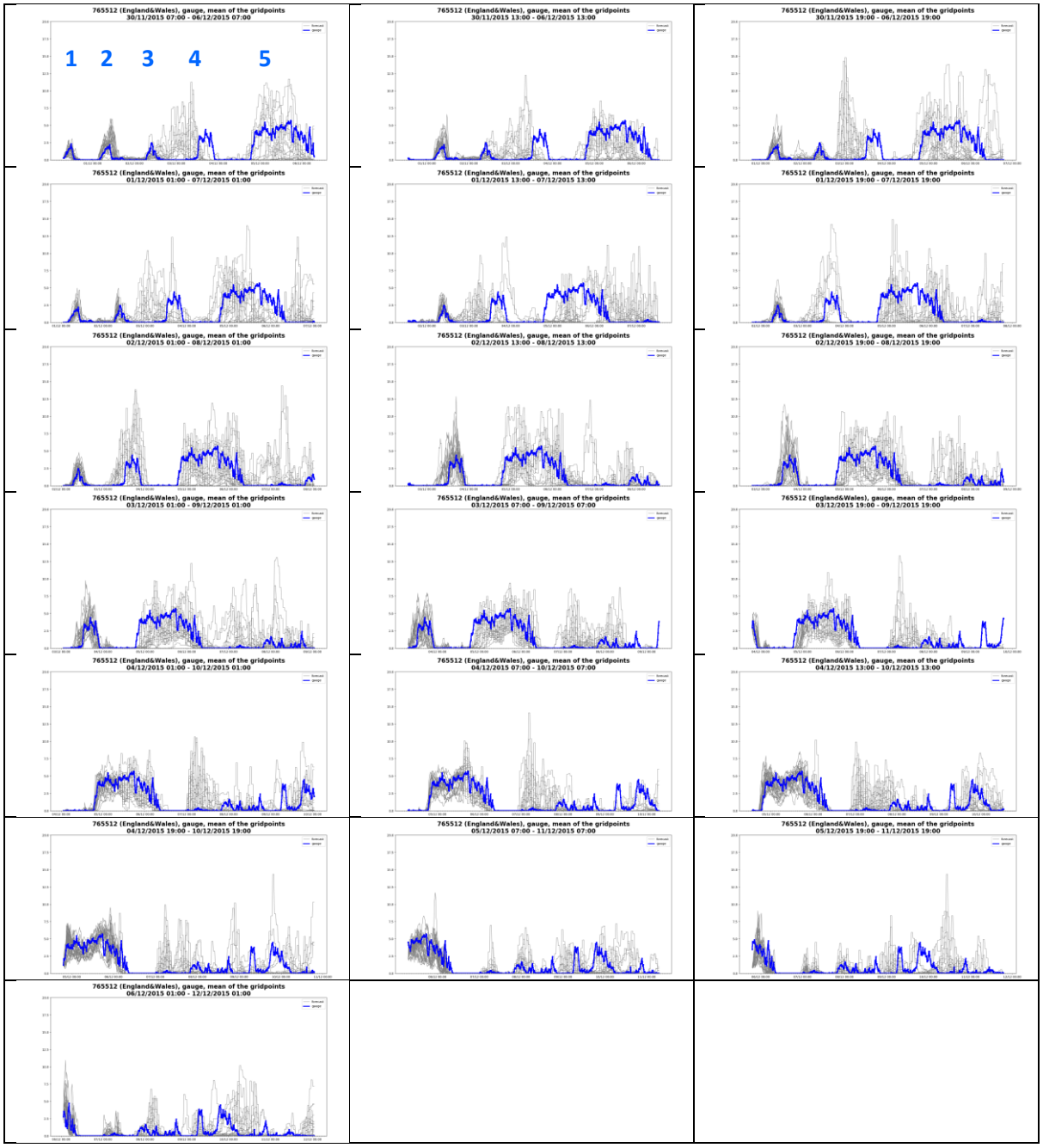


Figure 3 Time-series of catchment-mean precipitation for the ensemble members (grey) are compared to those from raingauges (blue). River Eden at Sheepmount for Storm Desmond.

The range of catchment-mean totals across ensemble members tightens with decreasing forecast lead-time and generally those from raingauges are encompassed by the ensemble spread.

For the entire event, the catchment-mean rainfall accumulations from the raingauges are generally higher than those from radar (not shown) with raingauge estimates tracking closer to the upper boundary of the ensemble values and the radar-rainfall estimates to the lower boundary.

The 99th percentile in-catchment values were also considered for the precipitation ensemble members and the raingauge and radar-rainfall accumulations. Some very large spot totals appear in the ensemble that do not feature in either the raingauge or radar-rainfall values. Generally, the 99th percentile time-series are noisier (see Figure 4 for examples) with the raingauge-derived values often larger than the radar-rainfall counterparts.

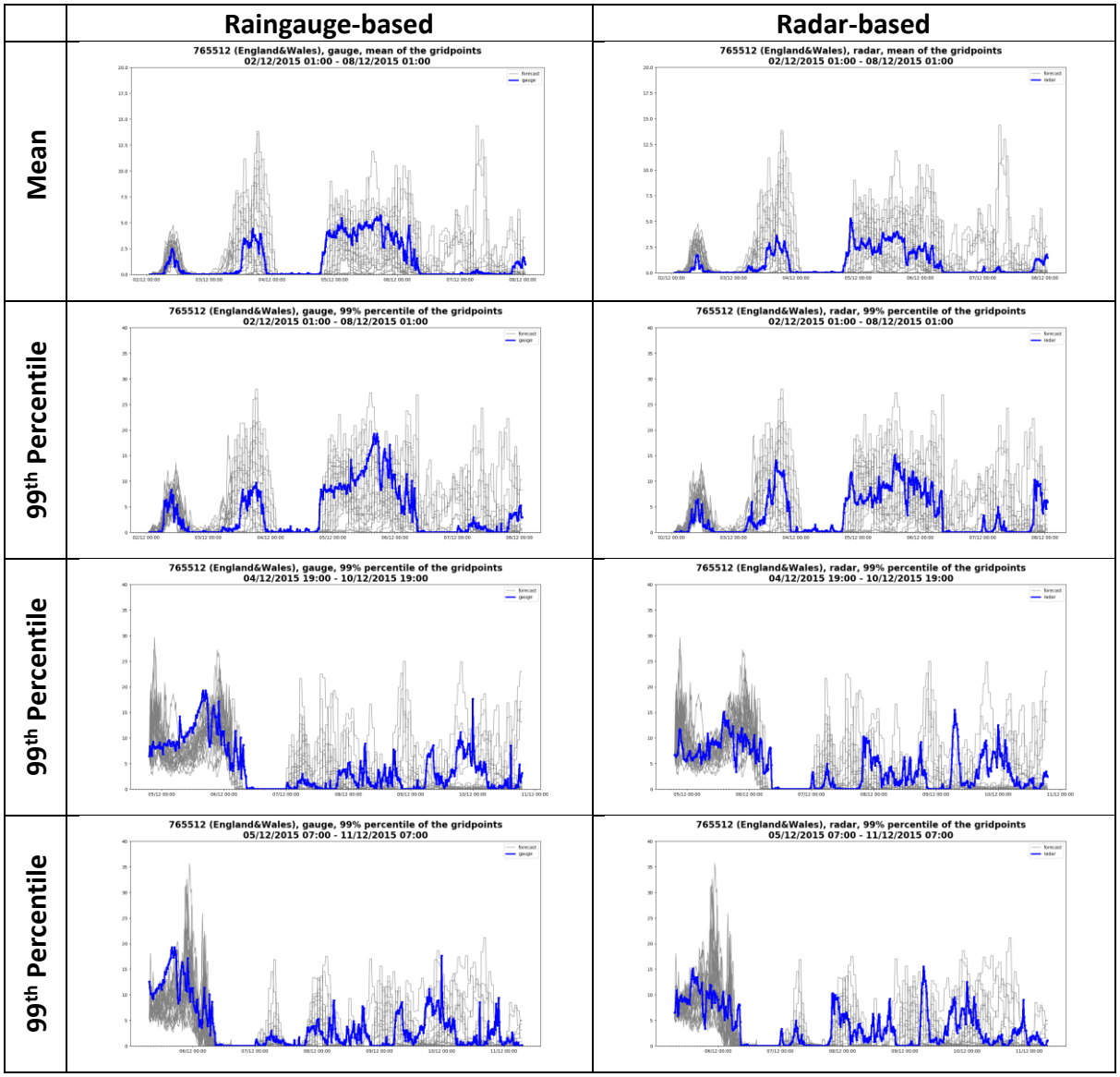


Figure 4 Comparison of catchment-mean precipitation for ensemble members (top row) and 99th percentile in-catchment values (lower rows) for the raingauge-based (left) and radar-based (right) catchment means for Storm Desmond. Note that the y-axis maximum for the bottom rows is different to the top row. Catchment is the River Eden at Sheepmount.

In Figure 4, the precipitation catchment-mean totals (top row) are compared to the 99th percentile in-catchment totals (lower rows). The left and right columns compare the raingauge and radar-derived values. In this instance the catchment-mean totals from radar are noticeable lower than those from raingauges. As can be expected, the 99th percentile in-catchment values are much larger than the catchment-means with the observed totals encompassed within the ensemble spread.

Comparing with the river flow results for the Eden at Sheepmount (Figure 5, as Figure 7.7 of the Phase 1 Report), the first two small peaks in precipitation are seen not to result in a noticeable river flow response. There is, however, a river flow response from the third and fourth precipitation events that were merged into a single event by the precipitation ensemble in the longer lead-time forecasts. This can be seen in the hydrographs as a spurious peak on 3 December for the first five forecasts presented

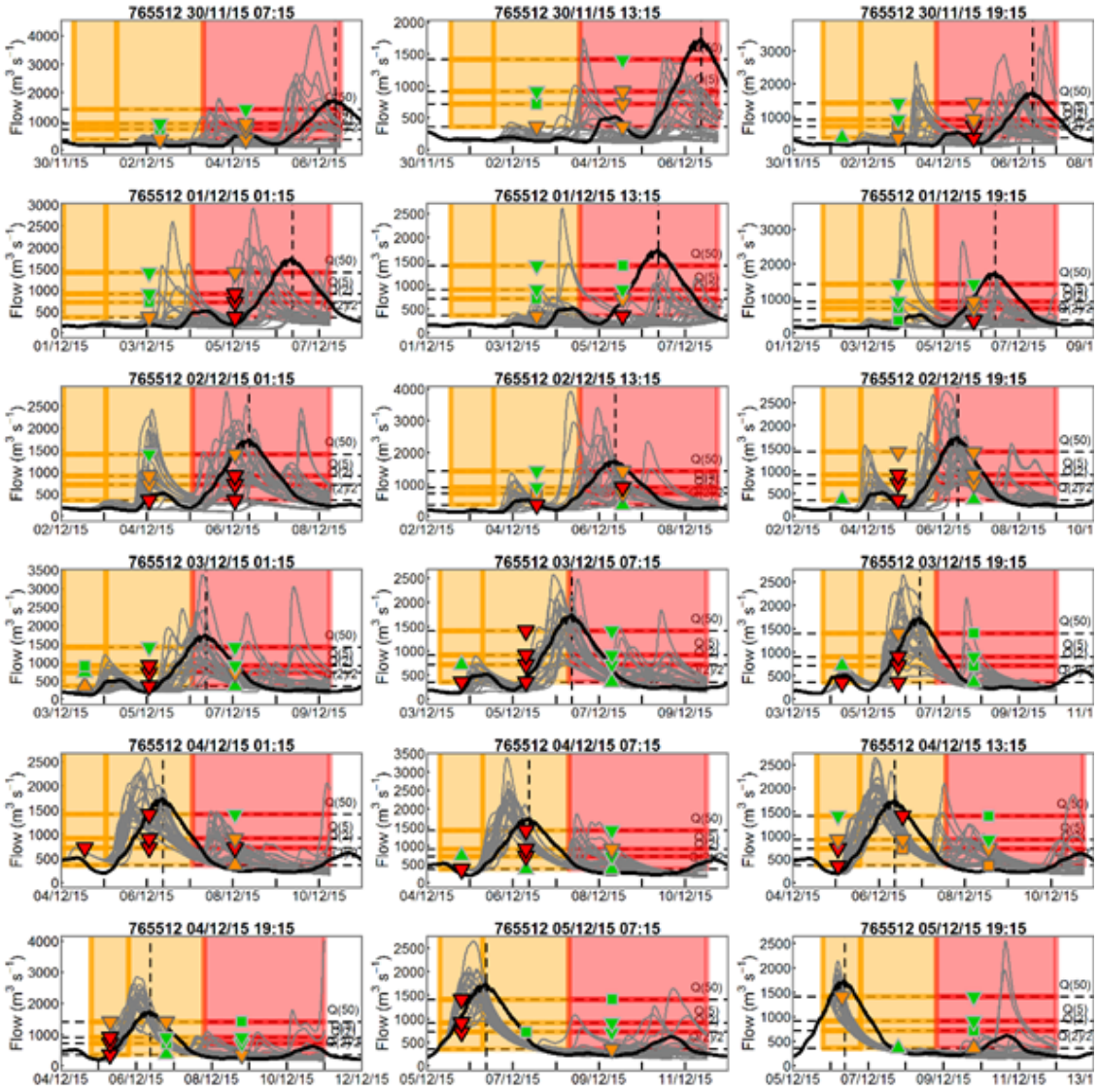


Figure 5 Ensemble river flow forecast hydrographs for the Eden at Sheepmount covering the peak river flow associated with Storm Desmond at 09:00 6 December 2015. Ensemble members are shown in grey and the observed river flow in black. The vertical dashed black line shows the time of peak flow at Sheepmount (09:00 6 December 2015). Symbol colours and shapes are defined in Section 7.2 of the Phase 1 Report. (This is Figure 7.7 of the Phase 1 Report.)

in Figure 3. Similarly, to the precipitation ensemble, there is high confidence in the river flow ensemble that this peak would occur. However, there are a small number of river flow ensemble members which give particularly high flows, exceeding the Q(50) threshold. The 99th percentile precipitation in-catchment values (not shown) provide a prolonged spell of intermittently heavy precipitation of between 20 and 40 mm. For later forecasts, the timing of the third peak is better captured by both the precipitation and river flow ensembles, although the magnitude of the peak is still around three times larger than observed. It is only when the forecasts start to be initialised within a few hours of this precipitation (fourth line of plots in Figure 5) that the peak's magnitude is well predicted. This highlights directly the impact of the precipitation ensemble forecast quality on the G2G river flow ensemble.

The fifth precipitation event results in the highest river flow response, with observed flows exceeding the Q(50) threshold. For most forecasts, the catchment-mean precipitation totals from the ensemble reach twice those of the observations as discussed above. For these forecasts (e.g. 07:00, 19:00, 01:00, 13:00) the peak is also overestimated by the majority of river flow ensemble members. However, for other forecast initialisations, the precipitation ensemble values are closer to, or below, those observed, and the river flow response is less than observed. *This again demonstrates the direct link between precipitation and river flow ensemble*. Interestingly, for the final six forecasts before the event end, the river flow ensemble still overestimates the peak magnitude, although it appears well captured by the catchment-mean precipitation values from the ensemble. However, when considering the 99th percentile in-catchment values for these forecasts, the precipitation ensemble is predicting much higher precipitation totals than observed. *This reinforces the importance of looking at other catchment precipitation values other than the mean when relating to extreme flood-producing events.*

Figure 6 considers the same time sequence but for the River Lune at Caton (724629). As for the Eden at Sheepmount, the raingauges detect five distinct episodes, whereas the precipitation ensemble shows only four, the third event occurring between the third and fourth peaks defined by the raingauge time-series. The fifth episode is well captured to begin with but the forecast for this does not remain as good for the Lune catchment throughout the time-window. There are several forecasts in the sequence where the largest (most prolonged) event is not captured as well. The ensemble also continues to struggle with the timing of the third and fourth peaks until both are contained within Day 1 of the forecast. Closer to the time, the catchment-means for the ensemble members are on the low side compared to those from the raingauges.

For the fifth peak, the raingauge-derived catchment-means seem to be closer to the upper boundary of the ensemble spread whereas those from radar are somewhat lower (not shown). In terms of the 99th percentile in-catchment values, the profile of rain throughout the event is somewhat different between the raingauge and radar, with the radar suggesting that the event lasts a little longer (also not shown). Consistent with the observed precipitation values being in the upper boundary of the ensemble spread at Caton, the river flow observations generally fall around the upper boundary of the river flow ensemble members. For some forecast initiations (e.g. 13:00 1 December) the precipitation ensemble values are much lower than observed, and the river flow ensemble drastically underestimates the peak: the ensemble performance is varying between initiations. This may be related to the weather model's ability to capture the local structure of the precipitation, for example due to orographic enhancement of precipitation seen for this event in Figure 2. This will be further investigated through the Phase 2 case studies.

The precipitation catchment-means and 99th percentile in-catchment values for the Lune at Caton are compared in Figure 7. For this particular forecast, where the onset of the main precipitation event was a little slow, the offset looks somewhat larger when compared to the radar-derived catchment-means. Interestingly when the 99th percentile in-catchment values are used, the timing offset looks less pronounced against both radar and raingauges totals. Overall, the radar and raingauge catchment-means are similar in magnitude but the 99th percentiles show greater differences, with the radar (in this instance) providing some larger totals compared to the raingauge values.

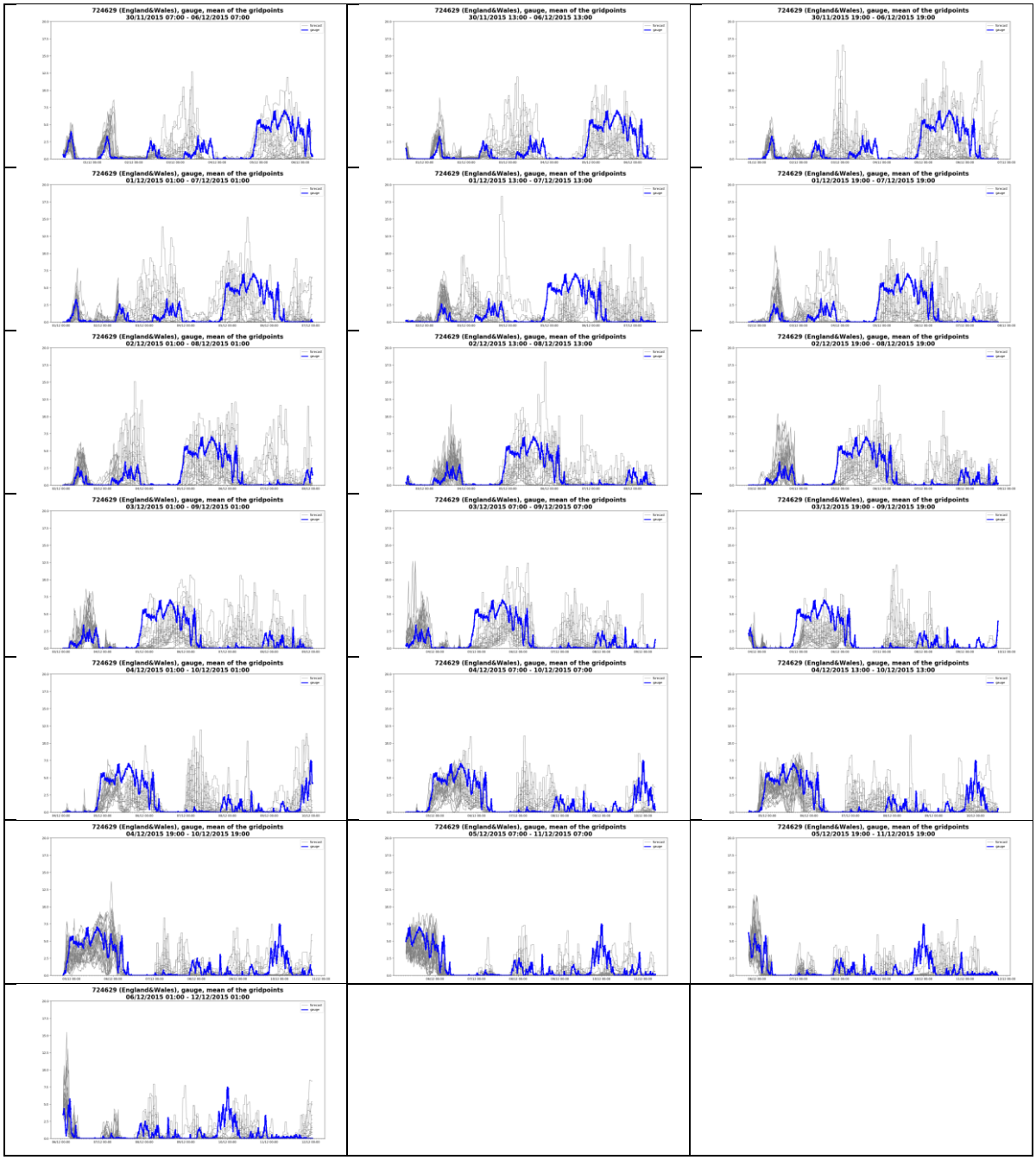


Figure 6 Time-series of catchment-mean precipitation for the ensemble members (grey) are compared to those from raingauges (blue). River Lune at Caton during Storm Desmond.

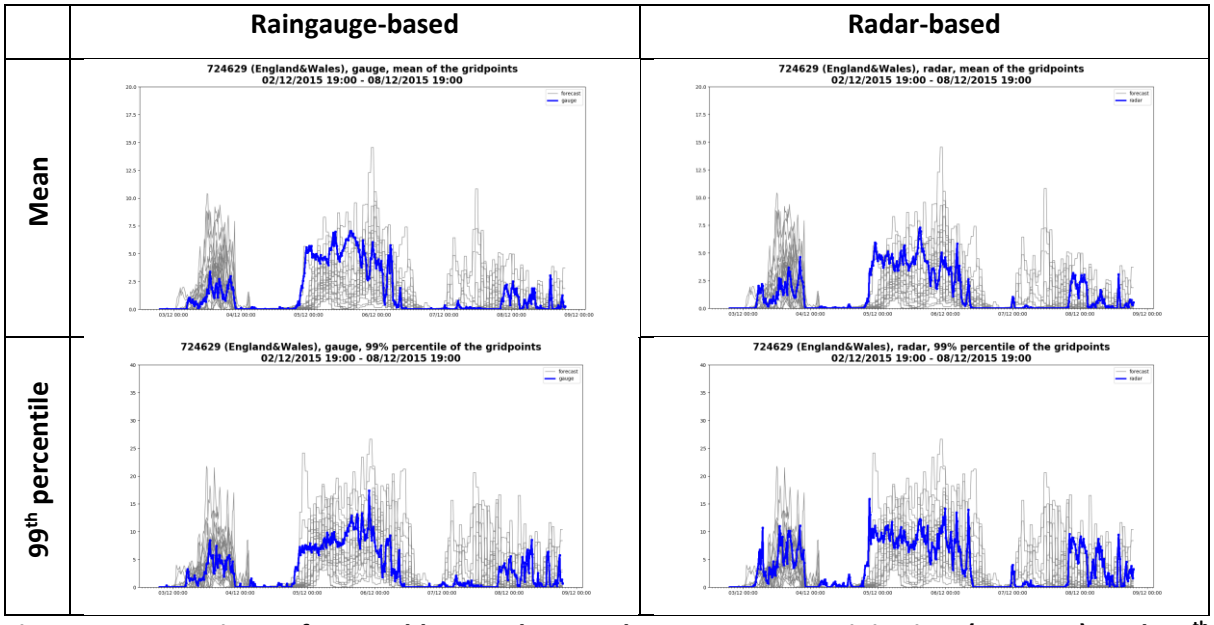


Figure 7 Comparison of ensemble member catchment-mean precipitation (top row) and 99th percentile in-catchment values (bottom row) for the raingauge-based (left) and radar-based (right) catchment means, for Storm Desmond. Note that the y-axis maximum for the bottom row is different to the top row. Catchment is the Lune at Caton (note different time to Figure 2 for the River Eden at Sheepmount).

Figure 8 shows another precipitation example, for the river gauging station on the River Dee at Mar Lodge (ID 234274, area 289 km^2) in the Grampians, Scotland. For this catchment it is harder to identify distinct precipitation episodes with catchment-means generally quite low. Broadly speaking there are three events: the first episode has a few large totals embedded, the second episode is very marginal, and the final prolonged event consists of what one could describe as steady rain. As for the other examples in this period, the ensembles produce a large episode of rain between the first and second observed events, whereas the first event in the weather prediction model is too short and does not reflect the observed intense bursts. The forecast for the first event improves with subsequent forecasts, both in terms of duration and intensity but again, as for the other locations, the ensemble continues to struggle with the period between the first and third events. The ensemble also struggles with the end of the third event, even at short forecast ranges, producing much more rain than was observed.

For this period the catchment-means from radar (not shown) produce much more precipitation overall and this signal is repeated and accentuated when considering the 99th percentile values. As discussed above, the catchment-mean precipitation values from the ensemble and raingauges are low, not suggesting extreme events, but the river flow peaks are high with the observations exceeding the Q(5) threshold. Hydrographs for this event at Mar Lodge are presented in Figure 9. Interestingly, the river flow ensemble members are generally much lower than the observed river flows, particularly for forecasts initialised before 07:00 3 December. It is possible that the peaks in precipitation are falling between the locations of raingauges, and thus not being captured by the raingauge gridded-rainfall. Another possible contributing factor is snowmelt. Figure 10 shows examples of snow measurements from one of the CEH COSMOS-UK sites at Glensaugh (365870 East, 780483 North) from ongoing work at CEH (Wallbank et al., in prep.). Although not situated in the immediate vicinity of the Dee at Park catchment, the Glensaugh site is at an altitude of around 400m and gives an indication as to the broader snow conditions across Scotland at this time. The snow is measured using an above-ground Cosmic Ray Neutron Sensor (blue), a SnowFox sensor (orange) and as modelled from the PACK

snowmelt model (with parameters as in the operational G2G Snowmelt module over Scotland) with observed precipitation at the site as input (grey). Interestingly, Figure 10 shows that the PACK model is significantly underestimating the amount of snow in this instance, suggesting a possible underestimation of the snowmelt contributing to the G2G ensemble output.

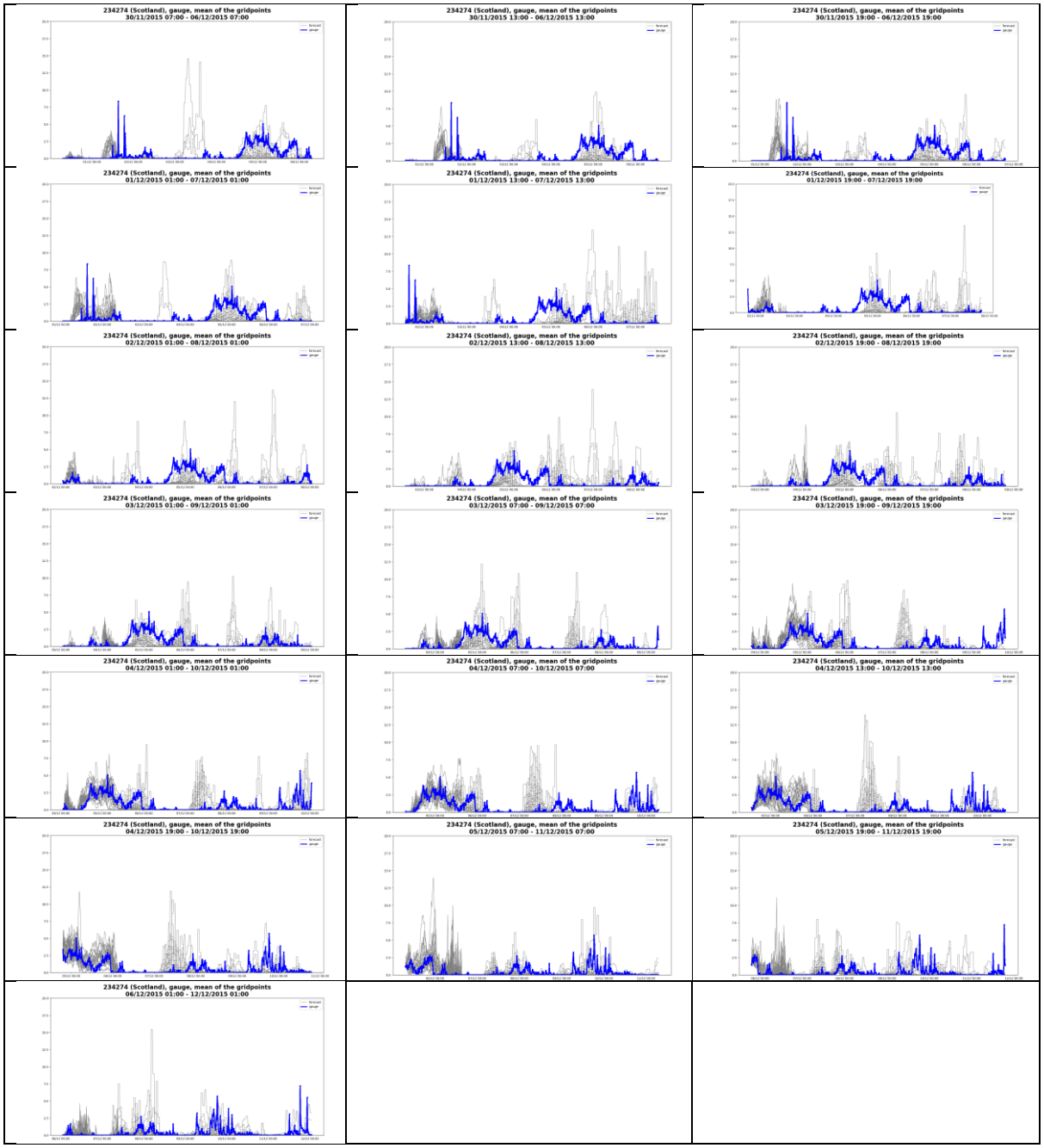


Figure 8 Time-series of catchment-mean precipitation for the ensemble members (grey) are compared to those from raingauges (blue). River Dee (Grampian) at Mar Lodge for Storm Desmond.

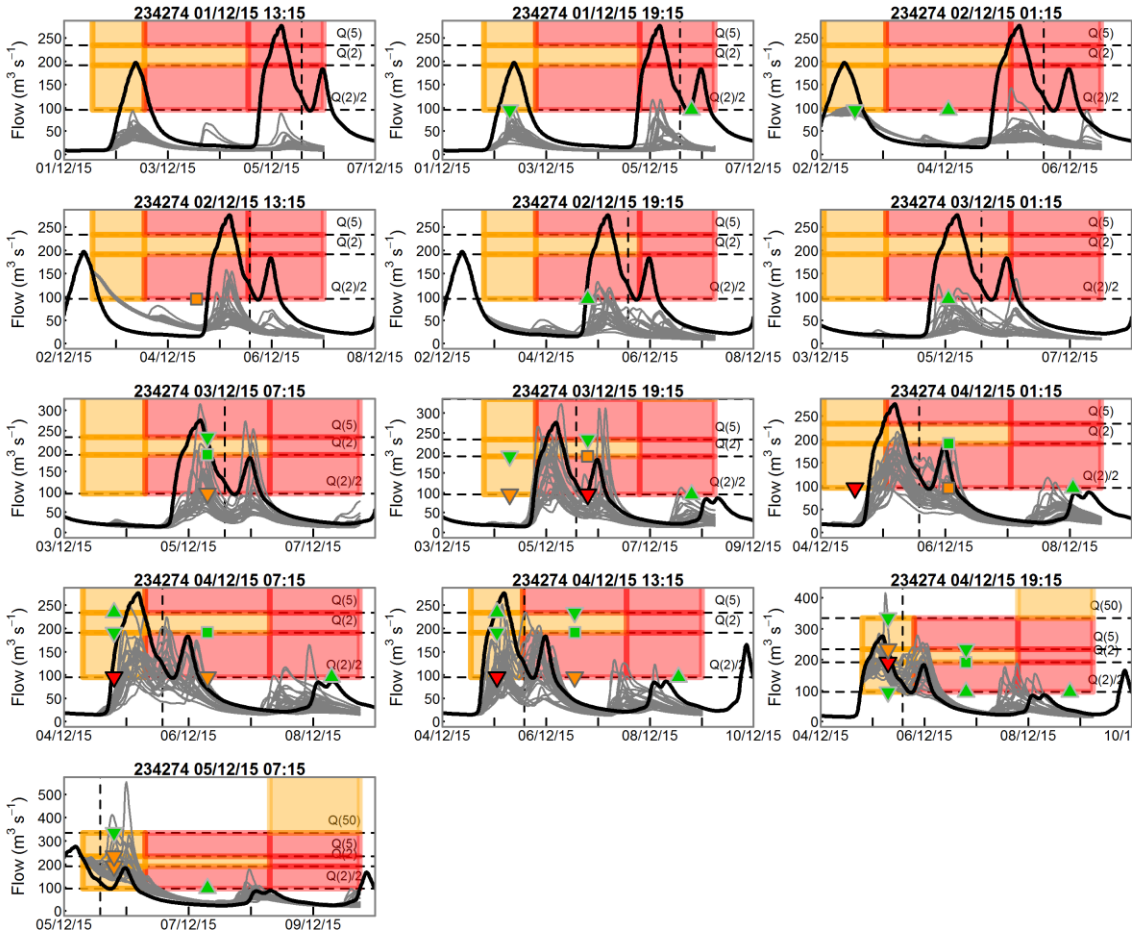


Figure 9 Ensemble river flow forecast hydrographs for the Dee at Mar Lodge covering the time 14:00 5 December 2015 during Storm Desmond. Ensemble members are shown in grey and the observed river flow in black. Symbol colours and shapes are defined in Section 7.2 of the Phase 1 Report.

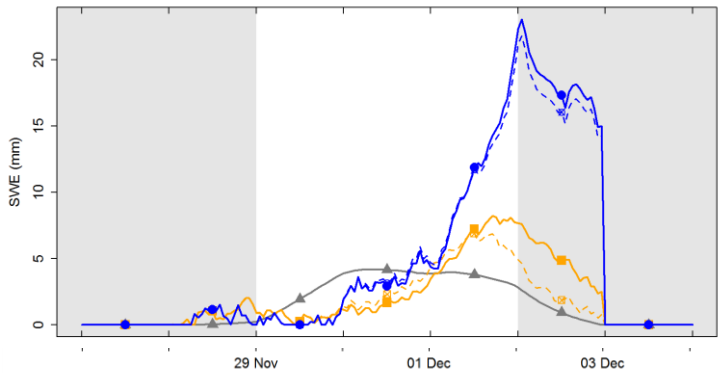


Figure 10 Snow cover measurements from the CEH COSMOS-UK Glensaugh site (NGR 365870 780483) during Storm Desmond. The snow is measured using an above ground Cosmic Ray Neutron Sensor (blue), a SnowFox sensor (orange) and as modelled from the PACK snowmelt model (with parameters as for the operational G2G Snowmelt module over Scotland) with observed precipitation at the site as input (grey). (Source: Wallbank et al. (in preparation)).

In Figure 11, the catchment-mean precipitation and 99th percentile are compared between the ensemble, radar and raingauge for a single forecast in the sequence. From Figure 8 it is clear that the forecasts really struggle for this catchment. This example shows a lot of deviation between the observed and ensemble traces. Events are offset from each other (first observed event), too short (second observed event), or just false alarms (between the second and third episodes). In this instance, the catchment-mean precipitation from radar signals a larger second episode compared to the raingauges and is further enhanced for the 99th percentile. In this instance, the radar totals are at times higher than those from the ensemble, especially for the second precipitation episode.

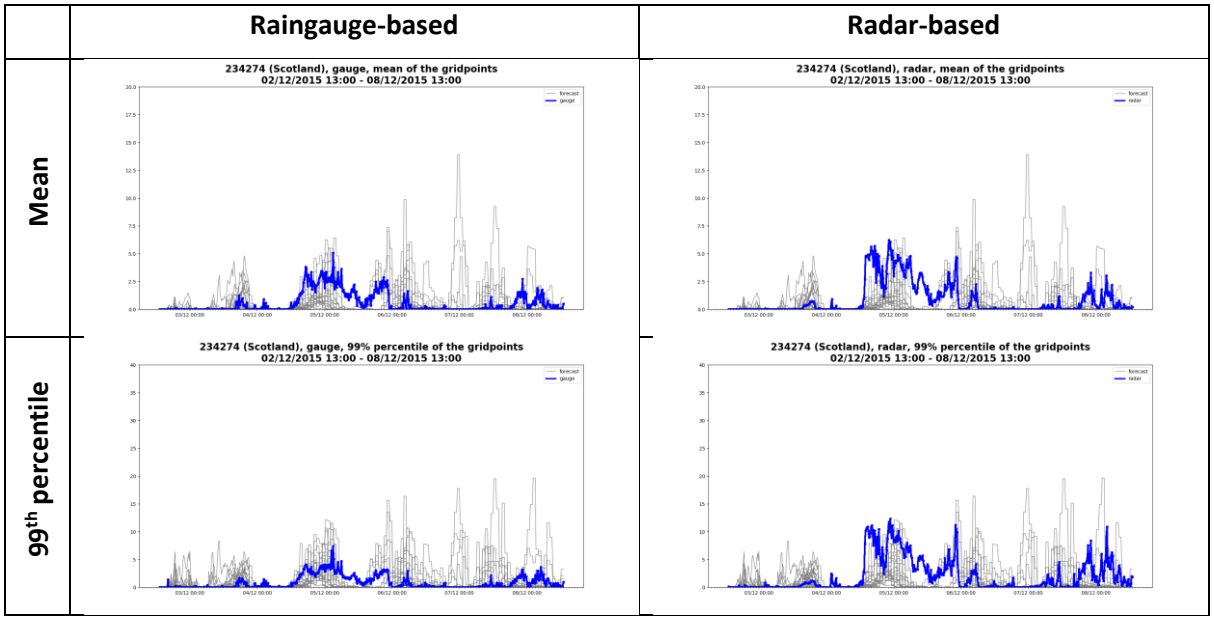


Figure 11 Comparison of ensemble member catchment-mean precipitation (top row) and 99th percentile in-catchment values (bottom row) for the raingauge-based (left) and radar-based (right) catchment means, for Storm Desmond. Note that the y-axis maximum for the bottom row is different to the top row. Catchment is the River Dee (Grampian) at Mar Lodge (note different time).

Figure 12 shows the Storm Desmond sequence of precipitation ensemble forecasts for the River Dee at Park (ID 234291, area 1834 km^2). Again, the precipitation traces are muted. Broadly speaking there are again three episodes (ignoring the initial blip). For the first two precipitation episodes the forecast is too fast. The first event in the initial forecasts (even though they are for relatively short lead-times) is not prolonged enough. The second event is too intense and too early. The third event is captured relatively well in terms of timing in the early forecasts (at the longest lead-times). The forecasts for this event are quite variable as the event approaches and the period beyond the third peak is quite irregular with continued mismatches between the intensity of the rain as well as the timing. Short-lead-time forecasts for the third episode are good with the ensemble catchment-mean totals encompassing the raingauge totals.

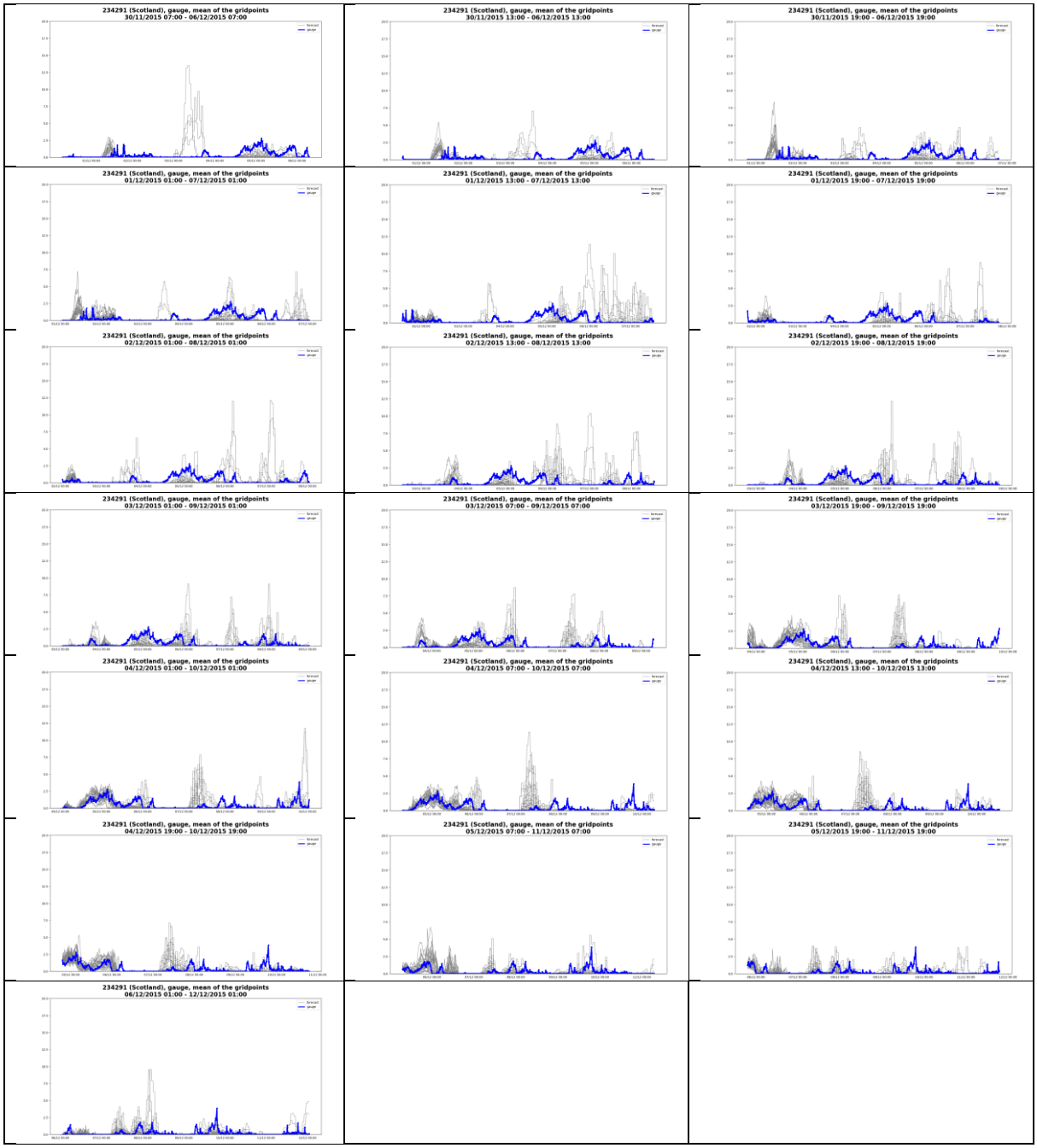


Figure 12 Time-series of catchment-mean precipitation for the ensemble members (grey) are compared to those from raingauges (blue). River Dee at Park during Storm Desmond.

The catchment-mean precipitation from radar tends to be somewhat higher than for raingauges (not shown). This signal is replicated when the 99th percentile in-catchment values are considered. The river flow hydrographs for the Dee at Park lead to similar conclusions as those discussed above for Mar Lodge and are not included here.

Under Phase 2, novel ways of displaying the rainfall and river flow ensemble information are to be explored, along with how the verification information can be incorporated. From the Phase 1 case studies and the work already carried out under Phase 2 (exploring the 99th percentile in-catchment values), it would already seem clear that the same form of rainfall ensemble envelope which includes the mean, median and 99th percentile would be very useful. These ideas will be taken forward in WP 4 and 5.

Comparing the catchment-mean precipitation for the ensemble, radar and raingauges and the in-catchment 99th percentile values, Figure 13 shows that the raingauge values are well matched for the Day 1 period with the radar values slightly larger and closer to the upper boundary of the ensemble envelope. The 99th percentile raingauge values are at the lower end of the ensemble envelope whereas the corresponding radar values are larger: at times larger than the largest ensemble values. Whilst there is arguably an offset in the second episode (based on this forecast time) the weather prediction model rainfall totals are closer to the radar-rainfall accumulations.

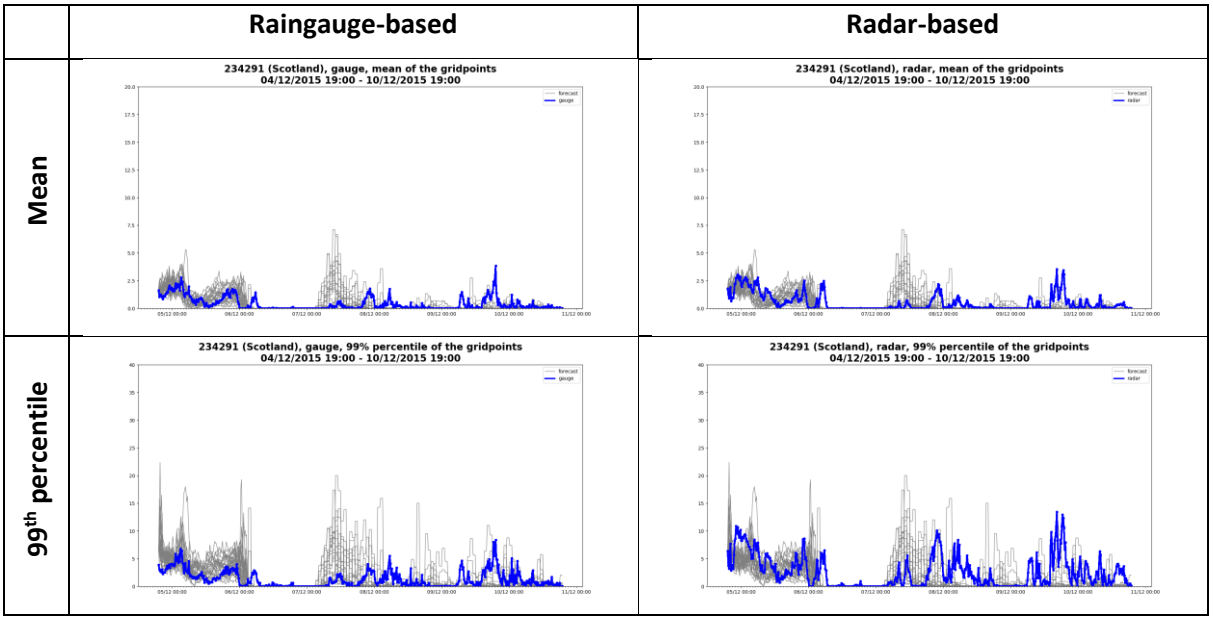


Figure 13 Comparison of ensemble member catchment-mean precipitation (top row) and 99th percentile in-catchment values (bottom row) for the raingauge-based (left) and radar-based (right) catchment means, for Storm Desmond. Note that the y-axis maximum for the bottom row is different to the top row. Catchment is the River Dee at Park and forecast start time is 19:00 4 December 2015.

For the Dee at Park catchment, it is worth showing a second example because it is rather peculiar. Figure 14 shows another forecast. In this instance there is a very large difference between the catchment-mean and in-catchment 99th percentile values, especially for raingauges, which show several large peaks that are not replicated in the 99th percentile radar-rainfall series. In this instance the radar and ensemble values are comparable. Furthermore, the raingauge series shows one longer precipitation episode whereas the radar suggests two distinct events, the second not that well forecast. By contrast, for the prolonged third event, the 99th percentile values from the ensemble compare favourably to those from raingauges whereas radar values are larger. As discussed above, it is possible that snowfall is affecting the measurement of precipitation in this case, with suggested snowfall around the start of December (Figure 10), coinciding with the timing of measured spurious peaks.

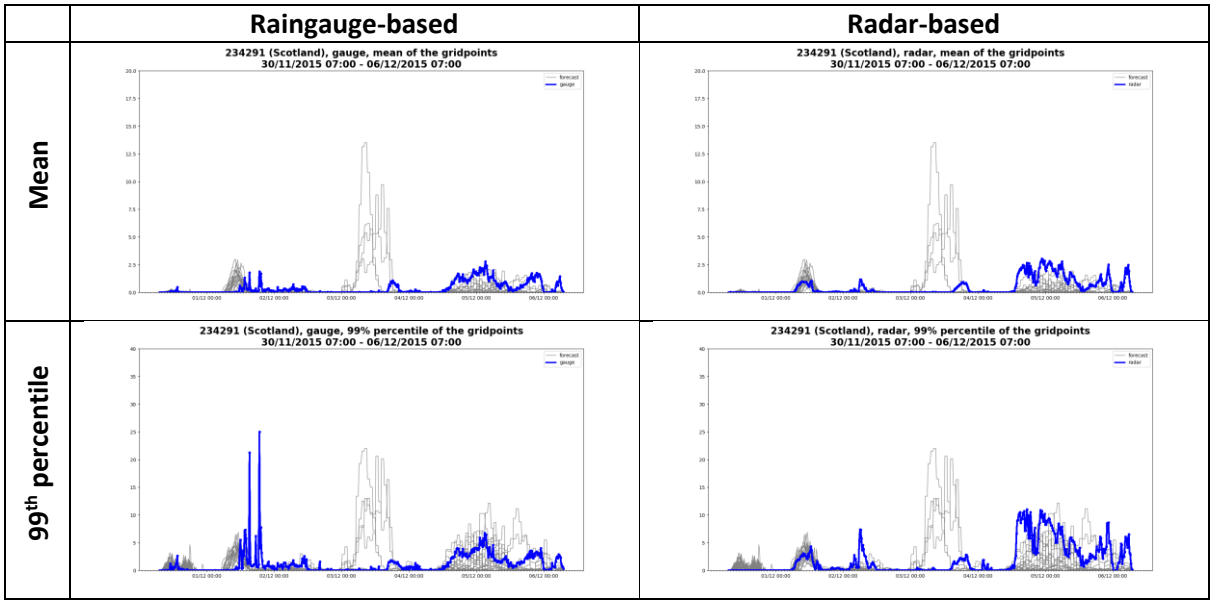


Figure 14 Comparison of catchment-mean precipitation for ensemble members (top row) and 99th percentile in-catchment values (bottom row) for the raingauge-based (left) and radar-based (right) catchment means, for Storm Desmond. Note that the y-axis maximum for the bottom row is different to the top row. Catchment is the River Dee at Park and forecast start time is 07:00 10 November 2015.

Another river gauging location on the Dee is at Polhollick (ID 234294, area 691 km^2). The forecast sequence shown in Figure 15 shows a very similar evolution. The forecast is too early in onset for the first precipitation episode. It has a poor grasp of the second event, in terms of intensity and timing. The third episode is better forecast, but again there is considerable variability in the forecast quality as the lead-time shortens. In this instance some of the middle forecasts are exceedingly poor (e.g. 13:00 1 December 2015). Beyond the third event the forecast is also very irregular with a lot of rain being forecast which does not materialise.

Over the entire sequence of forecasts the catchment-means for radar rainfall and the 99th percentile in-catchment values match better for the events beyond the main (third) peak but the forecasts are not particularly good overall (not shown). The river flow hydrographs for the Dee at Polhollick lead to similar conclusions as those discussed above for Mar Lodge and are not included here. The hydrographs for the Dee at Polhollick were included in the Phase 1 Report as Figure 7.11.

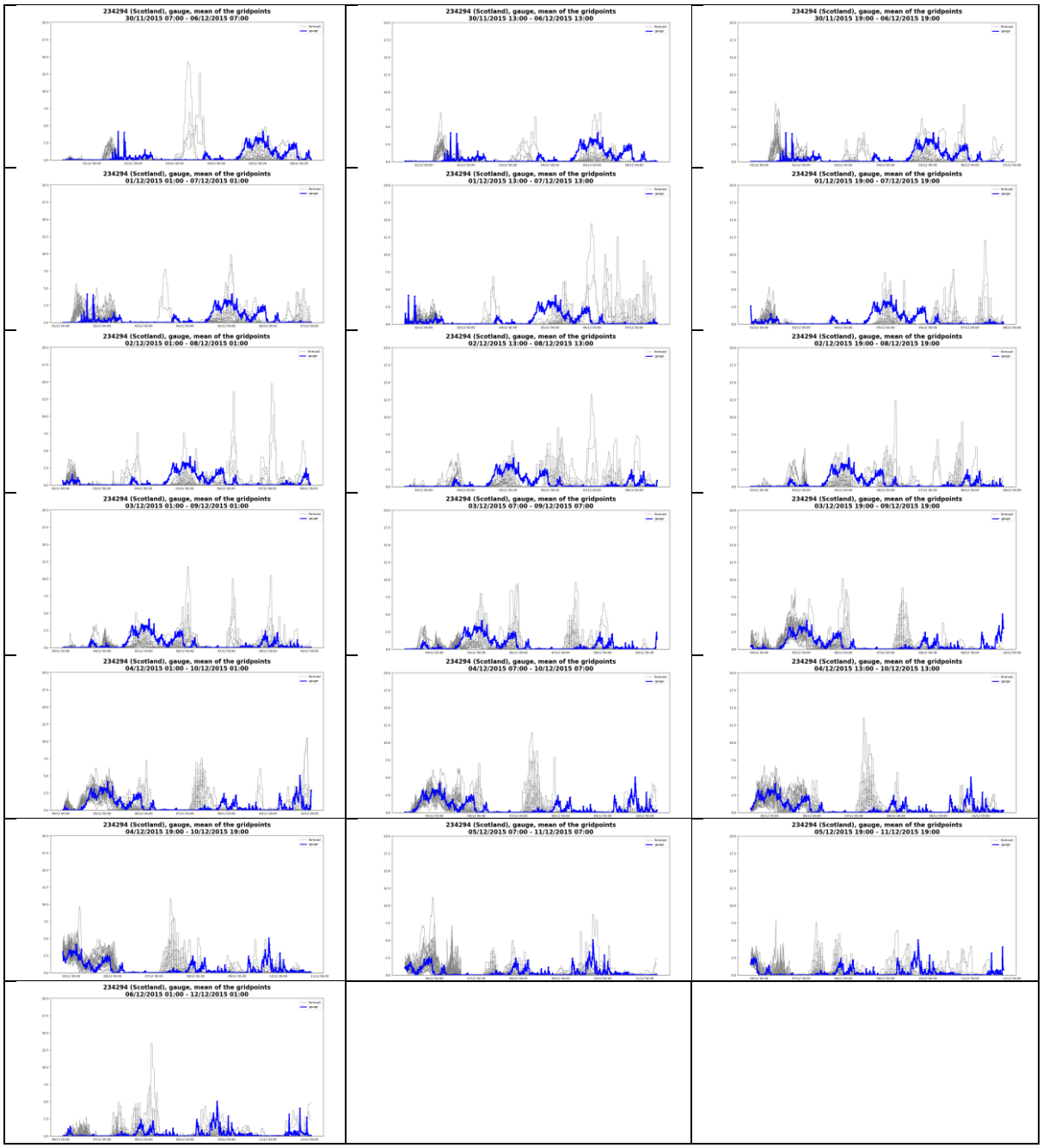


Figure 15 Time-series of catchment-mean precipitation for the ensemble members (grey) are compared to those from raingauges (blue). River Dee at Polhollick during Storm Desmond.

Figure 16 shows the comparison of a particular forecast between the catchment-means and the in-catchment 99th percentiles. In this instance the catchment-means for the raingauges and radar are at the high end of the ensemble range for the main (third) event. Beyond that, the forecast is poor. When considering the 99th percentile in-catchment values, there is poor correspondence between the forecast ensemble members and the observed values. The exception is the event beyond Day 5 for the radar 99th percentiles which shows fairly good correspondence.

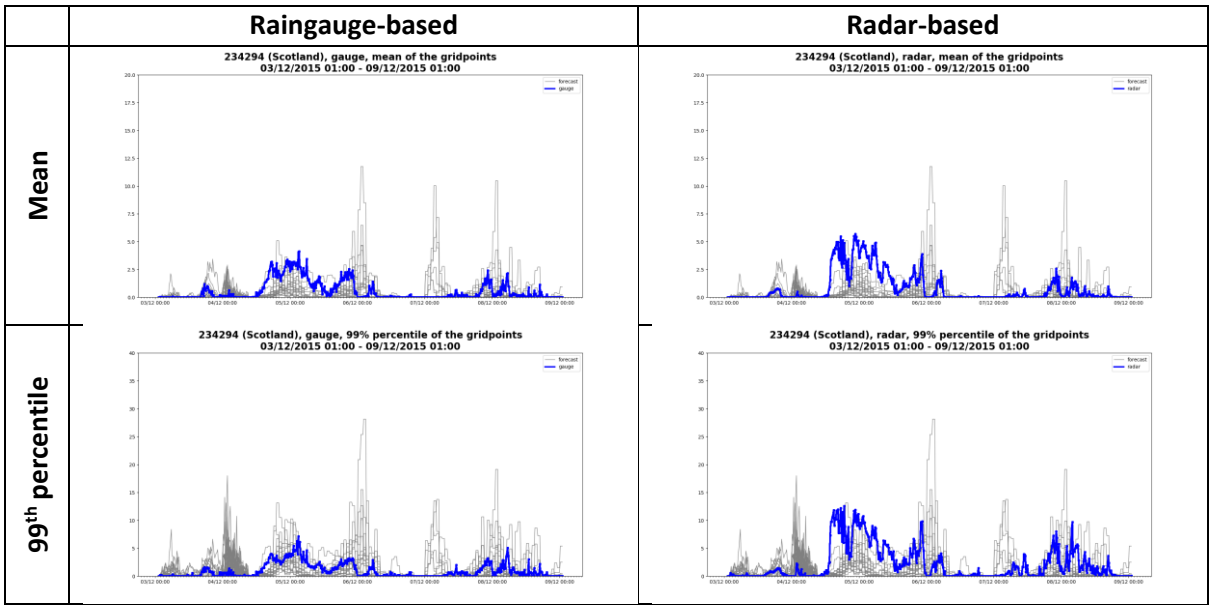


Figure 16 Comparison of catchment-mean precipitation for ensemble members (top row) and 99th percentile in-catchment values (bottom row) for the raingauge-based (left) and radar-based (right) catchment means, for Storm Desmond. Note that the y-axis maximum for the bottom row is different to the top row. Catchment is the River Dee at Polhollick (234294) and the forecast start time is 01:00 3 December 2015.

4. Case Study 2: Storm Eva (post storm)

Figure 17 shows the sequence of daily radar-rainfall accumulations for the period 21-25 December 2015 in the aftermath of Storm Eva.

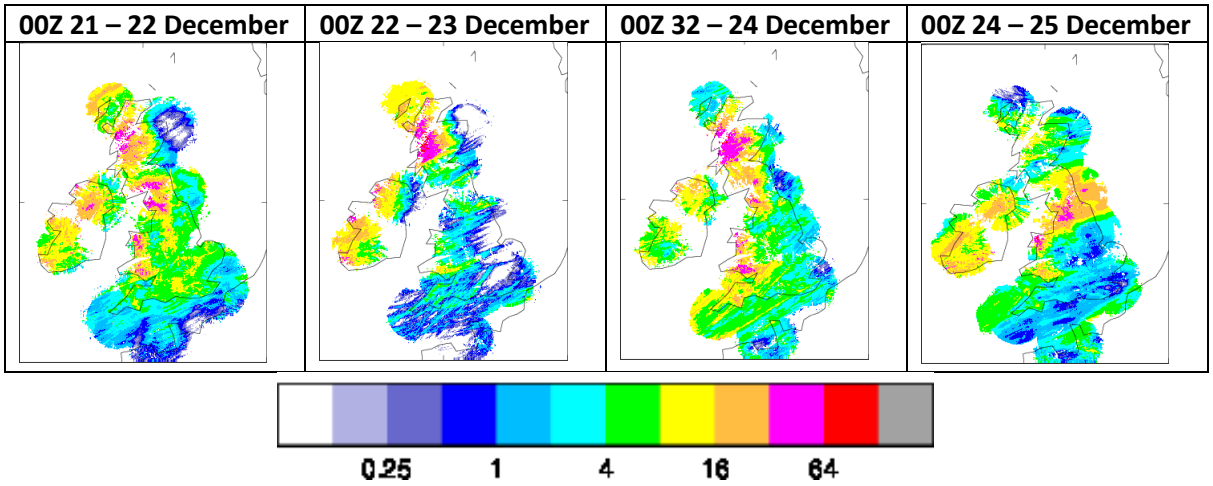


Figure 17 Sequence of daily radar-rainfall accumulations (mm) over the period 21 to 25 December 2015

The forecast sequence for Storm Eva (post storm) is between 19:00 21 December 2015 to 19:00 25 December 2015. Figure 18 shows time-series of catchment-mean precipitation of the ensemble members and raingauges for the Eden at Sheepmount. Again, three episodes or events can be identified from the first forecast panel. Broadly speaking this appears to be a good sequence of forecasts though forecasts for Days 4 to 6 can be variable. The forecast for the third peak is initially quite good, but then gets worse before it improves again. The catchment-mean values for the ensemble become too large before coming into the right range. Comparing the sequence based on the catchment-means and in-catchment 99th percentile values (not shown), there is reasonable correspondence for the radar and ensemble values, with little difference between the catchment-means.

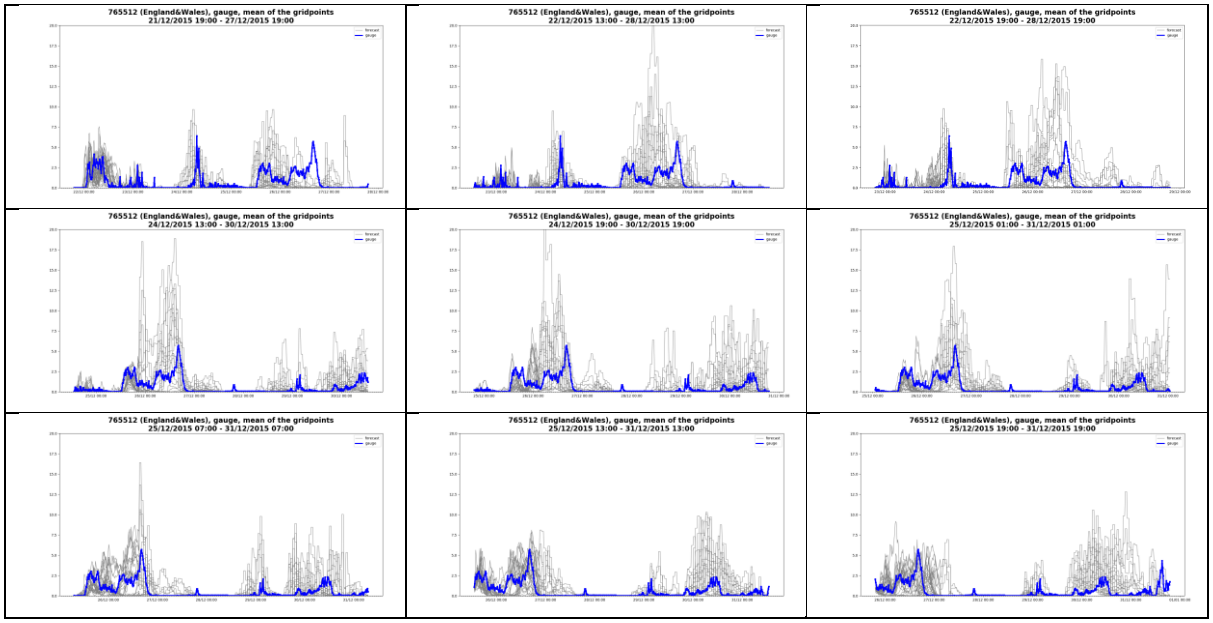


Figure 18 Time-series of catchment-mean precipitation for the ensemble members (grey) are compared to those from raingauges (blue). River Eden at Sheepmount for post Storm Eva period 19:00 21 December to 19:00 25 December 2015.

Figure 19 compares the catchment-mean and 99th percentile in-catchment values for a specific forecast. The second event begins too soon with the catchment-means for raingauges being slightly larger than the corresponding radar-rainfall (which is quite unusual). There is a big mismatch in intensities for the third episode. There are large mismatches in intensity when considering the 99th percentile in-catchment values where the weather prediction model produces some very large values, accentuating the timing issues. In this instance the 99th percentile values for radar appear larger (for the most part).

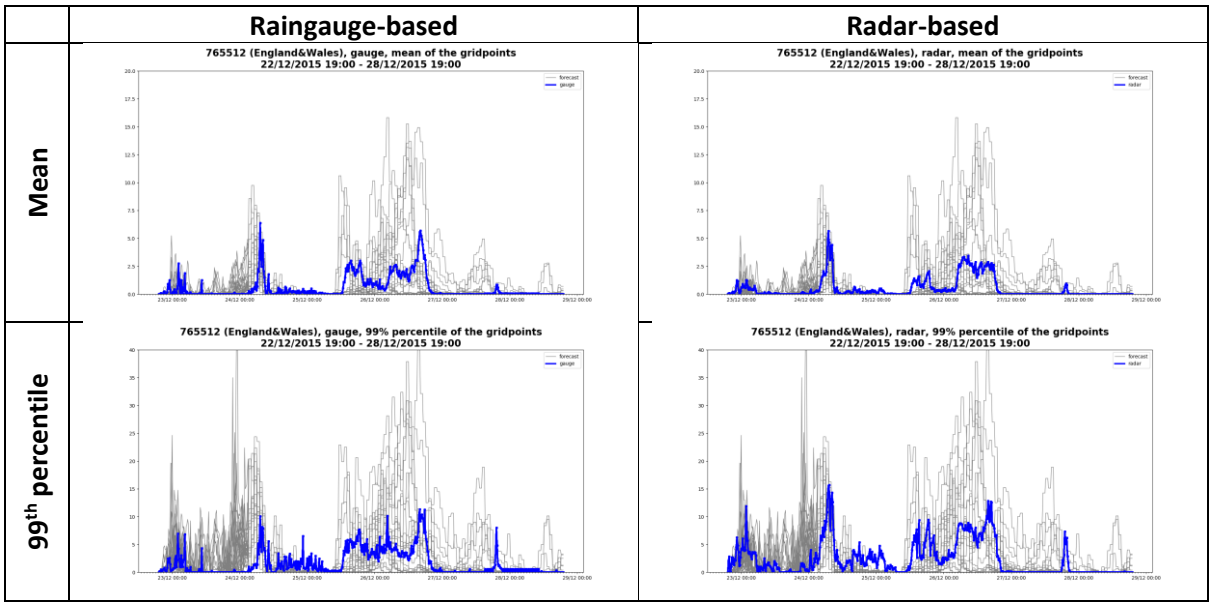


Figure 19 Comparison of catchment-mean precipitation for ensemble members (top row) and 99th percentile in-catchment values (bottom row) for the raingauge-based (left) and radar-based (right) catchment means. Note that the y-axis maximum for the bottom row is different to the top row. Catchment is the River Eden at Sheepmount and forecast start time is 19:00 22 December 2015 (post Storm Eva).

Figure 20 shows the forecast sequence for the River Lune at Caton. In this instance it is harder to define specific events based on the observed trace but, on the whole, the timings for the first three forecasts look relatively good with some intensity differences, especially for the "third" episode around four days into the forecast. These intensity differences persist as the forecast lead-time decreases.

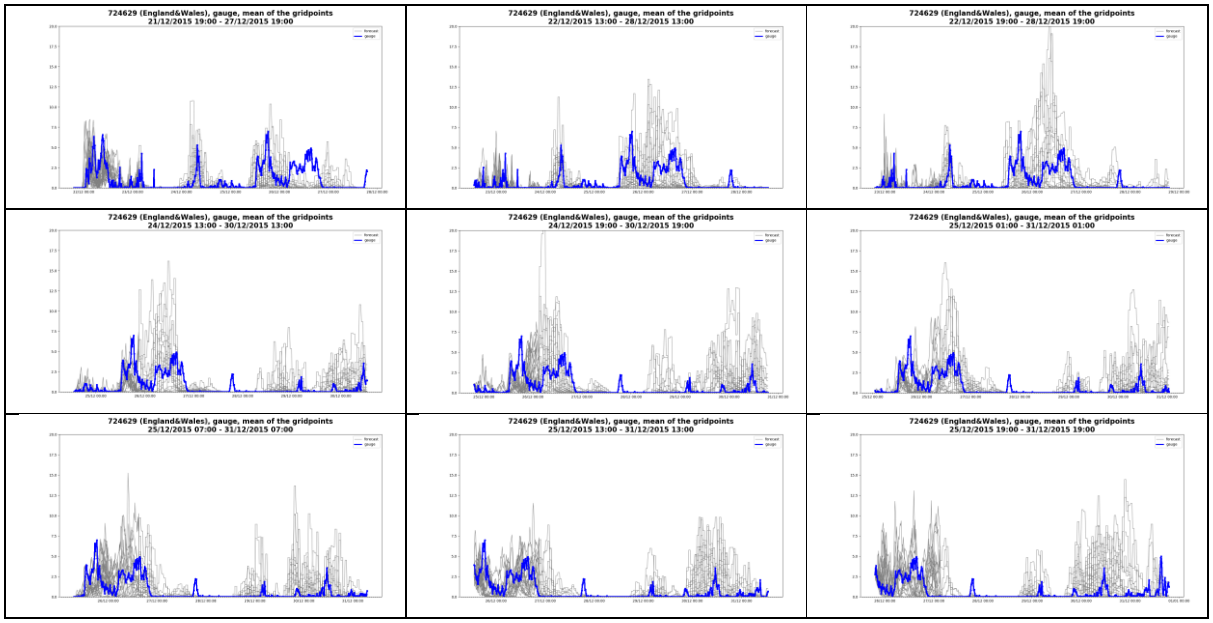


Figure 20 Time-series of catchment-mean precipitation for the ensemble members (grey) are compared to those from raingauges (blue). River Lune at Caton for the period 19:00 21 December to 19:00 25 December 2015 (post Storm Eva).

Catchment-means are comparable between the radar and the raingauges on the whole (not shown), whilst the traces based on the 99th percentile in-catchment values show the radar values are higher on the whole, though the raingauges can produce some interesting spikes.

Overall, for Storm Eva, both the Eden at Sheepmount and Lune at Caton catchments show an overestimation of precipitation by the ensemble, for both the catchment-mean and 99th percentile of in-catchment values. This also shows through in the river flow ensemble response for these catchments (Figures 21 and 22), with the peak around 27 December vastly overestimated with the majority of ensemble members easily crossing the Q(50) threshold, compared to the observed crossing of Q(2). This is particularly true for the earlier forecasts issued prior to 25 December (top two rows of Figures 21 and 22 for river flow and 18 and 20 for precipitation), but the overestimation continues in the run-up to the event. This again shows, as for Storm Desmond, how river flow ensemble performance can be directly linked to that of precipitation. This will be investigated further with the help of the Phase 2 case studies, including the possibility of directly mapping individual ensemble members.

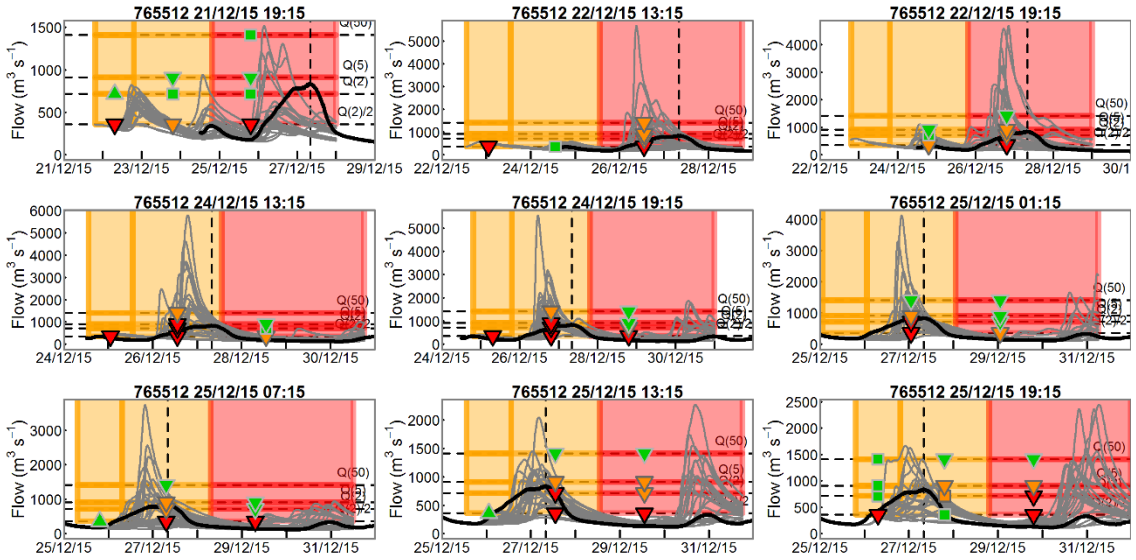


Figure 21 Ensemble river flow forecast hydrographs for the Eden at Sheepmount covering the time 09:00 27 December 2015 (post Storm Eva). Ensemble members are shown in grey and the observed river flow in black. The vertical dashed black line shows the time of peak flow at Sheepmount (08:00 27 December 2015). Symbol colours and shapes are defined in Section 7.2 of the Phase 1 Report. (This is Figure 7.14 of the Phase 1 Report.)

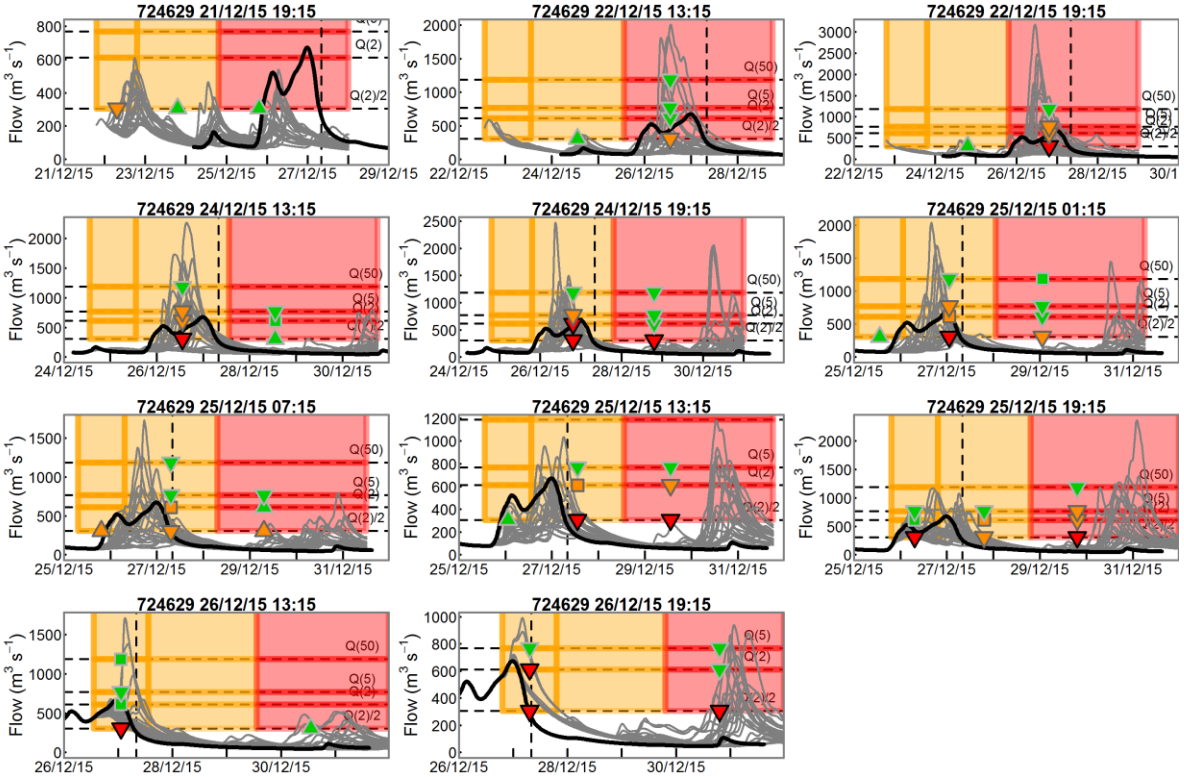


Figure 22 Ensemble river flow forecast hydrographs for the Lune at Caton covering the time 09:00 27 December 2015 (post Storm Eva). Ensemble members are shown in grey and the observed river flow in black. The vertical dashed black line shows the time of peak flow at Sheepmount (08:00 27 December 2015). Symbol colours and shapes are defined in Section 7.2 of the Phase 1 Report. (This is Figure 7.15 of the Phase 1 Report.)

5. Case Study 3: Storm Frank

Figure 23 shows the sequence of daily radar-rainfall accumulations for the period 26-27 December 2015 associated with Storm Frank.

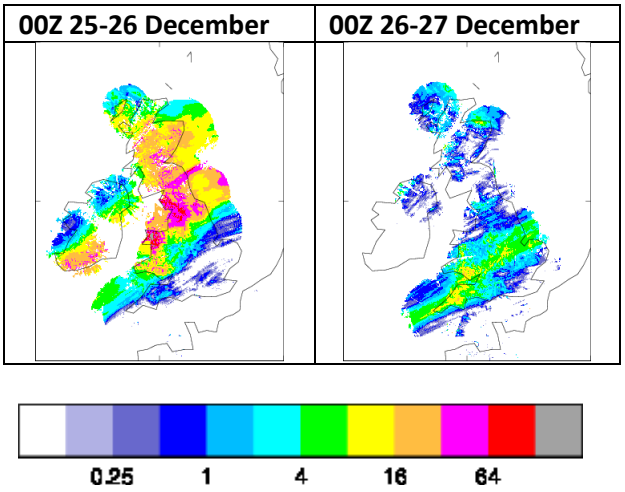


Figure 23 Sequence of daily radar-rainfall accumulations (mm) over the period 26 to 27 December 2015

Storm Frank is right at the end of the December 2015 time-series available from Phase 1. As a result, there are only two forecasts to evaluate: 13:00 and 19:00 26 December 2015. Because of their short duration, Figure 20 shows the time-series for all observation options to be considered for precipitation verification: catchment-means and 99th percentile in-catchment values for raingauges and radar.

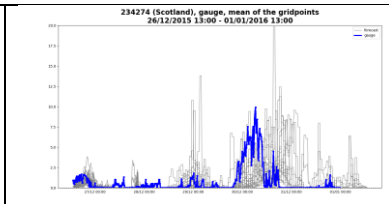
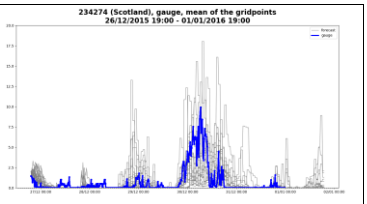
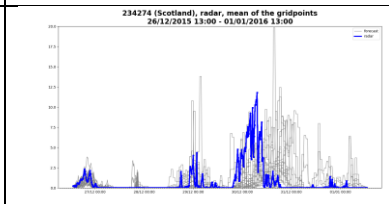
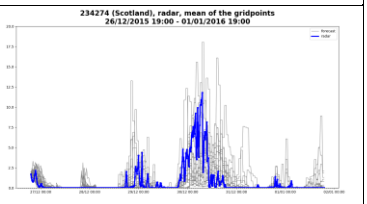
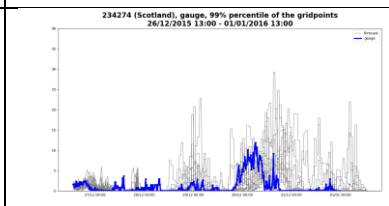
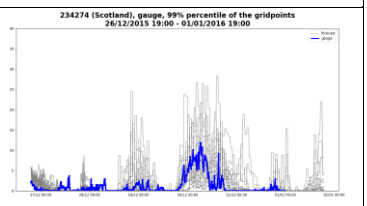
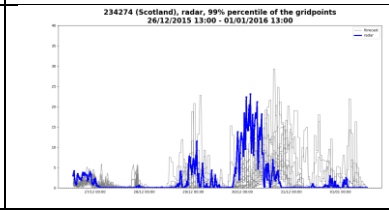
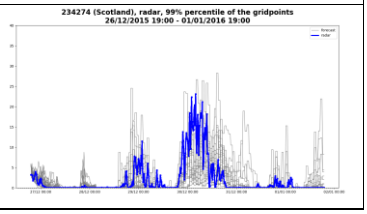
Catchment-mean, raingauge		
Catchment-mean, radar		
99th percentile, raingauge		
99th percentile, radar		

Figure 24 Time-series of catchment-mean precipitation and 99th percentile in-catchment values for the ensemble members (grey) are compared to those from raingauges and radar (blue). River Dee (Grampian) at Mar Lodge during Storm Frank.

Considering first the Dee at Mar Lodge catchment shown in Figure 24, there is one main precipitation event in the sequence which follows on from three smaller events. For the main rain event, the precipitation persists for longer in the ensemble members than in the observations (either raingauge or radar). Catchment-means from radar are somewhat higher overall than those from raingauges for this catchment. This is even clearer when looking at the in-catchment 99th percentile values, where the ensemble values appear to be closer in magnitude to the higher observations of the radar.

Figure 25 shows the same display for the River Dee at Park. For this catchment the forecasts are relatively good into Day 3, although the biggest precipitation peak in the period is more extended than it was in reality. Catchment-means appear to be in relatively good agreement. The 99th percentile in-catchment values show a lot more noise for the forecasts, with the radar values in this instance again higher than the raingauge values. In terms of the means the raingauge and radar values are broadly similar.

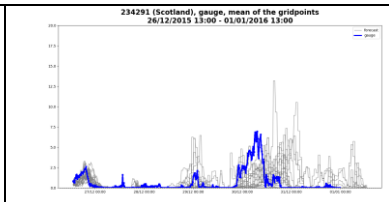
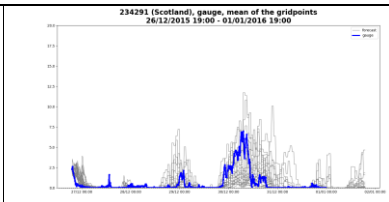
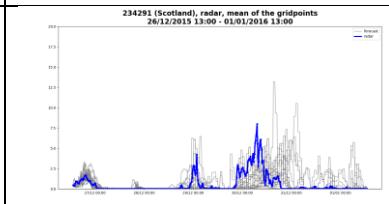
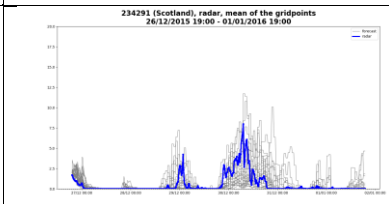
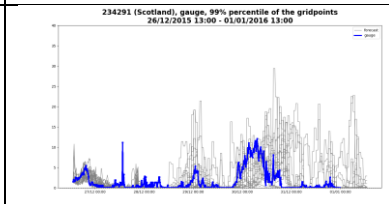
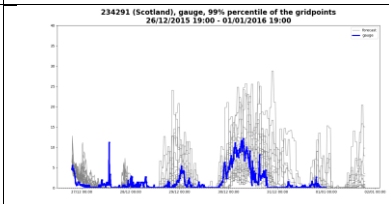
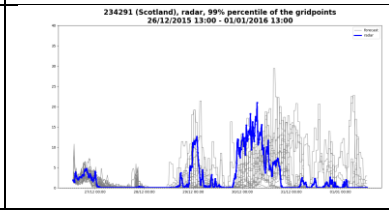
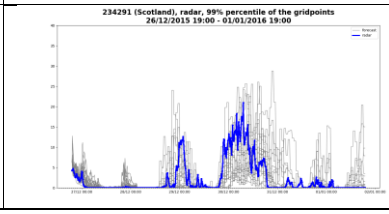
Catchment-mean, raingauge		
Catchment-mean, radar		
99th percentile, raingauge		
99th percentile, radar		

Figure 25 Time-series of catchment-mean precipitation and 99th percentile in-catchment values for the ensemble members (grey) are compared to those from raingauges and radar (blue). River Dee at Park during Storm Frank.

Finally, the Dee at Polhollick is considered in Figure 26. Similar to the Dee at Park, the main precipitation peak in the forecast is too broad, with the event extended beyond its observed end. For this catchment the catchment-means from the ensemble can be rather large. In terms of the 99th percentile in-catchment values the differences between the radar and raingauge values are again evident. Again, the raingauge values are prone to producing some spikes (possibly due to the presence of snow). The weather prediction model values are comparable (at best) to the radar values, but generally too large.

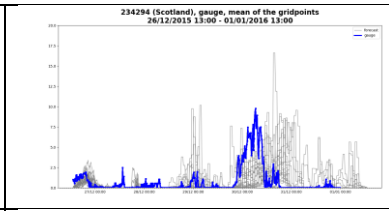
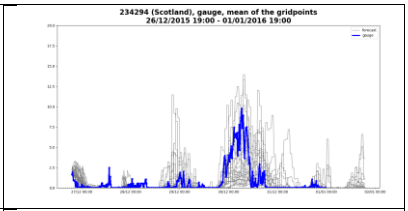
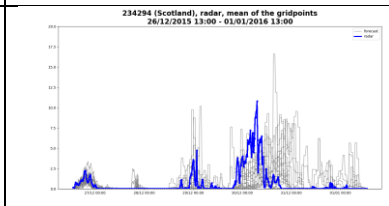
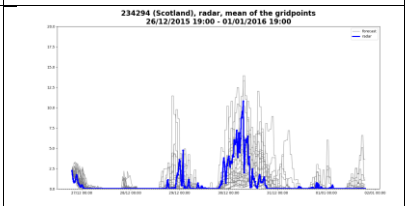
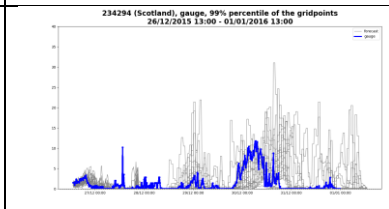
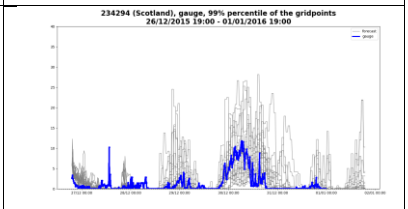
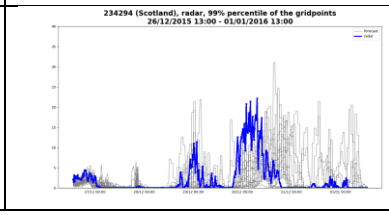
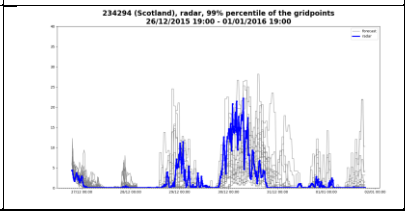
Catchment-mean, raingauge		
Catchment-mean, radar		
99th percentile, gauge		
99th percentile, radar		

Figure 26 Time-series of catchment-mean precipitation and 99th percentile in-catchment values for the ensemble members (grey) are compared to those from raingauges and radar (blue). River Dee at Polhollick during Storm Frank.

For the Storm Frank case study, river flow hydrographs for the two available forecasts for the Dee at Park, Polhollick, and Mar Lodge are shown in Figure 27. For all three catchments the observed river flow peak is large (much higher than the Q(50) threshold). As expected from the precipitation ensemble forecasts where the main precipitation is forecast late, there is a tendency for the main peak to be slightly delayed in the river flow response. Overall, the magnitude of the river flow peak seems reasonably well captured, again agreeing with the precipitation ensemble. Although it is difficult to assess fully, due to being at the end of the period of available data, there is also some indication that the extended precipitation in the ensemble forecasts (beyond its observed end) is resulting in a longer-duration river flow peak.

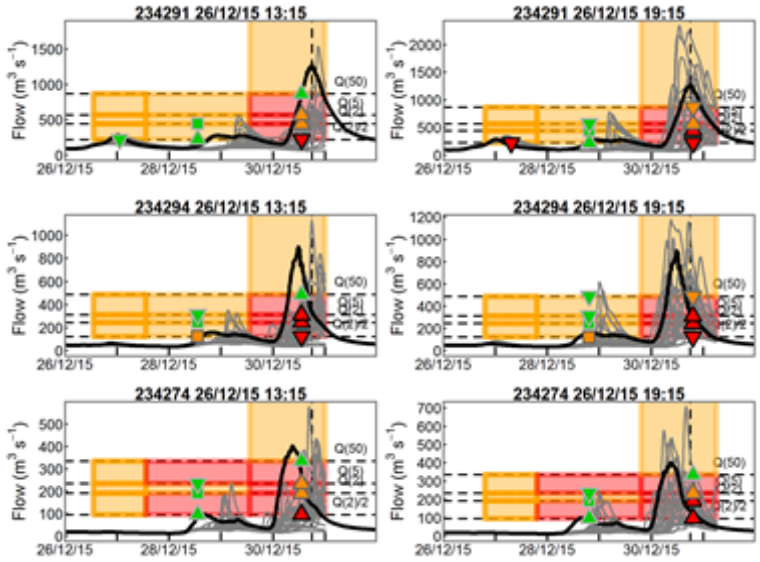


Figure 27 Ensemble river flow forecast hydrographs for the Dee at Park (top), Polhollick (middle) and Mar Lodge (bottom) covering the time 14:00 30 December 2015 (Storm Frank). Ensemble members are shown in grey and the observed river flow in black. The vertical dashed black line shows the time of peak flow at Park (14:00 30 December 2015). Colour-scale and symbols are as defined in sections 7.2.1 and 6.2.1 of the Phase 1 Report. (Source: Figure 7.17 of Phase 1 Report).

6. Summary

Considering the forecast evolutions shown here for storms Desmond, Eva and Frank it is clear that the different weather prediction model configurations have a part to play in terms of introducing biases. The high intensity bias from MOGREPS-G is often clear for the longer lead-times.

Stitching together the forecasts is also somewhat problematic, because there may be mismatches in the way the event is forecast between MOGREPS-UK and MOGREPS-G. Therefore, it is unlikely that a cleverer merging of forecasts in this time-range will cure the timing issues that are sometimes observed in the case studies over the Days 2-4 range. This can at least partially explain why the subjective forecast performance takes a dip in this range: that is, whilst there is a clearer bias in intensities beyond Day 4, the timing of events appears to be not too bad. Some effort should be invested in post-processing the ensembles, such as more sophisticated blending and merging and bias correction.

Overall, a high correspondence was generally found between the performance of the precipitation and river flow ensembles. In particular, when the precipitation ensemble miss-timed a peak, or significantly overestimated the magnitude of precipitation, this also showed through in the river flow response. This will be investigated further in the Phase 2 case studies, including the direct mapping of individual ensemble member behaviour. For the Scotland case studies, the influence of snow was identified as a possible reason for differences between the precipitation ensemble performance and river flow response. This was highlighted for catchments of the River Dee (Aberdeenshire) for the Desmond case study, where the river flow ensemble vastly under-predicted the river flow peaks, particularly for longer lead-time forecasts, and unrealistic spikes were observed in the raingauge data. As snow can affect the performance of both ensembles in a variety of ways and affect the quality of

precipitation observations available for verification, it is important that the presence of snow is kept in mind when interpreting verification results obtained for winter periods in Phase 2.

The precipitation observation types are also showing some interesting variations in effect on both precipitation and river flow verification performance. In Phase 2 the merged precipitation product will also be considered in relation to precipitation verification. It will interesting to see whether this product has added-value for precipitation and river flow forecasting.

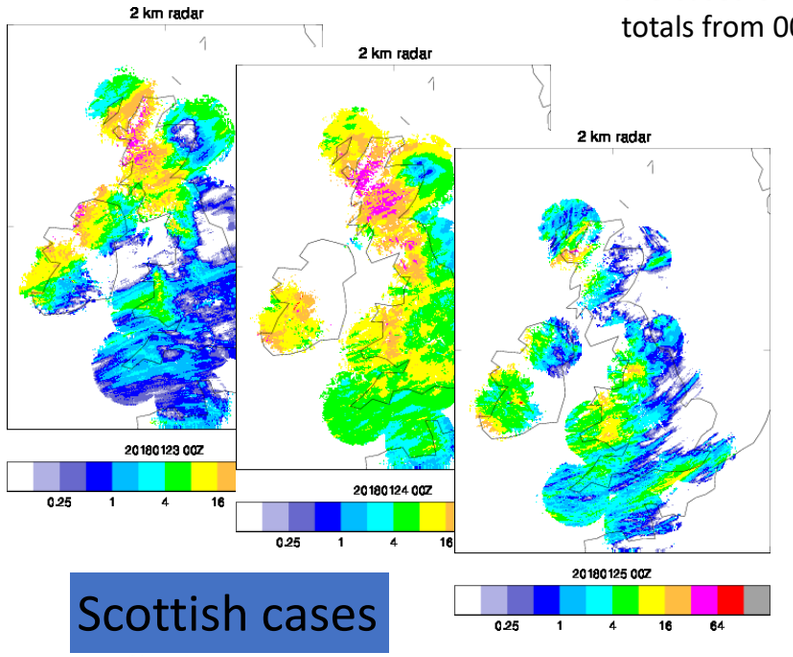
The outcomes from this case study analysis provide some useful pointers for Phase 2 work concerned with "Real-time displays". For example, it would seem sensible to provide a precipitation ensemble "envelope" defined by the catchment-mean (as the lower boundary) and the 99th percentile in-catchment value (as the upper boundary) for visualising the precipitation forecast uncertainty.

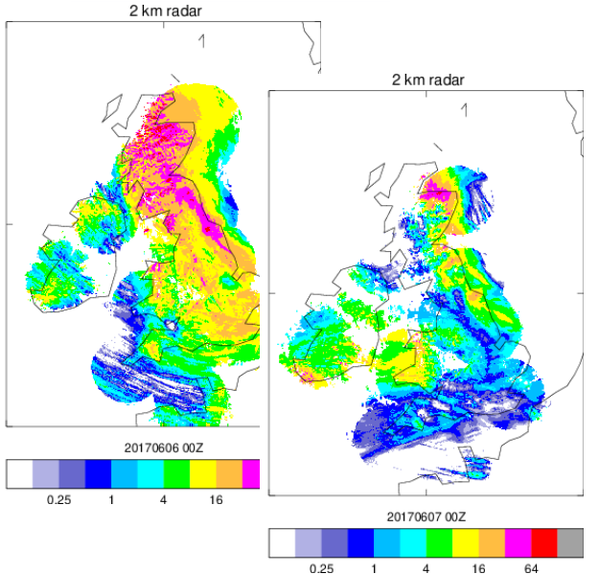
Rainfall and River Flow Ensemble Verification: Phase 2
Precipitation assessment of case studies
Final Report Appendix C.2

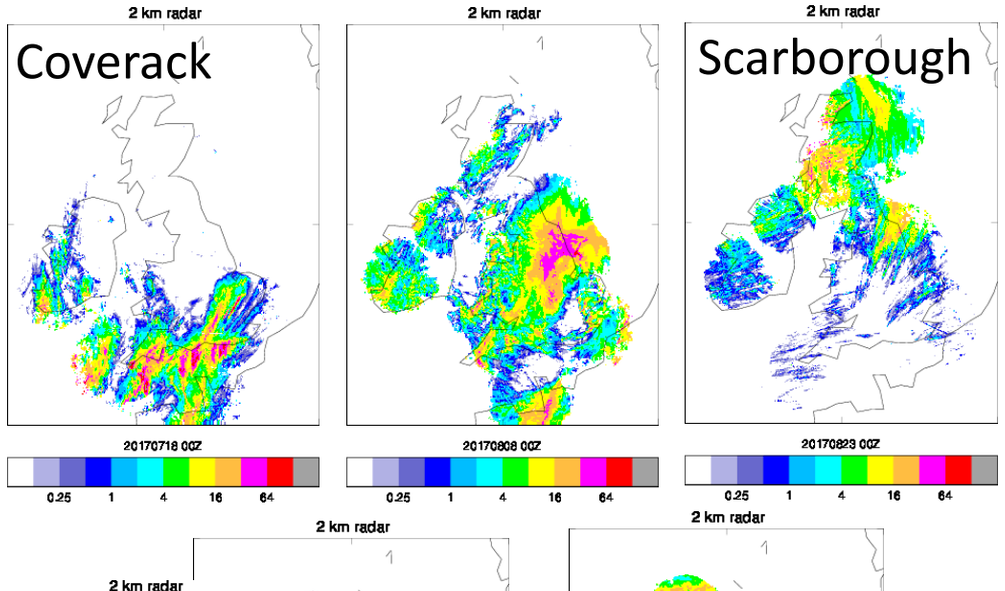
Gallery of case studies

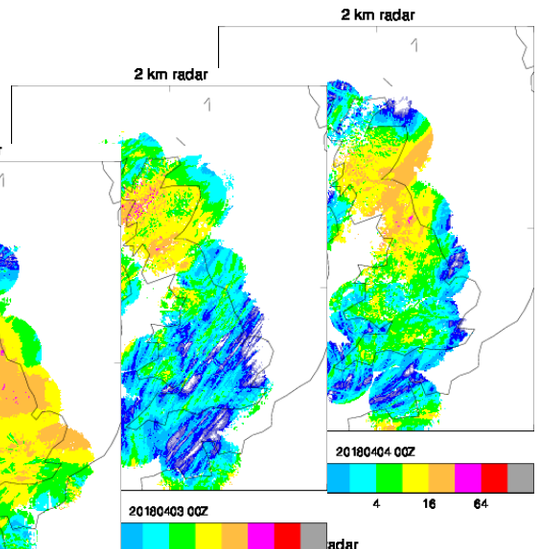
These “quick looks” were produced at the start of the project & are based on coarsened radar data (to 12 km) and represent daily totals from 00 UTC to 00 UTC only.

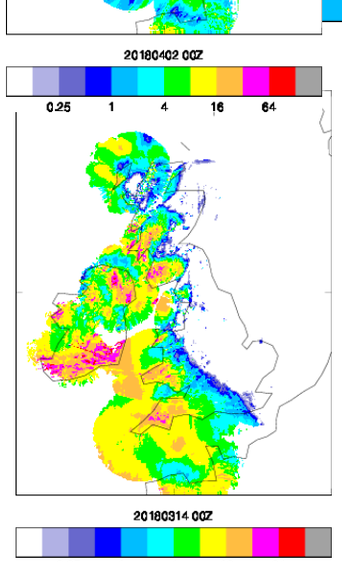
Scottish cases

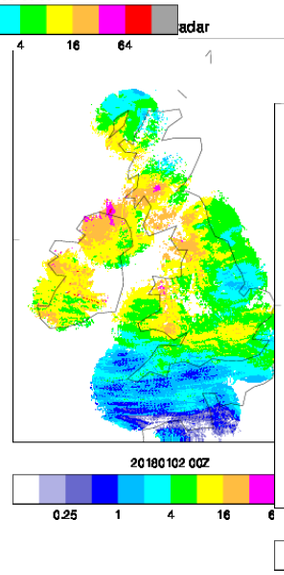


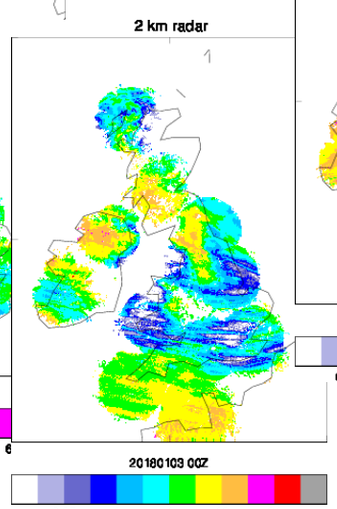


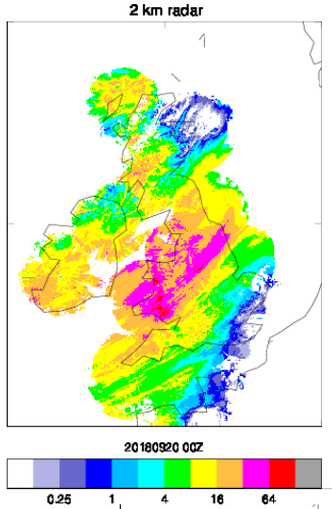


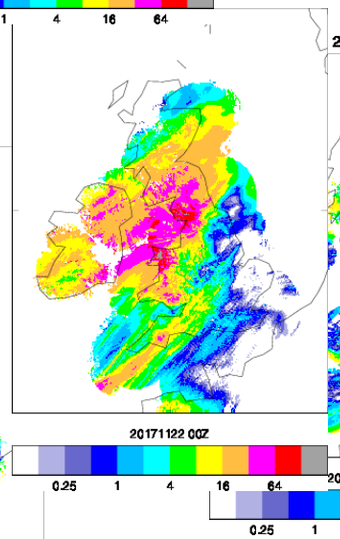


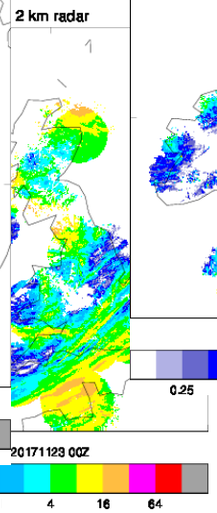


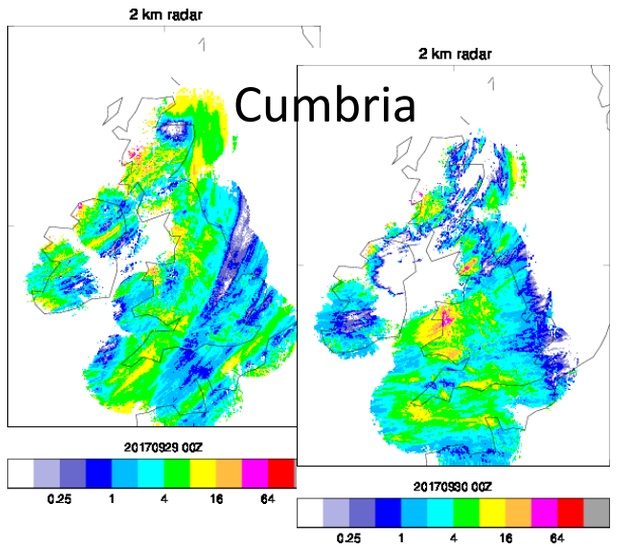


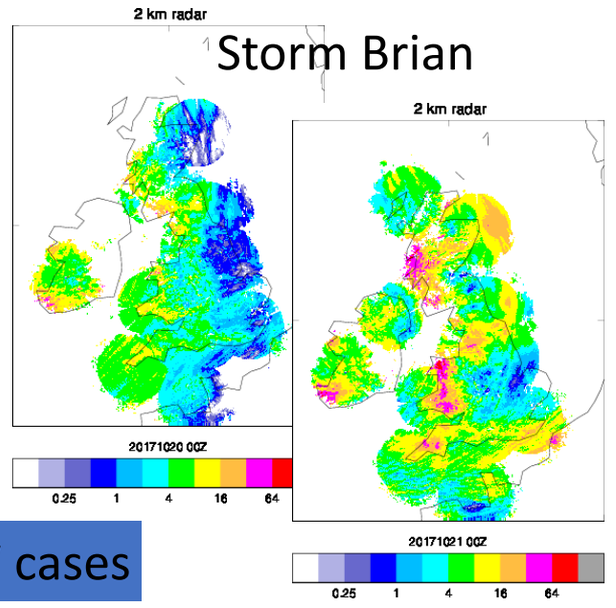




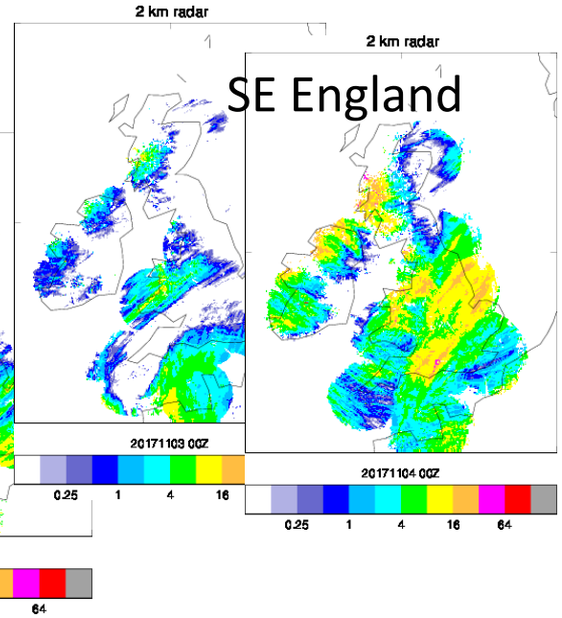








E&W cases



Case study approach

- Search for the maximum 24h precipitation across all catchments as a predictor for areas of interest in terms of flooding potential. A map indicating the catchment location and maximum rainfall is provided.
- For each forecast horizon (Day 1, Days 2-3 and Days 4-6) a single forecast initialisation was used to illustrate Time-window Probabilities (TWPs). Here the 99th percentile threshold probabilities are plotted.
- For Day 1 the identified catchment's precipitation hyetograph is also shown. The 99th percentile in-catchment value for each hour is plotted. Gauge data were used.
- The observation maps show a new depiction of the spot 99th percentile 24h accumulation value observed in the catchment as an indication of the heaviest rainfall (instead of the catchment mean). All three observation types are shown: raingauge, radar and merged product.
- Each case is provided with a commentary.
- A second slide for each case provides the context in terms of where the peak rainfall occurred as opposed to where the river flow response (if any) was observed. Hyetographs for these catchments (when different) are provided.

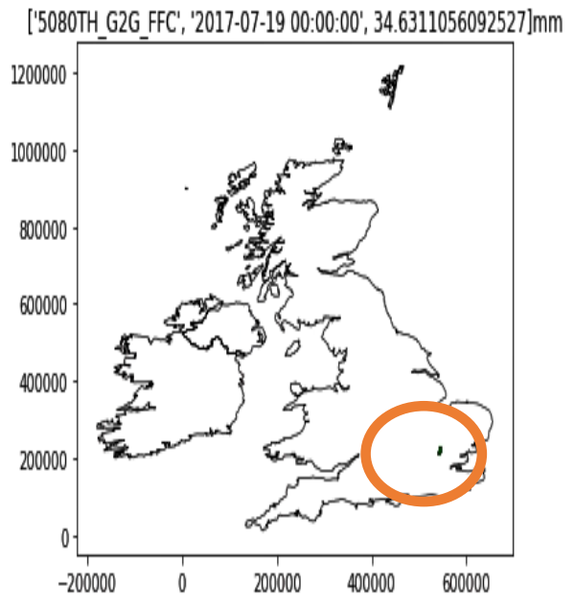
Case study	River flow catchments	24h peak rainfall catchments
18/07/2017		Walkham at Horrabridge (47118)
09/08/2017		Gypsey Race at Boynton (Boyntn1)
23/08/2017		Derwent (NE) at Low Marishes (MARISH1)
30/09/2017		Kent at Sedgwick (730511)
21/10/2017	Irwell at Irwell Vale (690140) Lune at Caton (724629) Wenning at Hornby (72452) Hindburn at Wray (724427) Wenning at Wennington (724326) Calder at Hebden Bridge (HEBDBR1) Calder at Todmorden (TODMDN1) Glaslyn at Beddgelert (065001_TG_1201 & PDM)	Glaslyn at Beddgelert (065001_TG_1201 & PDM)
3&4/11/2017		Moors River at Hurn Court (43214)
22 & 23/11/2017	Lune at Caton (724629) Wenning at Hornby (72452) Hindburn at Wray (724427) Wenning at Wennington (724326) Eden at Sheepmount (765512) Eden at Temple Sowerby (760502) Eden at Gt Musgrave Bridge (760112) Eden at Kirkby Stephen (760101) Glaslyn at Beddgelert (065001_TG_1201 & PDM)	Glaslyn at Beddgelert (065001_TG_1201 & PDM)
27/12/2017		Wensum at Costessey Mill (E19862)
02& 03/01/2018		Derwent at Portinscale (751007)
12-14/03/2018	Dove at Rocester Weir (4008) Dove at Hollinsclough (4033, & PDM) Torne at Auckley (4050)	Torne at Auckley (4050)
02-04/04/2018	Malton (Malton1) Derwent (NE) at Low Marishes (MARISH1) Riccal at Nunnington (Nunnington & PDM) Glaslyn at Beddgelert (065001_TG_1201 & PDM)	Glaslyn at Beddgelert (065001_TG_1201 & PDM)
20/09/2018		Taff at Fiddlers Elbow (057007_TG_504)

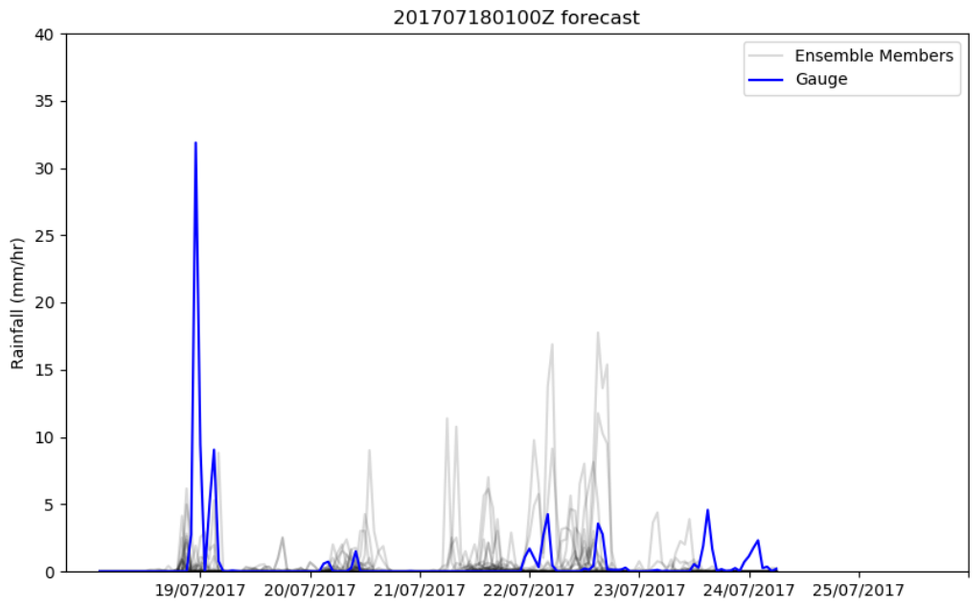
England & Wales

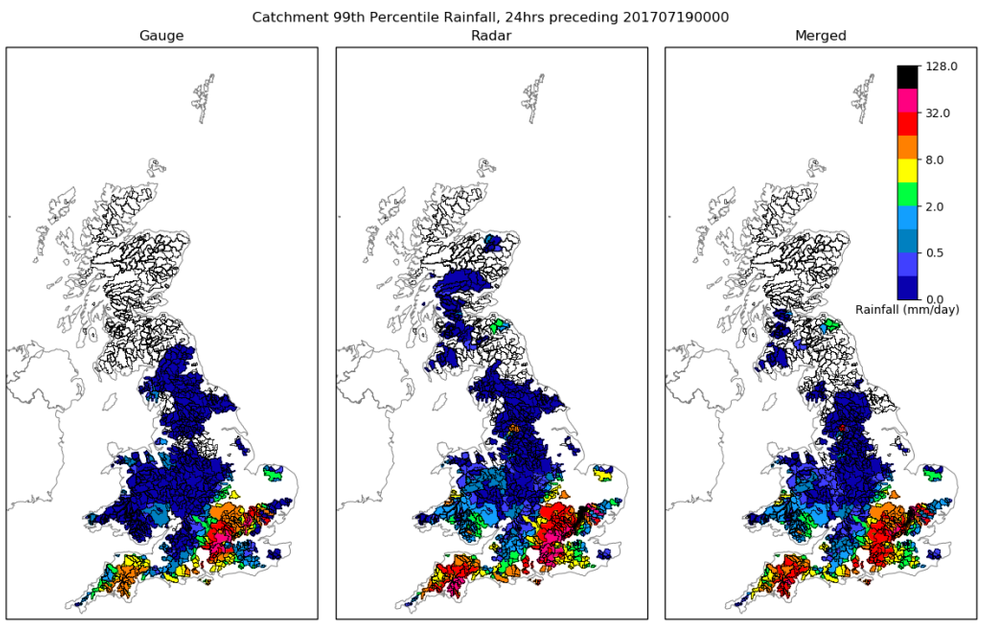
Coverack - 18 July 2017

Case Study Synopsis. Flash Flooding in Coverack with danger to life, helicopter rescues, 50 properties flooded, damaged roads and infrastructure.

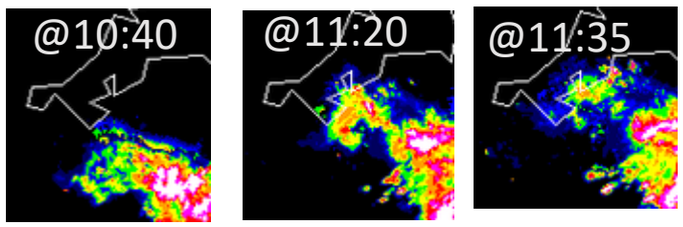
Max Catchment	Date/Time of Max	Rainfall (mm)
5080TH_G2G_FFC	2017-07-19 00:00:00	34.63





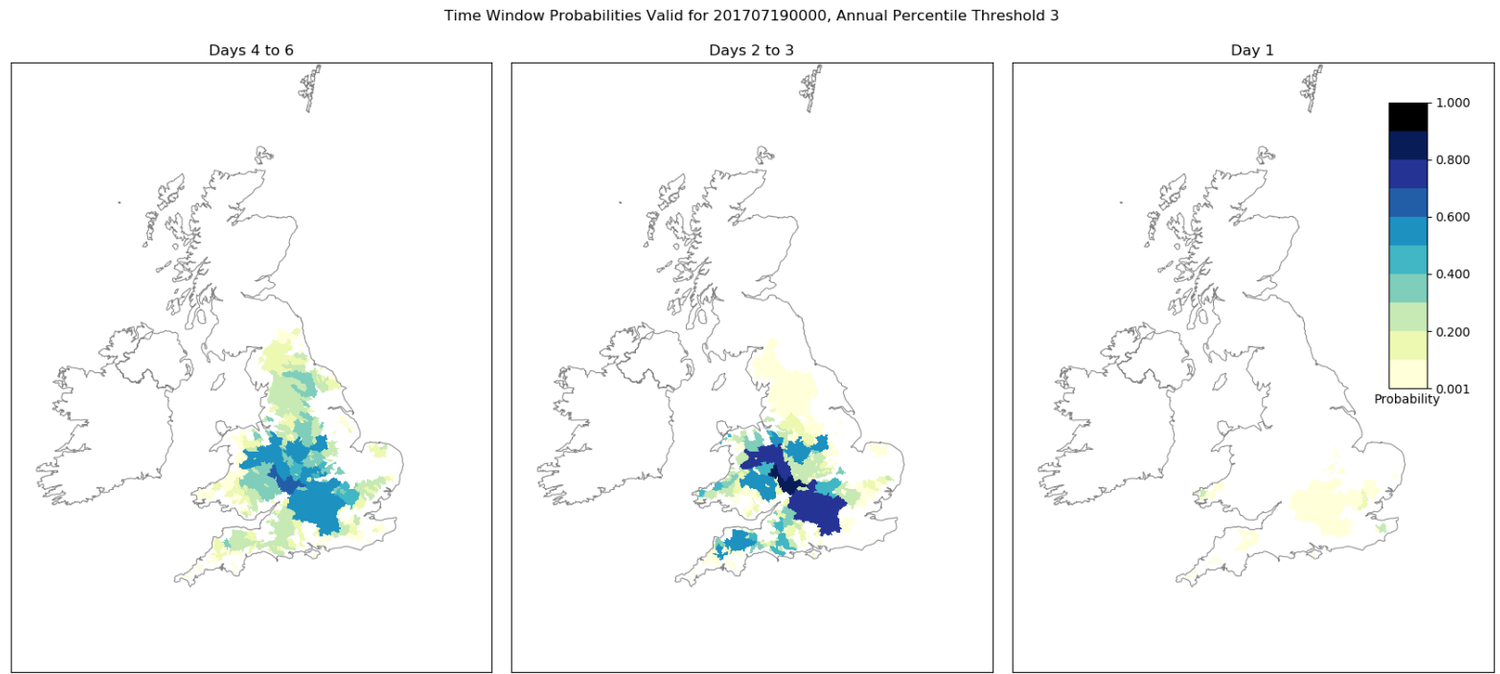


Radar provided a vital nowcasting tool with a clear advantage over other observation sources and weather models.



Summary

Coverack was a highly localised event, with no catchment with a river flow gauge in the vicinity so no catchment TWP was calculated. From precipitation observations, the nearest catchment was north-east of London around Essex/Hertfordshire. The forecast for this catchment captured the timing of rainfall but under-forecast the amount by around 3 times. Significant differences can be seen between observation sources. The largest differences of potentially 100mm (red v black) in around Essex/Hertfordshire. Large differences can also be seen in South Devon. TWPs through time windows Days 4-6 and Days 2-3 and focused on the southern Central England and Eastern Wales/West Midlands. TWPs for Day 1 show very few probabilities with low probabilities (0.1/0.2) in the region of the catchment with most rain. Non-catchment-based products did provide some indication of large precipitation totals in the vicinity of Coverack.



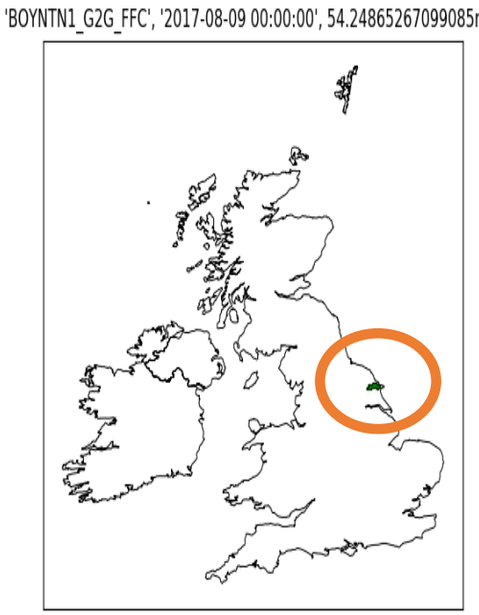
South East – 8 August 2017

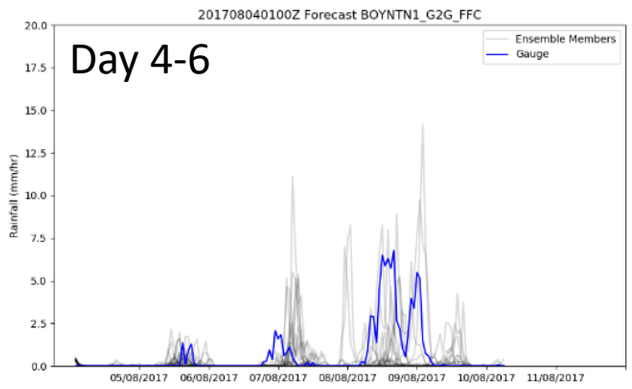
Case Study Synopsis. EFAS Flash Flood notification. Minor impacts from surface water flooding: Essex, Kent, Surrey, Greater London, Suffolk. No fluvial impacts recorded.

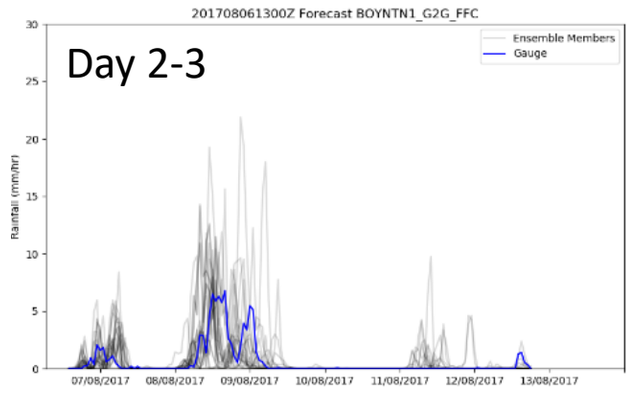
Summary

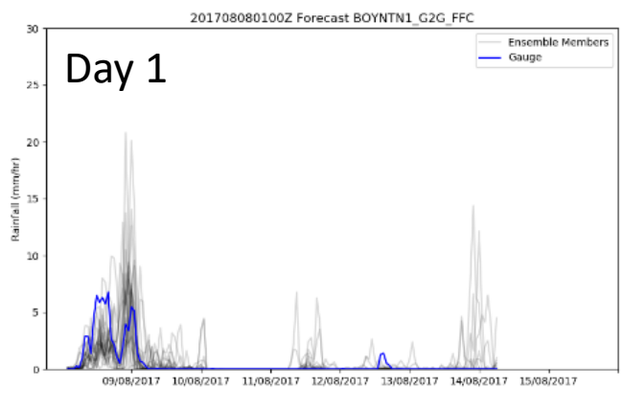
The maximum rainfall was picked up in a catchment north of the Humber with no impacts, though there is good agreement between the ensemble forecast and observed rain in this catchment.

Max Catchment	Date/Time of Max	Rainfall (mm)
BOYNTN1_G2G_FFC	2017-08-09 00:00:00	54.3

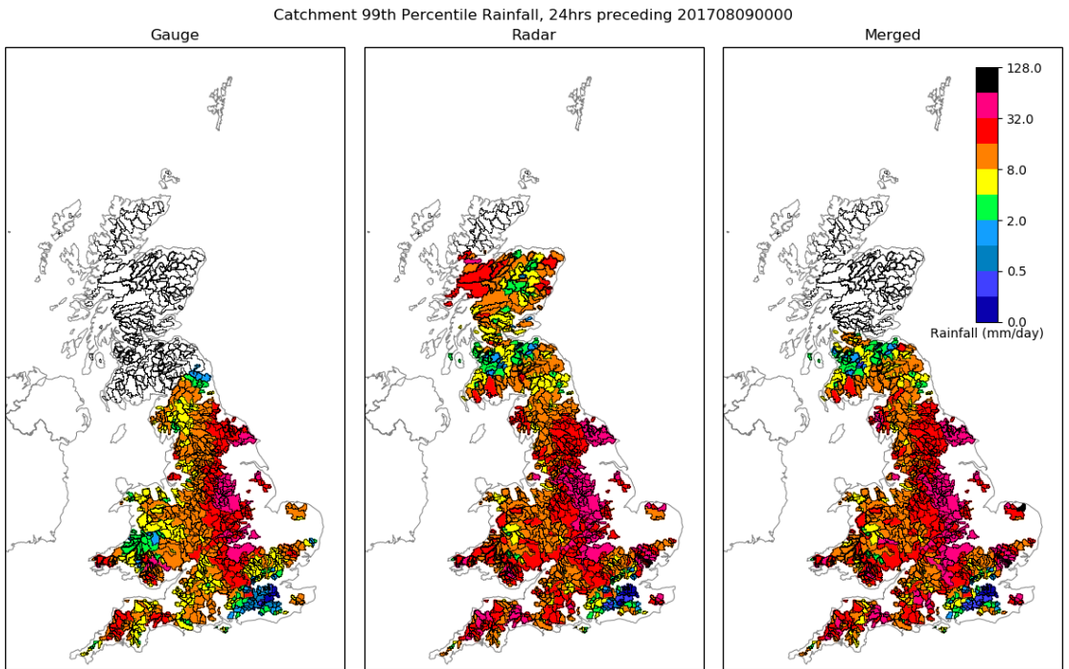


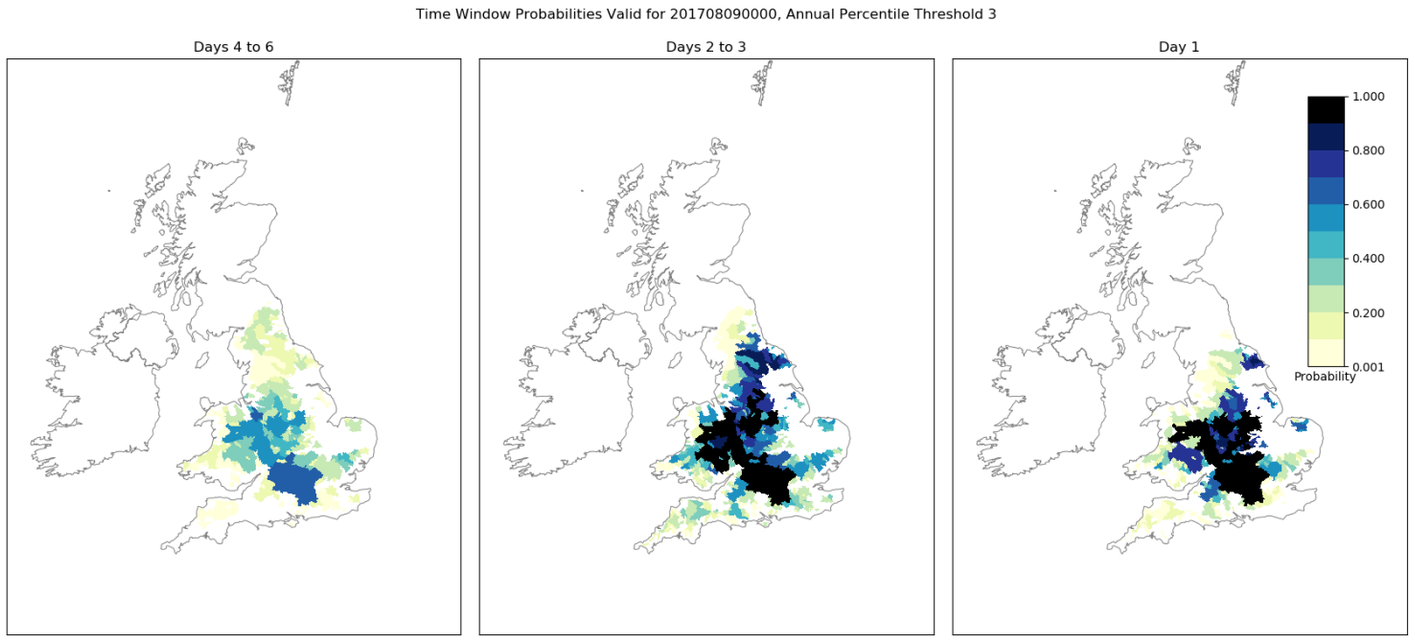






Time-Window Probabilities also identified some of the areas affected in the Day 4-6 window. The probability of an event increased into the Day 2-3 window but was lowered in the Day 1 window. TWPs did however identify the catchment that received the highest daily rainfall total from Days 4-6. The probabilities for this catchment increased window-by-window and was correctly given a high probability of an extreme rainfall event occurring. Peaks in observed rainfall are well correlated with ensemble members for the highest catchment. The observed rainfall is also largely within the ensemble spread.

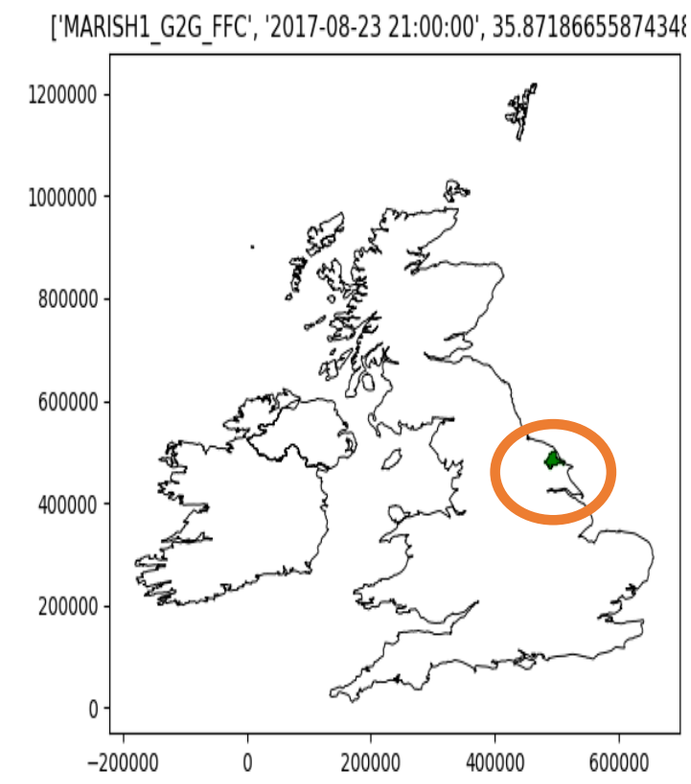


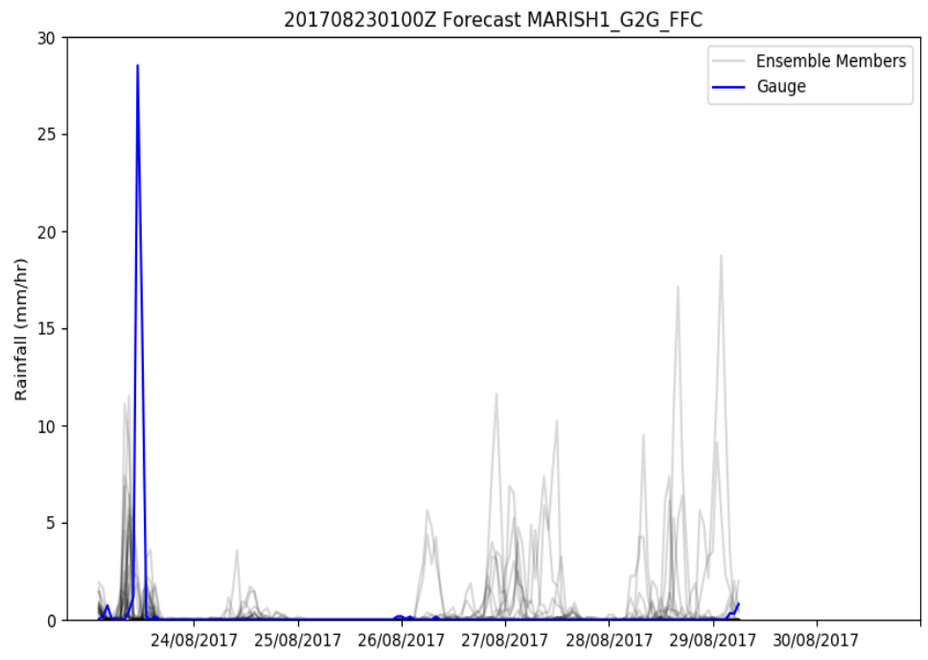


Scarborough – 23 August 2017

Case Study Synopsis. Convective event - no river impacts recorded. Flash flooding from surface water, causing travel disruptions. Significant impacts - North Yorkshire (Scarborough), minor for York and West.

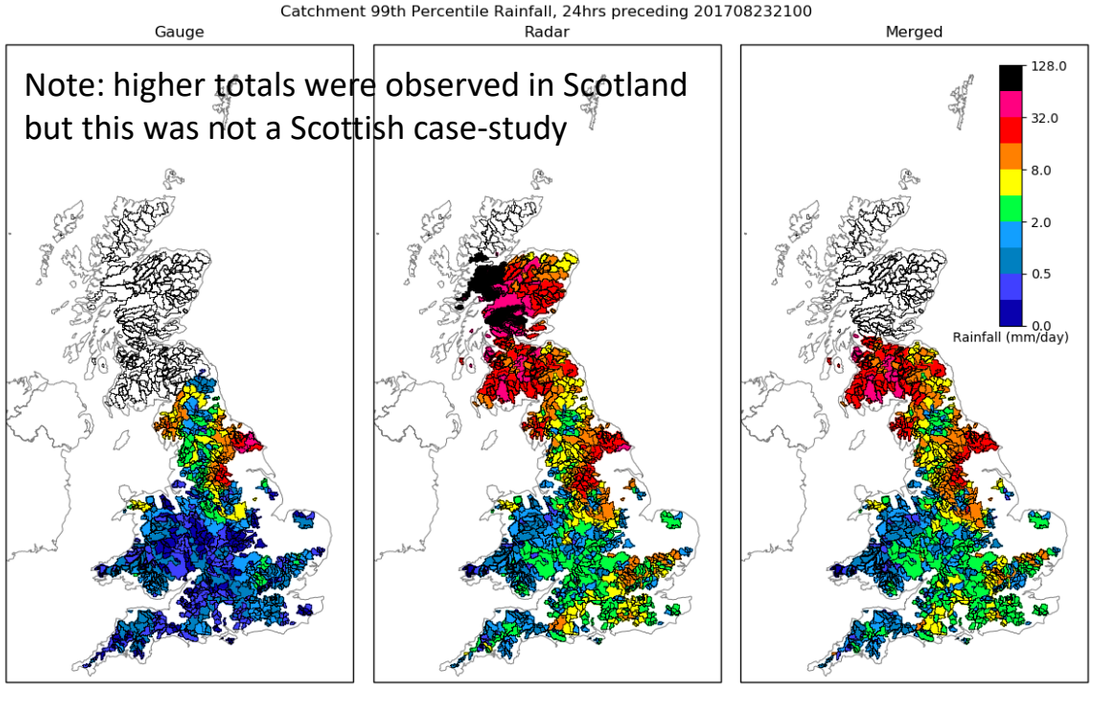
Max Catchment	Date/Time of Max	Rainfall (mm)
MARISH1_G2G_FFC	23/08/2017 21:00	35.87



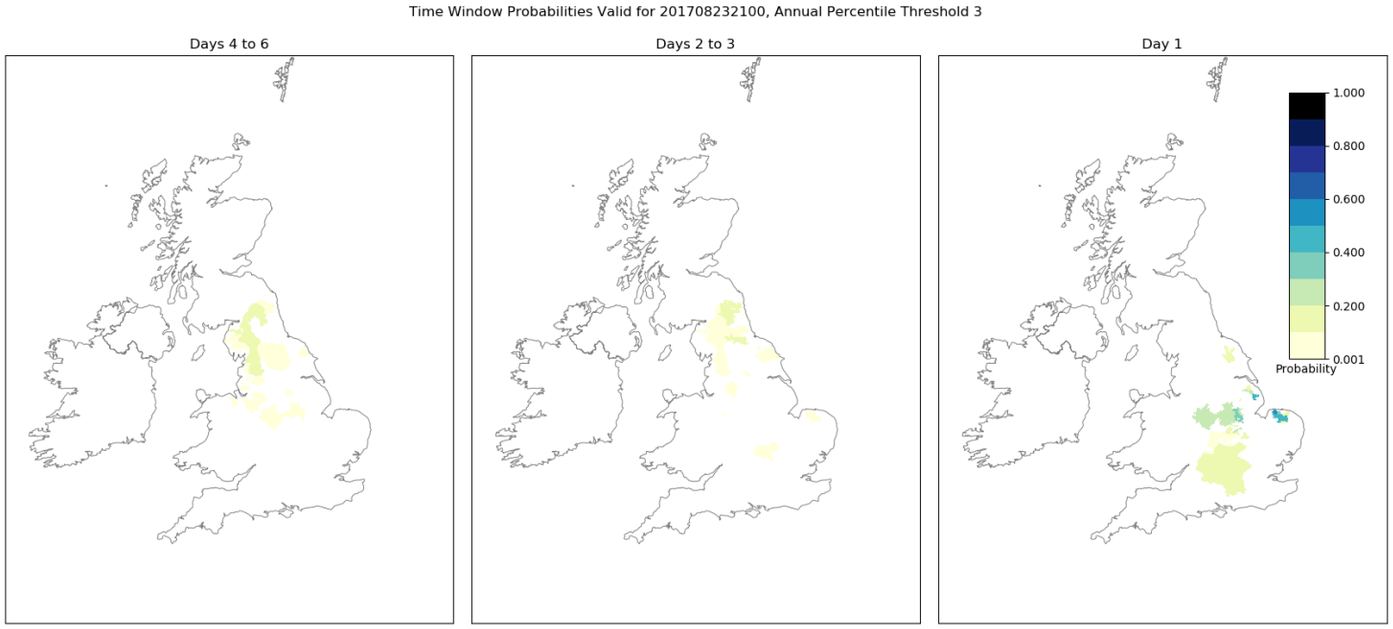


Summary

Very low TWPs for the 99th percentile TWPs at Days 4-6 and Days 2-3. The TWP increases into Day 1 but also moves to a neighbouring catchment inland. Other high probabilities are present around the Wash at longer lead-times. The forecast for the wettest catchment shows a peak in rainfall that corresponds with the peak in raingauge observations. The magnitude of this peak is approximately 3 times larger in the observed (gauge) than the highest ensemble member. Though there are some higher probabilities in the NE, this event was not that well captured by the weather model. Between the different observation sources, there are some large discrepancies across the South and South East. Approximately 8mm in Hertfordshire/Essex, Dorset and West Sussex.



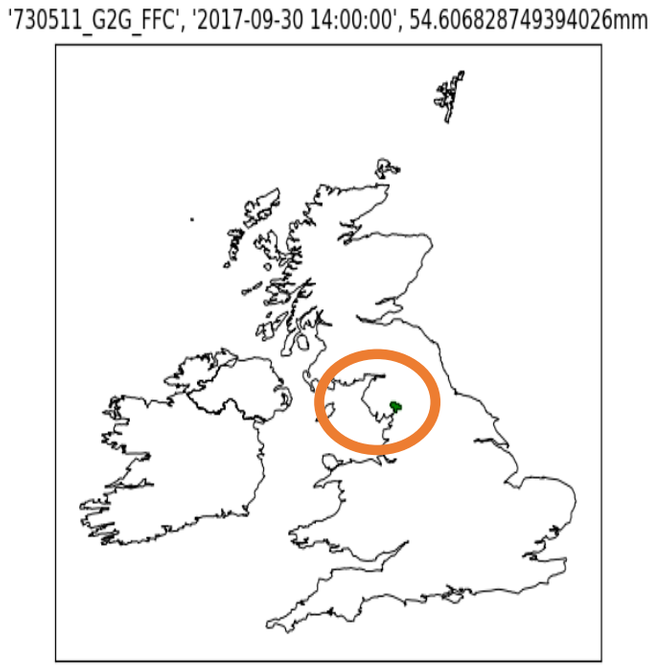
Note: higher totals were observed in Scotland but this was not a Scottish case-study

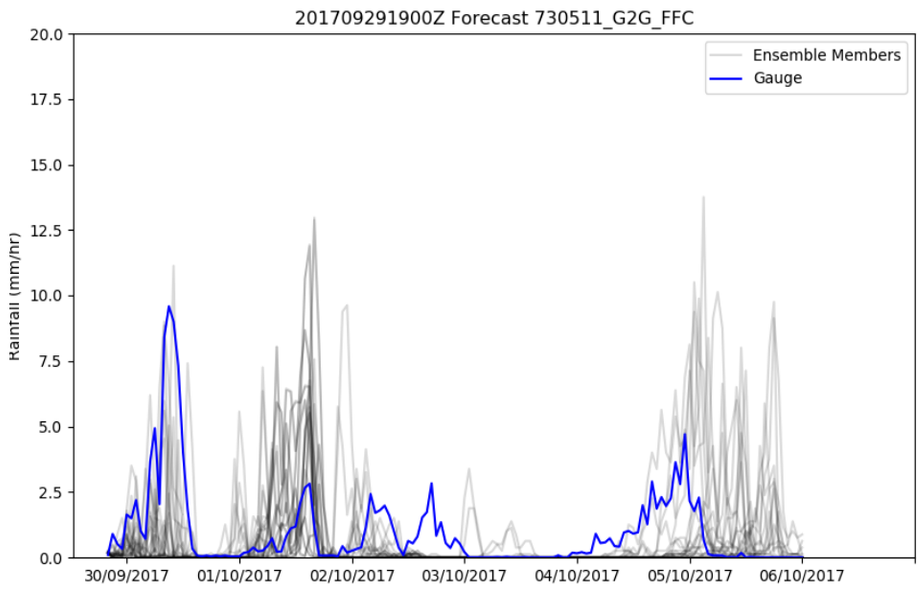


Cumbria – 30 September 2017

Case Study Synopsis. Narrow band of heavy rain over the south of Cumbria. 40-45 mm in 1 hour and 60+ mm in 2-2.5 hrs at Millom and Haverigg between 0800 and 1030 BST. A total of 150-200 properties affected by surface water flooding, largely in the Millom area and with a small number of flooded properties in Windermere and Haverigg.

Max Catchment	Date/Time of Max	Rainfall (mm)
730511_G2G_FFC	30/09/2017 14:00	54.6

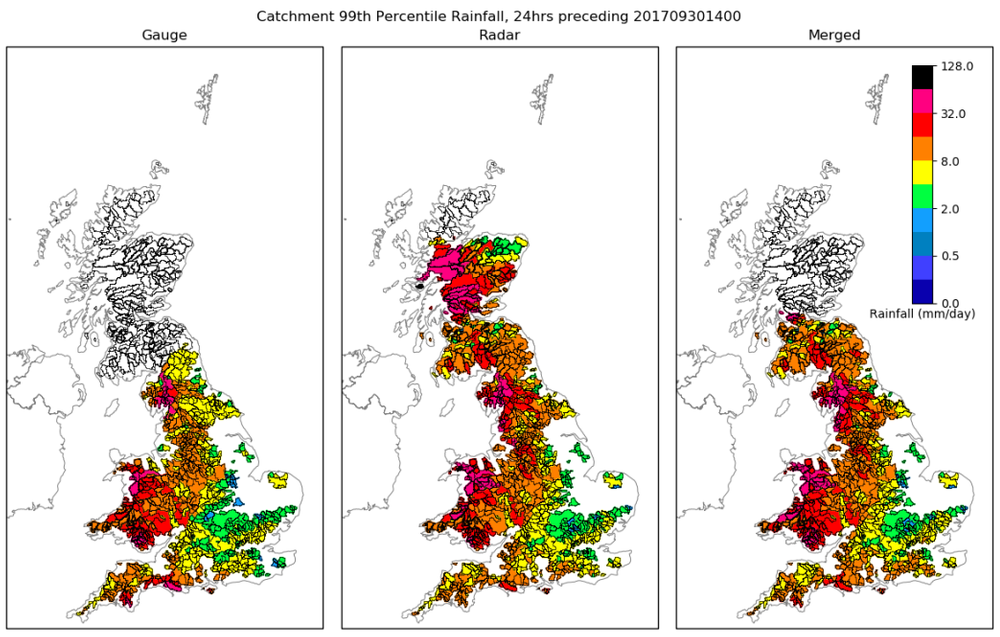


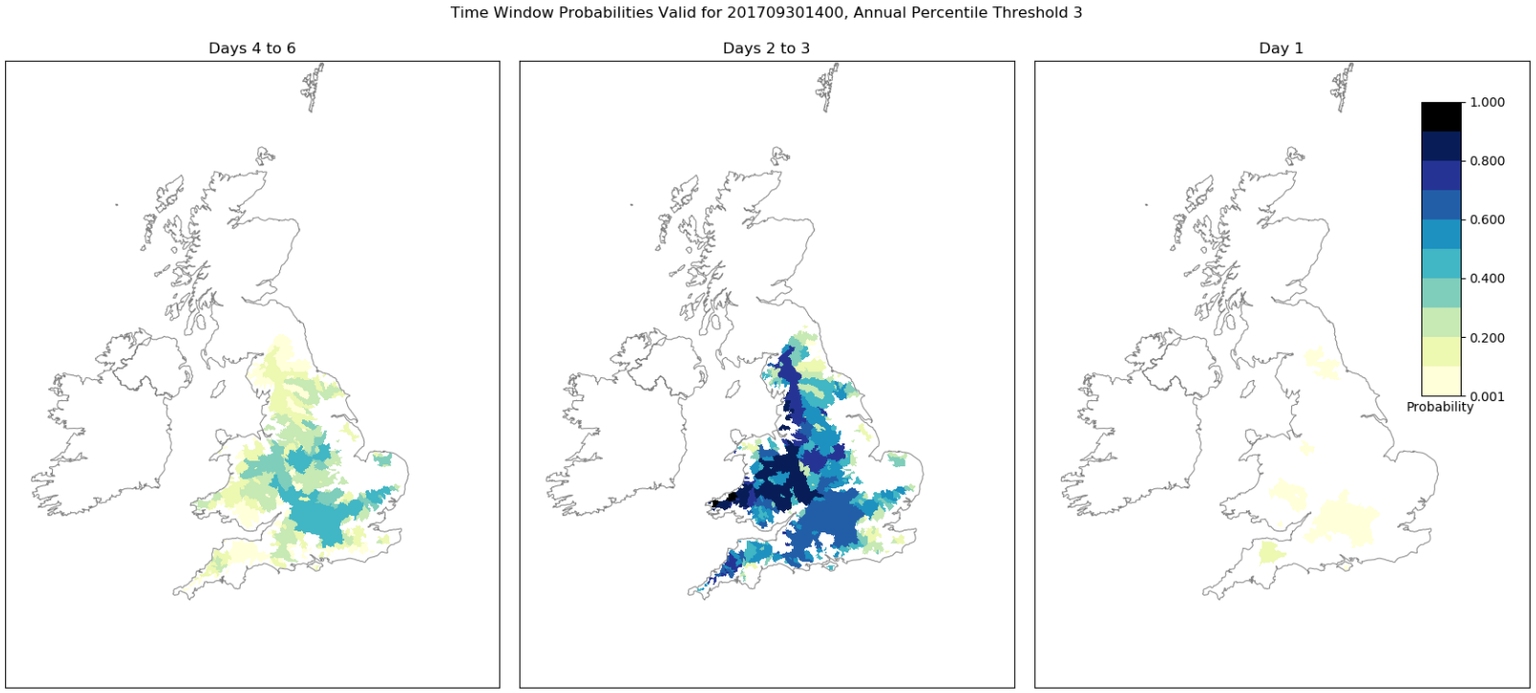


Summary

The catchment with the highest precipitation corresponds with case-study location. The forecast for this catchment seems accurate. Raingauge observations are within the ensemble spread and peaks are well correlated. TWPs identify Cumbria catchments having low probabilities (0.1-0.3) in Days 4-6 with higher probabilities focussed on the West Midlands and South. Probabilities rise across the England & Wales into Days 2-3. the highest probabilities now 0.8-1 in Wales, West Midlands and parts of Lancashire. The probability of an event occurring in the catchment of highest observed precipitation is 0.5. Probabilities in the Day 1 time-window drop significantly to a maximum of 0.3 for a group of catchments around Exmoor & the Quantocks.

All observation sources are largely in agreement. Biggest differences between radar and gauge around the Yorkshire Dales.



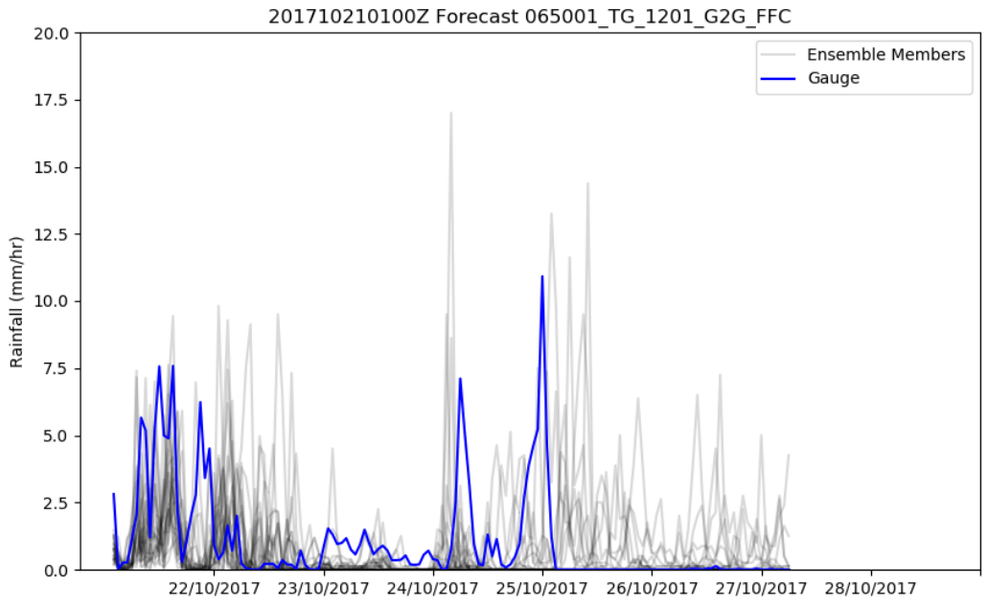


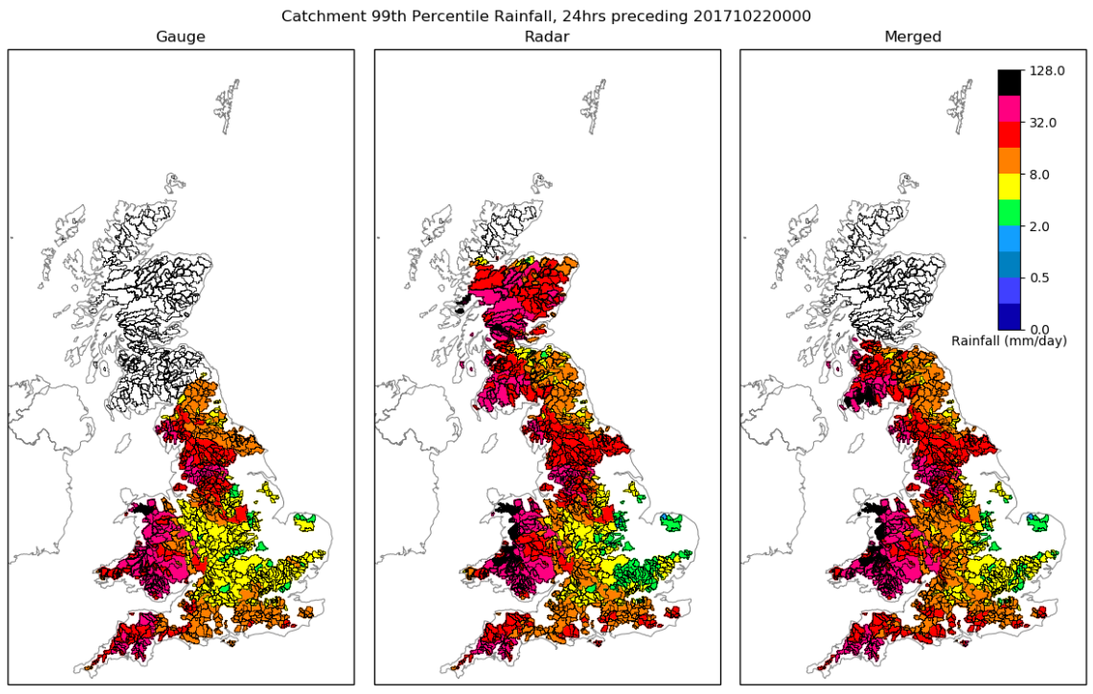
Storm Brian – 21 October 2017

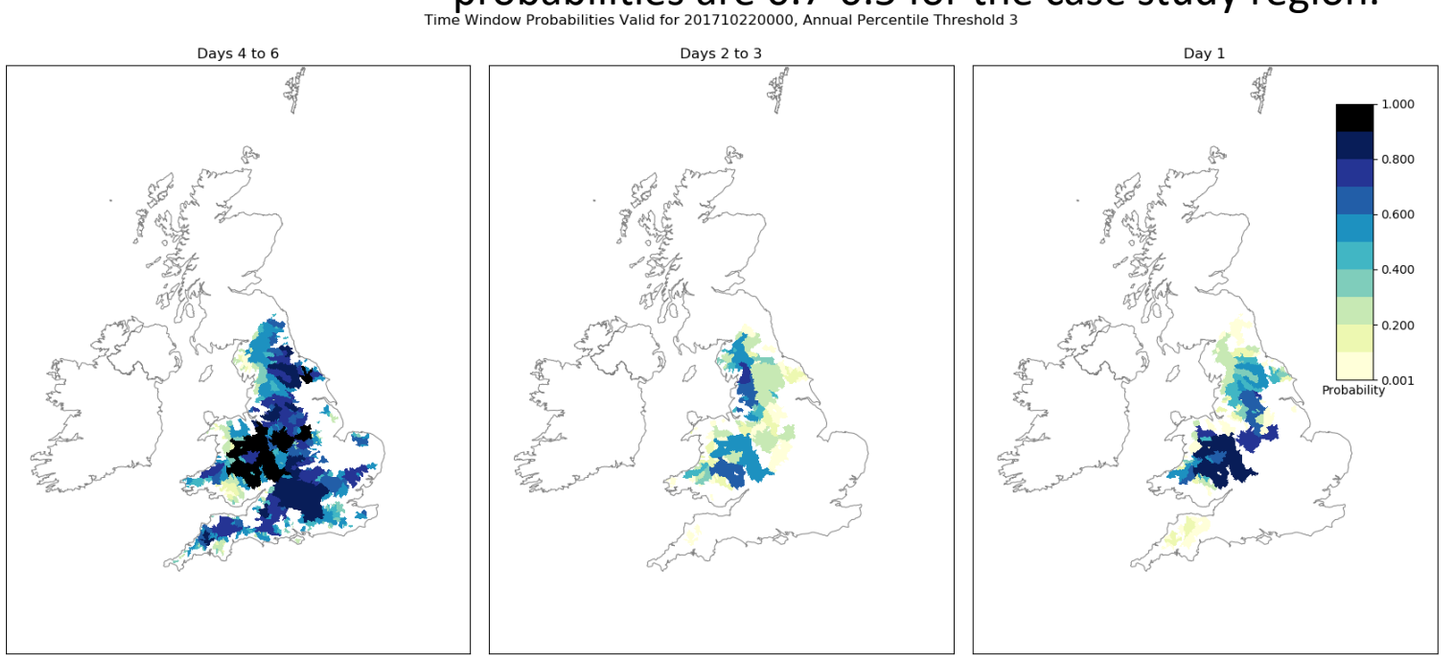
Case Study Synopsis. Missed SIG event. Flood Sirens in Todmorden, Hebden Bridge and Mytholmroyd. 4 properties (river), 1 industrial building (river), 1 pre-school (SW) and a bakery (SW) flooded in Rossendale plus 5 properties in Rawtenstall from river and 8 in Rawtenstall from surface water

Max Catchment	Date/Time of Max	Rainfall (mm)
065001_TG_1201	22/10/2017 00:00	70.9

Highest catchment rain Glaslyn at Beddgelert (065001_TG_1201 & PDM) in Wales. Impact region in W Yorkshire.





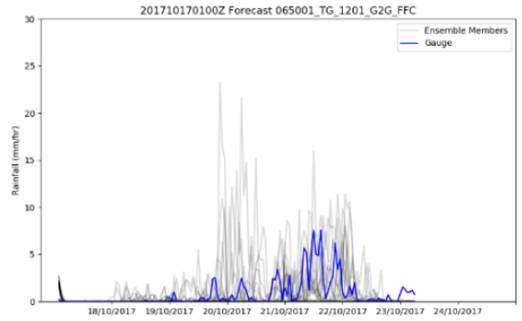
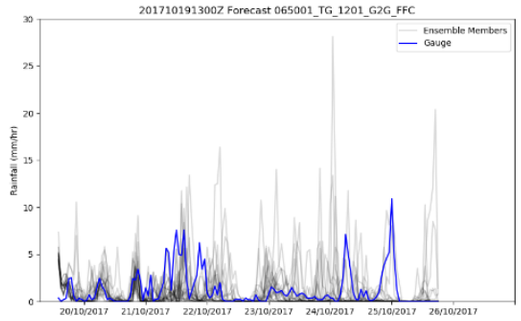
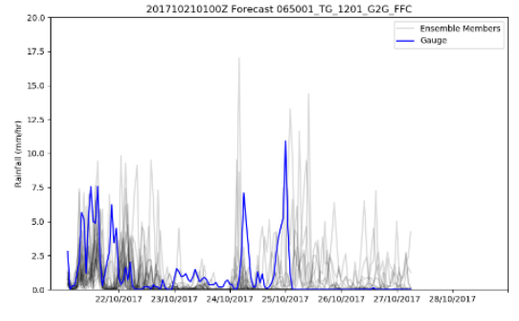
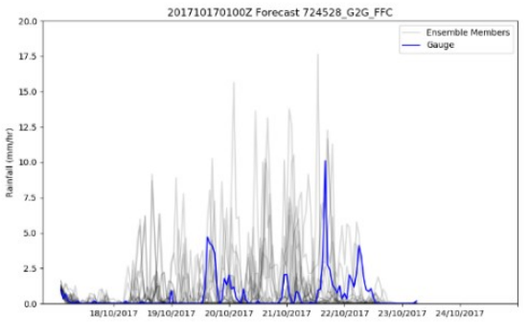
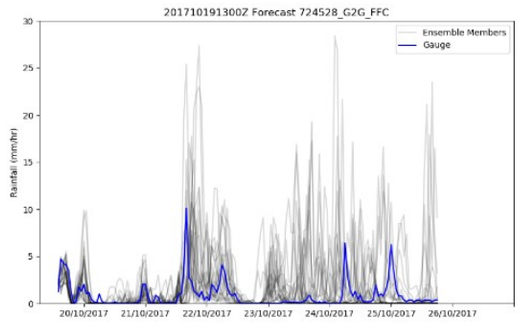
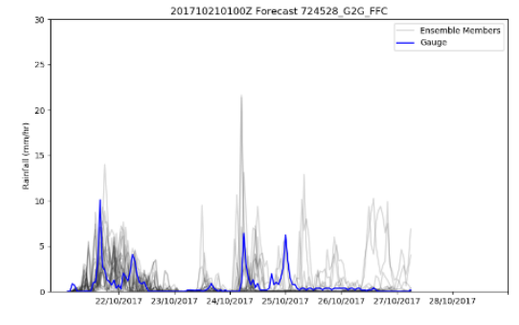
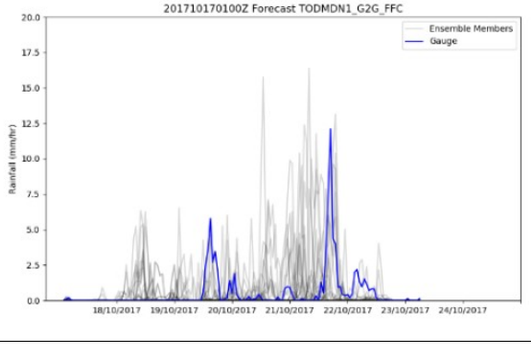
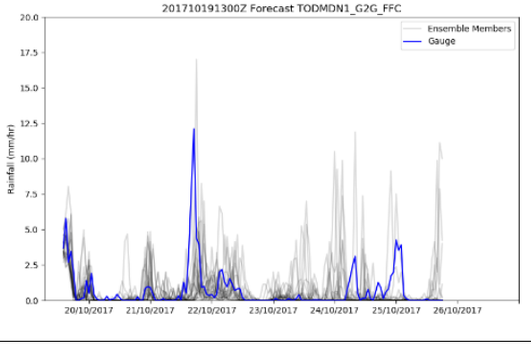
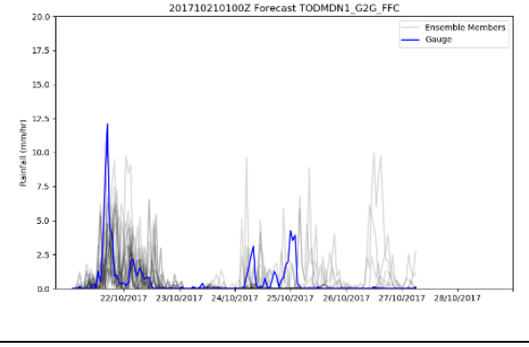


Summary

The catchment with the highest precipitation does not correspond with the river flow impacts. The forecast for the catchment captures the timing of the event well. An exception to this is a peak in the observations just before 22 Oct 00:00 occurring earlier than forecast. Other than this peak, raingauge observations are within the ensemble spread. The 3 observation sources largely agree on the distribution and amount of rainfall across England and Wales. Areas of the greatest discrepancies are in the SE, where the gauge totals are higher than radar. In the NE and W Wales the radar records higher totals. TWPs are high across England & Wales for Days 4-6, with the highest probabilities of 1 across Wales and the West Midlands. On Days 2-3 probabilities are much lower. The highest probability of 0.8 in Cumbria, with slightly lower probabilities still present in Wales. Probabilities in the case-study region are 0.7/0.3. In the Day 1 window the highest probabilities (0.9) are in similar areas to the Days 4-6 window. The probabilities are 0.7-0.5 for the case study region.

Storm Brian – 21 October 2017

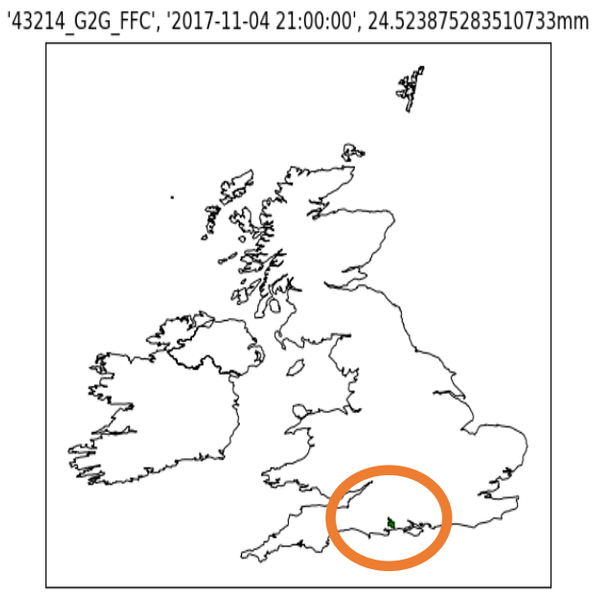
Temporal forecast evolution in highest precipitation catchment and those with river flow responses.

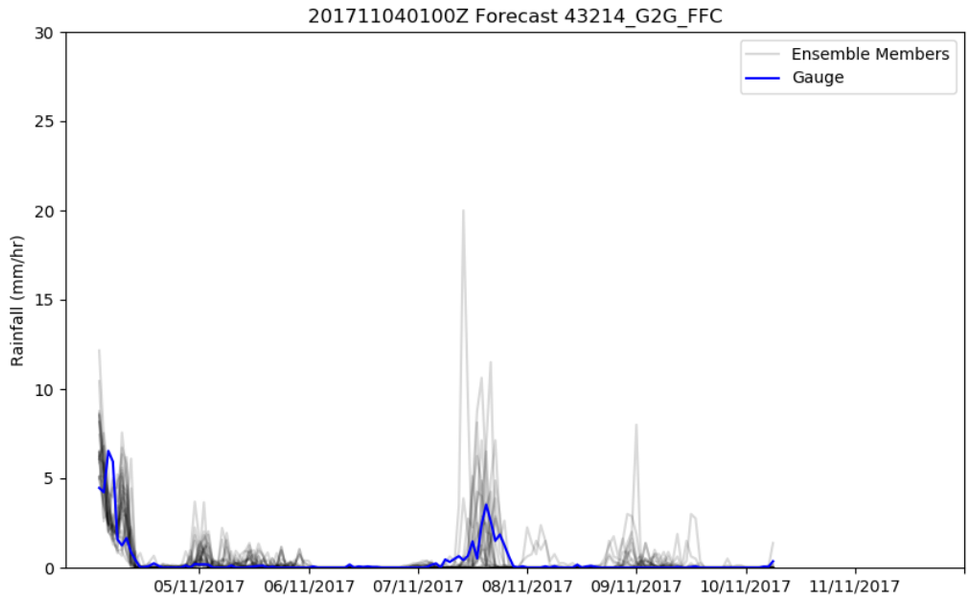
Catchment ID	Comments	Days 4-6	Days 2-3	Day 1
065001_TG_1201	Catchment as identified as having highest 24hr rainfall is not in case-study region.			
724528	Catchment is in case-study region. Forecast captures the magnitude of the event well. Consistent with 724427, 724326, 724629.			
TODMDN1	Consistent with impact location. Similar profile for HEBDBR1 – Obs outside of ensemble spread on Day 1 TW.			

SE England – 3-4 November 2017

Case Study Synopsis. Rainfall false alarm? No impacts noted. G2G Deterministic supported minimal impacts. Some suggestion of higher flows from G2G ensembles.

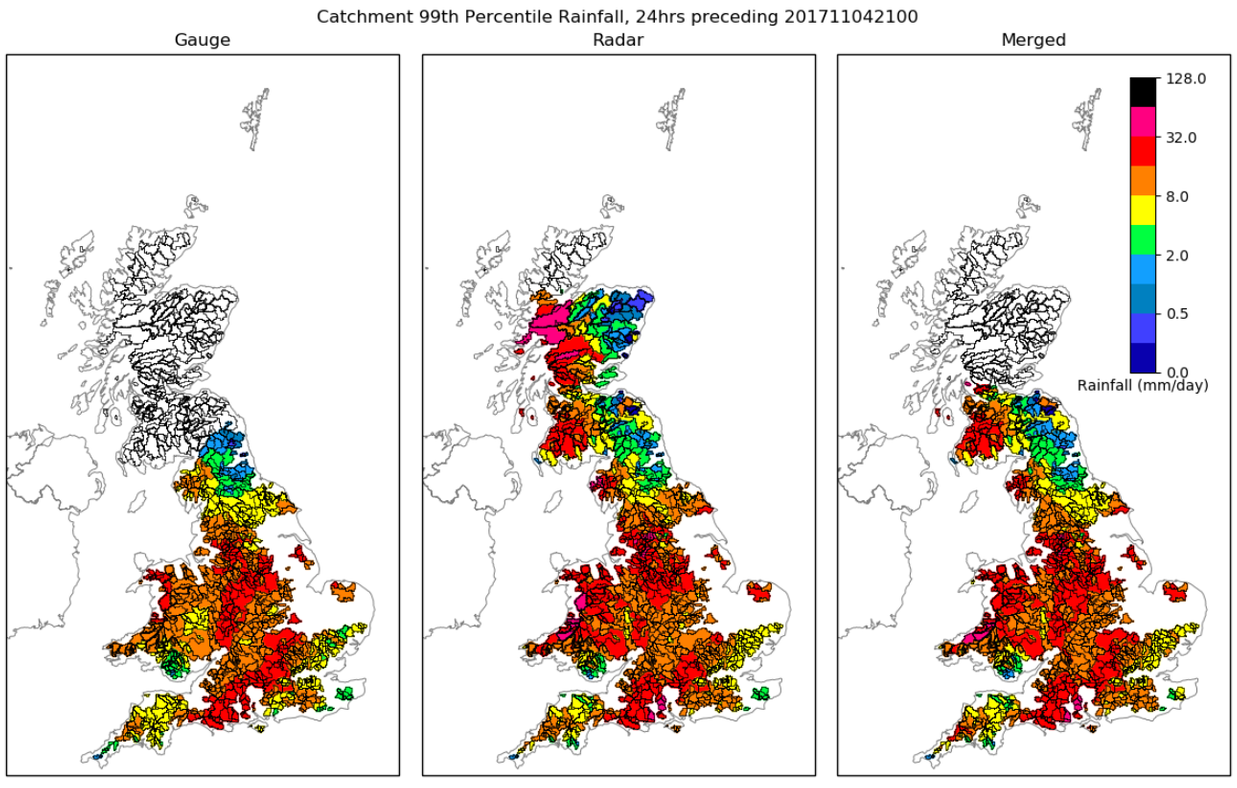
Max Catchment	Date/Time of Max	Rainfall (mm)
43214_G2G_FFC	04/11/2017 21:00	24.5

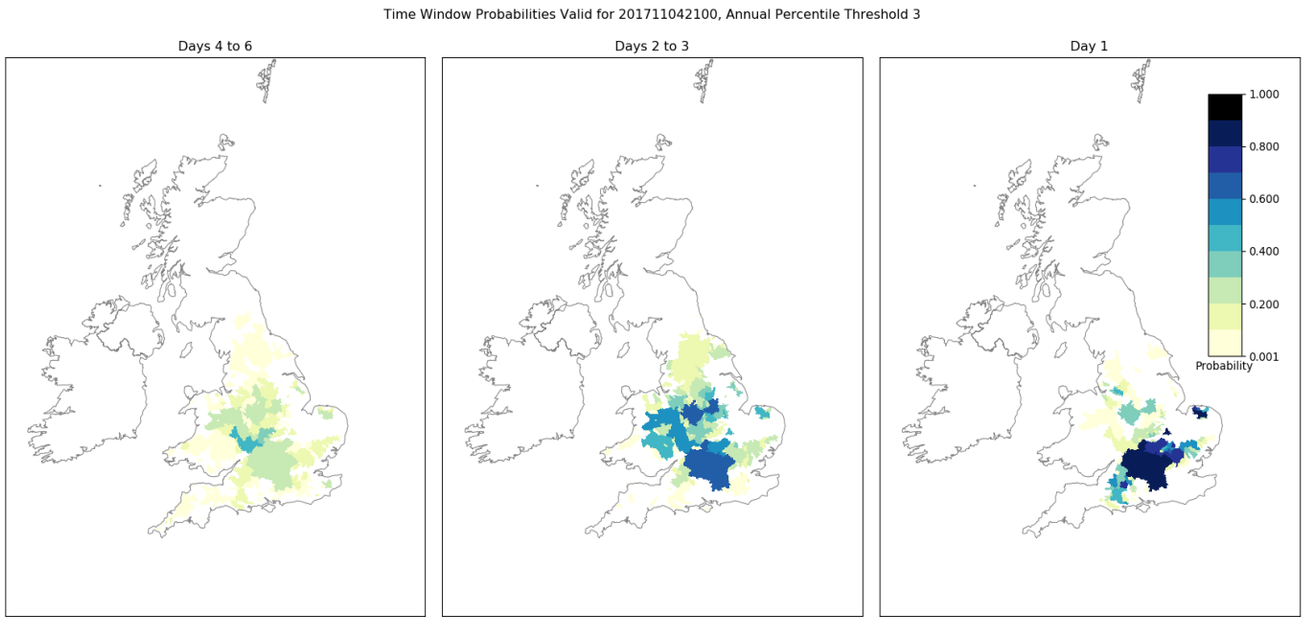




Summary

The catchment receiving the highest 24hr accumulation between 3 and 4 November was in the Bournemouth area. Similar amounts of rainfall were seen around the Chiltern Hills, Snowdonia, from West Midlands up to Manchester/Peak District and on the North Norfolk coast. Generally, the distribution of rainfall across England and Wales is consistent across observation types. The forecast for the highest catchment captures the general timing of the peak in rainfall but shows two peaks either side of the observed peak. This leaves the observed peak outside of the ensemble spread at that time. The TWPs identify a large proportion of C & S England, which received similar amounts of rainfall to the highest catchment in the Days 4-6 window with probabilities around 0.2-0.5. Probabilities increase into the Days 2-3 window. The greatest probabilities (~0.7) across central Southern England and the West Midlands. Into the Day 1 time-window, probabilities increase again across central Southern England to 0.9 along with an area on the north Norfolk coast. From the observations, these areas did receive similar amounts of rainfall to the highest catchment (Bournemouth region). Probabilities in this area were much lower, around 0.1 to 0.5.



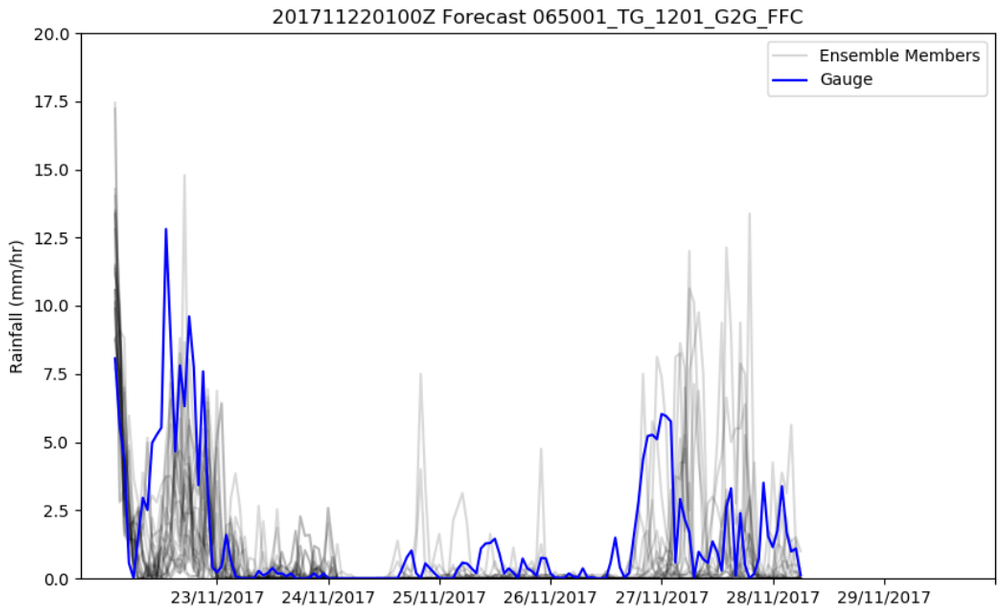


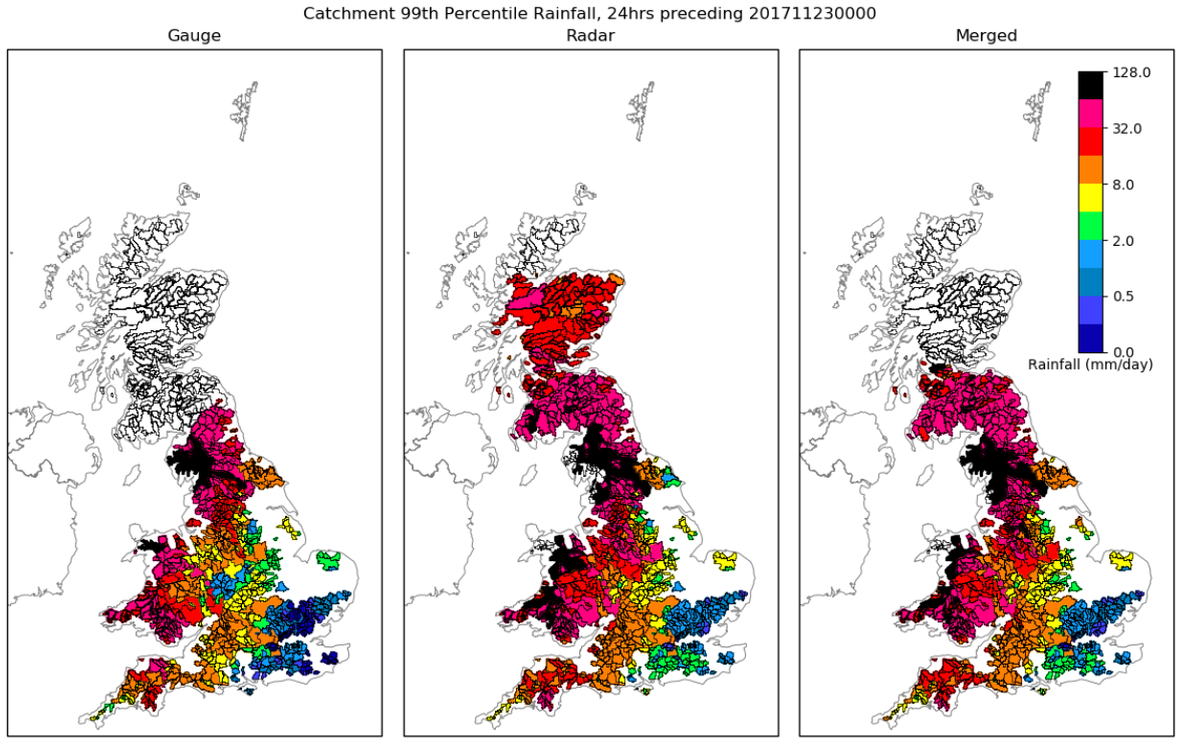
NW England & N Wales – 22-23 November 2017

Case Study Synopsis. Impacts on 23rd recorded as SIG over Cumbria, Lancashire for rivers and also minor for Cumbria, Anglesey, York, and Lancashire.

Max Catchment	Date/Time of Max	Rainfall (mm)
065001_TG_1201	23/11/2017 00:00	112.3

Highest catchment rainfall
Glaslyn at Beddgelert (065001_TG_1201 & PDM)
Snowdonia NW Wales



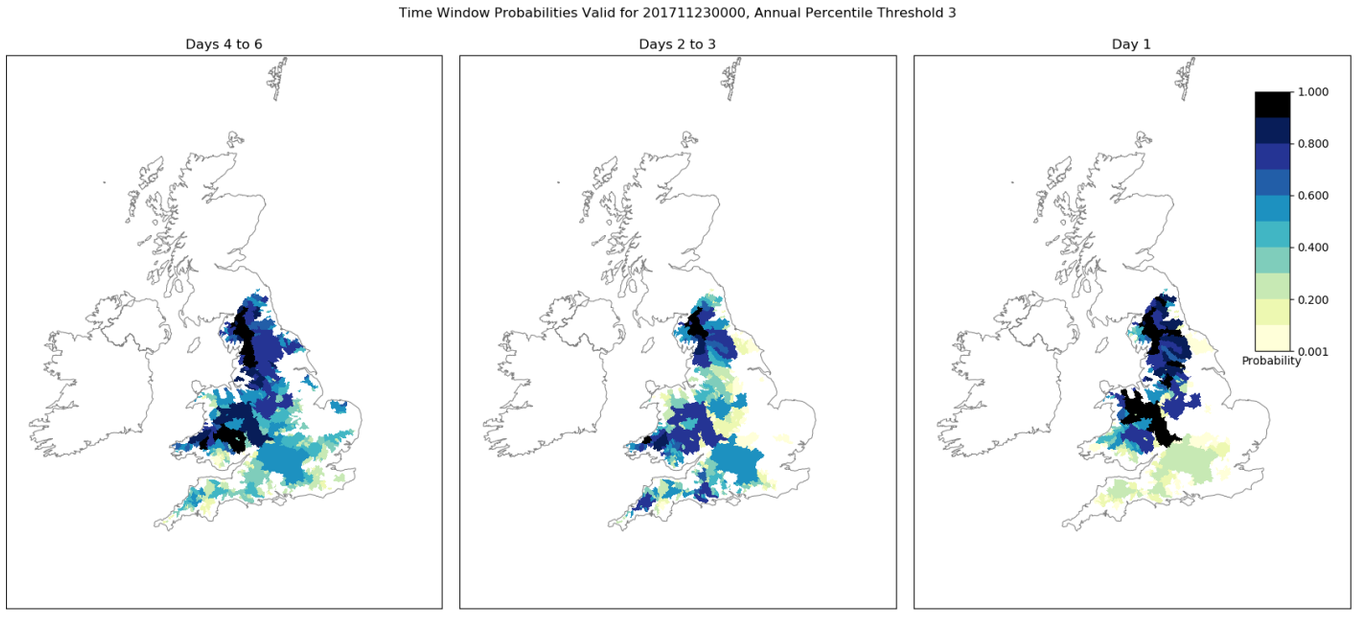


Summary

The catchment of the highest precipitation corresponds with case study location of N Wales. The forecast for this catchment captures the event reasonably. Raingauge observations are within the ensemble spread for most of 22 Nov except for the highest peak of the event.

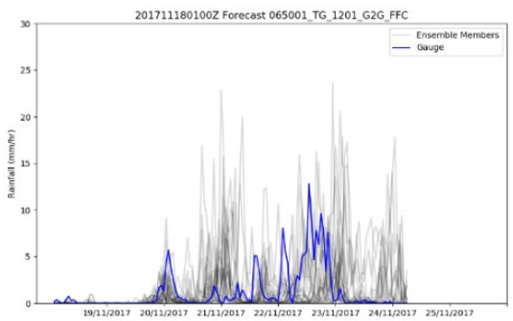
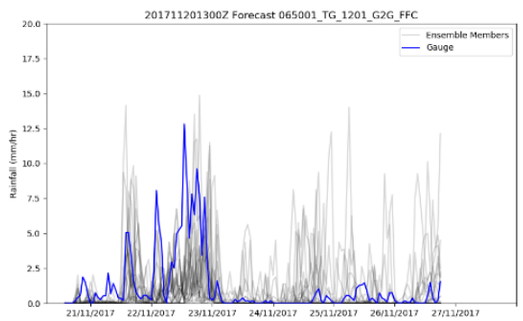
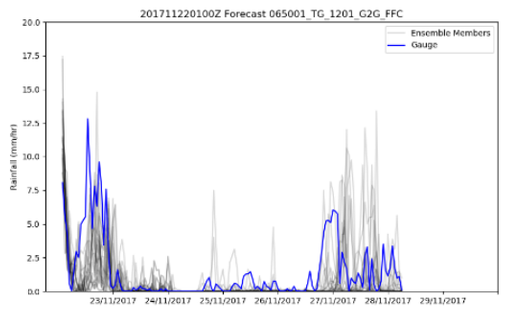
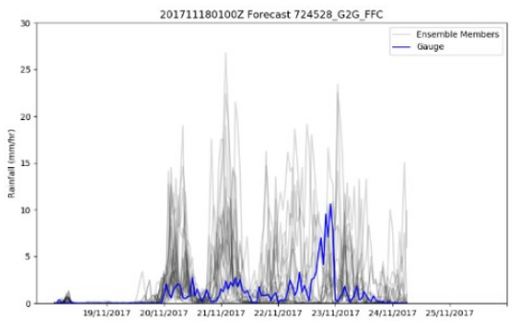
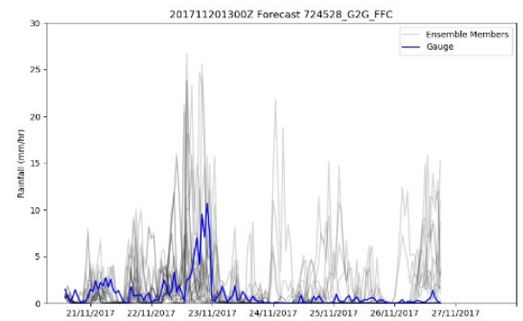
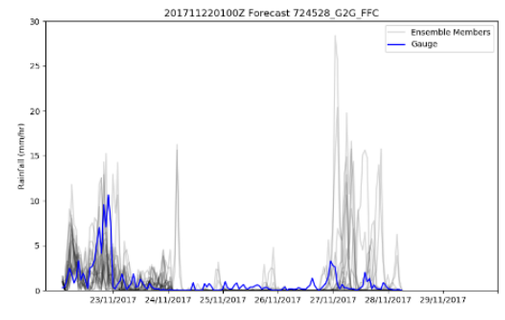
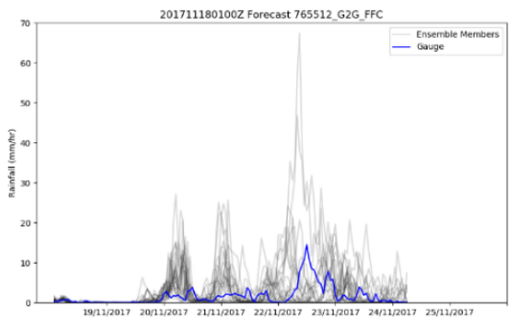
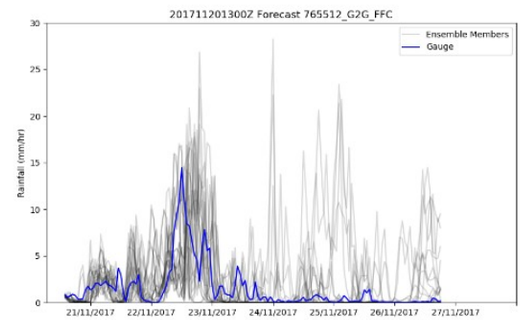
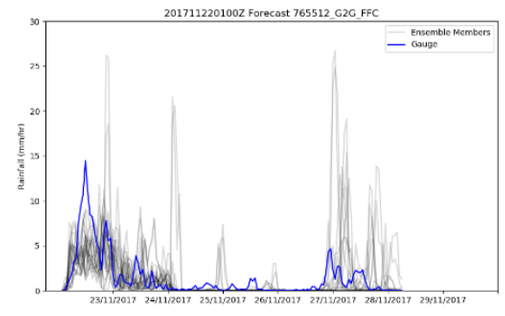
The 3 observation sources show a consensus on the distribution of rainfall. There are some discrepancies around magnitudes, particularly across the West Midlands.

TWPs show NW England with a probability of 1 through all 3 time-windows. In Days 4-6 probabilities of ~0.5 are spread across most catchments of England and Wales. In Days 2-3, probabilities become more focussed on western parts of England and Wales. Probabilities increase to 1 in the Day 1 window in N Wales, W Midlands and W & N Yorkshire.



NW England & N Wales – 22-23 November 2017

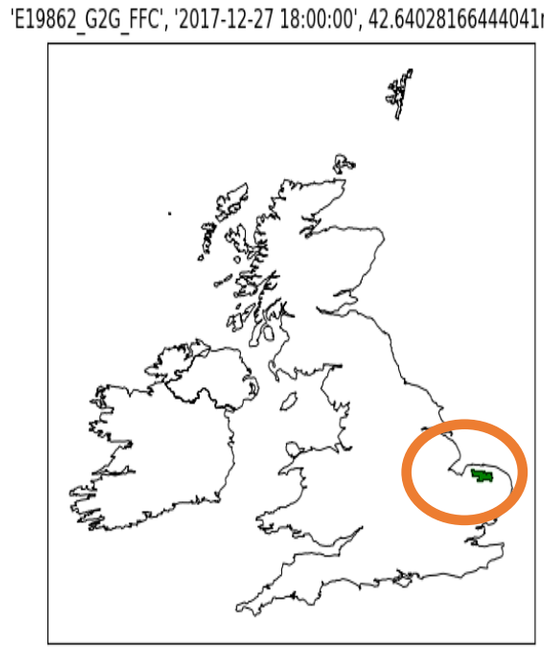
Temporal forecast evolution in highest precipitation catchment and those with river flow responses.

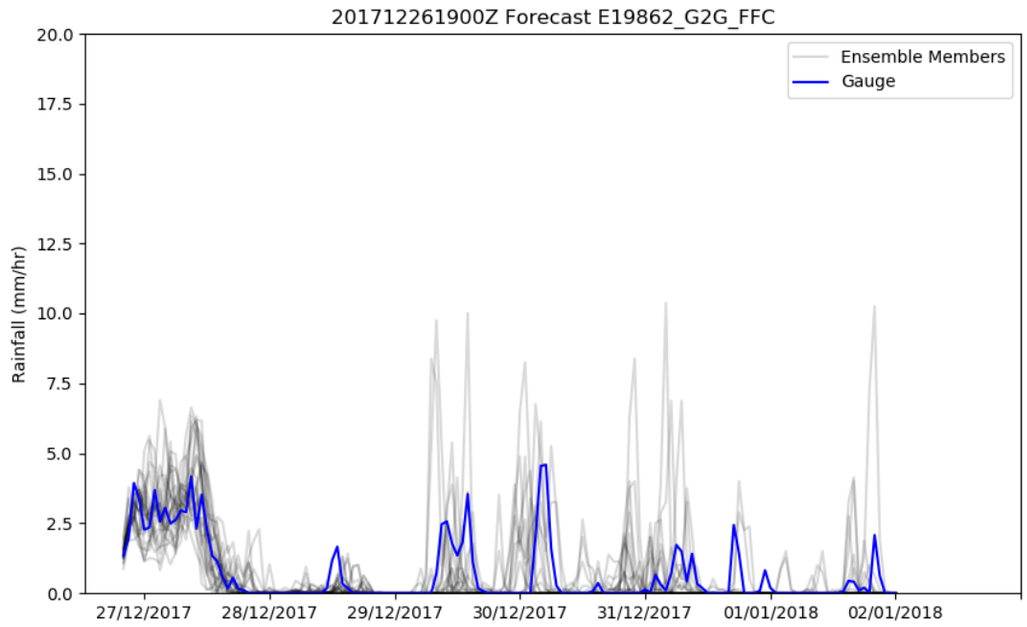
Catchment ID	Comments	Days 4-6	Days 2-3	Day 1
065001_TG_1201	Catchment as identified as having highest 24hr rainfall is in case-study region.			
724528	Consistent with 724427, 724326, 760112, 760101			
765512	Consistent with 760502 All or nothing ensemble spread at Days 4-6, well forecast at Days 2-3. Observed peak just outside the ensemble spread at Day 1.			

E & SE England – 27 December 2017

Case Study Synopsis. No river flood impacts noted. Minor impacts from surface water flooding: Midlands, SW and SE England.

Max Catchment	Date/Time of Max	Rainfall (mm)
E19862_G2G_FFC	27/12/2017 18:00	42.6



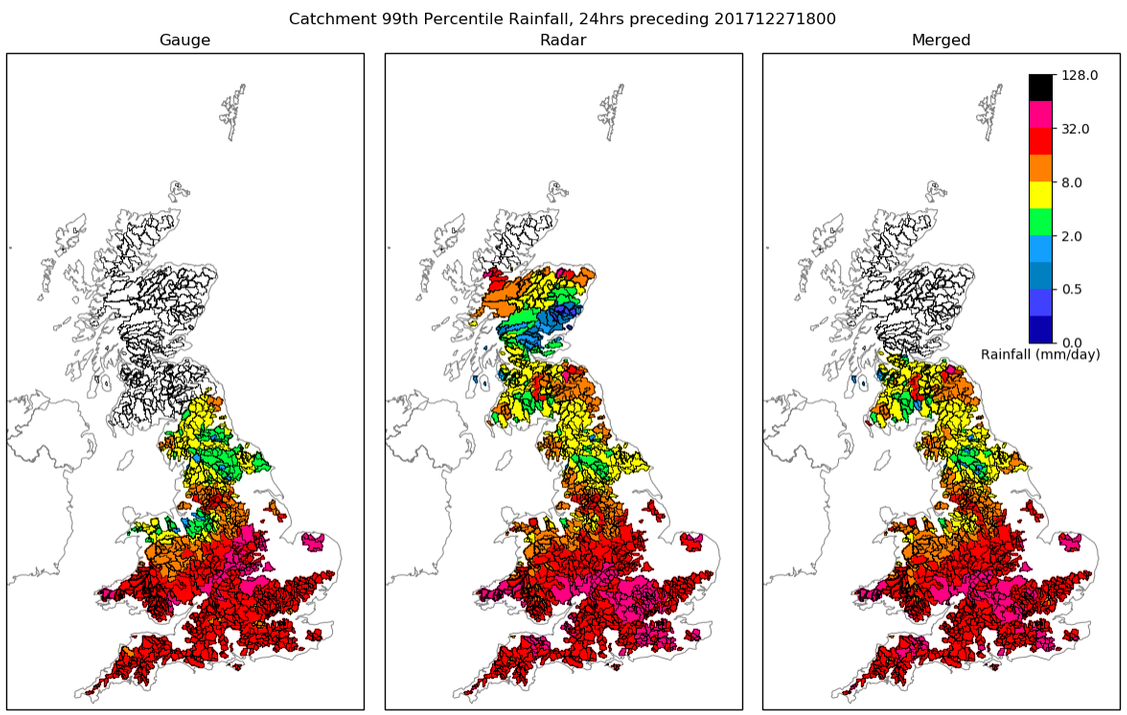


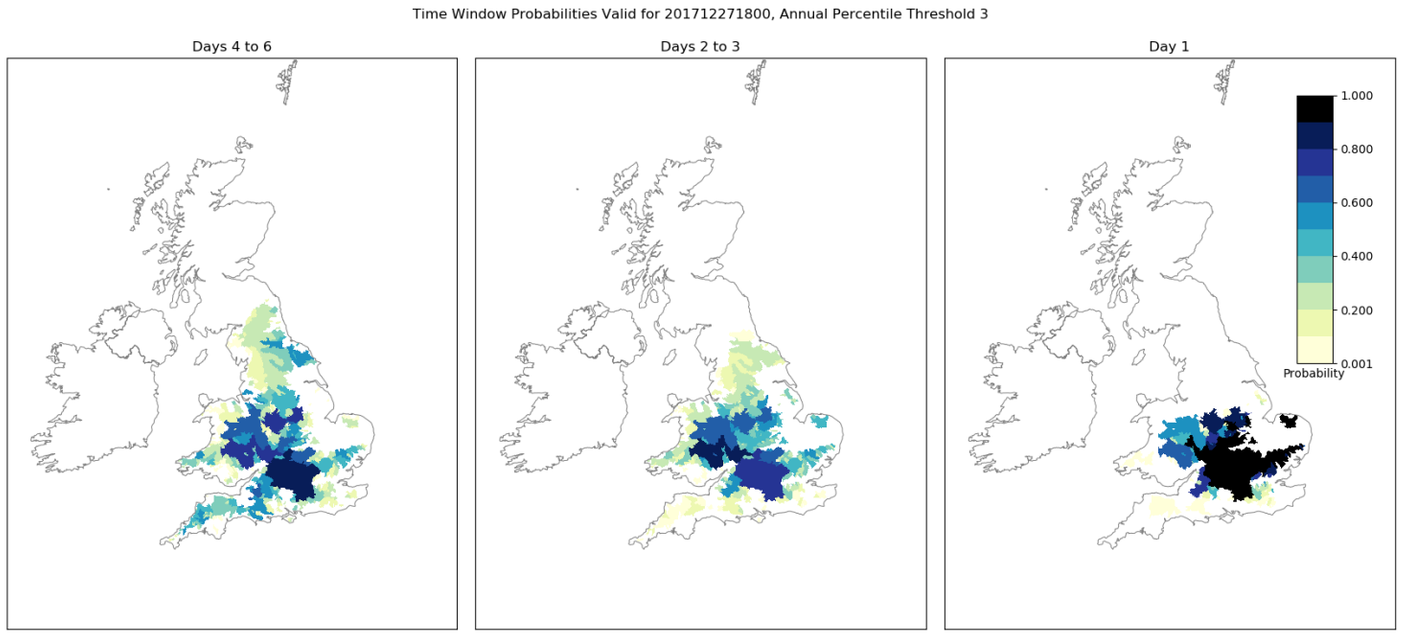
Summary

From the raingauge observations, the catchment receiving the most rainfall was in North Norfolk. Similar rainfall totals can be seen across the areas identified in the case-study synopsis. These vary across the observation sources. Raingauge observations show more rainfall around the Midlands and East Anglia whereas radar shows more in the SE, SW and South.

The forecast for the highest rainfall catchment is good and shows persistent rain. The raingauge observations are almost at the centre of the ensemble spread for the duration of the event.

TWPs identified the areas affected. In the Days 4-6 window probabilities of around 0.5 are spread across the UK, the highest being 0.9 in central southern England. In the Days 2-3 TWPs are of a similar magnitude but less widespread and focussed on SE Wales, West Midlands and central southern England. By Day 1 the highest probabilities have generally shifted eastward and increased to 1 across the South, SE and East.

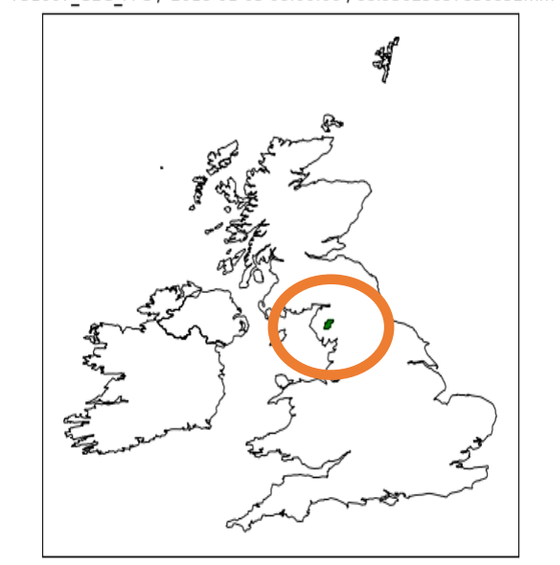


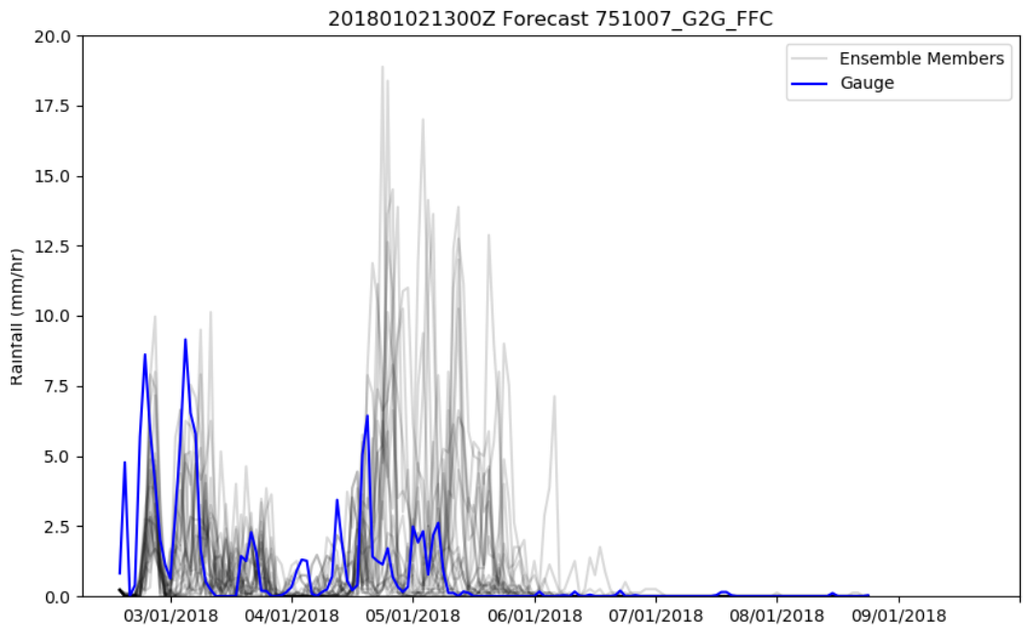


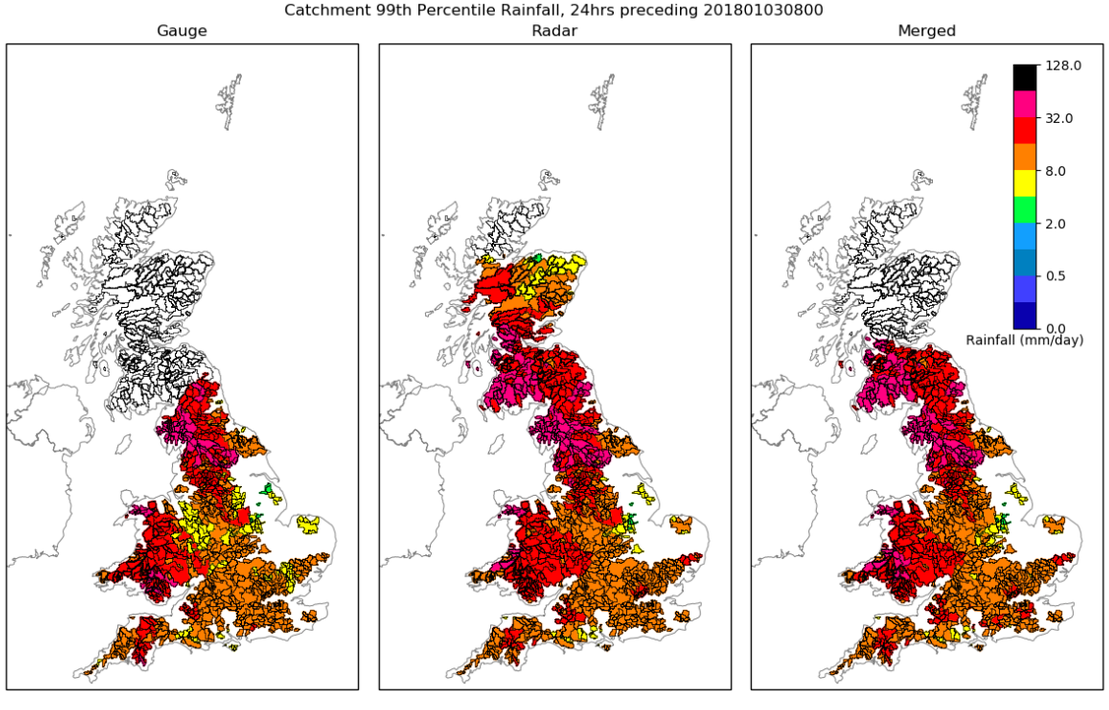
Storm Eleanor – 2-3 January 2018

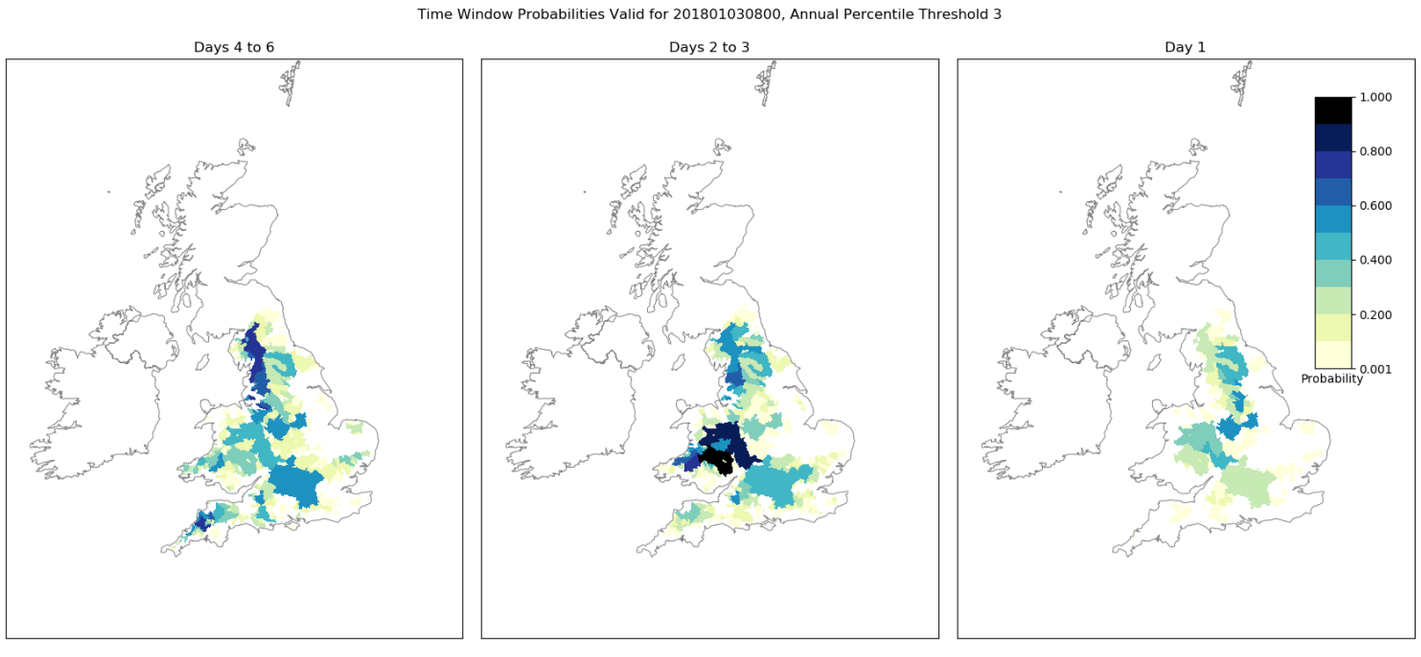
Case Study Synopsis. EFAS Flash Flood notification. No fluvial impacts recorded. Little response in G2G deterministic but larger response in G2G ensembles.

Max Catchment	Date/Time of Max	Rainfall (mm)
751007_G2G_FFC	03/01/2018 01:00	68.9









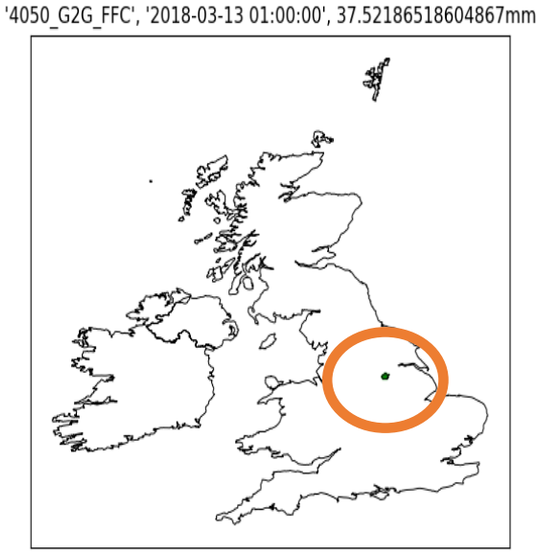
Summary

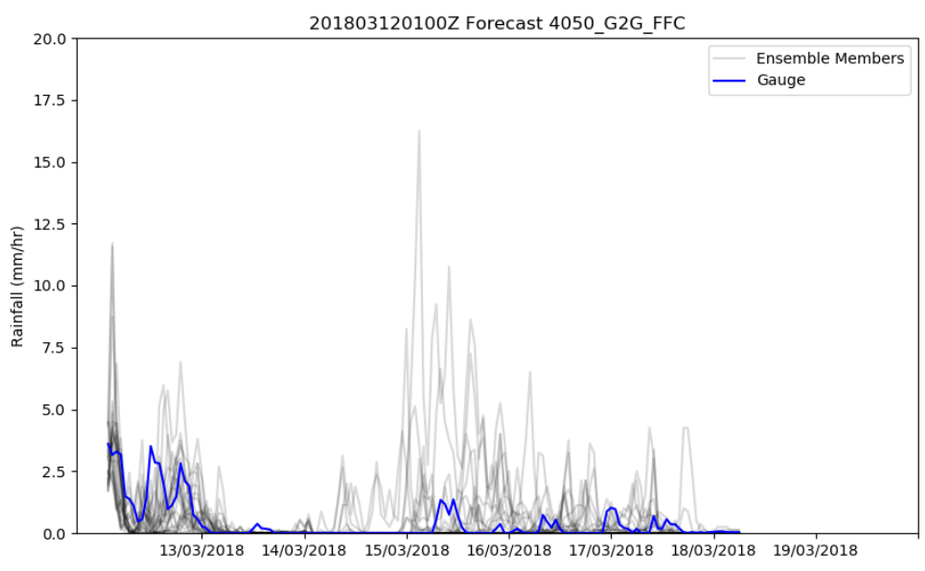
From raingauge observations, the catchment receiving the highest rainfall (~69mm) was in Cumbria/Lake District. The forecast for this catchment generally captures the magnitude of the event. Ensemble members show 2 peaks during the event but the raingauge observations show these peaks occurring slightly earlier. The observations show a ~5mm/h peak just slightly earlier than the 2 main peaks which isn't forecast. The 3 observation sources are largely in agreement on totals and distribution of rainfall. TWPs are highest,~0.7/0.8, in a region of Cumbria/Lancashire in Days 4-6. Probabilities decrease to 0.5 at the most in the region in Days 2-3. The highest TWPs of 1 are focussed on S Wales and West Midlands. By Day 1, probabilities fall across England and Wales. The highest probabilities of 0.6 are in the East Midlands/South Yorkshire. Probabilities around the highest catchment are around 0.3.

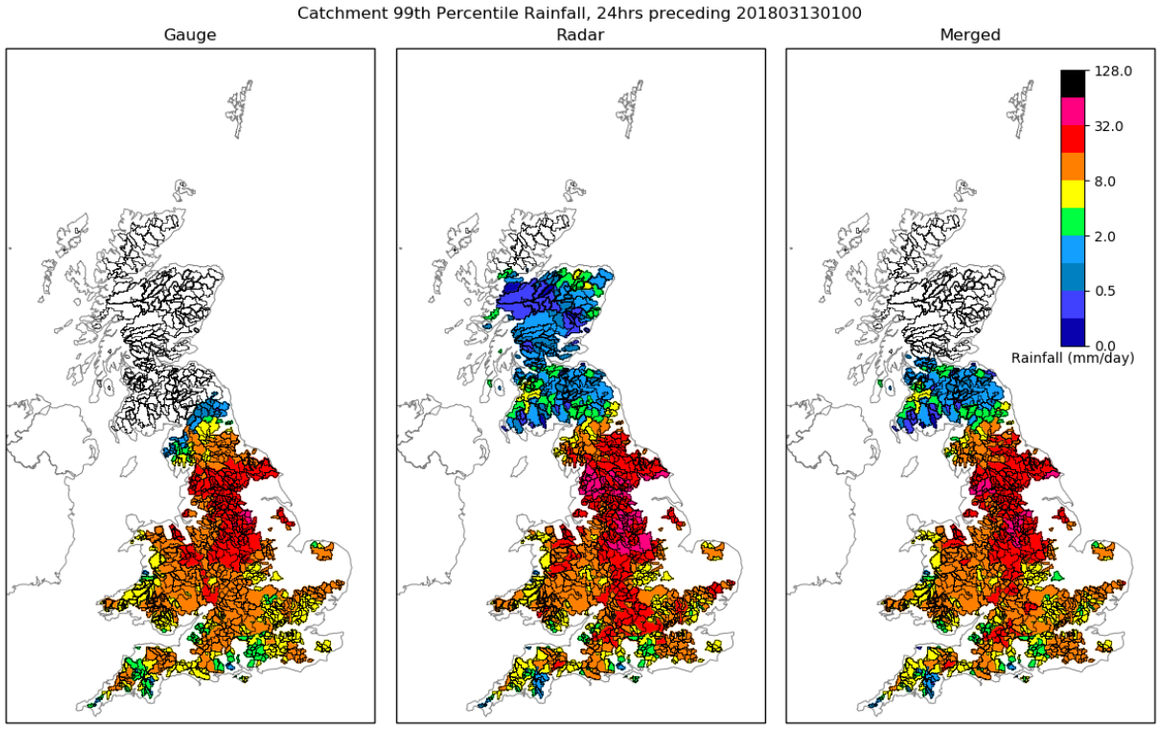
SE & SW England & Derbyshire – 12-14 March 2018

Case Study Synopsis. Widespread flooding around Burton, South Derbyshire. G2G gave poor advice on the fluvial flood risk in Staffs/High Peaks area. Unusually rapid response given the amount of rain, which lead to quite a few minor impacts from surface water and river flooding. Once the rain was in the gauges, G2G then significantly increased the response within the gridded MRDET through this area.

Max Catchment	Date/Time of Max	Rainfall (mm)
4050_G2G_FFC	13/03/2018 01:00	37.5

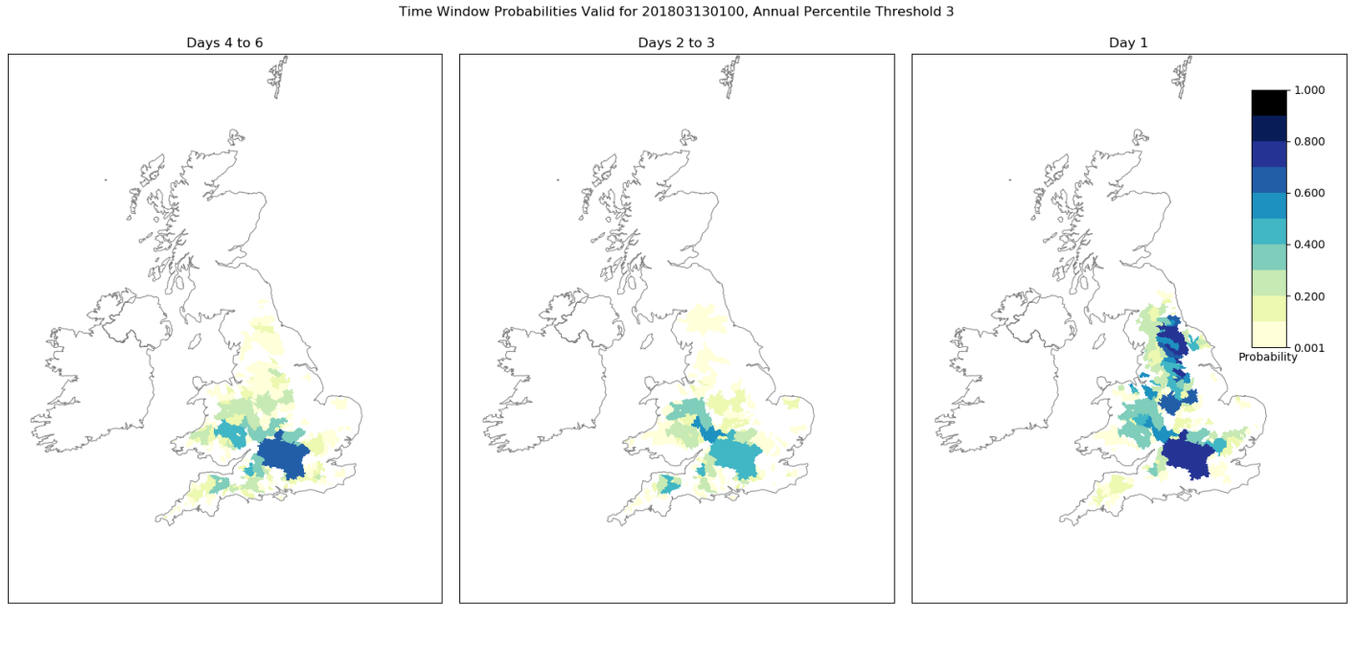






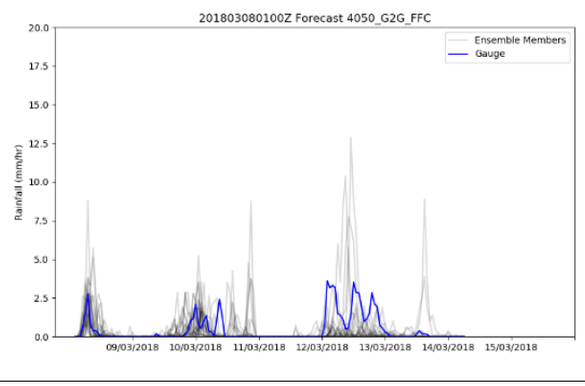
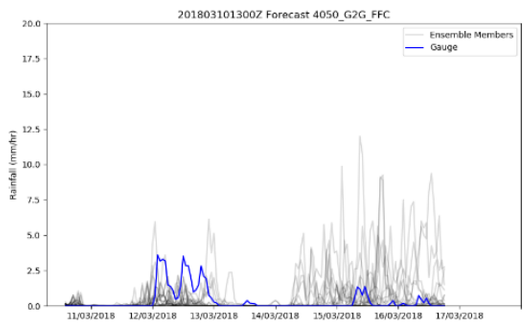
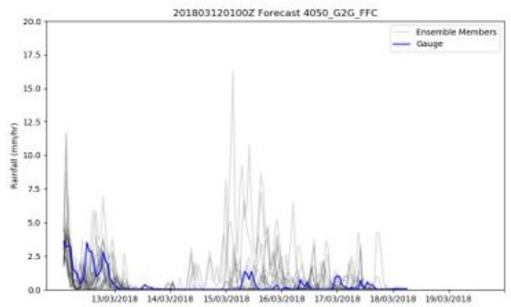
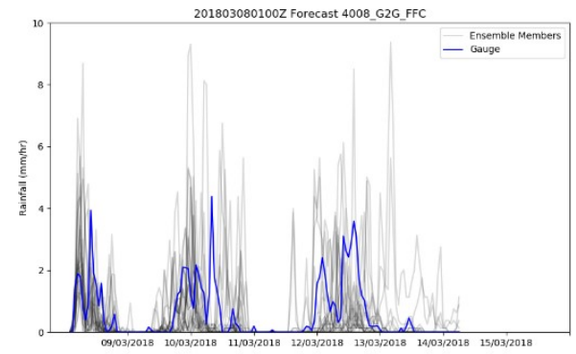
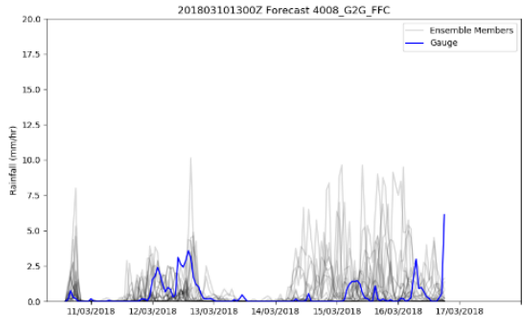
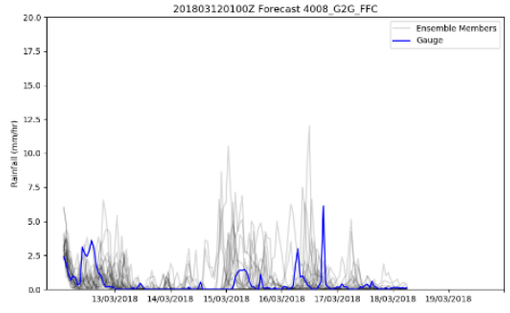
Summary

Catchment with the highest 24hr rainfall total in the region of the case-study synopsis location (Derbyshire/High Peaks). The forecast for this catchment captures the event well. Raingauge observations are almost entirely within the ensemble spread. The observation types are largely in agreement on the distribution of rainfall. Radar observations show higher rainfall totals than raingauges across the Peak District & Lancashire. TWPs focus on Southern England in Days 4-6 with the highest probability being 0.7. Probabilities in the Derbyshire/High Peaks region are around 0.3. In Days 2-3, probabilities decrease to 0.01/0.1, in the Derbyshire/High Peaks. The focus remains on Southern England and a small region around Exmoor with probabilities of between 0.4-0.6 at the most. Probabilities increase up to around 0.8 into Day 1. 0.8 is present in Central Southern England, W North York Moors, S Yorkshire.



SE &SW England & Derbyshire – 12-14 March 2018

Temporal forecast evolution in highest precipitation catchment and those with river flow responses.

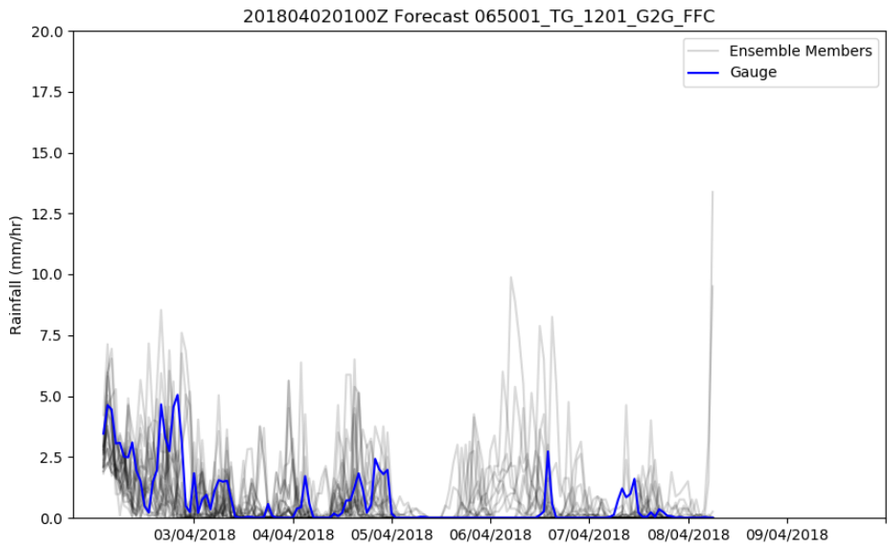
Catchment ID	Comments	Days 4-6	Days 2-3	Day 1
4050	Catchment as identified as having highest 24hr rainfall. Consistent with 4033 G2G & PDMs – except for Days 2-3 when obs are within ensemble spread			
4008				

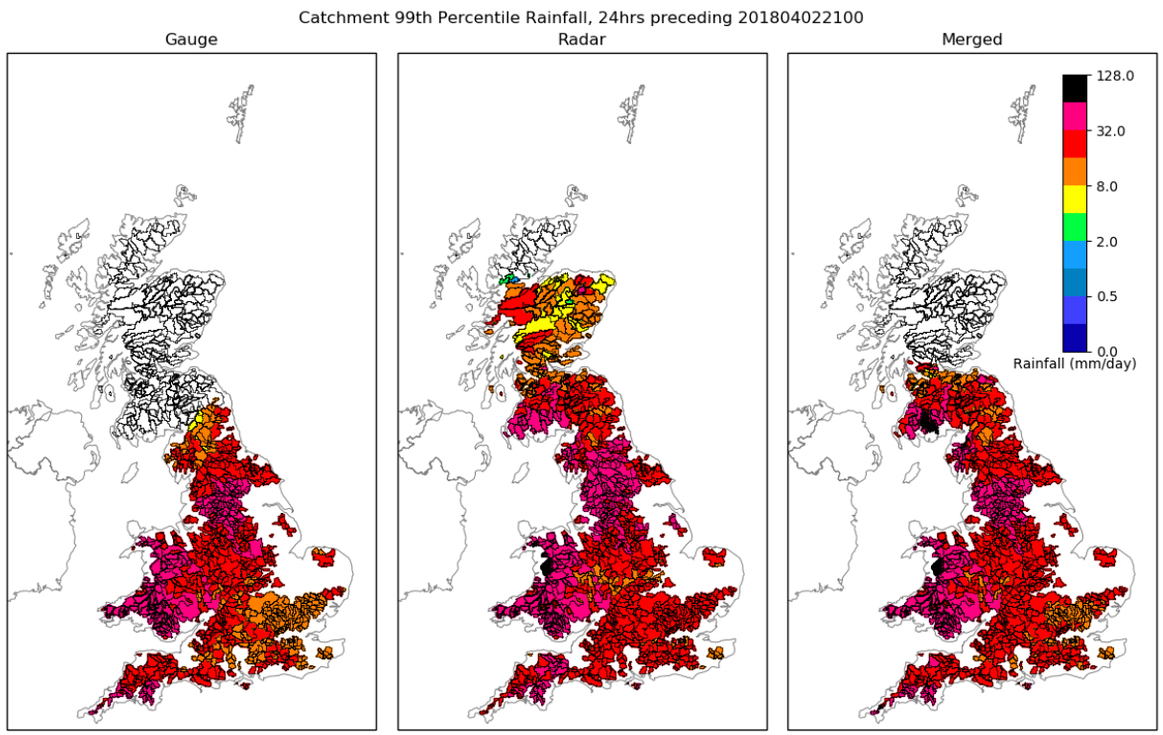
SW/Central/NE England & Wales – 2-4 April 2018

Case Study Synopsis. Minor river flooding impacts noted on 3 April: N and W Yorkshire and on 4 April in N Yorkshire, Durham and Tyne and Wear. Minor roads around Linton-on-Ouse closed due to flooding from small streams and high flows on River Ouse. Surface water flooding caused closures or partial closure of arterial A roads around Bishops Auckland (Durham) and Tyne & Wear area.

Max Catchment	Date/Time of Max	Rainfall (mm)
065001_TG_1201	02/04/2018 21:00	60.4

Highest rainfall catchment Glaslyn at Beddgelert (065001_TG_1201)

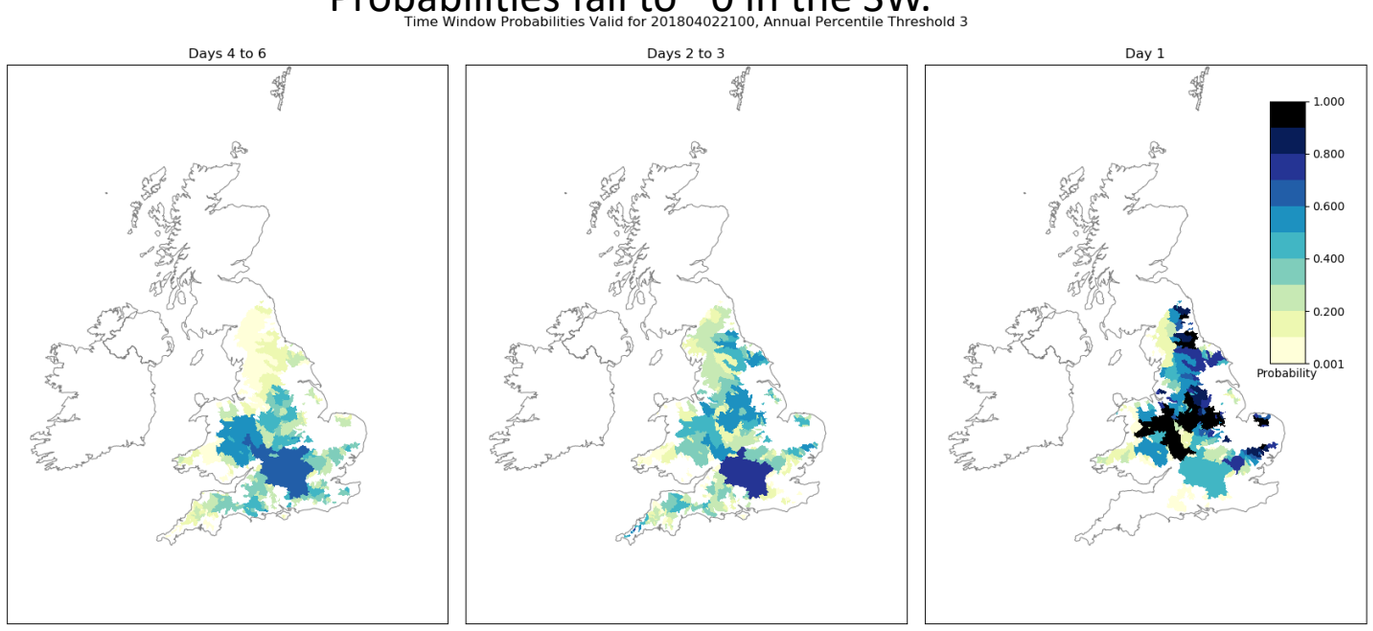




Summary

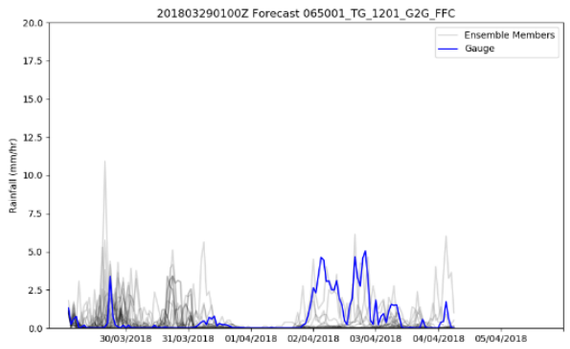
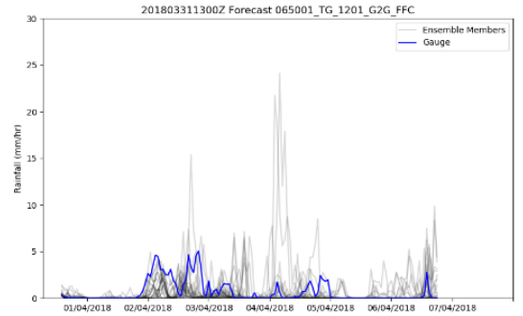
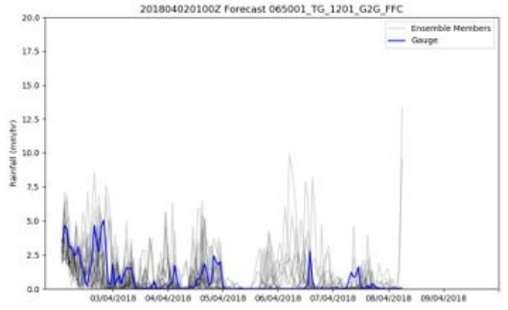
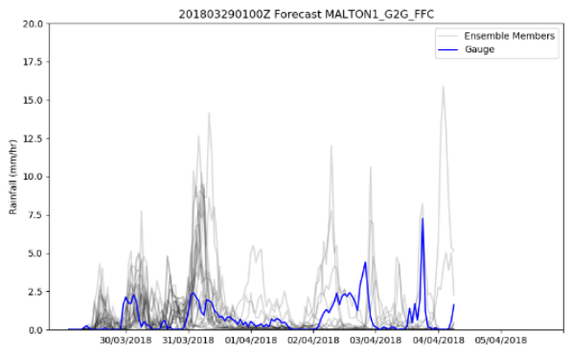
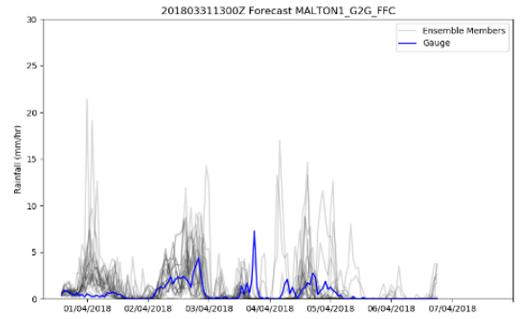
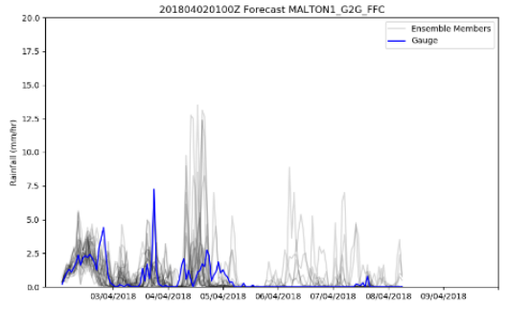
Catchment with the highest 24hr rainfall total in the region of the case study synopsis location (Wales). The forecast for this catchment captures the event well, observations are within the ensemble spread for all of the event. The ensemble spread is fairly large (0 to 7.5mm) for the latter parts of 2 April.

The observation sources are largely in agreement on the distribution of rainfall. Some small discrepancies can be seen in the rainfall totals, particularly in SE England where raingauge observations are lower than radar. A band across the West Midlands shows higher observations from gauges than radar. TWPs for Days 4-6 show the highest probabilities of 0.7 in central S England with similar probabilities into Wales, Midlands & South Coast. In the Days 2-3 window, probabilities become better correlated with the affected regions. Probabilities in the NE & Cornwall increase to ~0.6. In the Day 1 window, probabilities increase to 1 in affected areas of the NE, Wales and Central England. Probabilities fall to ~0 in the SW.



SW/Central/NE England & Wales – 2-4 April 2018

Temporal forecast evolution in highest precipitation catchment and those with river flow responses.

Catchment ID	Comments	Days 4-6	Days 2-3	Day 1
065001_TG_1201	Catchment as identified as having highest 24hr rainfall. Consistent with Glaslyn PDM			
MALTON1	Consistent with MARISH1, Nunnington G2G, F2581PDM			

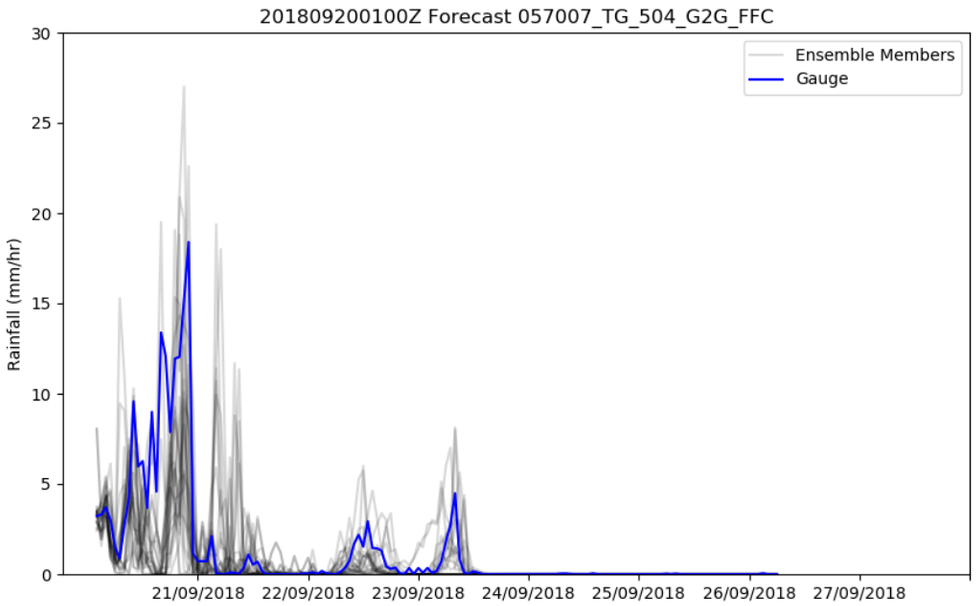
Storms Ali & Bronagh – 20 September 2018

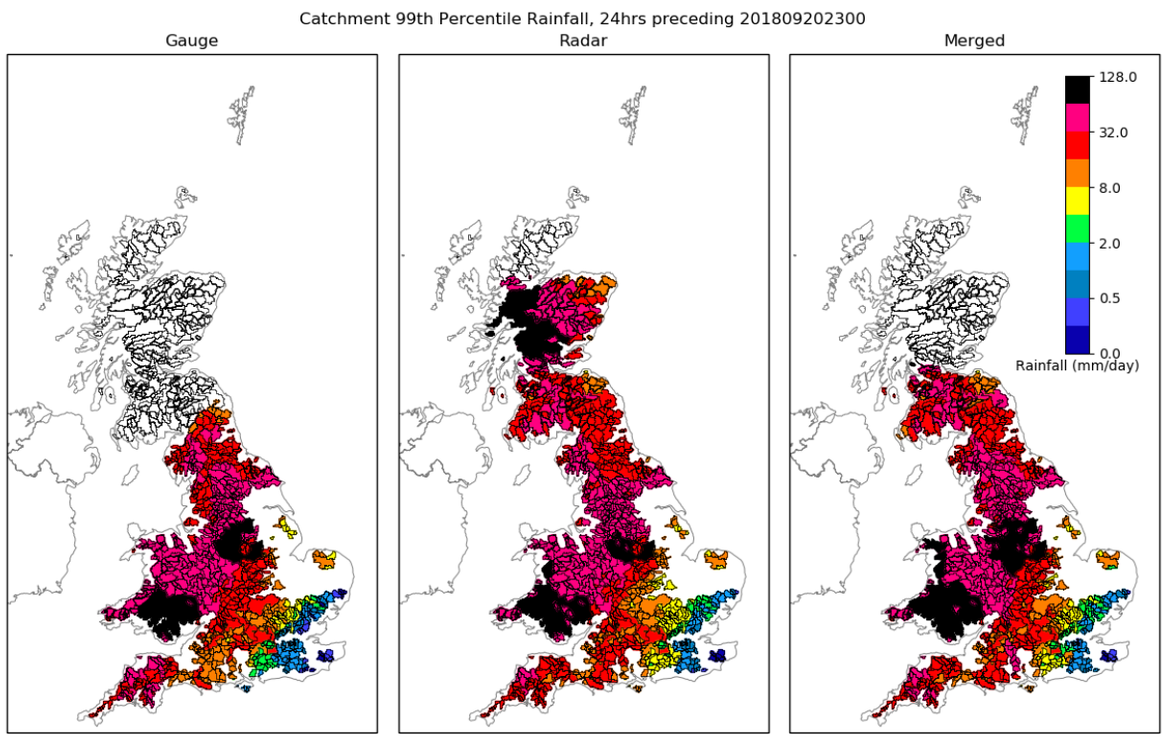
Case Study Synopsis. Primarily, surface water flooding. Transport disruption in Sheffield and Rotherham. Some road flooding in Wales (Pontypridd).

Max Catchment	Date/Time of Max	Rainfall (mm)
057007_TG_504	20/09/2018 23:00	142.6

Taff at Fiddlers Elbow (057007_TG_504)

S Wales





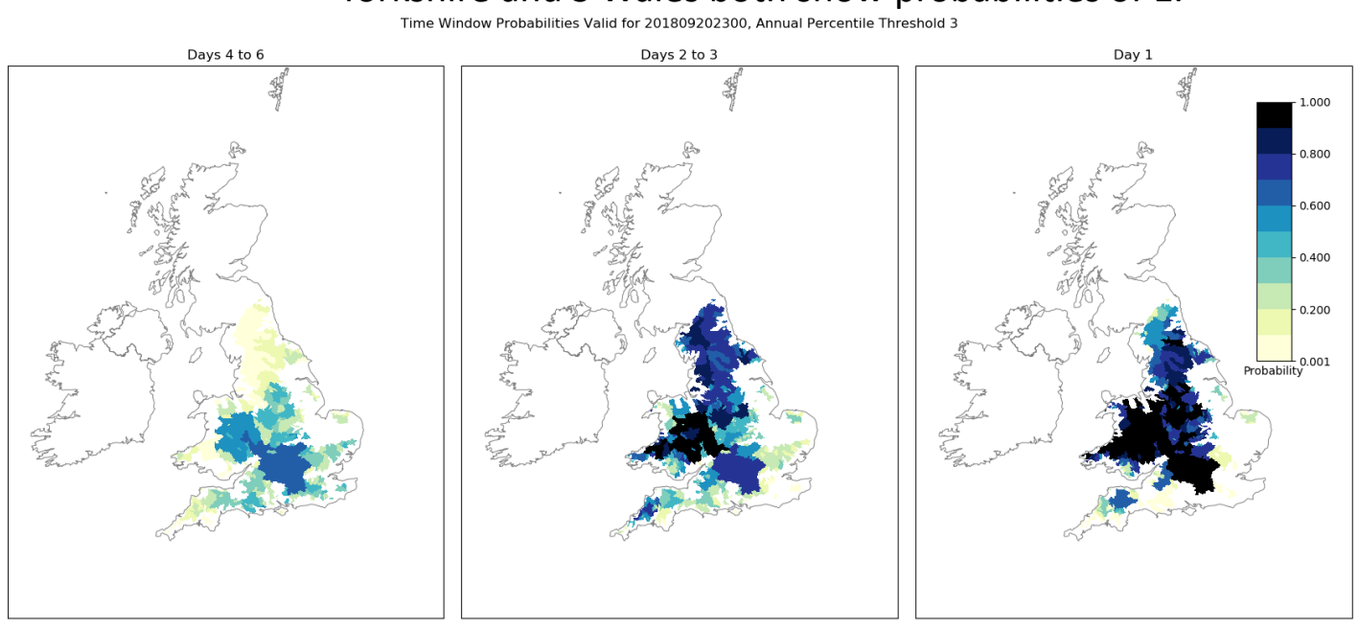
Summary

Highest rainfall catchment well correlated with case-study impact region of S Wales.

The forecast for this catchment captures the event reasonably well. Ensemble members tend to converge on 2 peaks in rainfall after an initial peak of just less than 5mm/hr. The raingauge observations largely follow this pattern and are within the ensemble spread. One difference is the raingauge observations do not fall to ~0 mm/hr around midday on 20 Jan. Raingauge observations fall but then continue to rise to the main peak of ~18mm/h.

The 3 observation sources are in good agreement on the distribution of rainfall. Some small discrepancies can be seen around the locations of the highest rainfall totals.

TWPs in Days 4-6 are focussed on central southern England and Eastern Wales with probabilities of 0.6-0.7. Probabilities of 0.5-0.6 can also be seen in south Yorkshire and parts of the Midlands. In Days 2-3, probabilities increase across the country. The biggest increases can be seen across northern England, SW Wales and N Devon/Cornwall. Probabilities of 1 are present across S Wales & W Midlands. Into Day 1, the prevalence of TWPs of 1 increases. The affected areas of S Yorkshire and S Wales both show probabilities of 1.

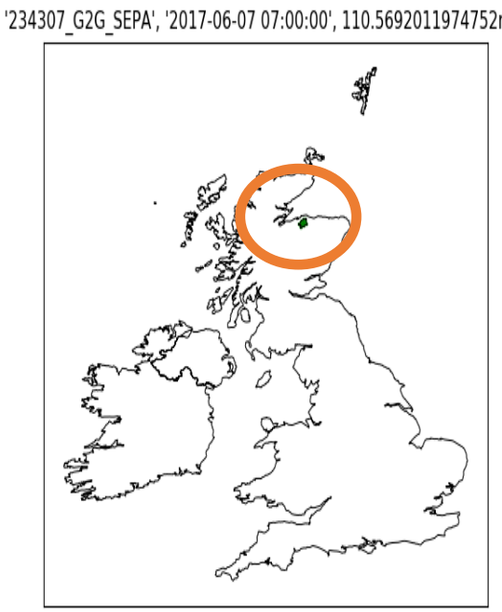


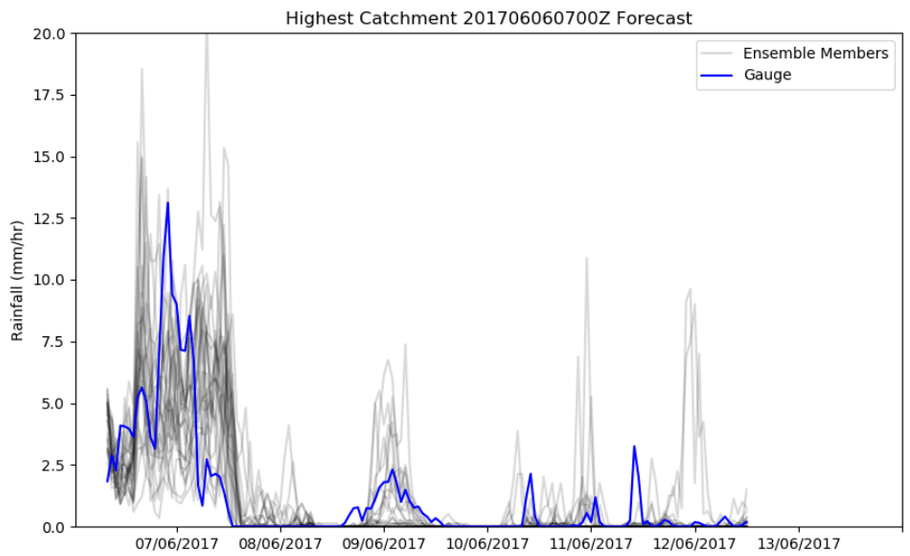
Scotland

6 & 7 June 2017 - Findhorn, Lossie and Nairn catchments

Case Study Synopsis. Hydrologically significant event. FGS 3x2 (low,sig). Flood defences at Forrest and Elgin prevented flooding in these places. River Nairn came close to overtopping defences.

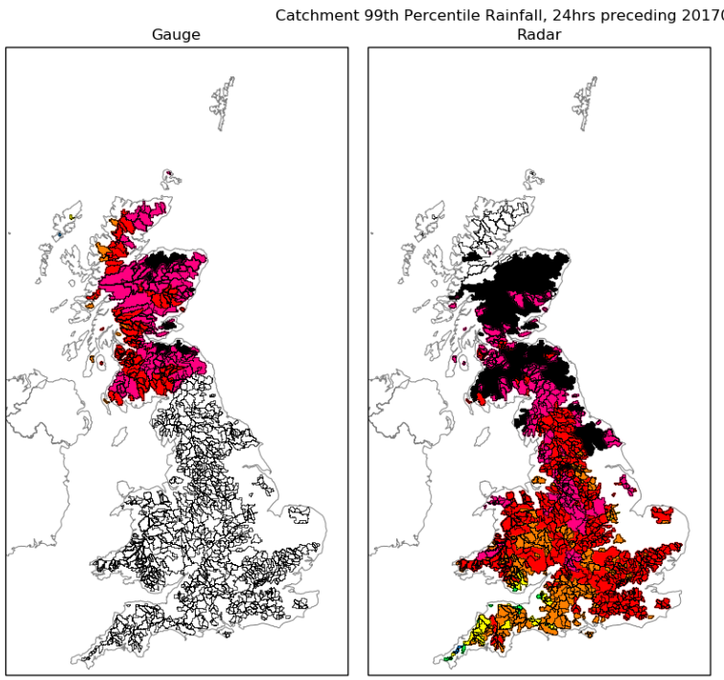
Max Catchment	Date/Time of Max	Rainfall (mm)
234307_G2G_SEPA	07/06/2017 07:00	110.6



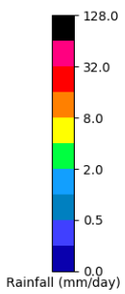


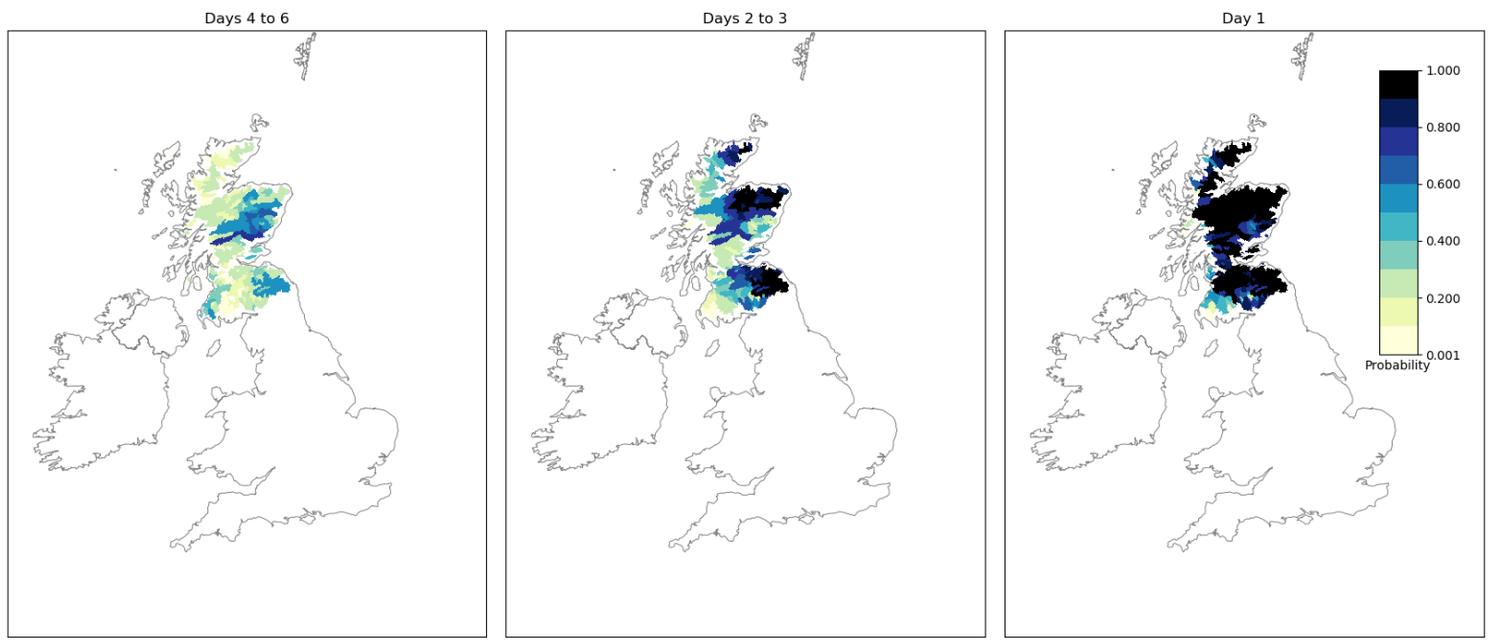
Summary

Catchment with the highest 24hr rainfall total is in the region of recorded impacts. The forecast for the highest catchment captures the general timing of the event. Raingauge observations are within the ensemble spread but this is large, between ~1.5 and 20mm at times. The forecast shows 2 peaks in rainfall with the observed rainfall occurring between them. Between the raingauge and radar observations, there is agreement on rainfall amounts in the case-study region. Outside of the region, the radar shows higher totals. Time-Window Probabilities are between 0.6 and 0.1 in the case-study region at Days 4-6. In the Days 2-3 window, probabilities increase to 1 across the region with similar probabilities to the SE of Scotland. By Day 1 probabilities of 1 are much more widespread around the case study region and SE.



Problem with Scottish merged file for this date

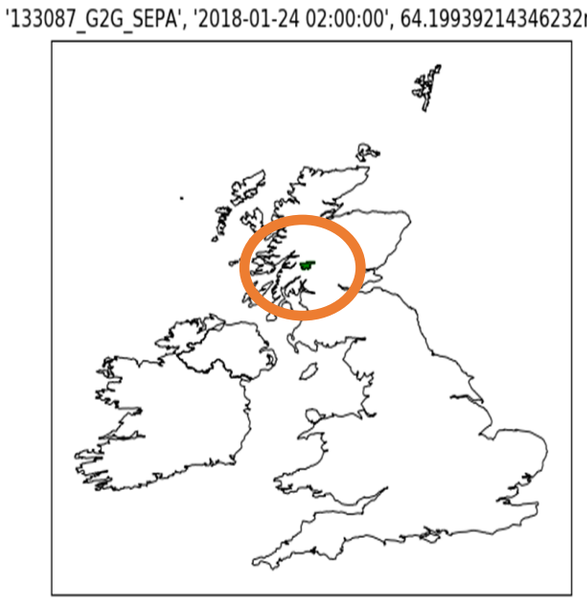


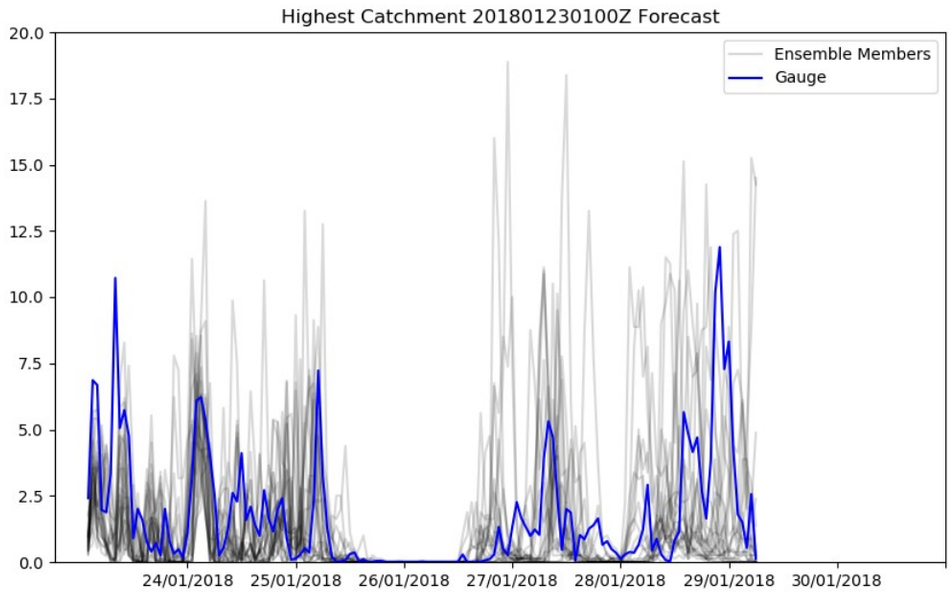


23 to 25 January 2018 - Scottish Borders

Case Study Synopsis. Several linked events. Snowmelt overnight 23-24 January.

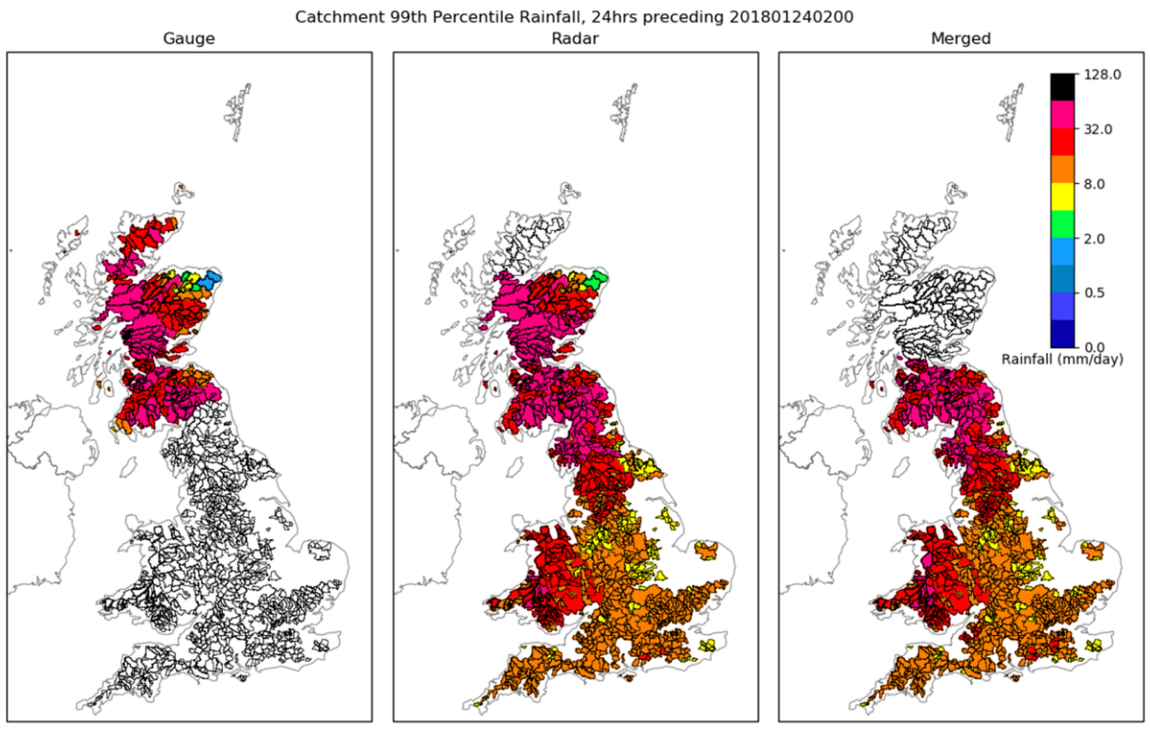
Max Catchment	Date/Time of Max	Rainfall (mm)
133087_G2G_SEPA	24/01/2018 02:00	64.2

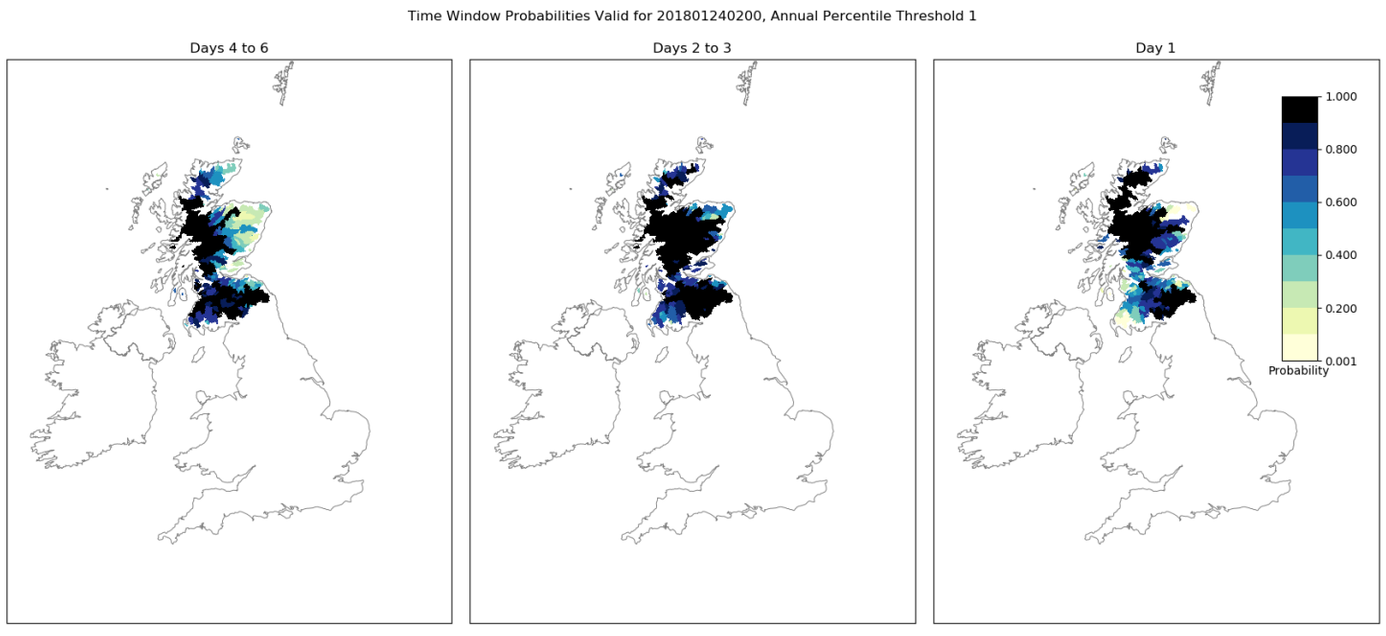




Summary

The catchment receiving the highest 24hr precipitation total was not in the region of the case-study impacts. Could suggest snowmelt was an important factor. The forecast for the highest rainfall catchment under-forecasts 2 peaks early on 23 Jan but captures their timing. After these peaks, forecast is accurate with raingauge observations within the ensemble spread. The different observation sources are mostly in agreement on rainfall totals and distribution. Differences are greatest along the SE and NE coasts where the raingauges give lower totals. TWPs identify areas of high rainfall from Days 4-6. In Days 4-6, probabilities are 1 to the W and S of Scotland and in some areas of the Scottish Borders. In Days 2-3, probabilities of 1 become much more widespread around the same areas. By Day 1, the high probabilities of 1 are focussed on the Scottish Borders and the West of Scotland (including the area around the catchment receiving the highest daily rainfall). Probabilities drop to ~0 on the NE coast, a region which received little rainfall on both the raingauge and radar obs.





Rainfall and River Flow Ensemble Verification: Phase 2
Case study analysis: hydrological impacts, rainfall and river flow time-series
Final Report Appendix C.3.1

1 Overall approach

1.1 Selection of case-studies and associated catchments

A list of potential case-studies was provided by the Project Board (FFC, EA, SEPA, NRW) based on operational experience and relevance. Associated catchments were then selected based on observed and/or modelled (G2G) river flow and rainfall, and consideration of any reported impacts. Note that the catchments with observed river flow available for this project are not the full set used by G2G operationally: this further constrained which catchments were selected. More details of the selection process are given below.

The selection of case-studies and associated catchments is based on the following information.

- G2G modelled river flows produced using observed rainfall, state-updating and flow-insertion
- Hydrological Summary for the United Kingdom (UKCEH, 2021)
- Guidance on case-study selection from the Project Board
- Locations of maximum rainfall (calculated by G2G domain (England & Wales or Scotland) for 24h raingauge accumulations)
- Flood Guidance Statement Verification (FGSV) Observed impacts summary

The main focus with regard to river flow is on the England & Wales and Scotland case-studies having river flooding impacts in order to demonstrate the prototype operational displays and their benefit. From a precipitation perspective, the focus is both on catchments which have the maximum 24h rainfall over the G2G domain being considered (England & Wales or Scotland) and also on catchments showing river flooding impacts. Note that some case-studies had limited or no noticeable river flow responses (for the catchments available) so only have catchments selected for a precipitation analysis.

Catchments used in the analysis are referred to by catchment name, G2G (NFFS/FEWS) ID, and G2G catchment area (noting that this may differ slightly from the NRFA catchment area). Where a PDM for a catchment is also available, the PDM (NFFS/FEWS) ID is also given.

1.2 Summary of associated documents and additional information

For reference, this document first provides a guide to the diagrams and displays used for case-study analysis (Section 2). This is followed by a discussion of the catchments considered for each case-study and the individual case-study analyses (Section 3). Then the key conclusions from across all case studies is given in Section 4. The full set of case-study plots can be found in the documents below.

River flow case-study plots documents, Appendix C.3

Precipitation case-study plot document, Appendix C.2

2 Presentation of case-study results including verification information in real-time displays

2.1 River flow diagrams and displays

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

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

An example diagram for analysing the ensemble hydrograph behaviour and threshold exceedance at a particular site is shown in Figure 2.1. For one forecast (selected, for example, to cover a specific time) the ensemble member hydrographs are plotted with one colour per ensemble member, with colours selected to match those used for the 24-member Storm Surge ensemble. If available (i.e. for post-event analysis) the observed flows are plotted in black to allow the ensemble performance to be visually assessed. Flow thresholds Q(2)/2, Q(2), Q(5) and Q(50) are indicated by horizontal black dashed lines, when exceeded by at least one ensemble member, or by the observations. If forecasts have been selected to analyse performance at a specific time of interest, this time is shown by a vertical black dashed line.

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

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

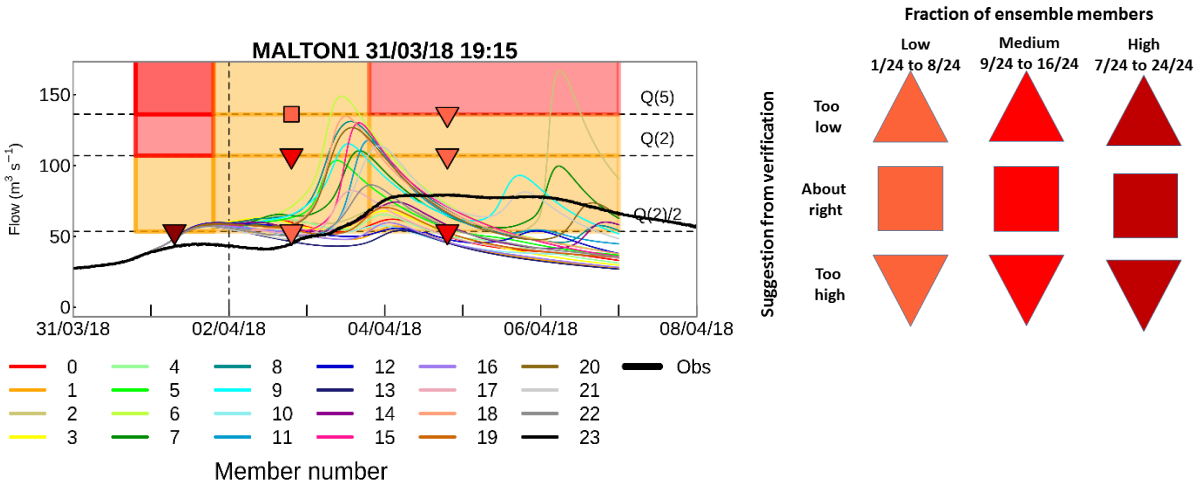


Figure 2.1 Example hydrograph used to place the ensemble threshold-exceedance in the context of the ensemble verification information. The catchment shown is Derwent at Malton (Malton1, NE England) for a forecast time-origin of 19:15 31 March 2018.

An example diagram for placing the ensemble dispersion in the context of climatological ensemble dispersion for a given site is given in Figure 2.2. Here, the Coefficient of Variation (CV) is used as a dimensionless measure of ensemble dispersion. It is defined as the ratio of the ensemble Standard Deviation (spread), σ, to the ensemble mean, $\bar{y}$, such that:

$$\mathrm{CV} = \frac{\sigma}{\bar{y}}. \qquad (1)$$

The Coefficient of Variation is calculated separately for each time-step in each forecast. For the individual forecast considered, the CV is plotted in red as a function of forecast lead-time in Figure 2.2. To calculate the climatological CV, the average is taken (separately at each forecast lead-time) of the CV values for all forecasts at the site of interest over the Phase 2 Period 1 (September 2017 to 31 August 2018). This is plotted in black. Thus, when the red line in Figure 2.2 is above the black line, the individual ensemble forecast is more spread than the reference climatology.

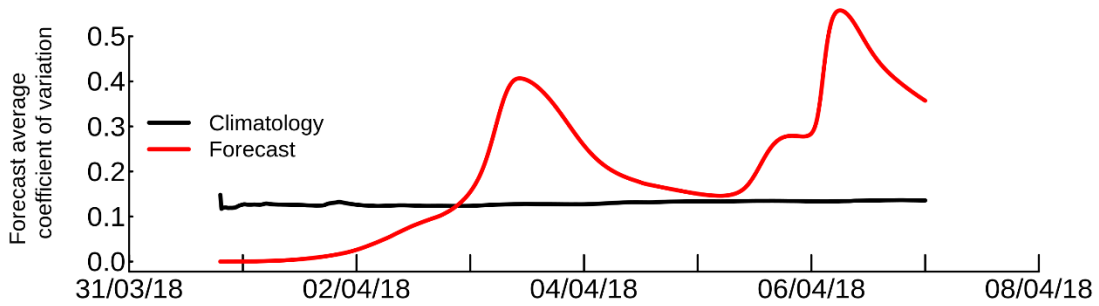


Figure 2.2 Example display of Coefficient of Variation of the ensemble forecast against forecast lead-time (given as the forecast time) for one ensemble forecast, placing the forecast ensemble spread in the context of climatological spread. The forecast time-origin is 19:15 31 March 2018.

To analyse the threshold-exceedance from a regional perspective, maps are drawn showing the threshold-exceedance for each site within a given region, for each threshold and lead-time range considered. An example for the North East of England region is shown in Figure 2.3. The symbol at each site indicates the direction of any correction suggested by the Reliability Diagram, and the colour

shows the fraction of ensemble members exceeding the threshold. The same symbols and colours are used as were used for the hydrographs in Figure 2.1.

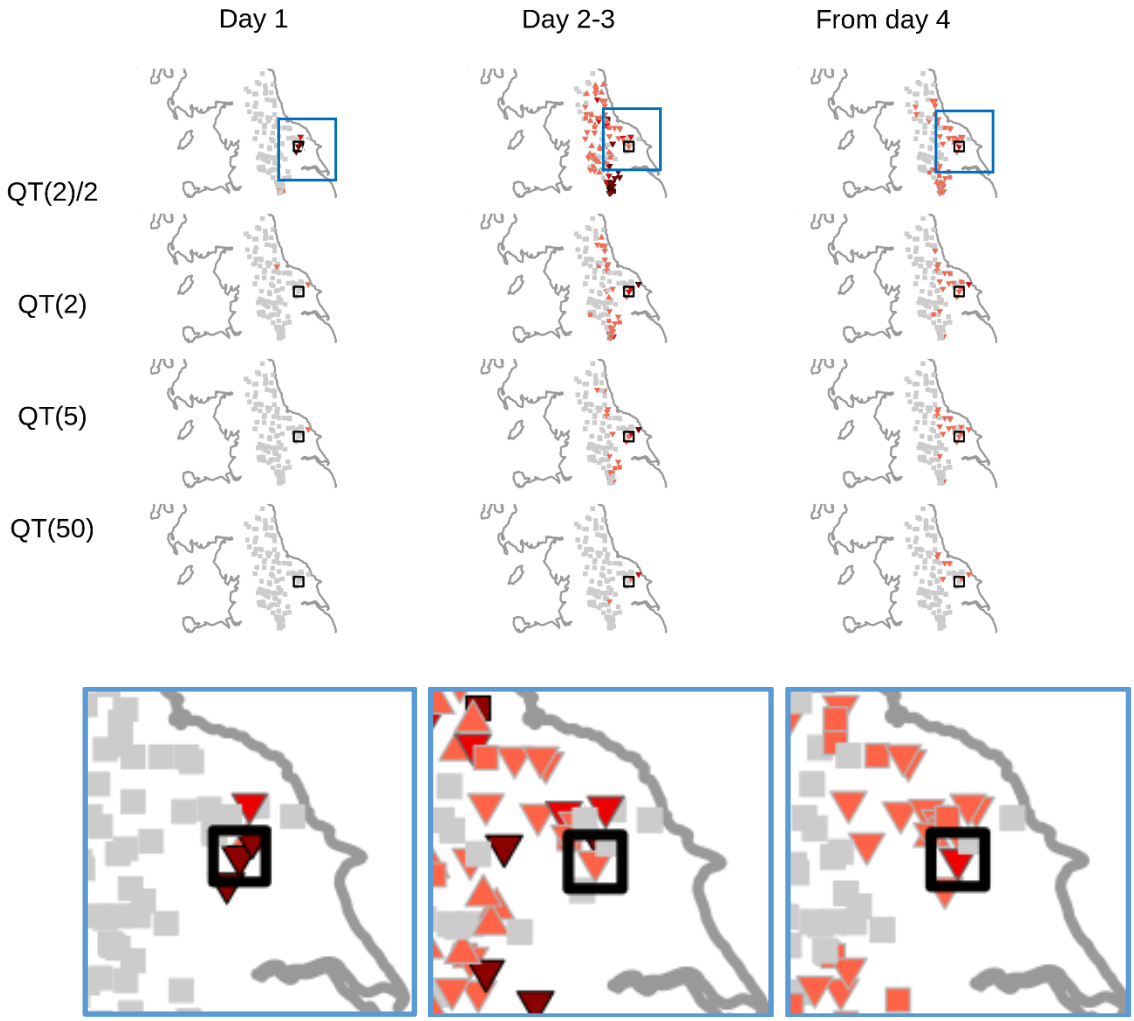


Figure 2.3 Example maps showing the variation in ensemble probability of river flow threshold-exceedance for a particular forecast over a given region (top panels) and zooming in on the catchment of interest (bottom panels). The time-origin of the forecast is 19:15 31 March 2018.

2.2 Precipitation diagrams and displays

Precipitation diagrams and displays for each case-study are presented and discussed in Appendix C.2.

To allow a direct comparison to be made between precipitation time-series and river flow hydrographs for catchments identified as having river flow impacts, additional catchment-precipitation time-series have been produced. These are included where appropriate in the analysis presented below in Section 3.

3 Analysis of case-studies and associated catchments

3.1 Case studies over England & Wales

The case-studies selected for England & Wales are analysed below in chronological order. Each analysis includes identifying features of the case-study, the associated catchments chosen, and key conclusions drawn from the results.

First, Table 1 lists the catchments considered for each case-study based on (i) the observed and/or modelled (G2G) river flow response (middle column), and (ii) the peak 24h raingauge rainfall total (right column). For consistency, throughout this report catchments are referenced as "River at Site Name (G2G ID)", where the G2G ID is the NFFS/FEWS ID. In Table 1, catchments that are indented are upstream of the preceding un-indented catchment. Note that the catchment with the peak 24h rainfall total does not necessarily overlap with the catchments selected based on river flow response, or with the areas reported as having rainfall or flood impacts.

Table 2 gives further catchment information for case-study catchments that have a river flow analysis. This includes National River Flow Archive (NRFA, nrfa.ceh.ac.uk) details if available. Note several catchments are used in more than one case-study and have been grouped accordingly and their locations indicated in Figure 1.

Table 1 Catchments used for case-studies over England & Wales. The catchments in the central column were analysed for both river flow and precipitation, the catchments in the right-hand column were analysed for precipitation only. The exception to this is the 27 December 2017 case-study catchment Boyd at Bitton (53131 & PDM)* which was only analysed for river flow.

Case-study	Catchments selected based on river flow response	Catchment with peak 24h rainfall total
2017		
18 Jul		Walkham at Horrabridge (47118)
9 Aug		Gypsey Race at Boynton (Boyntn1)
23 Aug		Derwent (NE) at Low Marishes (MARISH1)
30 Sep		Kent at Sedgwick (730511)
21 Oct	Irwell at Irwell Vale (690140) Calder at Hebden Bridge (HEBDBR1) Calder at Todmorden (TODMDN1) Glaslyn at Beddgelert (065001_TG_1201 & PDM)	Glaslyn at Beddgelert (065001_TG_1201 & PDM)
3-4 Nov		Moors River at Hurn Court (43214)
22-23 Nov	Lune at Caton (724629) Wenning at Hornby (72452) Hindburn at Wray (724427) Wenning at Wennington (724326)	Glaslyn at Beddgelert (065001_TG_1201 & PDM)

	Eden at Sheepmount (765512) Eden at Temple Sowerby (760502) Eden at Gt Musgrave Bridge (760112) Eden at Kirkby Stephen (760101) Glaslyn at Beddgelert (065001_TG_1201 & PDM)	
27 Dec	Boyd at Bitton (53131 & PDM)*	Wensum at Costessey Mill (E19862)
2018		
2-3 Jan		Derwent at Portinscale (751007)
12-14 Mar	Dove at Rocester Weir (4008) Dove at Hollinsclough (4033 & PDM) Torne at Auckley (4050)	Torne at Auckley (4050)
2-4 Apr	Derwent at Malton (Malton1) Derwent at Low Marishes (MARISH1) Riccal at Nunnington (Nunnington & PDM) Glaslyn at Beddgelert (065001_TG_1201 & PDM)	Glaslyn at Beddgelert (065001_TG_1201 & PDM)
20 Sept		Taff at Fiddlers Elbow (057007_TG_504)

Table 2 Catchment information for case-study catchments over England & Wales that have a river flow analysis.

Case-study	Site	G2G ID (NFFS/FEWS ID)	Region	River	G2G area (km^2)	NRFA ID	NRFA area (km^2)
21 Oct 17	Irwell Vale	690140	NW	Irwell	103	69022	101
	Hebden Bridge	HEBDBR1	NE	Calder	74		
	Todmorden	TODMDN1	NE	Calder	18		
21 Oct 17 22-23 Nov 17 2-4 Apr 18	Beddgelert	065001_TG_1201	WA	Glaslyn	68	65001	68.6
22-23 Nov 17	Caton	724629	NW	Lune	984	72004	983
	Hornby	724528	NW	Wenning	230	72807	232
	Wray	724427	NW	Hindburn	83		
	Wennington	724326	NW	Wenning	140	72009	142
	Sheepmount	765512	NW	Eden	2274	76007	2286.5
	Temple Sowerby	760502	NW	Eden	618	76005	616.4
	Gt Musgrave Bridge	760112	NW	Eden	225		
	Kirkby Stephen	760101	NW	Eden	68	76014	69.4
27 Dec 17*	Bitton	53131	SW	Boyd	48	53017	47.9

12-14 Mar 18	Rocester (Dove)	4008	MI	Dove	398	28008	399
	Hollinsclough	4033	MI	Dove	8	28033	8
	Auckley	4050	MI	Torne	128	28050	135.53
2-4 April 18	Malton	MALTON1	NE		1407		
	Low Marishes	MARISH1	NE	Derwent	468	27087	457.5
	Nunnington	Nunnington	NE	Riccal	53	27093	40

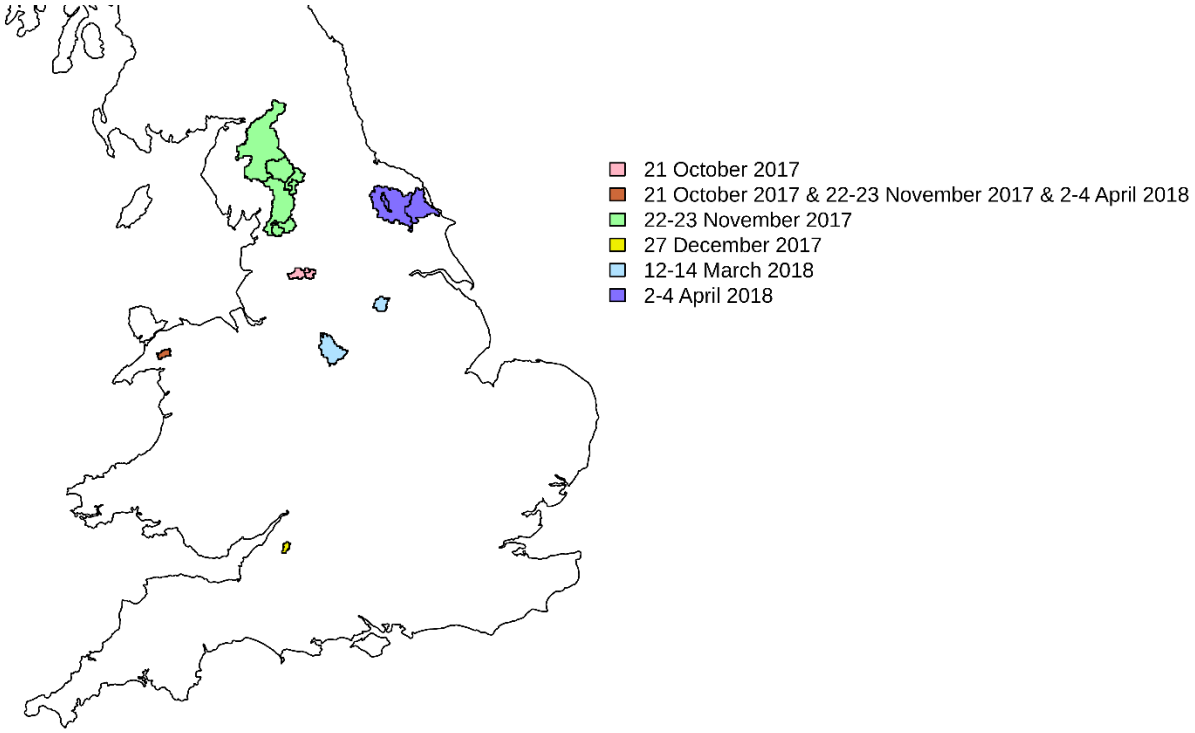


Figure 4 Case-study catchments over England & Wales that have a river flow analysis. Note that some catchments cover multiple case-studies as indicated by the key.

3.1.1 18 July 2017. Flash flood case

Flash flooding in Coverack. This location has no associated G2G gauged catchment so was not included in the analysis for this project. As the Coverack storm was a highly localised event, neighbouring gauged G2G catchments did not receive large precipitation totals and do not show a response in river flow. A response is seen for **Walkham at Horrabridge** (47118) which, though not located close to Coverack, was affected by heavy, quasi-stationary convective cells to the NW of Coverack, further along the line of convection. The locations of Coverack and Walkham at Horrabridge (47118) are shown in Figure 5 alongside a Hyrad display of 15 minute radar-rainfall accumulations (H19) for 17:15 18 July 2017.

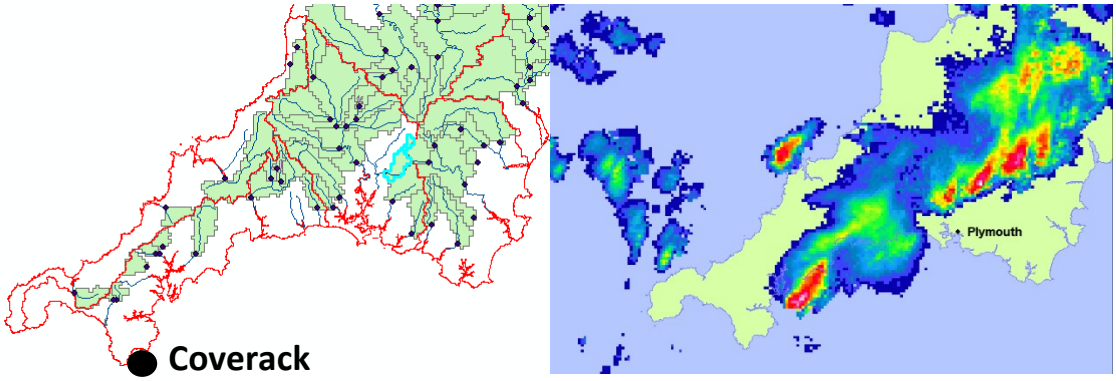


Figure 5 South West peninsula G2G gauged catchments (green infill) highlighting the locations of Walkham at Horrabridge (cyan catchment boundary) and Coverack (left map) and Hyrad display of 15 minute radar-rainfall accumulations (H19) for 17:15 18 July 2017 (right map).

A rainfall peak was identified in the Thames region. Catchment of the **Ash at Mardock** (5080TH) in Hertfordshire, a tributary of the River Lee and near the town of Wareside. No river flow response was seen and river flows were not considered for this case-study.

3.1.2 9 August 2017. Surface water flooding case. EFAS Flash Flood with no recorded fluvial impacts.

For this case-study the peak rainfall was identified in the North East of England for **Gypsey Race at Boynton** (Boyntn1), a chalk stream on the Yorkshire Wolds.

River flows were not considered for this surface water flooding case-study.

3.1.3 23 August 2017. Convective event with flash flooding from surface water, significant in Scarborough, minor for York and west.

For this case-study, the 24h rainfall peak was identified in the North East for the catchment **Derwent at Low Marishes** (MARISH1)

River flows were not considered for this surface water flooding case-study.

3.1.4 30 September 2017. Narrow band of heavy rain over the south of Cumbria caused surface water flooding (Millom, Windermere, Haverigg)

For this cast study, the 24h rainfall peak was identified for **Kent at Sedgwick** (730511)

Observed and modelled river flows (simulation-mode with observed rainfall, state-updating and flow-insertion) were examined for Cumbria. As no threshold-crossings were seen, *river flows were not considered for this surface water flooding case-study.*

3.1.5 21 October 2017. Disruption and SWF in west of country due to Storm Brian. River flow impacts in Lancashire at Rawtenstall alongside SWF impacts.

For this case-study, the FGSV Observed impacts summary identified river flooding of five properties in Rawtenstall. Nearby catchment of **Irwell at Irwell Vale** is considered here.

Irwell at Irwell Vale (690140)

For Irwell Vale, the ***observed river flows*** reached Q(2) on the evening of 21 October, but didn't cross this threshold. Initial ***Days 3-6*** forecasts showed low probabilities of crossing the Q(2)/2 threshold (although the peak timing was highly uncertain). Later forecasts (origin on 17 October) showed a low probability of Q(2) being exceeded, although there was some jumpiness between forecasts. ***Days 2-3*** forecasts consistently gave a low probability of Q(2)/2 being exceeded, with some members crossing

this threshold later on in the forecast, around 24 October. Although the majority of ***Day 1*** forecasts showed a low probability of crossing Q(2)/2 only, there were forecasts which suggested higher threshold crossings and one (origin 07:16 21 October) where the Q(2)/2 threshold was not exceeded by any members.

Overall there were a number of sites in this geographical area predicting threshold-crossings of the Q(2)/2 and Q(2) threshold around 21-22 October, even at longer lead-times. In general, the verification suggested that these probabilities might be too high, although several instances are seen of possibly too low, or about right, probability values. In general, there were two periods of high ensemble spread: around 00:00 22 October, and again around 24-25 October. This suggests possible uncertainty about secondary river flow peaks several days after the main case-study event. This uncertainty is also seen in the precipitation ensemble forecasts, for example initiating at 13:15 19 October 2017 as shown in Figure 6.

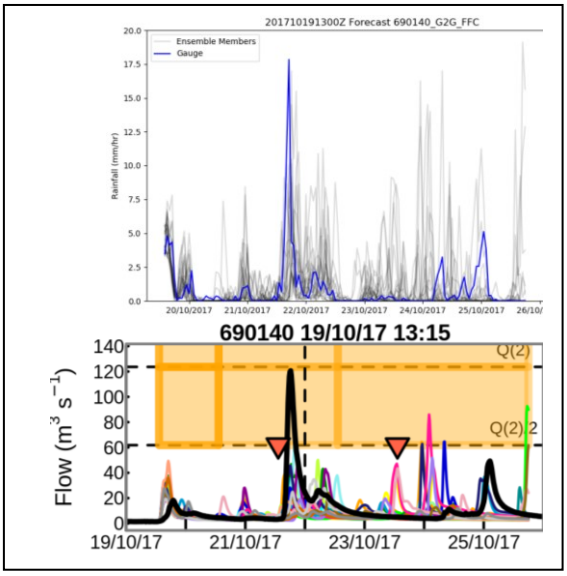


Figure 6 Example rainfall (raingauge blue and ensemble members grey) and river flow (colours as discussed in Section 2.1) time-series for the BMR forecast for Irwell at Irwell Vale (690140) initiated at 13:15 19 October 2017.

The FGSV also notes that there were flood sirens on the River Calder at **Hebden Bridge, Todmorden** and **Mytholmroyd**. These catchments border Irwell Vale but drain east to join the Yorkshire Ouse. As Mytholmroyd has no useable observed river flow data for this period, only Hebden Bridge and Todmorden are considered here.

Calder at Hebden Bridge (HEBDBR1)

Calder at Todmorden (TODMDN1)

Observed river flows crossed the Q(5) threshold late on 21 October at Todmorden, with a similarly timed peak crossing the Q(2)/2, and reaching the Q(2) threshold, at Hebden Bridge. The ensemble did not consistently capture non-zero probabilities of these thresholds being crossed until the final few days before the observed peak river flow. There were a small number of very early Days 3-6 forecasts (e.g. origins 13:15 16 October, 01:15 17 October (Figure 7, left panel)) which did show low probabilities of crossing the Q(5) threshold at Todmorden, with one ensemble member indicating a Q(50) crossing at Hebden Bridge. This suggests that that there was, initially, some indication from the ensemble that high flows were a possibility, although it was not until 20 October that there was any further indication

of these events being possible. Figure 7 shows example BMR forecasts initiated at 01:15 17 October 2017 and 01:15 21 October 2017 for both precipitation and river flow. It can be seen that the higher river flows for the earlier forecast relate to higher forecast precipitation values on 21 October. In contrast, the Day 1 forecast origin shows forecast precipitation values that were both delayed and lower than observed, leading to a much-lower river flow peak around 12 to 24 hours later than that observed.

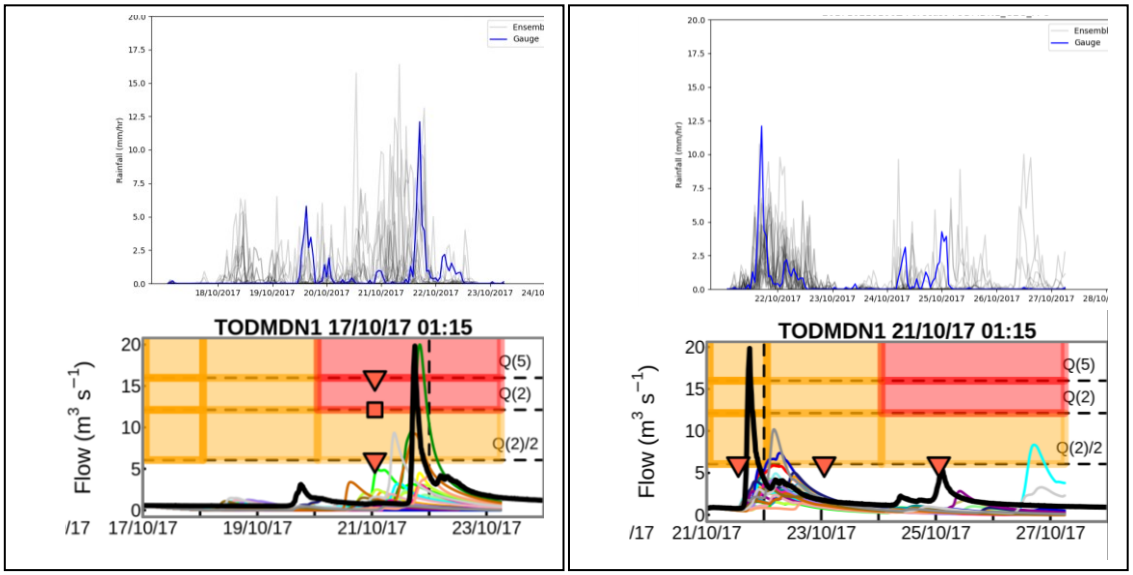


Figure 7 Example rainfall (raingauge blue and ensemble members grey) and river flow (colours as discussed in Section 2.1) time-series for the BMR forecasts for the Calder at Todmorden (TODMDN1) initiated at 01:15 17 October (left) and 01:15 21 October (right) in 2017.

Consideration of the regional threshold-crossing maps suggests that the ensemble as a whole was not predicting the high peaks for Days 2-3 forecasts at larger scales: there is not simply a spatial displacement of the highest predicted rainfall/river-flow values.

Interestingly the rainfall peak for this case-study occurs not in the North West of England, but instead in North Wales at Beddgelert. Whilst no river flow thresholds (Q(2)/2 or greater) are crossed at this site, river flow plots are included for consistency and comparison with the rainfall time-series.

Glaslyn at Beddgelert (065001_TG_1201). **PDM catchment** Glaslyn_001

For Beddgelert, observed river flows (just) crossed the Q(2)/2 threshold in the evening of 21 October, with a second, higher crossing of this threshold on 25 October. The first peak – the focus of this case-study, and where peak 24h rainfall was identified – was not forecast by any ensemble member after 18 October, although some Days 4-6 forecasts did have a low probability of this threshold being crossed. An example is shown for the 01:15 17 October forecast origin in Figure 8 (left panel) for both the G2G and PDM model output. Both models show a similar response to the forecast rainfall with some members predicting early threshold crossings linked to high forecast rainfall on the 20 October. For the 21 October peak, the precipitation-ensemble spread encompasses the raingauge observed precipitation values and this is reflected in the G2G and PDM ensembles also encompassing the observed river flows. The right-hand panel of Figure 8 shows a later Day 1 forecast origin for 01:15 21 October 2020. It can be seen that at this short lead-time the forecast precipitation values are much lower (noting the different axis on the precipitation plots) and the river flow peak is not captured.

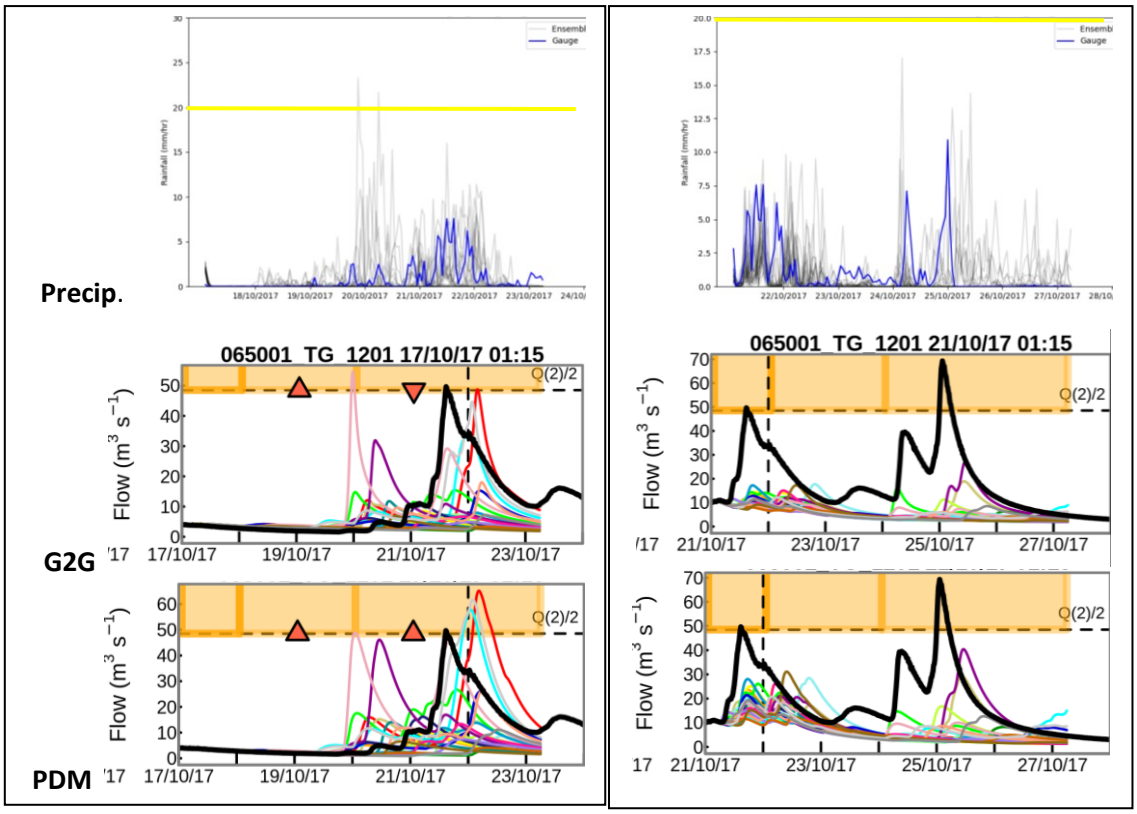


Figure 8 Example rainfall (raingauge blue, ensemble members grey, 20 mm h^{-1} shown in yellow), G2G and PDM river flow (colours as discussed in Section 2.1) time-series for the BMR forecasts for Glaslyn at Beddgelert (065001_TG_1201) initiated at 01:15 17 October (left) and 01:15 21 October (right) in 2017.

3.1.6 3 & 4 November 2017. No noticeable river response.

For this case-study, the 24h rainfall peak was identified at South West of England catchment **Moors River at Hurn Court** (43214). This differs from the locations indicated by the case-study events document of SE England. Further discussion of the geographical selection of catchments of interest from analysis of heavy precipitation is given in Appendix C.2.

River flows were not considered for this case-study.

3.1.7 22 & 23 November 2017. (Deep low pressure system brought heavy rainfall to northern and western Britain).

Of all the Phase 2 case-studies, this has the greatest river flow responses. The highest flows are seen in NW England, which recorded November highest peak flows for Lune and Eden catchments (UKCEH, 2021). To capture these events, the focus is on two nested sets of catchments: the Lune at Caton and upstream catchments, and the Eden at Sheepmount and upstream catchments. This gives a range of catchment sizes and characteristics.

Lune at Caton (724629, 983km^2)

- *Upstream* **Wenning at Hornby** (724528, 232km^2)
  - *Headwater* **Hindburn at Wray** (724427, 83km^2)
  - *Headwater* **Wenning at Wennington** (724326, 142km^2)

Observed river flows peaked close to, but did not cross, the Q(50) threshold in the early hours of 23 November 2017 at Caton. This reflects the earlier peaks (late on 22 November) at headwater

catchments Wray, where Q(5) was crossed, and Wennington, where Q(5) was reached but not crossed. The Q(5) threshold was also reached at Hornby, upstream of Caton. In this instance the **earlier *forecasts*** from Days 2-3 and 2-6 seemed to capture the event better that those in Day 1. This is particularly clear at Caton, where the majority of Day 1 forecasts had no ensemble members crossing the Q(5) threshold, although it is also seen for the other catchments considered. Examination of the hydrographs suggests that this is not a timing issue of the ensemble member peaks being "missed" by the shorter one-day time-window: rather, the ensemble member rain-rates are noticeably lower in Day 1 of the ensemble member forecasts. Analysis of the rainfall time-series confirms this interpretation, with forecasts closer to the peak event having considerably lower rainfall values. Two example forecasts for the Lune at Caton are shown in Figure 9; forecast origins 13:00 20 November 2020 (peak in Days 2-3) and 01:00 22 November 2020 (peak in Days 2-3)

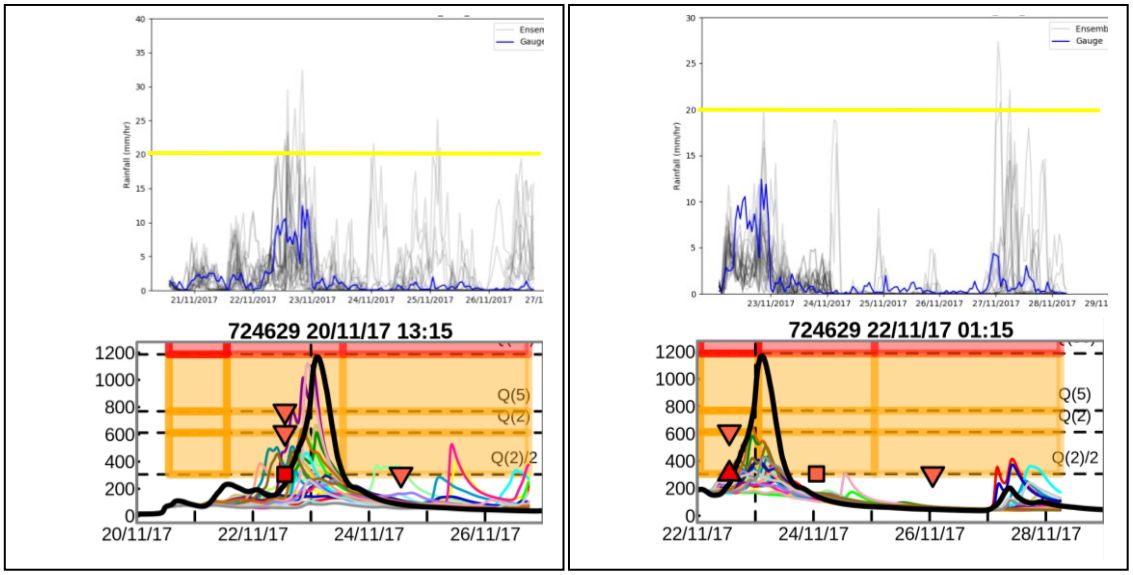


Figure 9 Example rainfall (raingauge blue, ensemble members grey, 20 mm h^{-1} shown in yellow) and river flow (colours as discussed in Section 2.1) time-series for the BMR forecasts for the Lune at Caton (724629) initiated at 13:15 20 November (left) and 01:15 22 November (right) in 2017.

Interestingly, the Days 3-6 (and to some extent Days 2-3) forecasts for the days following the event of interest also show a low chance of high thresholds (up to Q(50)) being crossed, when in this instance the observed river flows show only a small rise. This suggests, perhaps, that higher threshold-crossings are, in general, more common at the longer lead-times, though this cannot be shown in a case-study context.

Eden at Sheepmount (765512, 2286.5km^2)

- *Upstream* Eden at Temple Sowerby (760502, 616.4km^2)
  - *Upstream* Eden at Gt Musgrave Bridge (760112, 225km^2)
    - *Headwater* Eden at Kirkby Stephen (760101, 69.4km^2)

Overall, the River Eden ***observed river flow*** peaks for this case-study were lower than those seen for the River Lune. For the catchments considered, the highest peak was seen at Gt Musgrave Bridge where the Q(5) threshold was crossed, though flows at Temple Sowerby came close to crossing the Q(5) threshold. At the other two catchments considered, Sheepmount and Kirkby Stephen, observed flow peaks narrowly crossed the Q(2) threshold. Early ***forecasts*** for all catchments predicted higher

peak flows with low-medium certainty, with some ensemble members predicting very high peak flows, far above the Q(50) threshold. There was also a large degree of uncertainty around the peak-timing at these lead-times. Closer to the event the ensemble settled on Q(2) and Q(5) threshold-exceedance. Interestingly, like for the Lune catchments for this case-study, a number of ensemble members were predicting a second high peak around 26 November, which did not materialise in reality.

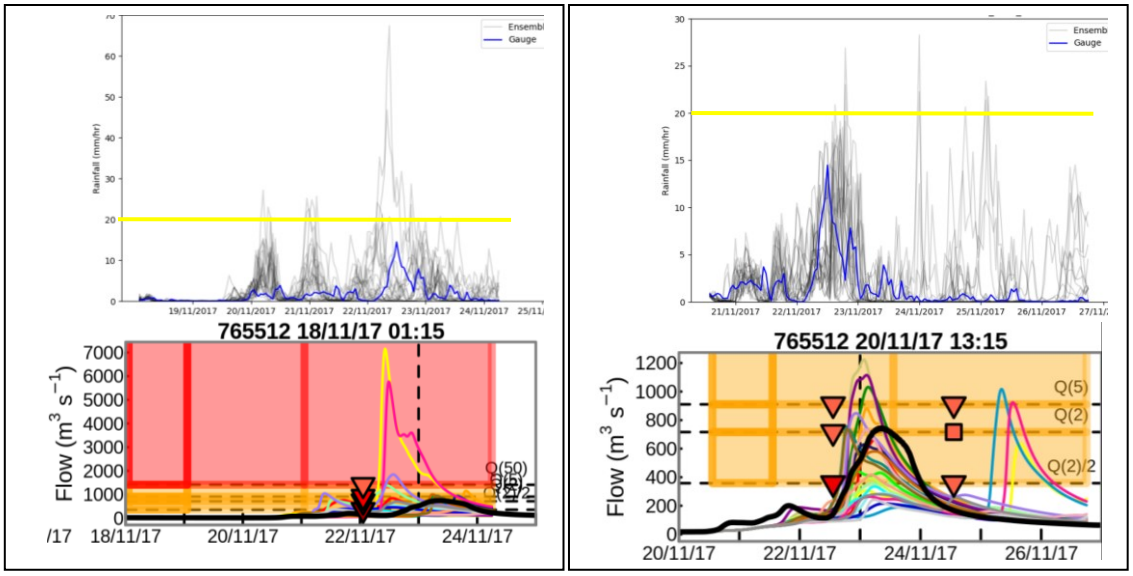


Figure 10 Example rainfall (raingauge blue, ensemble members grey, 20 mm h^{-1} shown in yellow) and river flow (colours as discussed in Section 2.1) time-series for the BMR forecasts for the Eden at Sheepmount (765512) initiated at 01:15 18 November (left) and 13:15 20 November (right) in 2017.

For this case-study, the 24h rainfall peak was located in North Wales, in the **Glaslyn at Beddgelert** catchment. This catchment is considered as an example for *both* G2G and PDM ensembles (NRW PDM catchment ID Glaslyn_001). Although flows are lower here than in NW England, observed flows are still high, crossing the Q(2) threshold in the evening of 22 November.

Glaslyn at Beddgelert (065001_TG_1201). **PDM catchment** Glaslyn_001

For this case-study, the ***observed river flows*** at Beddgelert crossed the Q(2) threshold in the evening of 22 November, with peak flows coming close to the Q(5) threshold. Similarly to the Lune catchments for this case-study (Figure 9), and the 21 October case-study for Beddgelert (Figure 8), the peak seemed to be better ***predicted*** by Days 4-6 forecasts, with ensemble member forecasts being, in general, too low for forecasts in Days 1-3. Again, this also matches with the precipitation analysis.

3.1.8 27 December 2017. Minor impacts from surface water flooding in Midlands, SW and SE England. No river flow impacts recorded.

For this case-study, the 24h rainfall peak was identified for Anglian catchment **Wensum at Costessey Mill** (E19862). Although this is predominantly a surface water flooding case-study, river flows close to the Q(2) threshold were seen for the PDM catchment Boyd at Bitton, and this case study was used as an example for that model.

Boyd at Bitton (53131) **PDM catchment** 530350

Figure 11 shows the PDM forecasts for all forecast initiation times covering this case-study. Overall, the verification statistics show this site performs well for low threshold-crossings as shown by the green background. A large range of predicted threshold-crossings are seen, with a small number of ensemble members indicating crossings well above the Q(50) threshold. It is thought these will be related to very large rainfall totals for these members, but this has not been analysed for this case-study.

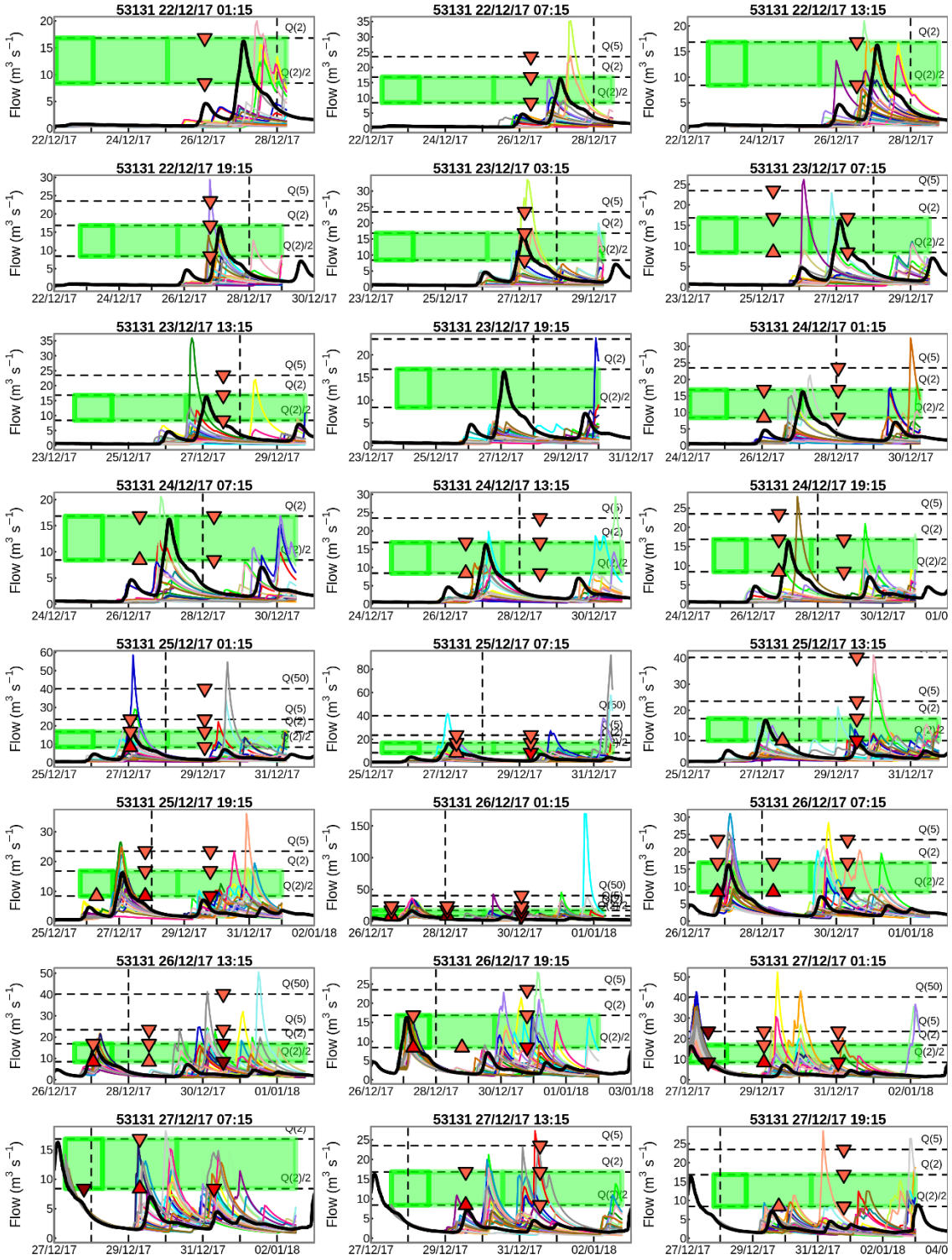


Figure 11 Example river flow (colours as discussed in Section 2.1) time-series for all PDM BMR forecasts for the Boyd at Bitton covering 00:00 28 December 2017.

3.1.9 2 & 3 January 2018 Storm Eleanor. No fluvial impacts recorded.

For this case-study, the 24h rainfall peak was identified in the North West for catchment **Derwent at Portinscale** (751007)

River flows were not considered for this surface water flooding case-study.

3.1.10 12-14 March 2018. Widespread flooding in South Derbyshire, with poor advice from G2G in the Staffs/High Peaks area.

Analysis of the G2G deterministic output with observed rainfall as input does not show any threshold-crossings or significant response in the Staffs/High Peaks area. To test whether this is the case when forecast rainfall is used instead, the following catchments were considered.

Dove at Rocester (4008)

- *Headwater* **Dove at Hollinsclough** (4033) also **PDM Hollinsclough** Zone 1 and 2

For Rocester, a double-peak is seen in the ***observed river flows***, with peaks around midday on 10 March and the late evening on 12 March. Both peaks cross the Q(2) threshold, although the second is noticeably larger. The ***G2G ensemble forecasts*** for this give some suggestion that the Q(2) threshold may be crossed, but this varies from forecast to forecast, even at short lead-times. For this case study, with two close-together peaks, the effects of timing uncertainty - and its relationship with the time-period used to calculate threshold-crossings - become apparent. There are, for example, ensemble members that capture only one peak in the Days 3-6 forecasts between the two observed peaks, and forecasts do "better" when the two peaks are considered in the same time-window.

For Hollinsclough, observed river flows remain low, well below the Q(2)/2 level, contrasting with the G2G ensemble member forecasts which show a low chance of crossing thresholds up to Q(5). This is seen throughout the period considered for this case-study, not just for the time surrounding the main case-study peak. This is shown for example forecast initiation times 13:15 10 March, 19:15 10 March, and 01:15 11 March 2018 in Figure 12. Although the performance of G2G for this catchment is generally too peaky, the peaks are normally of an appropriate magnitude (e.g. see the G2G Performance Summary), with a moderate negative bias seen overall, suggesting this is not the cause of the poor performance for this case-study event. For the 13:15 10 March 2018 forecast origin, Figure 12 also shows the BMR ensemble rainfall time-series. It can be seen that the rainfall forecast values are much higher than those observed, at similar times to when the large river flow forecast peaks are seen. This suggests that the poor hydrological model performance for this case-study is linked to the rainfall forecasts.

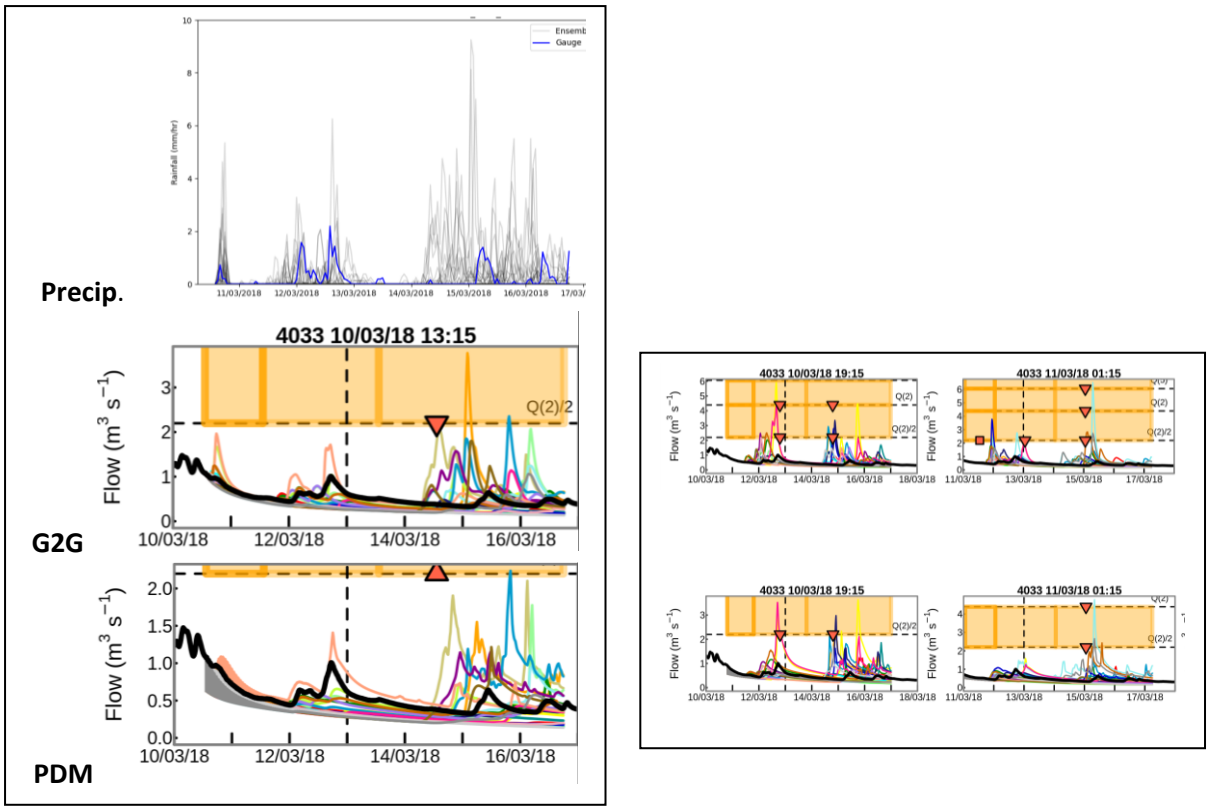


Figure 12 Example rainfall (raingauge blue, ensemble members grey, 20 mm h^{-1} shown in yellow), G2G and PDM river flow (colours as discussed in Section 2.1) time-series for the BMR forecasts for the Dove at Hollinsclough (4033) initiated at 13:15 10 March 2018. The right-hand panel shows additional river flow forecast origins 19:15 10 March and 01:15 11 March 2018.

The 24h Rainfall peak for this case-study was identified in the Midlands for catchment **Torne at Auckley** (4050, 141km^2) which shows both modelled (known rainfall and past flows) and observed river flows crossing the Q(2) threshold. This catchment is considered for both rainfall and river flow analysis.

Torne at Auckley (4050)

Observed river flows for Auckley reach, but do not cross, the Q(2) threshold. Overall the observed river flow peak is poorly predicted at longer lead-times with ensemble members either vastly overestimating the peak (e.g. forecast origin 07:15 7 March) or under predicted (e.g. forecast origin 01:15 9 March). Example forecasts of both rainfall and river flow for origins 01:00 8 March and 13:15 12 March 2018 are shown in Figure 13. It can be clearly seen that the over-estimation of the river flow peak is not directly linked to the magnitude of the rainfall peaks (the forecast rainfall being higher for the earlier initiation time), but rather to the duration of the rainfall.

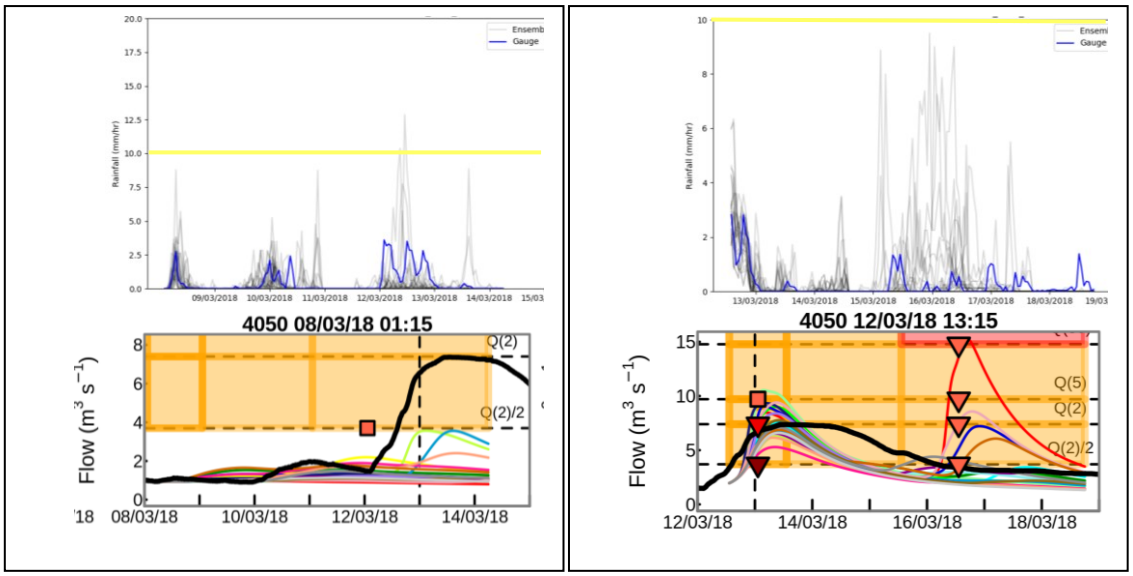


Figure 13 Example rainfall (raingauge blue, ensemble members grey, 10 mm h^{-1} shown in yellow) and river flow (colours as discussed in Section 2.1) time-series for the BMR forecasts for the Torne at Auckley (4050) initiated at 01:00 8 March and 13:15 12 March 2018.

3.1.11 2-4 April 2018. Minor river flooding impacts noted on 3 April in N and W Yorkshire and on 4 April in N Yorkshire, Durham and Tyne and Wear.

Derwent at Malton (Malton1)

- *Upstream* **Derwent at Low Marishes** (MARISH1)
- *Headwater* **Riccal at Nunnington** (Nunnington) **(PDM** Nunnington, F2581**)**

For headwater catchment Nunnington on the River Riccal, ***observed river flows*** peaked around 00:00 3 April, with the Q(2)/2 threshold being crossed. The larger Derwent catchments Low Marishes and Malton both showed broader observed peaks with the highest flows seen from 00:00 4 April. For Malton, the Q(2)/2 threshold was crossed in the early hours of 3 April. For Low Marishes, observed river flows crossed the Q(2)/2 threshold on 31 March, increasing to close to the Q(2) threshold by 4 April. Overall, ***ensemble forecasts*** for both Derwent catchments showed a low-medium chance of very high (up to Q(50)) river flow threshold-crossings, which were not observed for this case-study. The high ensemble member peaks persisted into the Days 2-3 forecasts.

For Nunnington, a number of ensemble members predicted an earlier peak than observed, particularly for forecasts over three days ahead. An example is given in Figure 14 (left) for forecast origin 01:15 29 March 2018. For early Days 4-6 forecasts, with the longest time-window for threshold-crossings, these peaks fell within one time-window (the same as the observed peak) and the guidance appeared good. However, for later forecasts some forecast peaks fell into the Days 2-3 window, while others (and the observed peak) fell into the Days 3-6 window, giving a more-mixed message. This highlights the effect of considering fixed time-windows, and the utility of looking at the hydrographs themselves alongside summary statistics. Comparing the precipitation and rainfall time-series in Figure 14, the false river flow peak in the Days 3-6 forecasts can be linked to high rainfall values persisting for much of the 31 March in a number of ensemble members. At these forecast lead-times very few rainfall ensemble members are predicting a high rainfall peak around the time of the observed peak (2-3 April 2018). This leads to a low probability of river-flows crossing the Q(2)/2

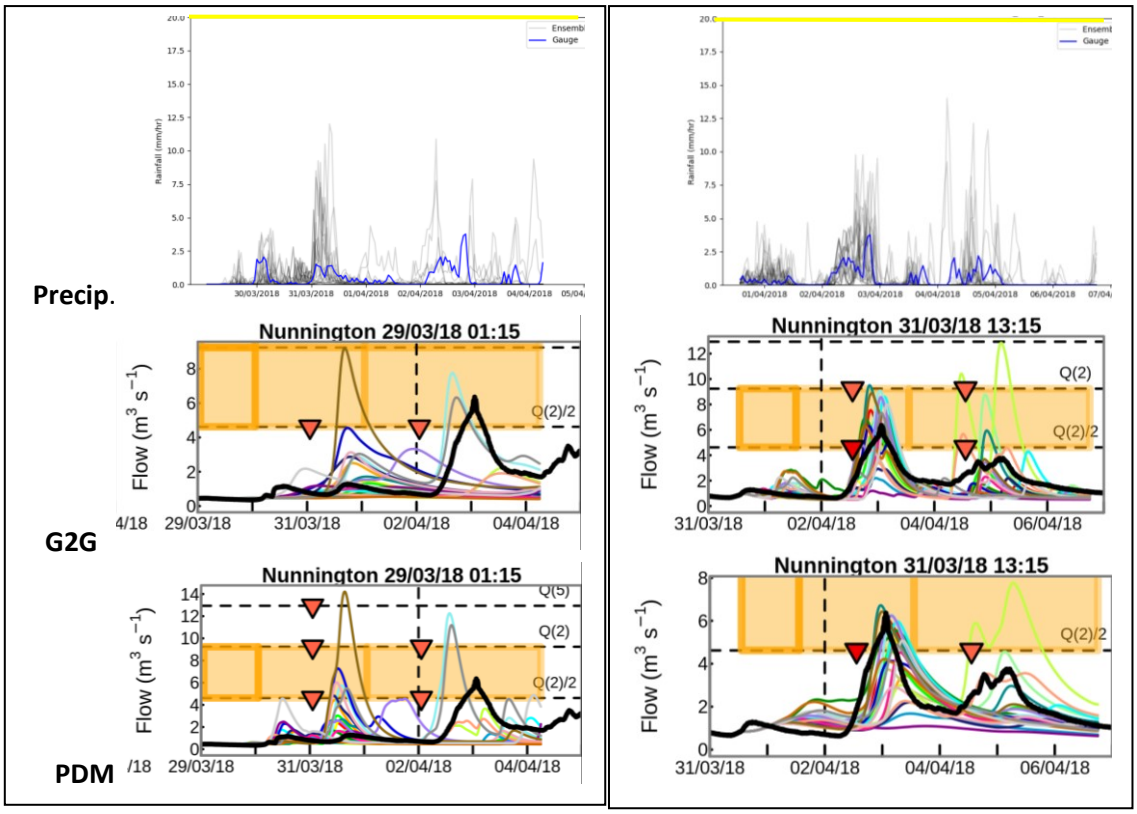


Figure 14 Example rainfall (raingauge blue, ensemble members grey, 20 mm h^{-1} shown in yellow), G2G and PDM river flow (colours as discussed in Section 2.1) time-series for the BMR forecasts for Riccal at Nunnington (Nunnington) initiated at 01:15 29 March (left) and 01:15 13:15 31 March (right) in 2018.

threshold at this time, both for G2G and PDM. Later forecasts (e.g. forecast origin 13:15 31 March 2018, Figure 14 right) show much higher probabilities of crossing the Q(2)/2 threshold, and for G2G, a low probability of crossing the Q(2) threshold.

Wales was also highlighted as an area affected by this case-study. This is where the 24h rainfall peak was identified, again in the catchment of **Glaslyn at Bedgellert**. For comparison, this is also included in the river flow analysis.

Glaslyn at Beddgelert. **PDM catchment** Glaslyn_001

Observed river flows for Beddgelert crossed the Q(2)/2 threshold late evening on 2 April. For all lead-times, ensemble forecasts were much lower than the Q(2)/2 threshold for G2G forecasts, although small peaks were seen on 2 April. No G2G ensemble members suggested a possible Q(2)/2 threshold-crossing. Forecasts from PDM showed slightly higher river flows as expected due to the use of a rainfall factor of 1.2 for this catchment, although the highest member peaks on 3 April still did not reach the Q(2)/2 threshold. However, for PDM there is one ensemble member with sufficiently high flows to suggest a Q(2)/2 threshold-crossing the following day. Comparison with the BMR rainfall time-series shows this second peak relates directly to an ensemble member with much higher precipitation values. An example is given in Figure 15 (left) for forecast origin 01:15 29 March 2018.

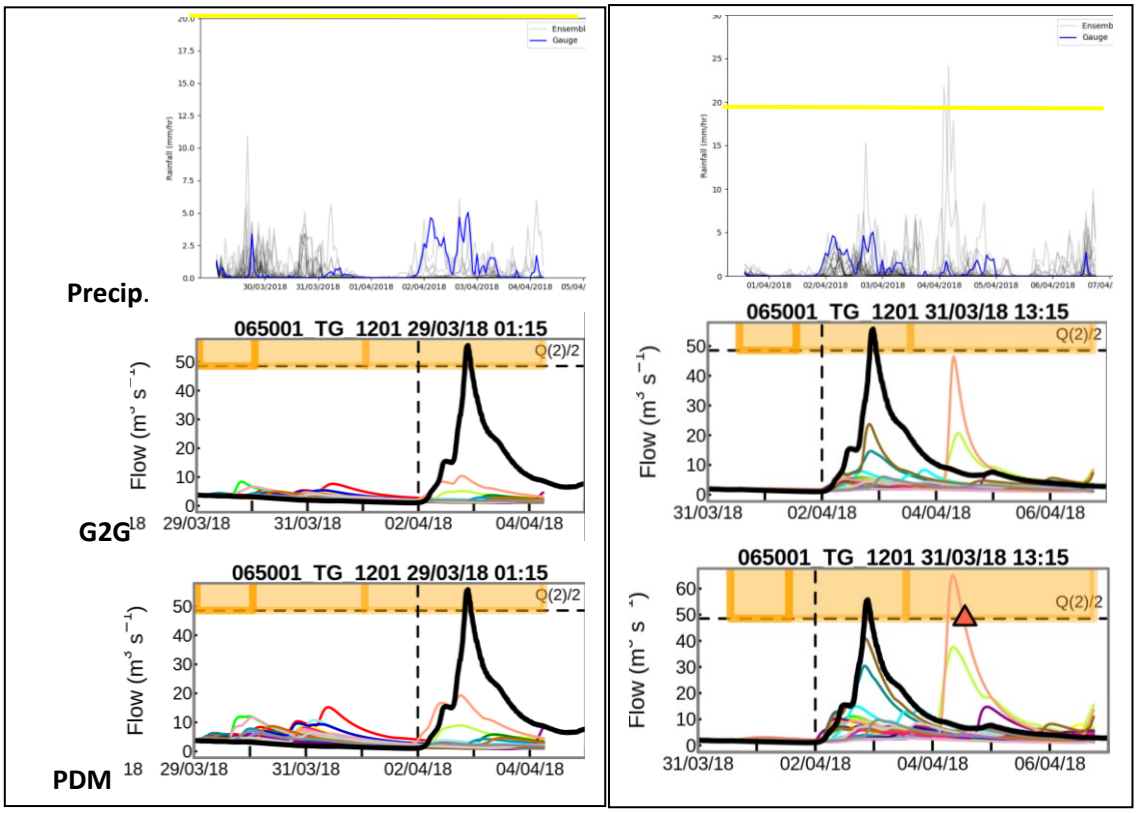


Figure 15 Example rainfall (raingauge blue, ensemble members grey, 20 mm h^{-1} shown in yellow), G2G and PDM river flow (colours as discussed in Section 2.1) time-series for Glaslyn at Beddgelert (065001_TG_1201) BMR forecasts initiated at 01:15 29 March (left) and 13:15 31 March (right) in 2018.

3.1.12 20 September 2018. Primarily surface water flooding case-study with transport disruption in Sheffield, Rotherham and Pontypridd.

River flows were not considered for this surface water flooding case-study.

3.2 Case-studies over Scotland

The two case-studies selected over Scotland are analysed below. Each analysis includes identifying features of each case-study, the associated catchments chosen and key conclusions drawn from the results.

First,

Table ***3*** lists the catchments considered for each case-study based on (i) the observed and/or modelled (G2G) river flow response (middle column), and (ii) the peak 24h raingauge rainfall total (right column). In

Table ***3***, catchments that are indented are upstream of the preceding un-indented catchment. Note that for the Scotland case-studies, as opposed to those for England & Wales, the catchment with the peak 24h rainfall is also included in the catchments used for the river flow analysis.

Table 4 then gives further catchment information for case-study catchments that have a river flow analysis. This includes National River Flow Archive (NRFA, nrfa.ceh.ac.uk) details if available. The locations of the catchments are mapped in Figure 16.

Table 3 Catchments used for case-studies over Scotland

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

Table 4 Catchment information for case-study catchments over Scotland

Case-study	Site	G2G ID (NFFS ID)	Region	River	G2G area (km^2)	NRFA ID	NRFA area (km^2)
7 Jun 17	Sheriffmills	234307	NW	Lossie	214	7003	216
	Wardend Bridge	234331	NW	Mosset Burn	29	7009	28.3
	Forres	234221	NW	Findhorn	781	7002	781.9
	Shenachie	234306	NW	Findhorn	416	7001	415.6
	Dunphail	234206	NW	Divie	165	7005	165
	Firhall	234218	NW	Nairn	312	7004	313
	Balnafoich	234164	NW	Nairn	130	7008	128.1
24 Jan 17	Sprouston	15012	SE	Tweed	3345	21021	3330
	Lindean	14990	SE	Ettrick Water	503	21007	499
	Brockhoperig	14987	SE	Ettrick Water	38	21017	37.5
	Deephope	14986	SE	Tima Water	30	21026	31
	Glen Orchy	133087	SW	Orchy	252	89003	251.2

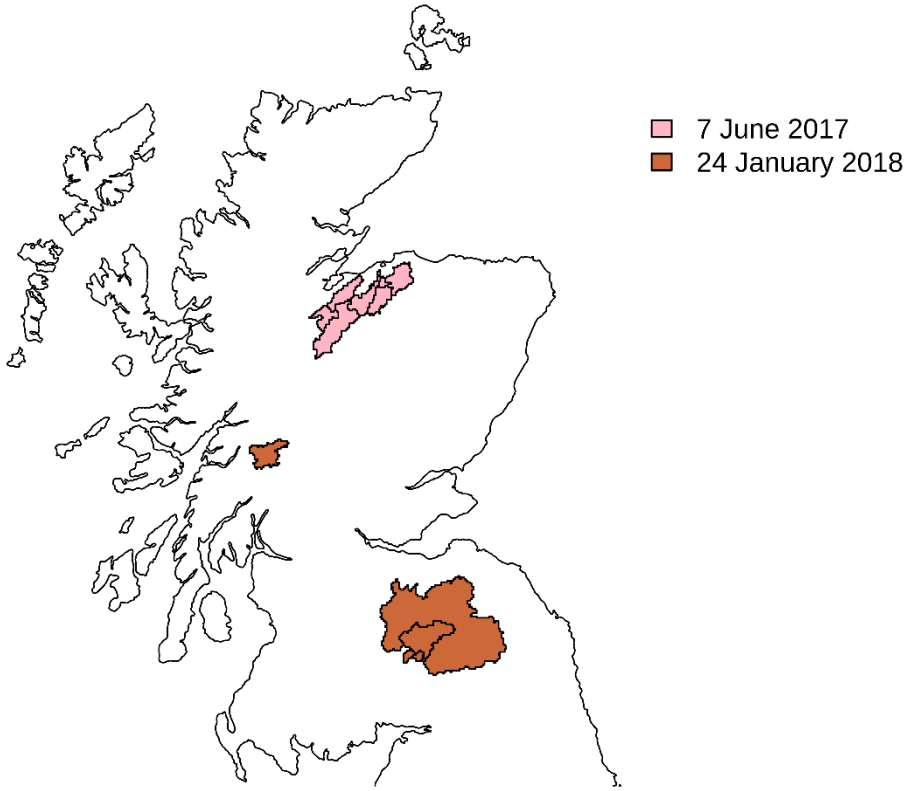


Figure 16 Case-study catchments over Scotland coloured by case study.

3.2.1 7 June 2017. Hydrologically significant event. Flood defences at Forres and Elgin prevented flooding in these places. River Nairn also affected.

This case-study showed high (Q(5) to Q(25)) river flow thresholds being exceeded in the Moray Region of North East Scotland. Catchments were selected on the three main rivers where significant flood events were noted – Lossie, Findhorn and Nairn – and also on the small Mosset Burn catchment. The peak 24h rainfall was recorded for the Lossie at Sheriffmills catchment.

Lossie at Sheriffmills (234307)

Mosset Burn at Wardend Bridge (234331 & **PDM** Mosset to Wardend Bridge)

Findhorn at Forres (234221)

- *Headwater* **Findhorn at Shenachie (**234306 & **PDM** Findhorn to Shenachie)
- *Headwater* **Divie at Dunphail** (234206 & **PDM** Lochindorb to Dunphail)

Nairn at Firhall (234218)

- *Headwater* **Nairn at Balnafoich** (234164)

Overall, all seven catchments considered for this case-study showed high river flow peaks. The highest ***observed river flow*** threshold-crossing was of the Q(50) threshold for the Mosset Burn at Wardend Bridge early on 7 June. However, the downstream Findhorn and Nairn catchments also showed the Q(5) threshold being clearly crossed as listed below.

Sheriffmills Q(5) evening 7 June
Wardend Bridge Q(50) early 7 June
Forres Q(5) early 7 June, **Shenachie** Q(2) evening 6 June, **Dunphail** Q(5) early 7 June
Firhall Q(5) early 7 June, **Balnafoich** Q(2) early 7 June.

For all catchments considered, initial ***ensemble forecasts*** of the case-study event beyond a lead-time of Day 3 showed little indication of river flow threshold-crossings above Q(2)/2. Although there was some variation from forecast-to-forecast, (e.g. Shenachie in Days 2-3) by Day 1, and late Days 2-3, the ensemble was giving a low-medium chance of threshold-crossings of the Q(5) and Q(50) threshold. Three example forecasts (Days 3-6, Days 2-3, and Day 1) are shown for the Shenachie catchment in Figure 17. Other sites lead to similar conclusions.

Figure 17 shows results for both G2G and PDM. Although the results are similar overall, with a direct correspondence between individual ensemble members, there are some differences in the presentation. Firstly, whereas the PDM forecasts continue for the full six-day BMR ensemble rainfall forecast duration, the G2G forecasts only extend out to ~4.5 days due to the requirement for air temperature data to run the G2G Snow Hydrology module. Secondly, the PDM forecasts do not show verification information above the Q(2)/2 threshold. As discussed in the Appendix B.2 Overall river flow verification summary, the PDM verification scores are calculated for single-catchments only, and there are too few events above the Q(2)/2 threshold to calculate verification scores. Of course, this case-study event in June 2017 does feature crossings above the Q(2)/2 threshold, and would allow scores to be calculated for higher verification thresholds if it were included in the overall verification period. However, as the verification period used in this instance is from 1 September 2017 to 31 August 2018, this is not the case. In a real-time forecasting system this might also be the case: there could be extreme river flow values which have not featured in the verification period used, and for which the verification can provide little guidance on ensemble forecast performance.

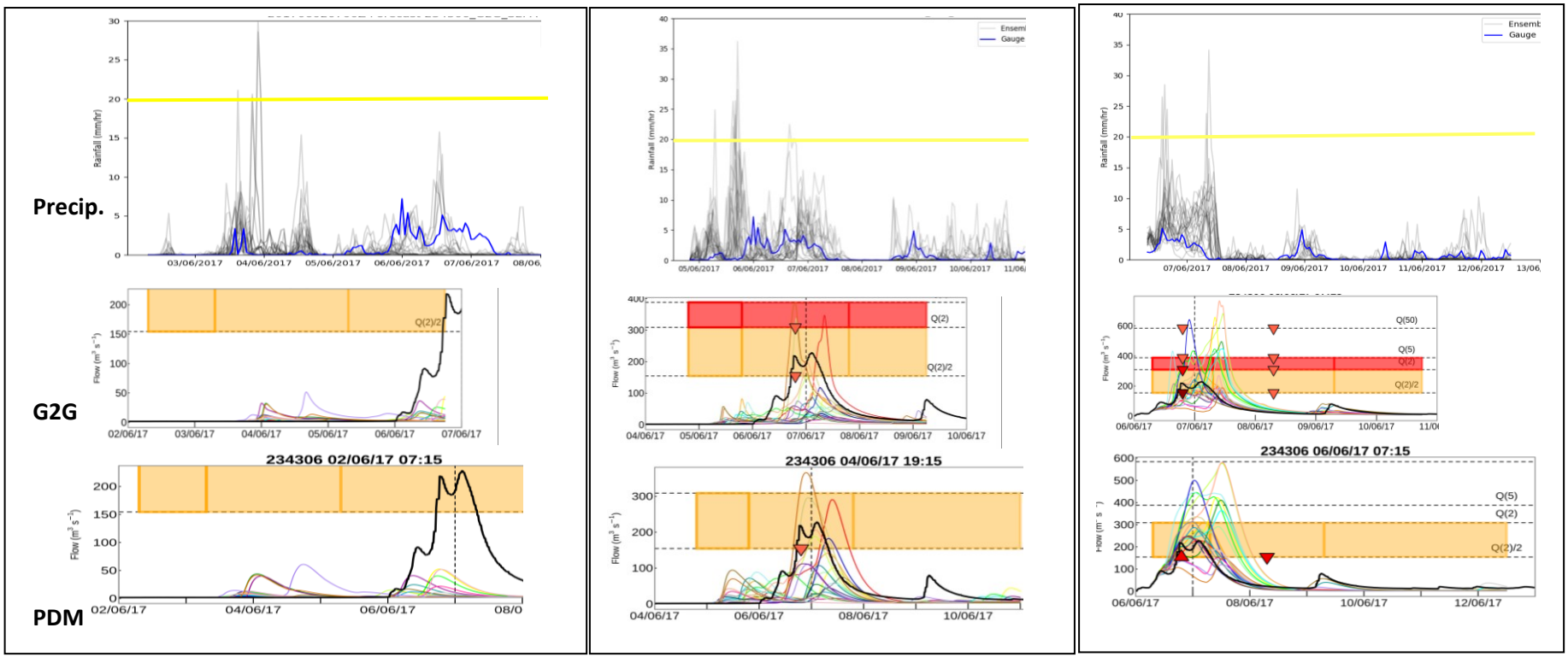


Figure 17 Example rainfall (raingauge blue, ensemble members grey, 20 mm h^{-1} shown in yellow), G2G and PDM river flow (colours as discussed in Section 2.1) time-series for Findhorn at Shenachie (234306) BMR forecasts initiated at 07:15 2 June (left), 19:15 4 June (middle) and 07:15 6 June (right) in 2017.

3.2.2 24 January 2018. Several linked events in the Scottish Borders. Snowmelt overnight from 23 to 24 January may have influenced this case-study.

For this case, observed river flows crossing thresholds Q(2) to Q(10) were noted in the highlighted Scottish Borders region. Example catchments on the River Tweed are considered here, with the highest threshold-crossings seen for the small headwater catchments.

Tweed at Sprouston (15012)

- *Upstream* **Ettrick Water at Lindean** (14990)
  - *Headwater* **Ettrick Water at Brockhoperig** (14987) & **PDM** Ettrick at Brockhoperig)
  - *Headwater* **Tima Water at Deephope** *(*14986)

For this case-study, all the catchments considered showed a double-peak in the observed river flow, with the two headwater catchments peaking around midday on 23 January (Q(2)/2) and again in the morning of 24 January (Q(2)). The larger downstream catchments showed a smaller fall between the peaks, particularly at Sprouston where the overall hydrograph shape was of one large peak which crossed the Q(5) threshold in the afternoon of 24 January.

Overall, there was little indication of the ensemble member forecasts capturing the double-peak characteristic. Overall, the G2G ensemble gave a reasonable indication of the observed threshold-crossings for the two larger catchments, but for the headwater catchments Brockhoperig and Deephope, many forecasts showed no chance of threshold-crossings, even at Q(2)/2. The Brockhoperig PDM model performed better, capturing a low chance of crossing the Q(2) threshold for all forecasts initiated after 07:15 18 January 2018, and a chance of crossing the Q(2) threshold for several Days 2-3 and Days 4-6 forecasts. An example is shown in Figure 18.

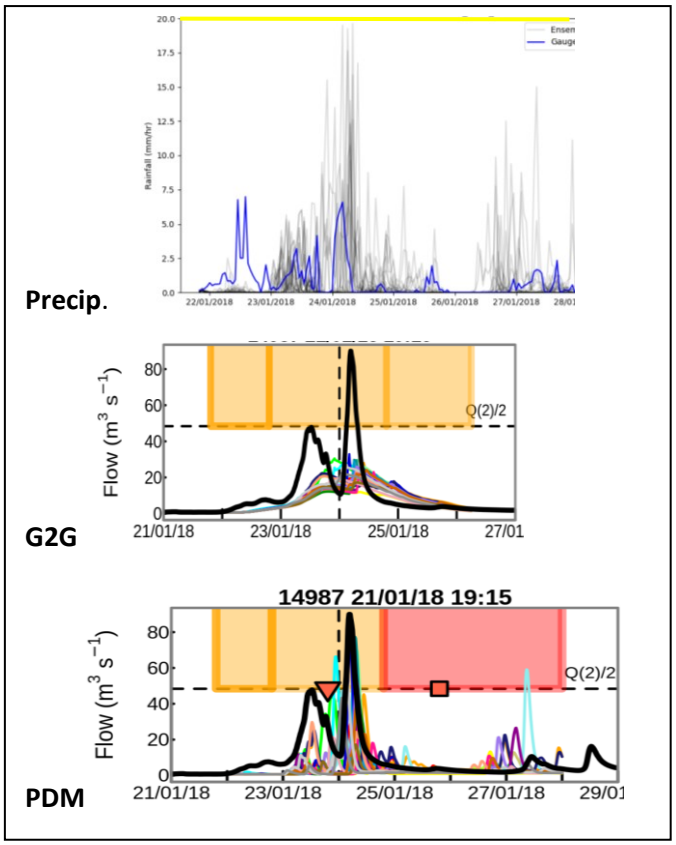


Figure 18 Example rainfall (raingauge blue, ensemble members grey, 20 mm h^{-1} shown in yellow), G2G and PDM river flow (colours as discussed in Section 2.1) time-series for Ettrick Water at Brockhoperig (14987) BMR forecasts initiated at 19:15 21 January 2018.

The 24h rainfall peak for this case-study was identified in the North West region for the Orchy at Glen Orchy catchment. This catchment also showed a high river flow response, with observed river flows crossing the Q(5) threshold and is considered here for both precipitation and river flow analysis.

Orchy at Glen Orchy (133087)

The ***observed river flows*** show a double-peak at Glen Orchy, the first, higher Q(5) peak occurring on the afternoon of 23 January, with the second, lower Q(2) peak occurring on the afternoon of 24 January. The ***G2G ensemble members*** tend to only forecast one peak, indicating a time around the second observed peak. Analysis of the rainfall ensemble forecasts suggested that forecasts underestimated the precipitation values for the first peak, with the second peak values tending to be overestimated as shown in Figure 19. Overall, G2G forecasts beyond Day 3 were better at capturing the observed peak river flows for this case-study. In particular, a number of Days 3-6 forecasts predicted a low chance of the Q(5) threshold being exceeded, whereas Days 2-3 forecasts generally only showed a low chance of the Q(2) threshold being crossed, and Day 1 forecasts the Q(2)/2 threshold. It is noted that river flow peaks could have been enhanced by snowmelt for this case as noted in the case-study description.

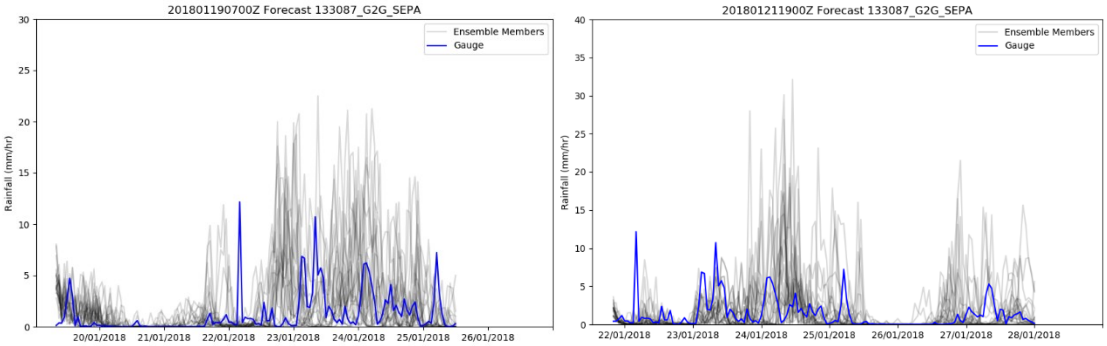


Figure 19 Example rainfall (raingauge blue, ensemble members grey, 20 mm h^{-1} shown in yellow), time-series for Orchy at Glen Orchy (133087) BMR forecasts initiated at 07:00 19 January (left) and 19:00 21 January (right) in 2018.

4 Conclusions from analysis of all case-studies

Analysis of the case-studies above leads to the following key conclusions.

- Useful information can be gained by **viewing together** the river flow and precipitation ensemble **time-series**
- Often a **direct link** can be made between the precipitation and river flow ensemble members for both PDM and G2G.
- Forecast performance does not necessarily improve with lead-time. For example, there are instances where longer lead-time forecasts perform better than those close to the event, or forecast performance varies between consecutive forecasts. This highlights the **advantage of looking at multiple forecast-origins** covering an event, not just the most-recent forecast.
- The **verification period chosen**, and **pooling method** for calculating verification statistics, **impact** on the thresholds where verification information can be provided.
- **One year** of data is **not sufficient** to calculate verification statistics **beyond the Q(2)/2** threshold for single sites, as exemplified here for PDM. Even for this threshold, single-site results from one verification year are highly noisy and **should be treated as demonstrative** instead of representative.
- In general, for a given catchment, better performance is seen for PDM than for G2G. This is expected from comparing a countrywide distributed model to a set of catchment-calibrated local models.

References

UKCEH (2021) Hydrological Summary for the United Kingdom. National River Flow Archive Catalogue of Monthly Summaries. https://nrfa.ceh.ac.uk/monthly-hydrological-summary-uk

Rainfall and River Flow Ensemble Verification: Phase 2
The Joint Coding Framework
Final Report Appendix D

1. Overview of Joint Coding Framework

This document provides an overview of the Joint Coding Framework developed in Phase 2 of the Rainfall and River Flow Ensemble Verification project. The Framework aims to ensure that the code developed under the Project is robust, consistent across both river flow and precipitation, and structured appropriately to allow future flexibility and operational implementation. The technical details of the Framework have been developed as a detailed flowchart combining the river flow and precipitation coding structures and data workflows. This is included at the end of this document in Figure 1.

Overall, the key principles underpinning the Joint Coding Framework are as follows.

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

Saving data. Once calculated, all verification score data and products derived from the raw precipitation and river flow fields will be saved in a standard, simple, human-readable format. This gives the flexibility of reading in such information using different systems/coding-languages in the future. Processed river flow and precipitation data will be saved in daily or monthly blocks, giving flexibility when verifying ensemble forecast performance over longer periods.

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

The following sections serve to summarise the Joint Coding Framework and its four stages of processing, calculating and output.

2. Stages of the Joint Coding Framework

There are four stages of the Joint Coding Framework.

Stage A. Initial processing

Stage B. Calculations involving the full 24 ensemble member data

Stage C. Calculations involving the binary observations and ensemble probabilities

Stage D. Final outputs.

These stages have been designed to maximise computational efficiency, whilst ensuring consistency across river flow and precipitation verification, and flexibility for future use. Further details of each stage are given below and are visualised in the detailed flowchart of the combined coding plans (Figure 1).

2.1 Stage A. Initial processing

This stage involves processing the observed and Best Medium-Range (BMR) ensemble precipitation grids to obtain the catchment values needed for precipitation verification, and to reformat the data

into the appropriate format for input into G2G. For the river flow, this stage also involves running the required G2G forecasts using the processed BMR forecast input. To maximise efficiency, the precipitation grids are processed once, covering the requirements of both the rainfall and river flow verification. The processing of precipitation grids, particularly those from the BMR ensemble (around 27 million 2 km resolution precipitation grids; 24 members run for 596 time-steps, 4 times a day, for 487 days), requires large CPU and memory resources. To facilitate this processing in a reasonable time-period (around 1 month) the data are processed in daily chunks, using the LOTUS cluster from the JASMIN HPC facility. Catchment values are output in NetCDF file format (1 file per date for observations, 1 file per forecast and member for the BMR ensemble) for later input into the precipitation verification (Stages B and Stage C).

The initial processing had been coded in python and, for compatibility, will run using either python 2 or python 3. Only common modules are used to simplify the code-use on multiple platforms.

2.2 Stage B. Calculations involving the full 24 ensemble member data

The most computationally expensive scores to calculate are those that require data from all 24 ensemble member forecasts. For this project, two scores fall into this category: the CRPS and Rank Histogram. Both these scores can be calculated on a subset of the data and later combined to give scores for the full verification period. This has two advantages: firstly, the computational requirements are reduced to a manageable level and, secondly, it provides the flexibility to consider different lengths of verification period at a later date. For both the river flow and precipitation verification, the CRPS and Rank Histogram are calculated separately for each day of data and for each forecast-origin time.

To enable the threshold-based scores to be calculated, the ensemble probability of threshold-crossing (river flow) or threshold-exceedance (precipitation) must first be calculated from the full 24 ensemble member data. For computational efficiency (so that the full ensemble data are only read once), this is completed at the same time as the CRPS and Rank Histogram calculations. The ensemble probability time-series are saved for input into the threshold-based score calculation code. The binary time-series of observed river flow and precipitation threshold-crossing/exceedance are also calculated and saved.

To ensure that the same score-calculation code is used for both river flow and precipitation verification, all scores are calculated using the R “verification” package. There are currently no equivalent packages or modules available in python. Note that the python and R programs are fully independent to simplify use of the code across multiple platforms, and to give flexibility for future applications.

2.3 Stage C. Calculations involving the binary observations and ensemble probabilities

Consistent with the CRPS and Rank Histogram, the threshold-based scores are calculated from the ensemble probability and binary observation time-series using the R “verification” package. However, as the computational demands are smaller for the threshold-based score calculations due to only the ensemble probabilities being read in (instead of data from all 24 members), these scores are calculated directly for the full verification period. Different methods of pooling the data are applied before calculating the threshold-based scores (e.g. pooling over all sites nationally, regionally or by catchment properties). The scores are calculated separately for each pooling option, and saved for future plotting and analysis.

2.4 Stage D. Final outputs

The final coding stage relates to the plotting of the score results and the production of summary diagrams and displays. There is no further data processing at this stage: all the data have been processed, and scores calculated and saved, in stages A, B and C. Where appropriate, the same plotting code has been used for both the river flow and precipitation verification diagrams for consistency across the different displays. In Phase 1 of the project, R code was used to produce the plots and diagrams. Python code has also has also been developed for creating catchment maps and verification diagrams.

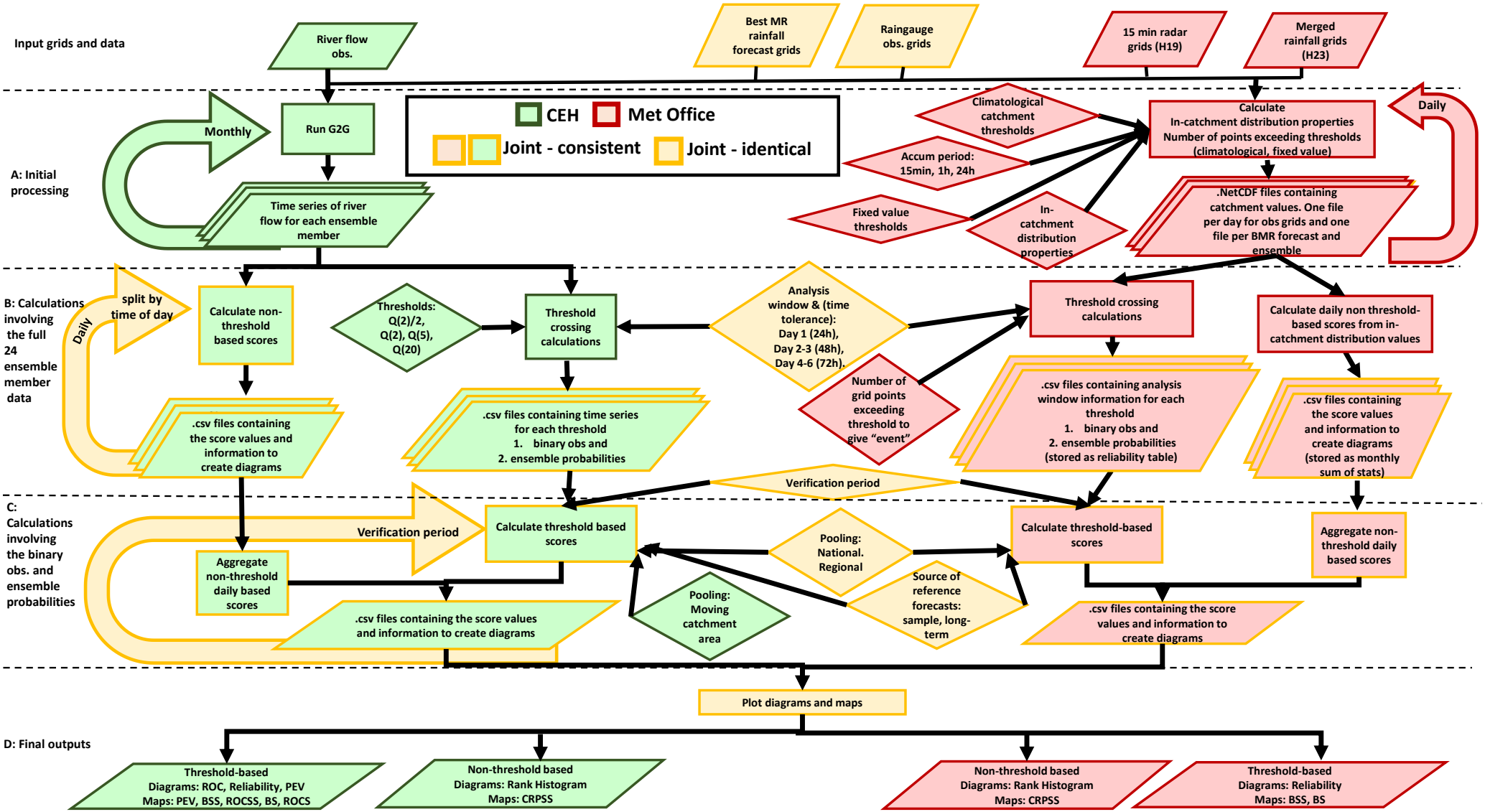
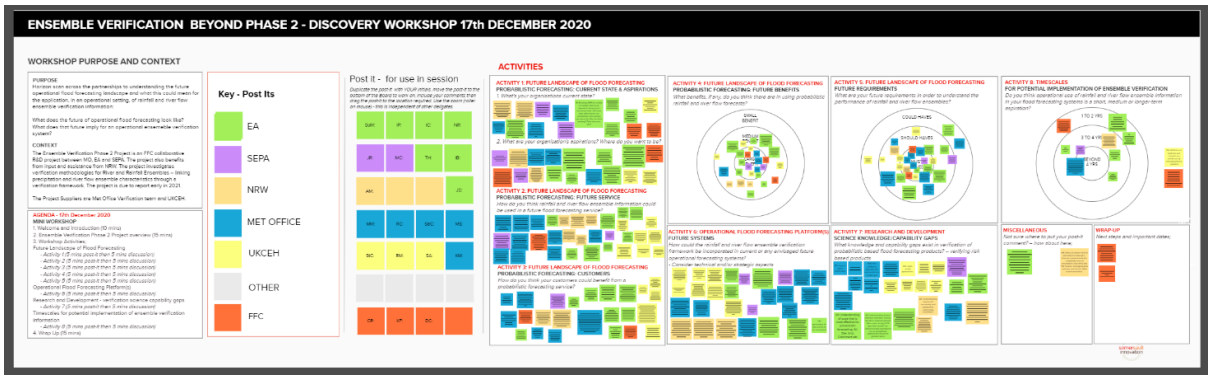


Figure 1 Detailed flowchart showing the technical details of the Joint Coding Framework, combining river flow and precipitation coding structure and data workflows.

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



FUTURE LANDSCAPE OF FLOOD FORECASTING

ACTIVITY 1: PROBABILISTIC FORECASTING: CURRENT STATE

a) What's your organisations current state?

Overall, people are not using ensembles (with a few exceptions) and people are not aware how best to use them. There are some emerging capabilities in systems to deal with ensembles but how to apply them to flood forecasting is poorly understood and there is no clear plan in many of the partner organisations.

ACTIVITY 1: PROBABILISTIC FORECASTING: ASPIRATIONS

b) What are your organisation's aspirations? Where do you want to be?

There is a clear and positive intention, at an organisational and at an individual level, that more use is made of probabilistic forecasting in the future. Indeed, there are emerging strategies in some of the represented organisations. For example, some want to underpin and improve their services in the future using probabilistic forecasting. There are varied aspirations on how to do this.

ACTIVITY 2: PROBABILISTIC FORECASTING: FUTURE SERVICE

How do you think rainfall and river flow ensemble information could be used in a future flood forecasting service?

There is a desire to improve decision making for flood forecasting by using ensembles combined with decision support tools or frameworks, perhaps particularly (but not exclusively) at longer lead-times. However, there is a sense from the responders that there is no coherent way on how to achieve this, with varied responses in how to apply ensembles to the forecasting problem. That said, there is a desire to understand uncertainty in an objective way and to use ensembles to provide a stable forecast flood risk narrative.

ACTIVITY 3: PROBABILISTIC FORECASTING: CUSTOMERS

How do you think your customers could benefit from a probabilistic forecasting service?

The responders thought the customers would benefit from a more 'honest' appraisal of confidence levels from a probabilistic based service. That would allow their customers to make better, more bespoke decisions and take more proportionate actions. They also thought a probabilistic based forecasting service will promote earlier discussions with customers and ultimately better outcomes, for example, around low confidence/high impact events. Efficiencies may be introduced such as using ensembles to inform automated, lower consequential decisions.

Dave Cox and Keith Fenwick, FFC

ACTIVITY 4: PROBABILISTIC FORECASTING: FUTURE BENEFITS

What benefits, if any, do you think there are in using probabilistic rainfall and river flow forecasts?

There is a consistent view from flood forecasting practitioners that using probabilistic rainfall and river flow forecasts will lead to medium to large future benefits. These may include; reduced subjectivity around confidence levels, reduce risk of missed impactful events, better operational decisions and ultimately reduced loss of life and improved damage mitigation.

ACTIVITY 5: FUTURE REQUIREMENTS

What are your future requirements in order to understand the performance of rainfall and river flow ensembles?

The future requirements should be considered by working in partnership across the partners and considering the whole forecasting chain with a clear purpose in mind. As the current state is set up largely for deterministic based flood forecasting services there is a need to understand where the greatest return in benefit lies with probabilistic forecasting. This extends to how the verification information is processed, delivered and visualised.

OPERATIONAL FLOOD FORECASTING PLATFORM(S)

ACTIVITY 6: FUTURE SYSTEMS

How could the rainfall and river flow ensemble verification framework be incorporated in current or any envisaged future operational forecasting systems?
- Consider technical and/or strategic aspects

Stakeholders need to recognise the importance of verification information and give it a high enough priority to provide sufficient resource and funding to implement within systems. Technical solutions are considered achievable but there's a prior need to fully understand the value of the verification data and where this should be best targeted. There is a need to set protocols or a 'minimum standard' on how a verification framework can be incorporated, including visualisation, into disparate systems. There is agreement that the processing of verification information is automated and should be integrated across rainfall and river flow. There are questions where the large volumes of data could be most efficiently hosted and a 'cloud' based approach integrated with FEWS is a potential solution.

RESEARCH AND DEVELOPMENT

ACTIVITY 7: SCIENCE KNOWLEDGE/CAPABILITY GAPS

What knowledge and capability gaps exist in verification of probabilistic based flood forecasting products? – verifying risk based products

Responders say there is a gap in knowledge in how best to apply ensemble forecasts to the flood forecasting problem. This stems from a lack of knowledge at what spatial and temporal scale ensembles are most effective. Verification of ensembles applied to extreme events is challenging with many variables needing to be accounted for.

TIMESCALES

ACTIVITY 8: FOR POTENTIAL IMPLEMENTATION OF ENSEMBLE VERIFICATION

Do you think operational use of rainfall and river flow ensemble information in your flood forecasting systems is a short, medium or longer-term aspiration?

Overall, the respondents comments placement suggested a medium term (3-4 years) aspiration of using rainfall and river flow ensemble information in the their flood forecasting systems. However, two responders (one FFC and one UKCEH) suggested it spanned short, medium and longer term aspirations as models, systems and capabilities evolve. One responder suggested beyond 4 years.

Dave Cox and Keith Fenwick, FFC