Hydrol. Earth Syst. Sci. Discuss., 11, 3811–3855, 2014
www.hydrol-earth-syst-sci-discuss.net/11/3811/2014/
doi:10.5194/hessd-11-3811-2014

Hydrology and Earth System Sciences
Open Access
Discussions

Alternative configurations of Quantile Regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River: a comparison

P. López López[1,2,*,**], **J. S. Verkade**[2,3,4], **A. H. Weerts**[2,5], **and D. P. Solomatine**[1,3]

[1]UNESCO – IHE Institute for Water Education, Delft, the Netherlands
[2]Deltares, Delft, the Netherlands
[3]Delft University of Technology, Delft, the Netherlands
[4]Ministry of Infrastructure and the Environment, Water Management Centre of the Netherlands, River Forecasting Service, Lelystad, the Netherlands
[5]Wageningen University and Research Centre, Wageningen, the Netherlands
[*]now at: Utrecht University, Utrecht, the Netherlands
[**]currently at: Deltares, Delft, the Netherlands

Received: 21 March 2014 – Accepted: 21 March 2014 – Published: 7 April 2014

Correspondence to: P. López López (patricia.lopez@deltares.nl)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

The present study comprises an inter-comparison of different configurations of a statistical post-processor that is used to estimate predictive hydrological uncertainty. It builds on earlier work by Weerts et al. (2011, herinafter referred to as WWV2011), who used the Quantile Regression technique to estimate predictive hydrological uncertainty using a deterministic water level forecast as a predictor. The various configurations are designed to address two issues with the WWV2011 implementation: (i) quantile crossing, which causes non-strictly rising cumulative predictive distributions, and (ii) the use of linear quantile models to describe joint distributions that may not be strictly linear. Thus, four configurations were built: (i) the "as is" implementation used by WWV2011, (ii) a configuration that implements a non-crossing quantile technique, (iii) a configuration where quantile models are built in Normal space after application of the Normal Quantile Transform, and (iv) a configuration that builds quantile model separately on separate domains of the predictor. Using each, four re-forecasting series of water levels at fourteen stations in the Upper Severn River were established. The quality of these four series was inter-compared using a set of graphical and numerical verification metrics. Intercomparison showed that reliability and sharpness vary across configurations, but in none of the configurations do these two forecast quality aspects improve simultaneously. Further analysis shows that skills in terms of Brier Skill Score, mean Continuous Ranked Probability Skill Score and Relative Operating Characteristic Score is very similar across the four configurations.

1 Introduction

Forecasting may reduce but can never fully eliminate uncertainty about the future. Hydrological forecasts will always be subject to many sources of uncertainty including those originating in the (estimated) initial state, boundary conditions, model structure and model parameters. Informed decision making may benefit from estimating the remaining uncertainties. A number of research studies suggest that enclosing predictive uncertainty estimates indeed leads to benefits to end users (Krzysztofowicz, 2001; Collier et al., 2005; Verkade and Werner, 2011; Ramos et al., 2013; Dale et al., 2014).

In the literature, various approaches to estimate predictive uncertainty have been presented. One of those is the use of meteorological ensemble forecasts, where initial atmospheric conditions are perturbed to yield an ensemble of atmospheric forecasts. These can be routed through a hydrological model, thus yielding an ensemble of hydrologic model forecasts which provides insight into the sensitivity of hydrological model results to various possible weather scenarios. Increasingly, hydrologic forecasting systems are including these ensemble predictions in the forecasting routines. An overview of applications and best practices was given by Cloke and Pappenberger (2009). More recent applications include the Environment Agency's National Flood Forecasting System – NFFS – (Schellekens et al., 2011) and the US National Weather Service's Hydrologic Ensemble Forecast Service HEFS (Demargne et al., 2014).

A second approach is statistical post-processing. Estimating predictive uncertainty through statistical post-processing techniques comprises an analysis of past, "observed" predictive uncertainty to build a model of future predictive uncertainties. It can be used as either an alternative or additional step to hydrological ensemble forecasting. In many hydrological forecasting applications, postprocessing is used in combination with deterministic forecasts (but it can also be applied to ensemble hydrologic forecasts if available; see, for example, Reggiani et al., 2009 and Verkade et al., 2013). A historical record of past forecasts and their verifying observations is then used to build a model of forecast error. (Note that other configurations are possible, but this one is the most straightforward and common one.) On the assumption that this error will be similar in future cases, the error model is then applied to newly produced deterministic forecasts, thus producing an estimate of predictive hydrological uncertainty. This estimate then includes uncertainties originating in both the atmospheric forecasts as well as those in the numerical simulation of streamflow generation and routing processes. Several hydrologic post-processors have been described in the scientific

HESSD
11, 3811–3855, 2014

literature, including the Hydrological Uncertainty Processor (HUP – Krzysztofowicz and Kelly, 2000), Bayesian Model Averaging (BMA – Raftery, 1993), the Model Conditional Processor (MCP – Todini, 2008), UNcertainty Estimation based on local Errors and Clustering (UNEEC – Solomatine and Shrestha, 2009), the Hydrologic Model Output Statistics (HMOS – Regonda et al., 2013) and Quantile Regression (QR – Weerts et al., 2011). The present paper focuses on the latter technique.

Quantile Regression (QR, Koenker and Bassett Jr., 1978; Koenker and Hallock, 2001; Koenker, 2005) aims to describe a full probability distribution of the variable of interest (the predictand), conditional on one or more predictors. Contrary to some of the other post-processors (such as HUP or BMA), QR requires few prior assumptions about the characterization of the model error. While it was originally developed for applications in the economic sciences, it has since been introduced into environmental modelling and climate change impact assessment (e.g. Bremnes, 2004; Nielsen et al., 2006). The technique has been applied in various research studies as a post-processing technique to estimate predictive hydrological uncertainty, including those de Weerts et al. (2011), Verkade and Werner (2011) and Roscoe et al. (2012). In each of these applications, the quantiles of distribution of the model error are estimated using single valued water level or discharge forecasts as predictors.

Weerts et al. (2011, hereinafter referred to as wwv2011) describe an implementation of QR for the Environment Agency in the UK. The "Historic Forecast Performance Tool" (HFPT; Sene et al., 2009) makes use of QR to estimate a predictive distribution of future water levels using the deterministic water level forecast as a predictor. The wwv2011 configuration of QR includes a transformation into Gaussian space using the Normal Quantile Transform (Krzysztofowicz and Kelly, 2000; Montanari and Brath, 2004; Bogner and Pappenberger, 2011). The problem of quantile crossing was addressed by omitting the domain of the predictor where the crossing occurred from the QR procedure and instead, in that domain, imposing a prior assumed distribution of the predictand.

HESSD
11, 3811–3855, 2014

The results of the wwv2011 analysis were verified for reliability and showed to be satisfactory. However, this verification was unconditional in the sense that only the full available sample of paired (probabilistic) forecasts and observation was assessed for reliability. When the HFPT was further tested (Vaughan, 2012), it was noticed that the probabilistic forecasts did not perform equally well in high flow conditions. One of the contributions of the present paper consists of a conditional analysis of forecast skill. Forecast skill is assessed for progressively higher flood levels, in terms of commonly used verification metrics and skill scores. These include Brier's probability Score, the Continuous Ranked Probability Score and corresponding skill scores as well as the Relative Operating Characteristic Score.

The configuration of QR in wwv2011 included two elements that, in the present paper, are explored in additional detail. These steps are (i) the technique for avoiding crossing quantiles and (ii) the derivation of regression quantiles in Normal space using the Normal Quantile Transform (NQT).

In wwv2011, crossing quantiles were avoided by manually imposing a distribution of the predictand in the domain of the predictor where crossing occurred. Since designing and implementing that particular configuration, an alternative technique for estimating non-crossing quantile regression curves has emerged (Bondell et al., 2010). As the latter technique requires less manual interference by the modelers, the present paper investigates whether implementation thereof yields estimates of predictive uncertainty that are of equal or higher quality.

In wwv2011, QR was applied using first degree polynomials, i.e. describing the distribution of the predictand as a linear function of the predictor. This, of course, assumes that the joint distribution of predictor and predictand can be described in linear fashion. To facilitate this, both marginal distributions (of forecasts and of observations) were transformed into Normal or Gaussian domain using the NQT. The joint distribution was subsequently described in Normal space using linear regression quantiles, and then back-transformed into original space. The resulting regression quantiles are then no longer linear. While this procedure yielded satisfactory results, there is no requirement

HESSD
11, 3811–3855, 2014

on the part of QR of either the marginal or joint distributions to be marginally or jointly Normal distributed. Also, the transformation and especially the back-transformation impose additional assumptions on the marginal distributions and can thus be problematic. Hence a justified question is whether this transformation to and from Normal space actually yields better results. In the present paper, this is tested by comparing multiple configurations of QR: derivation of regression quantiles in original space and in Normal space. As an additional step, a piecewise linear configuration is tested, where the domain of the predictor is split into several, mutually exclusive and collectively exhaustive domains, on each of which the regression quantiles are calibrated.

The objective of this work is to thoroughly verify uncertainty estimates using the implementation of QR that was used by wwv2011, and to inter-compare forecast quality and skill in various, differing configurations of QR. The configurations are (i) "classical" QR, (ii) QR constrained by a requirement that quantiles do not cross, (iii) QR derived on time series that have been transformed into the Normal domain, and (iv) a piecewise linear derivation of QR models. A priori, it is expected that imposing a non-crossing requirement yields results that are at least as good as those of the "classical" implementation of QR, and that derivation in Normal space and piecewise linear derivation each constitute a further improvement in quality and skill compared to derivation in original space.

The novel aspects and new contributions of this work include the thorough verification of an earlier implementation of QR, the application of the non-crossing QR to this particular case study and the exploration of techniques for ensuring that joint distributions can be described using linear QR models.

This paper first describes the approach, materials and methods, including the study basin, the hindcasting process, the analysed QR configurations and the verification process. Subsequently, results and analysis are presented. The paper ends with conclusions and discussion.

HESSD
11, 3811–3855, 2014

2 Approach, materials and methods

The present study consists of an experiment in which verification results of four differently configured post-processors (each based on the Quantile Regression technique) are inter-compared. By the varying configurations, two potential issues are addressed: quantile crossing and possible non-linearity of the joint distribution of predictor and predictand.

2.1 Study basin: Upper Severn River

The Upper Severn Basin (Fig. 1) serves as the study basin for the present study. Its predominantly hilly catchment extends from the Welsh Hills at Plynlimon to the gauge at Welshbridge in Shrewsbury and is approximately 2284 km^2 large. Lake Vyrnwy (Vyrnwy River) and Llyn Clywedog (Clywedog River) are two reservoirs located in the headwaters of the catchment. The Upper Severn includes rock formations classified as non-aquifers as well as loamy soils characterised by their high water retention capacity. Annual precipitation varies with topography from 700 to 2500 mm (EA, 2009). Flooding occurs relatively frequently, with major floods occurring in autumn 2000, February 2002, February 2004, summer 2007, fall 2012 as well as at the time of writing this manuscript, early 2014. To manage flood risk, the UK environment Agency developed the River Severn Catchment Flood Management Plan in 2009. Flood risk is also managed through the Midlands Flood Forecasting System (MFFS), which is based on the Delft-FEWS forecast production system (Werner et al., 2013). The Upper Severn configuration in MFFS consists of a sequence of numerical models for modeling of rainfall–runoff (MCRM; Bailey and Dobson, 1981), hydrological routing (DODO; Wallingford, 1994) and hydrodynamical routing (ISIS; Wallingford, 1997) processes as well as an internal MCRM error correction procedure based on the Autoregressive Moving Average (ARMA) technique. The input data for MFFS consists of Real Time Spatial (RTS) data (observed water level data, rain gauge data, air temperature and catchment average rainfall data), Radar Actuals, Radar Forecast, and Numerical Weather Prediction data. This input data is provided by the UK Meteorological Office.

The uncertainty models are used to estimate predictive uncertainty at fourteen hydrological stations on the Upper Severn River, each having different catchments characteristics. Figure 1 shows a map with the forecasting locations and their basins. Table 1 summarizes some key hydrological data.

2.2 Hindcasting process

The uncertainty models (Sect. 2.3) are derived using a joint historical record of observations and forecasts. The latter is acquired through the process of reforecasting or hindcasting. For this, a standalone version of the forecast production system MFFS is used. Prior to every forecast, the models are run in historical mode over the previous period to produce an estimate of internal states (groundwater level, soil moisture deficit, snow water equivalent, snow density, etc). In this historical mode, models are forced with observed precipitation, evapotranspiration and temperature. The system is subsequently run in forecast mode twice daily, with forecast issue times of 08:00 and 20:00 UTC, with a maximum lead time of 48 h. The selected reforecasting period is from 1 January 2006 through 7 March 2013. Of this period, the period up to 6 March 2007 is used to "spin up" the models. The remaining six years are used for the calibration and validation of the uncertainty models.

2.3 Uncertainty models

In the present study, predictive uncertainty is modelled using Quantile Regression (QR). The basic configuration is simple, and identical across all cases: the predictive distribution of future observed water levels is modeled as a series of quantiles, each estimated as a linear function of a single predictor which is the deterministic water level forecast. Four different configurations are inter-compared. Configuration Zero (QR0) constitutes the most straightforward case, where QR is applied "as is", i.e. in its most

HESSD
11, 3811–3855, 2014

basic form, in which no attempt is made to avoid crossing quantiles and no transformation or piecewise derivation is applied. Configuration One (QR1) addresses the problem of the crossing quantiles using the technique proposed by Bondell et al. (2010). If quantile crossing problem does not occur, this technique provides the same estimates as in the base scenario. Because of this, it is also applied to the remaining configurations. In some cases, the joint distribution of forecasts and observations is not best modelled using linear quantile regression models across the full domain of the predictor. However, by applying a transformation or by modelling sub-domains of the predictor, linear models may be used nonetheless. This is what is done in Configurations Two (QR2) and Three (QR3) respectively. The configurations are each described in detail in the following four subsections; for reference, they are also listed in Table 2. As the non-crossing quantiles are applied to Configurations Two through Four, the comparison in the present paper is effectively between these three latter configurations (QR1, QR2 and QR3).

The joint distribution of forecasts and their verifying observations is based on the EA archives of water level observations and on the forecasts from the hindcasting procedure. The available record is cross-validated through a leave-one-year-out cross-validation analysis. From the six years' worth of forecasts that are available for calibration and validation, five are used for model calibration and the single remaining year is used for model validation. Subsequently, another year is chosen for validation and the calibration period then comprises the remaining five years. This is repeated until all six years have been used for validation.

Uncertainty models are developed for each combination of lead time and location separately. While the forecasts have a maximum lead time of 48 h with one-hour intervals, the QR models are derived on a limited number of lead times, namely for 1 h lead time and then 3 through 48 h lead time with 3 h increments. The leave-one-year-out cross validation procedure yields approximately 3760 observation-forecast pairs for every combination of lead time and location.

HESSD
11, 3811–3855, 2014

2.3.1 QR0: Quantile Regression

Quantile Regression (Koenker and Bassett Jr., 1978; Koenker and Hallock, 2001; Koenker, 2005) is a regression technique for estimating the quantiles of a conditional distribution. As the sought relations are conditional quantiles rather than conditional means, quantile regression is robust with regards to outliers. Quantile Regression does not make any prior assumptions regarding the shape of the distribution; in that sense, it is a non-parametric technique. It is, however, highly parametric in the sense that for every quantile of interest, parameters need to be estimated.

In the present work, Quantile Regression is used to estimate lead time n specific conditional distributions of water level,

$$\phi_n = \{H_{n,\tau_1}, H_{n,\tau_2}, \ldots, H_{n,\tau_T}\} \tag{1}$$

where T is the number of quantiles τ ($\tau \in [0,1]$) considered. If T is sufficiently large and the quantiles τ jointly cover the domain $[0,1]$ sufficiently well, we consider ϕ_n to be a continuous distribution. Here, $T = 25$ and $\tau \in \{0.02, 0.06, \ldots, 0.98\}$,

$$\phi_n = \{H_{n,\tau=0.02}, H_{n,\tau=0.06}, \ldots, H_{n,\tau=0.98}\}. \tag{2}$$

We assume that (cf. wwv2011), separately for every lead time n considered, for every quantile τ there is a linear relationship between the water level forecast S and its verifying observation H,

$$H_{n,\tau} = a_{n,\tau} S_n + b_{n,\tau} \tag{3}$$

where $a_{n,\tau}$ and $b_{n,\tau}$ are the slope and intercept from the linear regression. Quantile Regression allows for finding the parameters $a_{n,\tau}$ and $b_{n,\tau}$ of this linear regression by, through a process of linear programming, minimising the sum of residuals,

$$\min \sum_{j=1}^{J} \rho_{n,\tau}(h_{n,j} - (a_{n,\tau} s_{n,j} + b_{n,\tau})) \tag{4}$$

HESSD
11, 3811–3855, 2014

where ρ is the quantile regression function for the τth quantile, h_j and s_j are the jth paired samples from a total of J samples, and $a_{n,\tau}$ and $b_{n,\tau}$ the regression parameters from the linear regression between water level forecast and observation. By varying the value of τ, the technique allows for describing the entire conditional distribution of the dependent variable H.

In the present work, solving Eq. (4) was done using the `quantreg` package (Koenker, 2013) in the R software environment (R Core Team, 2013). Figures 2–4 give a graphical overview of the resulting quantiles. These plots are discussed in Sect. 3.

2.3.2 QR1: non-crossing Quantile Regression

A potential problem with using QR for derivation of multiple conditional quantiles is that quantiles may cross, yielding predictive distributions that are not, as a function of increasing quantiles, monotonously increasing. `wwv2011` have addressed this issue by assuming a fixed error model in the domain of the predictor where there is the danger of quantiles crossing. In the present research study, the technique proposed by Bondell et al. (2010) is used. This technique imposes a non-crossing restriction on the solution of Eq. (4). Without this restriction, the solution to the proposed optimization problem is identical to that of classical quantile regression, i.e. to the models derived using QR0. The technique is freely available online (NCSU Statistics, 2010) and is coded in the statistical computing language R (R Core Team, 2013).

2.3.3 QR2: Quantile Regression in normal space

In this configuration, timeseries of water level observations and water level forecasts are first transformed into the Normal domain. This results in timeseries that are marginally Normal distributed. Subsequently, Quantile Regression models are calibrated using the non-crossing quantiles technique. Posterior to the derivation of QR models, the variables are back-transformed into original space. The rationale for using

HESSD

11, 3811–3855, 2014

the transformation is that the joint distribution of transformed timeseries appears to be more linear, and can thus be better described by linear conditional quantiles.

The Normal Quantile Transformation (NQT) is a quantile mapping or cdf-matching technique that matches the (empirical or modeled) cdf of the marginal distributions with a standard normal cdf. Here, the empirical cdf of the marginal distributions is used. Thus, the variables are mapped to a standard normal distribution,

$$H_{\mathrm{nqt}} = Q^{-1}(F(H)) \tag{5}$$
$$S_{\mathrm{nqt},n} = Q^{-1}(F(S_n))$$

where $F(\cdot)$ is the Weibull plotting position of the data point considered. The equivalent of Eq. (3) then becomes

$$H_{\mathrm{nqt},n,\tau} = a_{n,\tau} S_{\mathrm{nqt},n} + b_{n,\tau} \tag{6}$$

which is solved by minimising the sum of residuals,

$$\min \sum_{j=1}^{J} \rho_{n,\tau}(h_{\mathrm{nqt},n,j} - (a_{n,\tau} s_{\mathrm{nqt},n,j} + b_{n,\tau})). \tag{7}$$

Posterior to the analysis in normal space, the variables are back-transformed to original space using a reversed procedure,

$$H = Q(F(H_{\mathrm{nqt}})) \tag{8}$$
$$S_n = Q(F(S_{\mathrm{nqt},n})).$$

Back-transformation is problematic if the quantiles of interest lie outside of the range of the empirical distribution of the untransformed variable in original space. In those cases, assumptions will have to be made on the shape of the tails of the distribution. Some authors have chosen to parameterize the distribution of the untransformed variable and use those statistical models for the back-transformation (see, for example,

HESSD
11, 3811–3855, 2014

Krzysztofowicz and Kelly, 2000). In the present study, this matter is treated through a linear extrapolation on a number of points in the tails of the distribution which was the solution chosen by Montanari (2005) and by wwv2011.

2.3.4 QR3: piecewise linear Quantile Regression

In an effort to try and use linear quantile models to describe a joint distribution that may be slightly nonlinear in nature, Van Steenbergen et al. (2012) applied linear models to partial domains of the predictor. They found the resulting distributions to be both more reliable and sharper. Multiple, mutually exclusive and collectively exhaustive domains were identified based on a visual inspection of the data. As this selection more or less coincided with two splits at the 20th and 80th percentile, thus three sub-domains were defined, comprising 20, 60 and 20 % of the data respectively.

2.4 Verification strategy

To understand and inter-compare the performance of different QR configurations, an extensive verification of forecast quality skill was carried out. Forecasts were assessed for reliability and sharpness, and a number of metrics were calculated. The latter include the Brier Score (BS), the mean Continuous Ranked Probability Score (CRPS) and area beneath the Relative Operating Characteristic (ROCA).

Reliability was assessed using reliability diagrams, that plot the relative frequency of occurrence of an event vs. the predicted probability of event occurrence. Proximity to the 1 : 1 diagonal, where observed frequency equals predicted probability, indicates higher reliability. Sharpness was explored by determining the width of the centred 80 % interval of the predictive distributions; the full sample of these widths is shown by means of an empirical cumulative distribution function (ecdf). The Brier Score (Brier, 1950) is defined as the mean squared error of a probabilistic forecast of a binary event. The mean CRPS (Hersbach, 2000) is a measure of the squared probabilistic error in the forecasts across all possible discrete events. The area beneath the Relative Operating

HESSD
11, 3811–3855, 2014

Characteristic is a measure of the forecasts' ability to discriminate between the exceedence and non-exceedence of a threshold, for example a flood threshold. A detailed description of these measures with their mathematical formulation can be found in Appendix A.

To allow for comparison across different locations, BS, CRPS and ROCA are expressed as skill, thus becoming Brier Skill Score (BSS), Continuous Ranked Probability Skill Score (CRPSS) and the Relative Operating Characteristic Score (ROCS),

$$\text{skill} = \frac{\text{score} - \text{score}_{\text{ref}}}{\text{score}_{\text{perfect}} - \text{score}_{\text{ref}}} \tag{9}$$

where score is the score of the system considered, $\text{score}_{\text{ref}}$ is the score of a reference system and $\text{score}_{\text{perfect}}$ is the highest possible score. Skill scores range from $-\infty$ to 1. The highest possible value is 1. If skill = 0, the system's score is as good as that of the reference system. If skill < 0 then the system's score is less than that of the reference system. In the case of BSS and CRPSS, the reference score comprises that of the sample climatology; in case of the ROCS, the reference score is the ROCA associated with "randomness", i.e. $\text{ROCA}_{\text{ref}} = 0.5$.

As the post-processor is intended to be used in flood forecasting, forecast skill is not only assessed for the full available sample of forecast, observation pairs, but also for subsets of high and extreme events. These subsets are defined by the climatological probability of non-exceedence P of the observation. For example, $P = 0.95$ denotes the sub-sample of forecast, observation pairs where the observation falls in the top 5 % of observations. Increasing the value of P from 0 (i.e. the full available sample) to a value close to 1 thus gives an indication of forecast performance for high events.

By construction, sample size for the computation of every verification metric varies with the climatological probability of non-exceedance P considered (Fig. 5). Increasing the value of P means lower sample size. Sampling uncertainties of the verification metrics were explored by bootstrapping. The minimum sample size was set to 50. The applied resampling method estimates the sampling distribution of each verification

HESSD
11, 3811–3855, 2014

score. Here, the 5th and 95th percentiles of those distributions are shown. These thus constitute the centred 90 % confidence intervals.

Verification metrics were calculated using the Ensemble Verification System (Brown et al., 2010).

3 Results and analysis

Results were produced for each of the fourteen locations listed in Table 1 and all of the lead times were considered. For practical reasons, the present section includes results for a limited number of lead times and locations only: 24 h, 36 h and 48 h lead times at Llanyblodwel, Pont Robert and Welshbridge. This combination thus comprises forecasting locations with varying sizes of contributing area. Pont Robert is located upstream, Llanyblodwel somewhere in the middle, and Welshbridge at the very outlet of the Upper Severn Basin.

3.1 Uncertainty models

Uncertainty models for the three locations are shown in Figs. 2–4. All scatter plots show observed water levels on the vertical axis vs. water level forecasts on the horizontal axis. Each of the figures consists of a matrix of multiple panels, with rows showing the four configurations considered and columns showing various lead times. Note that across configurations, the scattered pairs are identical. On the scatter plots, a summary of the estimated uncertainty models is superimposed, consisting of a selection of quantiles only: $\tau \in \{0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99\}$. Note that these quantiles were derived for plotting purposes only, and do not necessarily coincide with the quantiles derived for verification. In the plots, the QR-estimated quantiles are shown in grey with the exception of the median quantile which is shown in blue.

From the figures (Figs. 2–4), a few general observations can be made. All scatter plots show that there is an obvious correlation of forecasted and observed water levels,

HESSD
11, 3811–3855, 2014

although in none of the combinations of location and lead time, all forecasts are equal to the observations. Spread of the forecast, observation pairs increases with increasing lead time. At zero lead time, the error correction technique ensures that modeled (i.e. simulated or forecasted) water levels are equal to the water level observation, hence at issue time there is no forecasting uncertainty. With increasing lead time, this uncertainty increases. The location with largest lag time (Welshbridge) shows spread that is more concentrated around the 1 : 1 diagonal than the other locations that have smaller contributing areas and shorter lag times. The location and slope of the quantiles show that in most cases, spread is modeled to be very small at low predicted values of the forecast, and increases with increasing value of the forecast.

The figures show how the uncertainty models, each based on a different configuration of Quantile Regression, differ from one another. Configurations 0 and 1 appear to be very similar. They differ only in those instances where the former configuration would lead to quantile crossing but are identical otherwise, which was indeed anticipated. Configurations 2 (derived using NQT transform) and 3 (piecewise linear approach) are quite different from the first two configuration, but not dissimilar to one another. In these configurations, the quantiles are not a linear function of the water level forecast, that is, not along the full domain. Note that this non-linearity constituted the very reason why these configurations were included in the analysis. Both models often – but not always – show a very small spread at lowest water level forecasts, followed by an increasing spread. At high water level forecasts however, spread no longer increases and sometimes decreases. This means that sharpness of the resulting probability forecasts then no longer reduces with increasing values of the water level prediction; sometimes it even increases.

From the pairs and the models, we can see that at both Llanyblodwel and Pont Robert, the deterministic forecast has a tendency towards underforecasting, i.e. to underestimate future water levels. This underforecasting is corrected for by the uncertainty models, that thus include a bias correction by resulting in a median forecast that is higher than the deterministic forecast. The joint forecast, observation distribution for Welshbridge shows that there is much less obvious underforecasting, or overforecasting for that matter.

3.2 Hydrographs

Hydrographs are shown in Fig. 6 at Welshbridge for a flood event that took place late 2011 and early 2012. The multiplot panel is composed by three columns representing three different lead times; 24, 36 and 48 h, and four rows for each of the four QR configurations. Each of these plots shows time in the horizontal axis, approximately 3 months and water level in the vertical axis. Deterministic forecast water level (green line), observations (blue dots), median quantile (light blue) and centered 50, 80 and 90 % confidence intervals are included (in shades of grey). Across the configurations for a particular lead time, water level observations and deterministic forecasts are identical.

From the plots, a number of observations can be made, each consistent with what was to be expected based on the QR models. Uncertainty increases with lead time, as is shown by the widest intervals at highest lead times, and vice versa. The deterministic forecast tends to underestimate water level observations. With increasing lead time, underforecasting increases. At 48 h lead time for high water levels the deterministic forecast provides a higher underestimation than for low and medium water levels, which is consistent with QR models shown in Fig. 4.

The probabilistic forecasts resulting from configurations 0 and 1 are quite similar to one another. They both show highest uncertainty at higher deterministic water level forecasts. Configuration 2 does not show this behaviour; at higher deterministic forecasts, probabilistic forecasts are sharper. Again, this is consistent with the QR model plots in Fig. 4. Configuration 3 results in forecasts whose width in the top 20 % of forecasts varies only slightly (at 24 h lead time) or almost not at all (at 36 and 48 h lead times) with the value of the predictor.

From a visual inspection, it appears that the median quantile obtained with the four QR Configurations improves the deterministic forecast. QR Configurations 0 and 1 provide a median quantile with a minor improvement. Differences between the median

HESSD
11, 3811–3855, 2014

quantile of QR Configuration 2 and the deterministic forecast are the lowest ones. QR Configuration 3 median quantile reproduces with the highest accuracy water level observations, including high, medium and low values.

3.3 Verification

3.3.1 Reliability and sharpness

Figures 7 and 8 show reliability diagrams for the full data sample and for the forecasts whose verifying observation falls in the top 10 % of observations ($P = 0.90$), respectively. When looking at the full available sample, the diagrams show reasonably high reliability: most plotting points are very near, or on the 1 : 1 diagonal. At 24 h leadtime, there is some underforecasting but this is no longer the case at the longer leadtimes shown.

At $P = 0.90$, forecasts are considerably less reliable. At all locations and at all leadtimes, there is considerable underforecasting at all but the tails of the predictive distributions. This overforecasting is more pronounced for the smaller basins, and vice versa. Forecasts from QR0 and QR1 are equally (un-)reliable. When comparing these to forecasts from QR2 and QR3, there is no configuration that yields more, or less, reliable forecasts across all cases. QR3 forecasts are nearly always among the least unreliable forecasts, although in many cases this is a shared position with varying other configurations.

Figures 9 and 10 show the distribution of width of the centred 90 % predictive intervals for the full available sample ($P = 0$) and the top 10 % of observations only ($P = 0.90$), respectively. The figures show that sharpness reduces with increasing leadtime as well as with increasing basin lag time. Inter-comparison of sharpness between the different cases shows that for the full sample (Fig. 9) there is little if any difference between the four configurations, and virtually none between QR0 and QR1. Forecasts for events that are more extreme ($P = 0.90$) show larger differences. Again, QR0 and QR1 yield forecasts of more or less equal width, but there are some differences

HESSD
11, 3811–3855, 2014

between these configurations and QR2 and QR3. These differences increase with increasing lead time and increasing basin lag time. At Welshbridge, QR2 yields sharpest forecasts, followed by QR3.

The old(-ish) adage has it that probabilistic forecasts should strive for sharpness subject to reliability (Gneiting et al., 2005). Unconditionally, both sharpness and reliability are more or less similar across four configurations. At $P = 0.90$ however, some forecasts are sharper than others but at the expense of reliability. On balance, usefulness of these forecasts may be equal. The trade-off between probability of detection and probability of false detection can be seen as a measure of this; the derived ROCS is analysed in the next section.

3.3.2 Skill scores

Figures 11–13 present the skill scores computed for probabilistic forecast verification. These plots show BSS, CRPSS and ROCS (vertical axes; each score on a new row) vs. the magnitude of the verifying observation, as a function of the observation which is expressed by its climatological probability of non-exceedance P (horizontal axes) for various lead times (columns). In each of the plots, results are shown for four QR configurations considered. To give an indication of the uncertainty in the estimation of metrics, median as well as 10 and 90 % estimates are shown.

From the figures, some general observations can be made. First of all, skills are mostly positive, with the exception of BSS and ROCS at the tails of P. Furthermore, skills deteriorate with increasing lead time, increase with increasing basin size and vary with the observation.

The Brier Skill Score (BSS) as a function of P has a concave, inverse U shape curve. BSS is lowest – sometimes even negative – at the tails of P and highest near median P. This is because BSS is calculated using event probabilities; and extreme events, whether low or high, are more difficult to correctly predict than non-extreme events. In terms of difference across the configurations: these are very limited. Only at the low tail do these become apparent, but often the differences are not significant.

HESSD
11, 3811–3855, 2014

Contrary to BSS and ROCS, CRPSS is a smooth, continuous measure that factors skill across all possible thresholds for each paired sample. This different formulation is reflected in its behaviour with increasing value of the observation. For short lead times, CRPSS is approximately constant. With increasing lead time, a small dip in CRPSS values is detected close to the median P. At nearly all lead times, the four QR configurations show very similar skill. The only exception is the highest lead times (48 h), in which QR Configuration 3 outperforms the remaining cases.

ROCS is a binary event skill with a similar formulation to BSS. However, ROCS values do not show the same pattern than BSS. ROCS is largely constant for the whole climatological distribution of the observations, as it can be seen at Welshbridge in Fig. 11. Pont Robert (Fig. 11) and Llanyblodwel (Fig. 12) present lower skill for the top half of the observations. Forecast quality decreases with increasing lead time, as it happens with BSS and CRPSS. No significant differences can be pointed out among the analysed QR Configurations.

4 Summary, conclusions and discussion

The research described in this paper had two objectives: (i) to extensively verify the estimates of predictive uncertainty for Upper Severn Basins that were produced using the Quantile Regression post-processing technique as described by wwv2011; (ii) to address two issues with the "as is" implementation of linear models of QR: (a) invalid predictive distributions due to the crossing quantile problem; (b) the description of slightly non-linear joint distributions by a linear QR model.

The verification of forecast quality builds on the verification that was carried out in an earlier paper (Weerts et al., 2011). In the present paper, multiple metrics and skill scores are presented. Also, a "conditional verification" was carried out, i.e. verification was done for a large number of sub-sets of available data, each representative for increasingly higher events. Verification showed that, unconditionally, in terms of all skills and metrics, forecast skill is positive. However, the analysis also shows that forecast quality and skill decreases with increasing value of the event.

The two issues described above were addressed by implementing several techniques, thus arriving at four configurations of Quantile Regression. The problem of crossing quantiles was addressed by adopting the non-crossing quantiles technique that was proposed by Bondell et al. (2010). This resulted in near-identical sharpness, reliability and skill. From a forecaster's point of view, the technique constitutes a methodological improvement as the post-processor will no longer produce invalid predictive distributions as a result of crossing quantiles, at no noticeable extra computational expense. The problem of linearly describing joint distributions of forecasts and observations that may not be linear in nature was addressed by two different approaches. The transformation to the Normal space attempted to produce a joint distribution that is "more linear". The piecewise linear derivation approach constitutes dividing the data into sub-samples on which the joint distribution is linear.

The intercomparison shows that none of the four Quantile Regression configurations consistently outperforms the others. Sharpness and reliability may vary across configurations, but there none results in a more favourable combination of the two. In terms of BSS, CRPSS and ROCS, the four configurations yield comparable forecast quality.

Addressing the problem of the non-linearity of the joint distributions by the solutions proposed in the present paper has not resulted in higher skill. Either the data was sufficiently linear for the techniques not to be required, or the techniques have not performed to expectation. In any case, a skill improvement does not provide a rationale for derivation of Quantile Regression models in Normal space as was done by wwv2011.

While none of the configurations has a proven higher skill, there may be alternative reasons for choosing one over the other. If the post-processors are used in operational forecasting systems, the forecasters will have to be able to explain to an end user how predictive uncertainty was estimated. Hence more complicated configurations will be less likely to be used. Also, forecasts have to be consistent with forecasters' beliefs

HESSD
11, 3811–3855, 2014

(Murphy, 1993), hence the post-processor will have to fit with the forecasters' perceptual model of forecasting error.

What would be a promising route to try and improve the skill of the estimates of predictive uncertainty that are produced by Quantile Regression? There are multiple possible answers here. First of all, there may be merit in adding predictors, i.e. by further conditioning forecast error on additional available variables. These could, for example, include the internal state variables of a model (dry or wet) and/or available observations at upstream locations. This route was taken by Solomatine and Shrestha (2009) in their UNEEC approach, and by Dogulu et al. (2014). Both compare a more complex UNEEC approach to QR and found improvement in skill. Stratification of the post-processing depending on different seasons or water level ranges could represent another alternative configuration. Another option would be to fully investigate additional configurations of the piecewise linear approach. For example, c-means or K-means clustering would allow for partitioning data to be used to build several regression models.

Appendix A

Verification metrics

For ease of reference, the probabilistic verification metrics used in this study are briefly explained. Further details can be found in the documentation of the Ensemble Verification System (Brown et al., 2010) as well as in reference works on forecast verification (Wilks, 2006; Jolliffe and Stephenson, 2012).

A1 Reliability diagrams

One desired property of probabilistic forecasts is that the predicted probabilities coincide with observed relative frequencies. Here, reliability diagrams are shown that, for each lead-time separately (indicator n is omitted from below equatons), plot the relative

HESSD
11, 3811–3855, 2014

frequency of non-exceedence of the estimated quantiles f_τ of the predictive distribution vs. the probability of non-exceedence τ,

$$f_\tau = \frac{\sum_{j=1}^{J} I_{\tau,j}}{J} \tag{A1}$$

where $I_{\tau,j}$ is an indicator variable $I_{\tau,j} = \begin{cases} 1 & \text{if } S_{\tau,j} < H_j; \\ 0 & \text{if } S_{\tau,j} \geq H_j \end{cases}$ that is determined for all j of J pairs of forecasts S and observations H.

A2 Sharpness

Sharpness is indicated by the width of the centred 80 % interval of the predictive distribution,

$$w_j = S_{\tau=0.90,j} - S_{\tau=0.10,j}, \tag{A2}$$

for all J forecasts. Again, sharpness is determined for each lead time n separately and the lead time indicators have been omitted from above equation. The combined record $w_{j=1,2,\ldots,J}$ is shown as an empirical cumulative distribution function.

A3 Brier score and Brier skill score

The (half) Brier Score (BS, Brier, 1950) measures the mean square error of J predicted probabilities that Q exceeds q,

$$\text{BS} = \frac{1}{J}\sum_{j=1}^{J}\{F_{X_j}(q) - F_{Y_j}(q)\}^2, \tag{A3}$$

where $F_{X_j}(q) = \Pr[X_j > q]$ and $F_{Y_j}(q) = \begin{cases} 1 & \text{if } Y_j > q; \\ 0 & \text{otherwise} \end{cases}$.

HESSD
11, 3811–3855, 2014

The Brier Skill Score (BSS) is a scaled representation of forecast quality that relates the quality of a particular system BS to that of a perfect system $BS_{perfect}$ (which is equal to 0) and to a reference system BS_{ref},

$$\begin{aligned} \mathrm{BSS} &= \frac{\mathrm{BS}-\mathrm{BS}_{\mathrm{ref}}}{\mathrm{BS}_{\mathrm{perfect}}-\mathrm{BS}_{\mathrm{ref}}} \\ &= \frac{\mathrm{BS}-\mathrm{BS}_{\mathrm{ref}}}{0-\mathrm{BS}_{\mathrm{ref}}} = \frac{\mathrm{BS}_{\mathrm{ref}}-\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} \\ &= 1-\frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} \end{aligned} \tag{A4}$$

BSS ranges from $-\infty$ to 1. The highest possible value is 1. If BSS = 0, the BS is as good as that of the reference system. If BSS < 0 then the system's Brier Score is less than that of the reference system.

A4 Mean continuous ranked probability score and skill score

The Continuous Ranked Probability Score (CRPS) measures the integral square difference between the cumulative distribution function (cdf) of the forecast $F_X(q)$, and the corresponding cdf of the observed variable $F_Y(q)$, averaged across J pairs of forecasts and observations,

$$\overline{\mathrm{CRPS}} = \frac{1}{J}\int_{-\infty}^{\infty}\{F_X(q)-F_Y(q)\}\mathrm{d}q. \tag{A5}$$

The Continuous Ranked Probability Skill Score (CRPSS) is a scaled representation of forecast quality that relates the quality of a particular system $\overline{\mathrm{CRPS}}$ to that of a perfect system $\overline{\mathrm{CRPS}}_{perfect}$ (which is equal to 0) and to a reference system $\overline{\mathrm{CRPS}}_{ref}$,

HESSD
11, 3811–3855, 2014
Alternative configurations of Quantile Regression
P. López López et al.

$$\mathrm{CRPSS} = \frac{\overline{\mathrm{CRPS}} - \overline{\mathrm{CRPS}}_{\mathrm{ref}}}{\overline{\mathrm{CRPS}}_{\mathrm{perfect}} - \overline{\mathrm{CRPS}}_{\mathrm{ref}}} = \frac{\overline{\mathrm{CRPS}} - \overline{\mathrm{CRPS}}_{\mathrm{ref}}}{0 - \overline{\mathrm{CRPS}}_{\mathrm{ref}}} = \frac{\overline{\mathrm{CRPS}}_{\mathrm{ref}} - \overline{\mathrm{CRPS}}}{\overline{\mathrm{CRPS}}_{\mathrm{ref}}} = 1 - \frac{\overline{\mathrm{CRPS}}}{\overline{\mathrm{CRPS}}_{\mathrm{ref}}} \tag{A6}$$

A5 Relative operating characteristic score

The relative operating characteristic (ROC; Green and Swets, 1966) plots the hit rate or probability of detection (PoD) vs. the false alarm rate or probability of false detection (PoFD). PoD and PoFD are calculated using the elements of a contingency table, which is valid for a single probabilistic decision rule, and are defined as follows

$$\mathrm{PoD} = \frac{\#\ \mathrm{hits}}{\#\ \mathrm{observed\ events}} = \frac{h}{o} \qquad \mathrm{PoFD} = \frac{\#\ \mathrm{false\ alarms}}{\#\ \mathrm{events\ not\ observed}} = \frac{f}{o'}. \tag{A7}$$

The ROC score measures the area under the ROC curve (AUC) after adjusting for randomness, i.e.

$$\mathrm{ROCS} = 2 \times (\mathrm{AUC} - 0.5). \tag{A8}$$

Acknowledgements. The authors would like to acknowledge the Environment Agency for provision of the data and use of the standalone NFFS system required to do the analyses described in the present manuscript.

HESSD
11, 3811–3855, 2014

Alternative configurations of Quantile Regression

P. López López et al.

References

Bailey, R. and Dobson, C.: Forecasting for floods in the Severn catchment, J. Inst. Water Eng. Sci., 35, 168–178, 1981. 3817

Bogner, K. and Pappenberger, F.: Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system, Water Resour. Res., 47, W07524, doi:10.1029/2010WR009137, 2011. 3814

Bondell, H. D., Reich, B. J., and Wang, H.: Noncrossing Quantile Regression curve estimation, Biometrika, 97, 825–838, doi:10.1093/biomet/asq048, 2010. 3815, 3819, 3821, 3831

Bremnes, J. B.: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output, Mon. Weather Rev., 132, 338–347, doi:10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2, 2004. 3814

Brier, G.: Verification of forecasts expressed in terms of probability, Mon. Weather Rev., 78, 1–3, 1950. 3823, 3833

Brown, J. D., Demargne, J., Seo, D.-J., and Liu, Y.: The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations, Environ. Modell. Softw., 25, 854–872, 2010. 3825, 3832

Cloke, H. and Pappenberger, F.: Ensemble flood forecasting: a review, J. Hydrol., 375, 613–626, 2009. 3813

Collier, C., Cross, R., Khatibi, R., Levizzani, V., Solheim, I., and Todini, E.: ACTIF best practice paper – the requirements of flood forecasters for the preparation of specific types of warnings, ACTIF International Conference on Innovation Advances and Implementation of Flood Forecasting Technology, Tromsø, Norway, 17–19, 2005. 3813

Dale, M., Wicks, J., Mylne, K., Pappenberger, F., Laeger, S., and Taylor, S.: Probabilistic flood forecasting and decision-making: an innovative risk-based approach, Nat. Hazards, 70, 159–172, 2014. 3813

Demargne, J., Wu, L., Regonda, S., Brown, J., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J., and Zhu, Y.: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, B. Am. Meteorol. Soc., 95, 79–98, doi:10.1175/BAMS-D-12-00081.1, 2014. 3813

Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H., and Shrestha, D. L.: A comparative analysis of predictive uncertainty estimation using Quantile Regression and UNEEC methods on several contrasting catchments, Hydrol. Earth Syst. Sci., in preparation, 2014. 3832

EA: Environment Agency: River levels: Midlands, available at: http://www.environment-agency.gov.uk/homeandleisure/floods/riverlevels/, last access: 1 October 2013. 3840

Gneiting, T., Raftery, A., Westveld, A., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Mon. Weather Rev., 133, 1098–1118, 2005. 3829

Green, D. M. and Swets, J. A.: Signal Detection Theory and Psychophysics, John Wiley & Sons Inc., New York, 1966. 3835

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecast., 15, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000. 3823

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: a Practitioner's Guide in Atmospheric Science, John Wiley & Sons, The Atrium, Southern Gate, Chichester, West Sussex, UK, 2012. 3832

Koenker, R.: Quantile Regression, Cambridge University Press, 2005. 3814, 3820

Koenker, R.: `quantreg`: Quantile Regression, R package version 5.05, available at: http://CRAN.R-project.org/package=quantreg, last access: 1 October 2013. 3821

Koenker, R. and Bassett Jr., G.: Regression Quantiles, Econometrica, 1, 33–50, 1978. 3814, 3820

Koenker, R. and Hallock, K.: Quantile Regression, J. Econ. Perspect., 15, 143–156, 2001. 3814, 3820

Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, J. Hydrol., 249, 2–9, 2001. 3813

Krzysztofowicz, R. and Kelly, K.: Hydrologic Uncertainty Processor for probabilistic river stage forecasting, Water Resour. Res., 36, 3265–3277, doi:10.1029/2000WR900108, 2000. 3814, 3823

Marsh, T. and Hannaford, J.: UK Hydrometric Register, Hydrological Data UK Series, Centre for Ecology and Hydrology, Wallingford, UK, 1–210, 2008. 3840

Montanari, A.: Deseasonalisation of hydrological time series through the Normal Quantile Transform, J. Hydrol., 313, 274–282, 2005. 3823

Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall–runoff simulations, Water Resour. Res., 40, W01106, doi:10.1029/2003WR002540, 2004. 3814
Murphy, A.: What is a good forecast? An essay on the nature of goodness in weather forecasting, Weather Forecast., 8, 281–293, 1993. 3832
Nielsen, H. A., Madsen, H., and Nielsen, T. S.: Using Quantile Regression to extend an existing wind power forecasting system with probabilistic forecasts, Wind Energy, 9, 95–108, 2006. 3814
R code software: available at: http://www4.stat.ncsu.edu/~bondell/software.html, last access: 1 October 2013, NC State 20 University Department of Statistics (NCSU Statistics), North Carolina, USA, 2010. 3821
R Core Team: R: a Language and Environment for Statistical Computing, available at: http://www.R-project.org/, R Foundation for Statistical Computing, Vienna, Austria, 2013. 3821
Raftery, A. E.: Bayesian model selection in structural equation models, Sage Foc. Ed., 154, 163–163, 1993. 3814
Ramos, M. H., van Andel, S. J., and Pappenberger, F.: Do probabilistic forecasts lead to better decisions?, Hydrol. Earth Syst. Sci., 17, 2219–2232, doi:10.5194/hess-17-2219-2013, 2013. 3813
Reggiani, P., Renner, M., Weerts, A., and Van Gelder, P.: Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system, Water Resour. Res., 45, W02428, doi:10.1029/2007WR006758, 2009. 3813
Regonda, S. K., Seo, D.-J., Lawrence, B., Brown, J. D., and Demargne, J.: Short-term ensemble streamflow forecasting using operationally-produced single-valued streamflow forecasts – a Hydrologic Model Output Statistics (HMOS) approach, J. Hydrol., 497, 80–96, 2013. 3814
Roscoe, K. L., Weerts, A. H., and Schroevers, M.: Estimation of the uncertainty in water level forecasts at ungauged river locations using Quantile Regression, Int. J. River Basin Manage., 10, 383–394, 2012. 3814
Schellekens, J., Weerts, A. H., Moore, R. J., Pierce, C. E., and Hildon, S.: The use of MOGREPS ensemble rainfall forecasts in operational flood forecasting systems across England and Wales, Adv. Geosci., 29, 77–84, doi:10.5194/adgeo-29-77-2011, 2011. 3813
Sene, K., Weerts, A., Beven, K., Moore, R., Whitlow, C., and Young, P.: Risk-based probabilistic fluvial flood forecasting for integrated catchment models? Phase 1 Report, Science Report SC080030/SR1, available at: http://evidence.environment-agency.gov.uk/FCERM, Joint

HESSD
11, 3811–3855, 2014

Defra/Environment Agency Flood and Coastal Erosion Risk Management Research and Development Programme, Rio House, Waterside Drive, Aztec West, Almondsbury, Bristol, BS32 4UD, 2009. 3814

Solomatine, D. P. and Shrestha, D. L.: A novel method to estimate model uncertainty using machine learning techniques, Water Resour. Res., 45, W00B11, doi:10.1029/2008WR006839, 2009. 3814, 3832

Todini, E.: A model conditional processor to assess predictive uncertainty in flood forecasting, Int. J. River Basin Manage., 6, 123–137, 2008. 3814

Van Steenbergen, N., Ronsyn, J., and Willems, P.: A non-parametric data-based approach for probabilistic flood forecasting in support of uncertainty communication, Environ. Modell. Softw., 33, 92–105, 2012. 3823

Vaughan, M.: Probabilistic flood forecasting (Environment Agency), 85th European Study Group with Industry, 16–20 April 2012, University of East Anglia, Norwich, 2012. 3815

Verkade, J. S. and Werner, M. G. F.: Estimating the benefits of single value and probability forecasting for flood warning, Hydrol. Earth Syst. Sci., 15, 3751–3765, doi:10.5194/hess-15-3751-2011, 2011. 3813, 3814

Verkade, J. S., Brown, J., Reggiani, P., and Weerts, A.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, J. Hydrol., 501, 73–91, doi:10.1016/j.jhydrol.2013.07.039, 2013. 3813

Wallingford: Wallingford Water, a flood forecasting and warning system for the river Soar, Wallingford Water, Wallingford, UK, 1994. 3817

Wallingford: HR Wallingford, ISIS software, available at: http://www.isisuser.com/isis/, last access: 1 October 2013, HR Wallingford, Hydraluic Unit, Wallingford, UK, 1997.. 3817

Weerts, A. H., Winsemius, H. C., and Verkade, J. S.: Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales), Hydrol. Earth Syst. Sci., 15, 255–265, doi:10.5194/hess-15-255-2011, 2011. 3812, 3814, 3830

Werner, M., Schellekens, J., Gijsbers, P., van Dijk, M., van den Akker, O., and Heynert, K.: The Delft-FEWS flow forecasting system, Environ. Modell. Softw., 40, 65–77, doi:10.1016/j.envsoft.2012.07.010, 2013. 3817

Wilks, D.: Statistical Methods in the Atmospheric Sciences, Academic Press, San Diego, California, USA, 2006. 3832

HESSD
11, 3811–3855, 2014

Table 1. Hydrometeorological and topographical information of analysed catchments at Upper Severn River (adapted from EA, 2013, and Marsh and Hannaford, 2008).

Station name	River	Basin area [km^2]	Elevation [m AOD]	Mean annual rainfall [mm]	Mean flow [m^3 s^{-1}]	Highest river level recorded [m]	Basin lag time [h]
Caersws	Severn	–	119	–	–	3.69	8–10
Abermule	Severn	580	83	1291	14.58	5.26	13–17
Buttington	Severn	–	62	–	–	5.5	8–10
Montford	Severn	2025	52	1184	43.3	6.96	10–15
Welshbridge	Severn	2025	47	–	–	5.25	15–20
Vyrnwy Weir	Vyrnwy	94.3	226.34	1951	4.24	1.8	2–5
Pont Robert	Vyrnwy	675	100	–	–	3.07	5–9
Meifod	Vyrnwy	675	81	–	–	3.67	7–10
Llanymynech	Vyrnwy	778	62	1358	21.08	5.19	3–6
Bryntail	Clywedog	49	212.05	2026	2.4	1.61	2–4
Rhos Y Pentref	Dulas	52.7	178.49	1313	1.45	2.42	1–3
Llanerfyl	Banwy	–	151	–	–	3.5	3–5
Llanyblodwel	Tanat	229	77.28	1267	6.58	2.68	7–10
Yeaton	Perry	108.8	61.18	767	1.6	1.13	15–20

HESSD
11, 3811–3855, 2014

Table 2. QR configurations used in the present study.

Identifier	Description
QR0	Classical Quantile Regression – Base scenario
QR1	Quantile Regression constrained by a non-crossing quantiles restriction
QR2	Quantile Regression, constrained by a non-crossing quantiles restriction, on the transformed data into Normal domain through Normal Quantile Transformation (NQT)
QR3	Piecewise linear derivation of Quantile Regression, constrained by a non-crossing quantiles restriction

HESSD
11, 3811–3855, 2014

Alternative configurations of Quantile Regression

P. López López et al.

Table A1. Contingency table.

	Event observed	Event NOT observed	$\sum$
Warning issued	hits h	false alarms f	w
Warning NOT issued	missed events m	quiets/correct negatives q	w'
$\sum$	o	o'	N

HESSD
11, 3811–3855, 2014

Alternative configurations of Quantile Regression

P. López López et al.

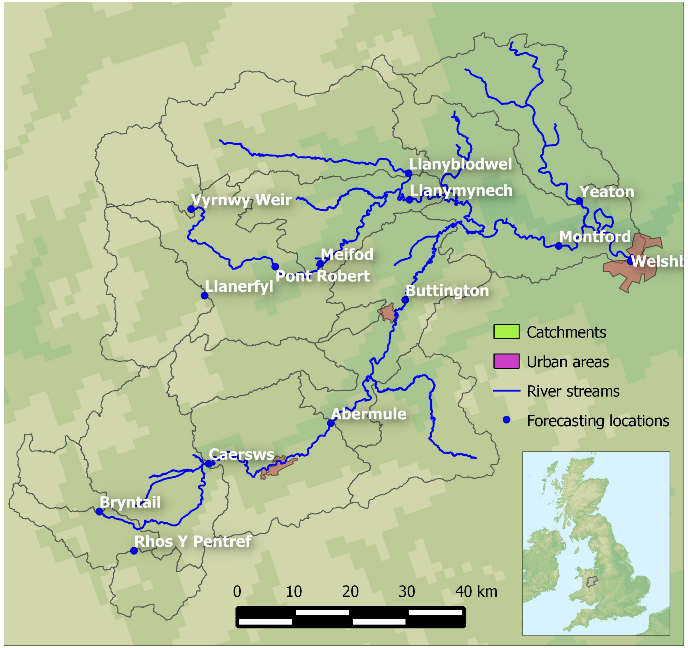

Fig. 1. The Upper Severn Basin including the 14 forecasting locations used in the present study. Note that the smallest river streams are not shown in the stream network. (The digital elevation model is made available by the European Environment Agency on a Creative Commons Attribution License; http://www.eea.europa.eu/data-and-maps/data/digital-elevation-model-of-europe.)

HESSD
11, 3811–3855, 2014

Alternative configurations of Quantile Regression

P. López López et al.

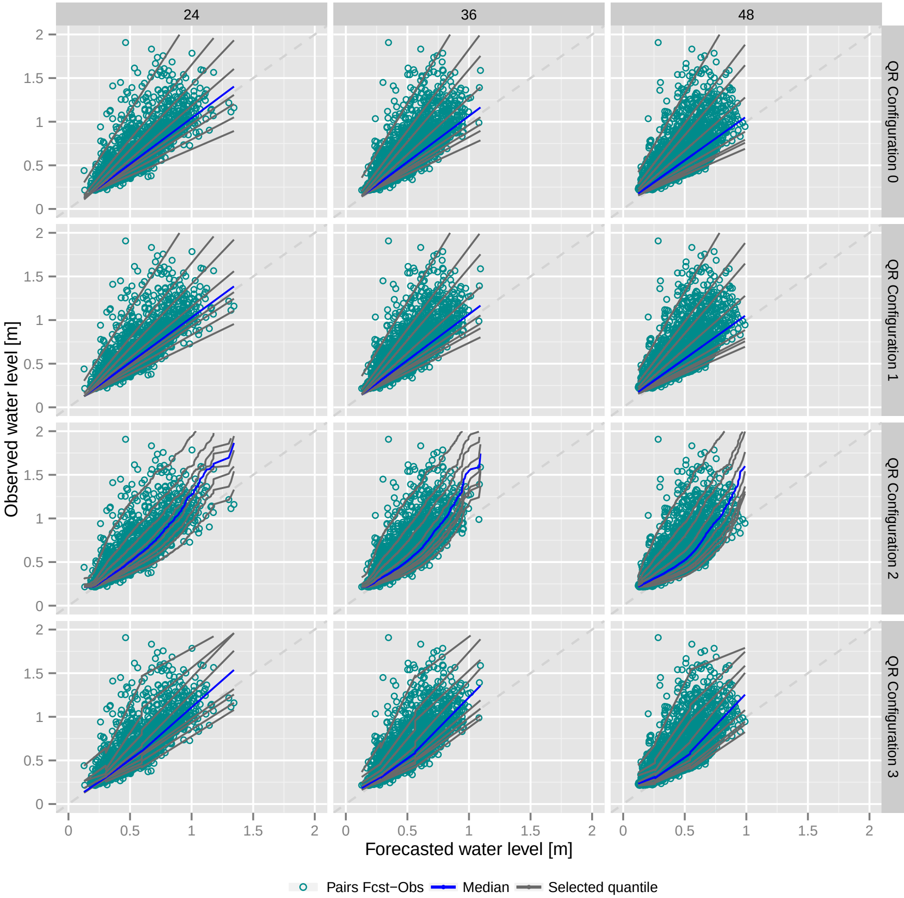

Fig. 2. Quantile Regression models for Llanyblodwel. Rows show the four different configurations; columns show different lead times.

HESSD
11, 3811–3855, 2014

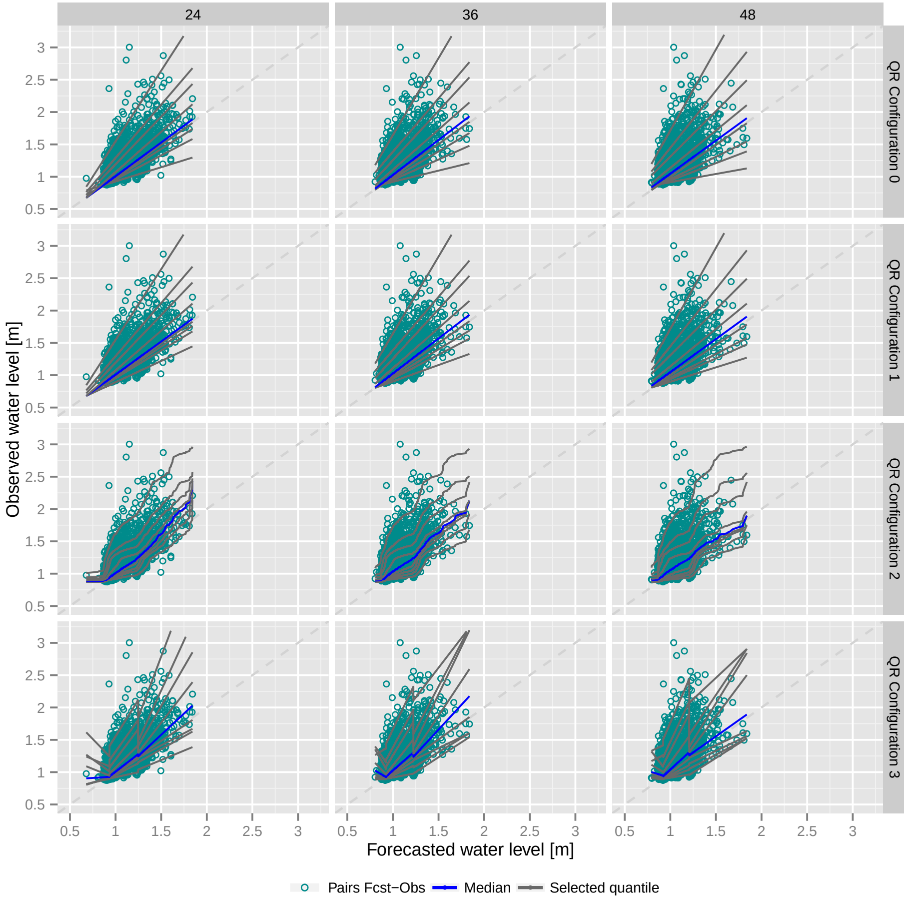

Fig. 3. Quantile Regression models for Pont Robert. Rows show the four different configurations; columns show different lead times.

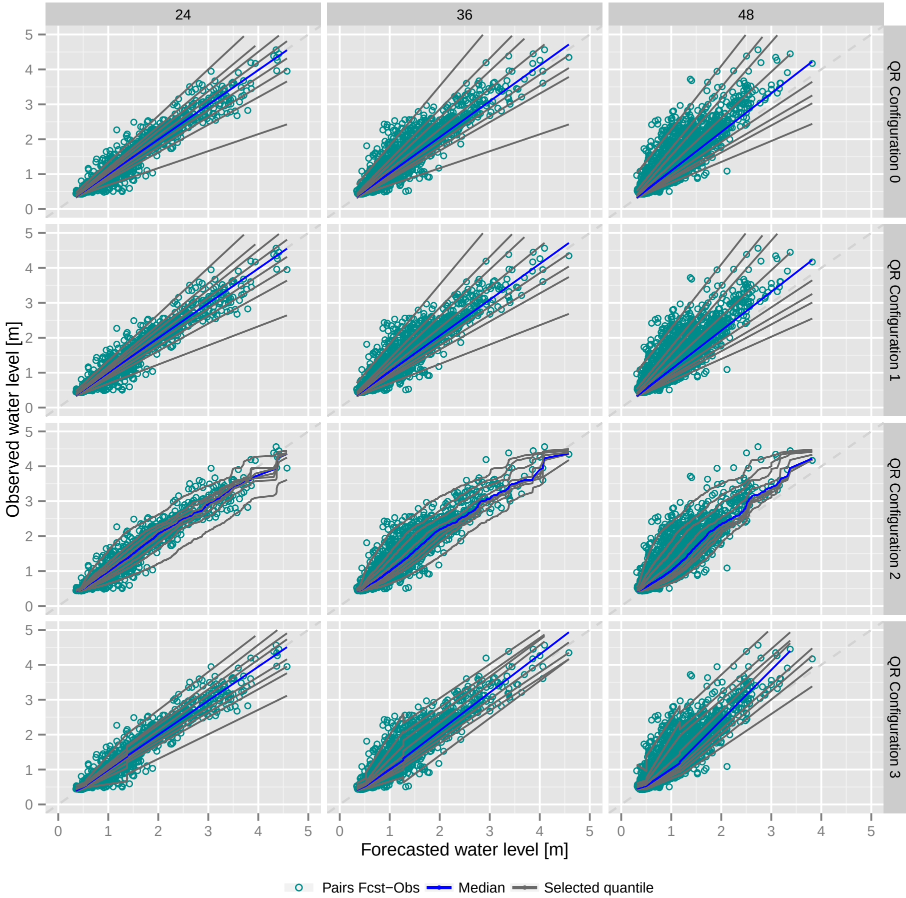

Fig. 4. Quantile Regression models for Welshbridge. Rows show the four different configurations; columns show different lead times.

HESSD
11, 3811–3855, 2014

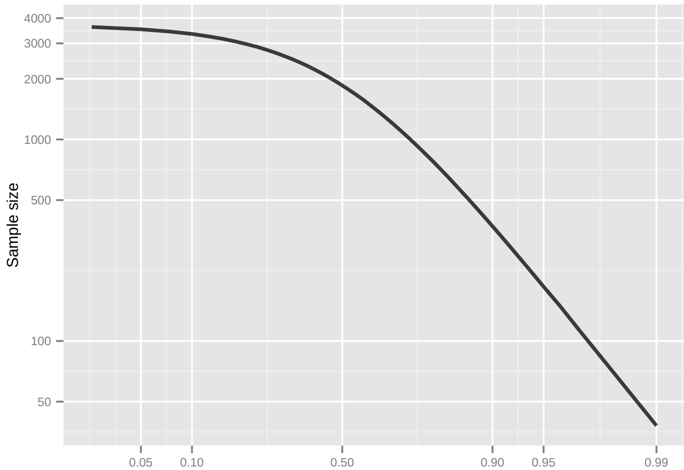

Fig. 5. Sample size as a function of the climatological probability of non-exceedence P.

HESSD
11, 3811–3855, 2014

Alternative configurations of Quantile Regression

P. López López et al.

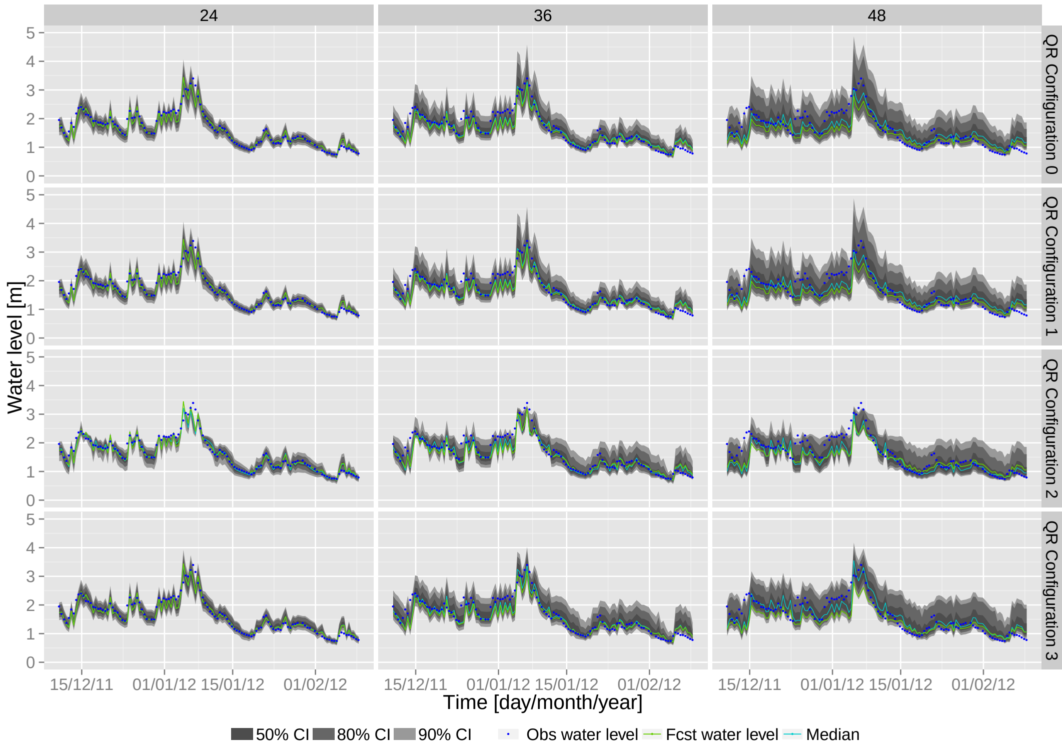

Fig. 6. Hydrographs of late 2011 and early 2012 events at Welshbridge.

HESSD
11, 3811–3855, 2014

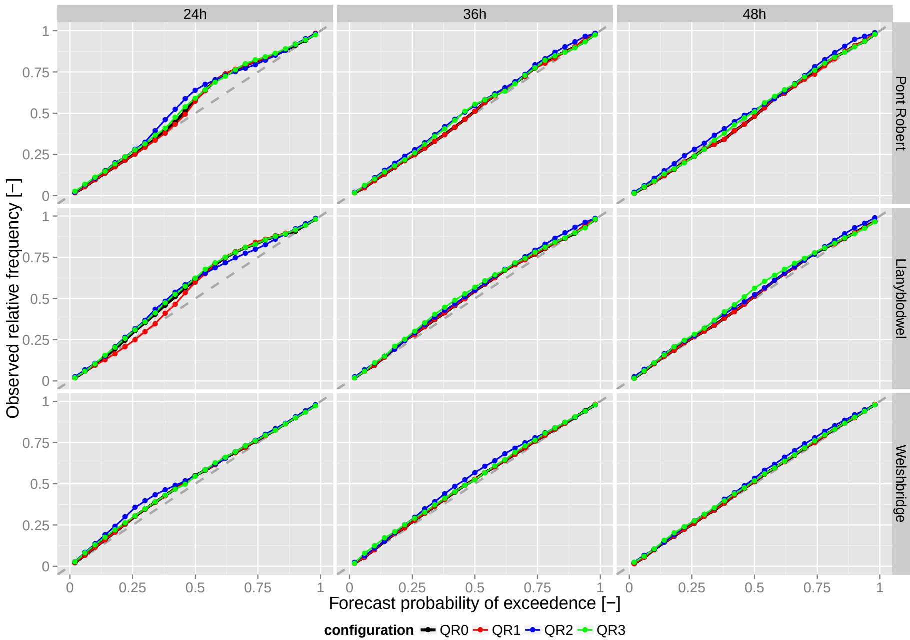

Fig. 7. Full sample reliability plots.

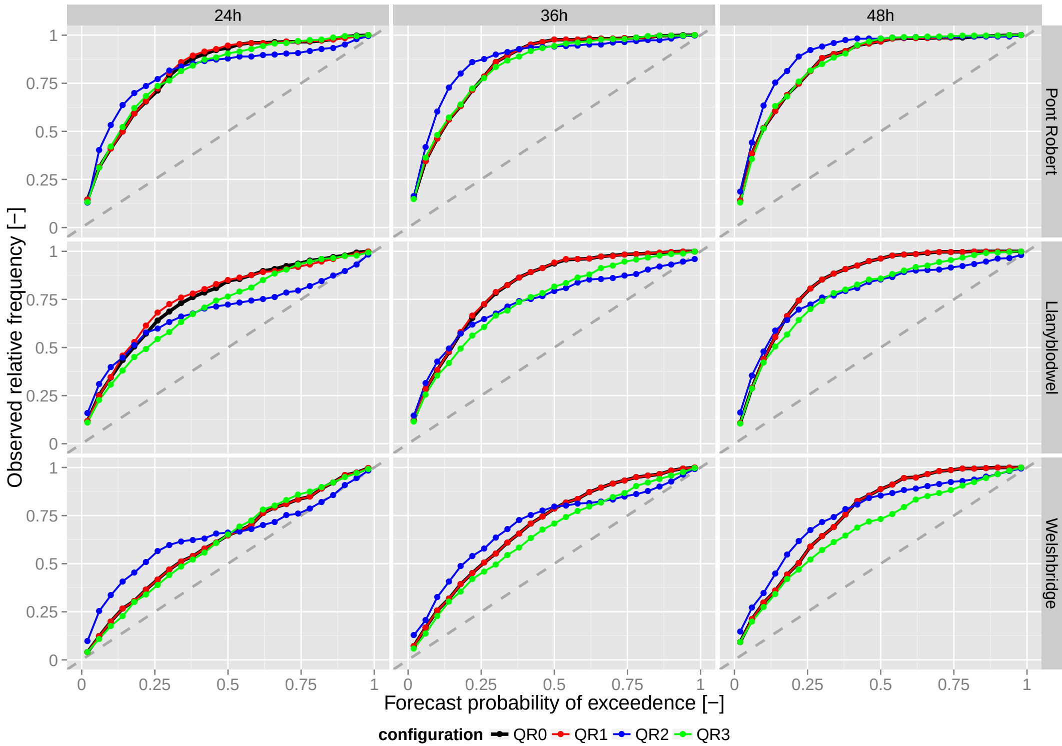

Fig. 8. Reliability plots for the forecasts associated with the top 10 % observations ($P = 0.90$).

HESSD
11, 3811–3855, 2014

Alternative configurations of Quantile Regression

P. López López et al.

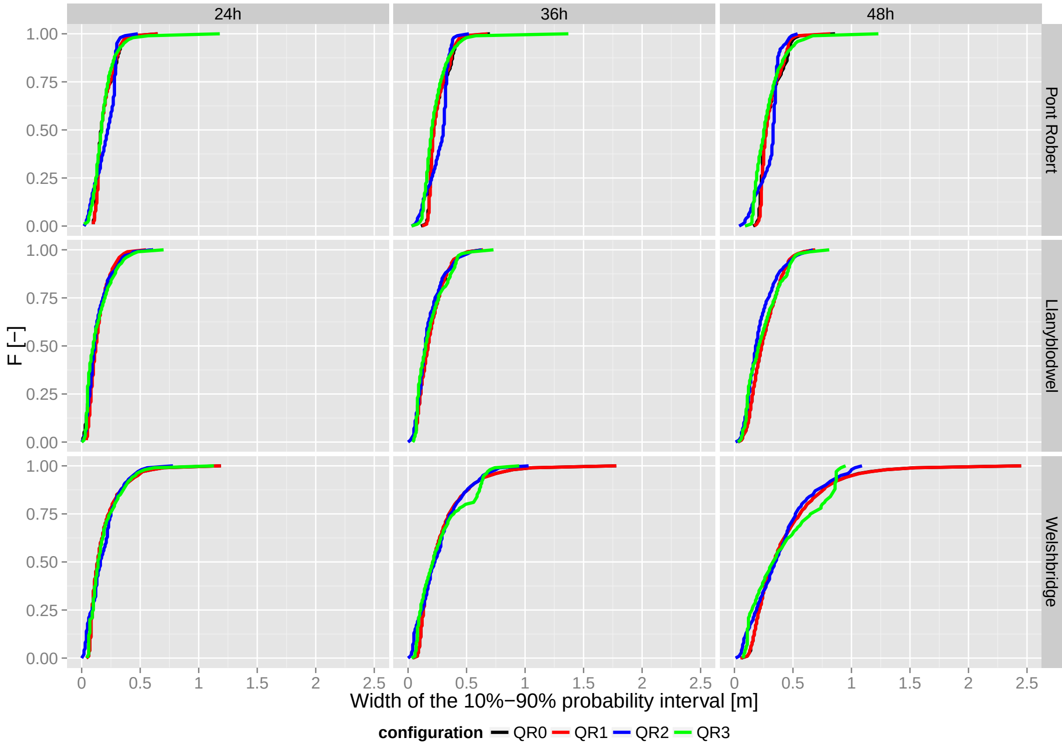

Fig. 9. Empirical cumulative distribution function of the centred 80 % confidence interval of the predictive distributions.

HESSD
11, 3811–3855, 2014

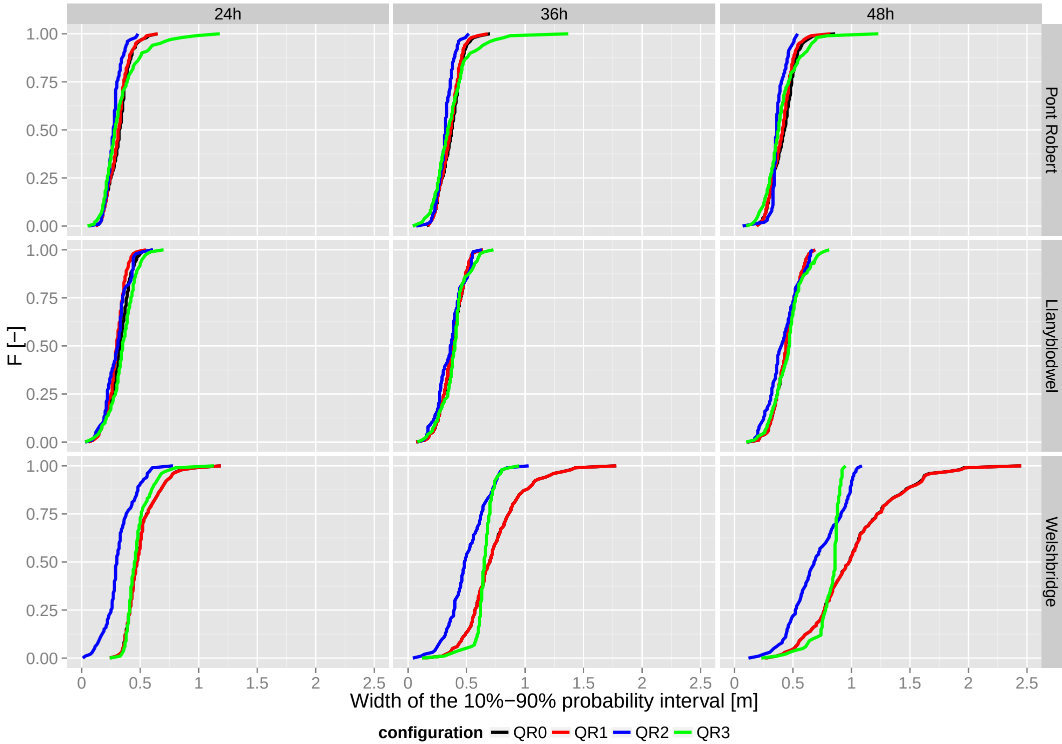

Fig. 10. Empirical cumulative distribution function of the centred 80 % confidence interval of the predictive distributions associated with the top 10 % observations.

HESSD
11, 3811–3855, 2014

Alternative configurations of Quantile Regression

P. López López et al.

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀ ▶
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

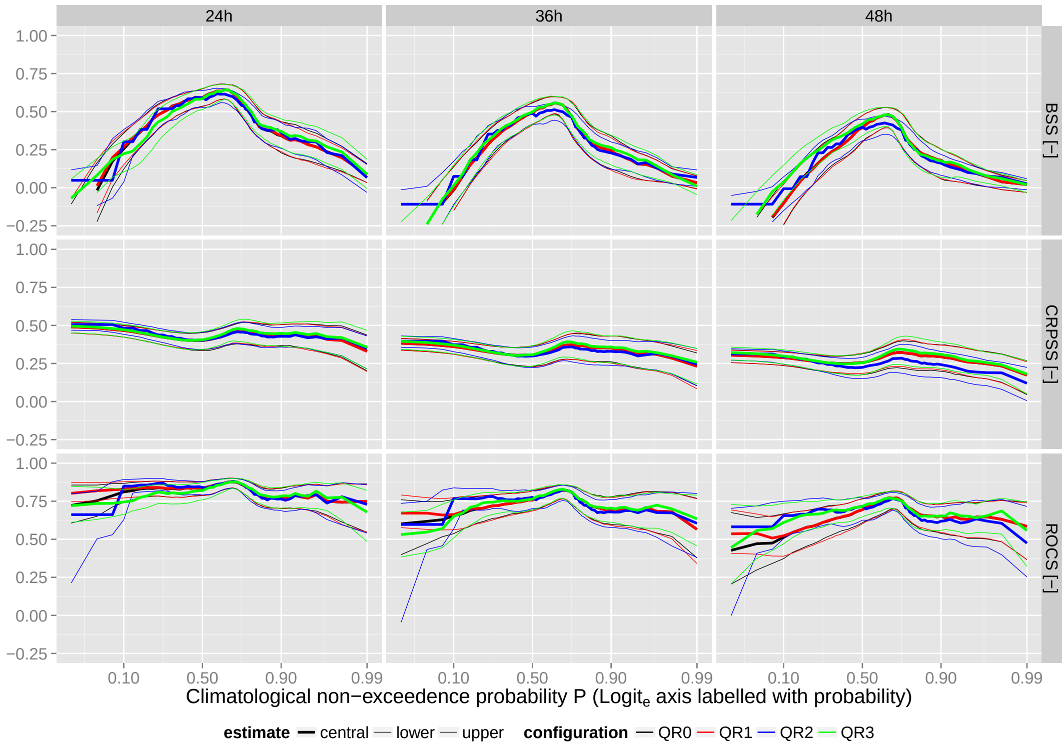

Fig. 11. Verification results for water level forecasts at Pont Robert station (5–9 h lag time). In the rows, three different metrics are shown; from top to bottom these are the Brier Skill Score (BSS), the mean Continuous Ranked Probability Skill Score (CRPSS) and the Relative Operative Characteristic Score (ROCS). Columns show various leadtimes: 24, 36 and 48 h.

HESSD
11, 3811–3855, 2014

Alternative configurations of Quantile Regression

P. López López et al.

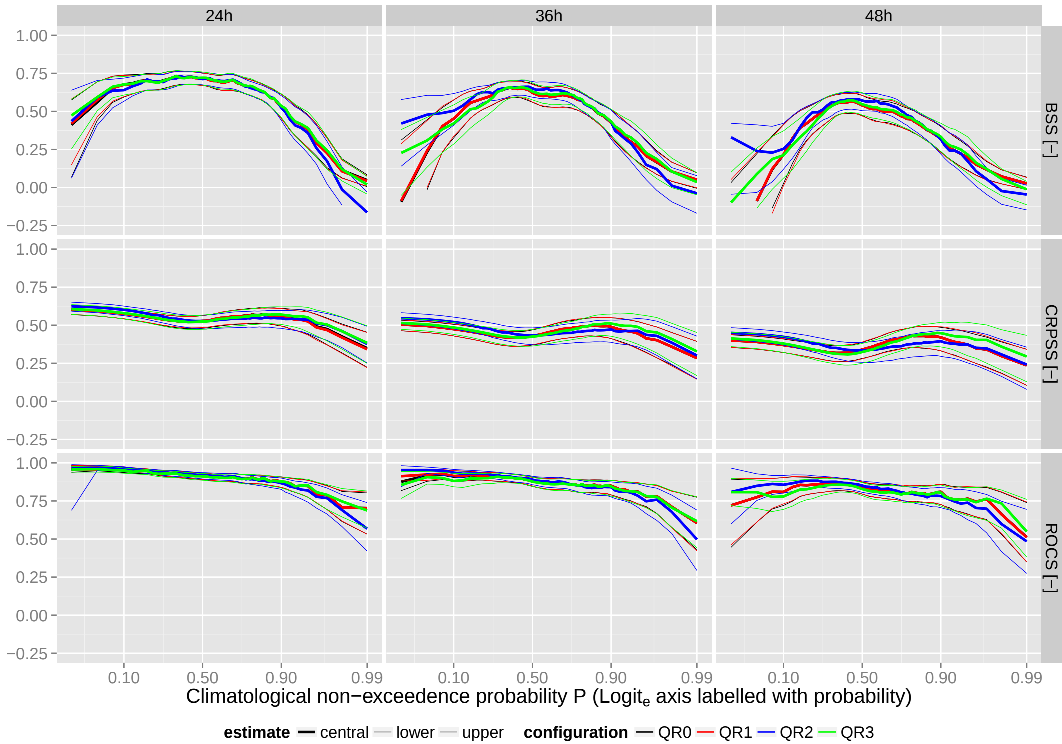

Fig. 12. Verification results for water level forecasts at Llanyblodwel station (7–10 h lag time). In the rows, three different metrics are shown; from top to bottom these are the Brier Skill Score (BSS), the mean Continuous Ranked Probability Skill Score (CRPSS) and the Relative Operative Characteristic Score (ROCS). Columns show various leadtimes: 24, 36 and 48 h.

HESSD
11, 3811–3855, 2014

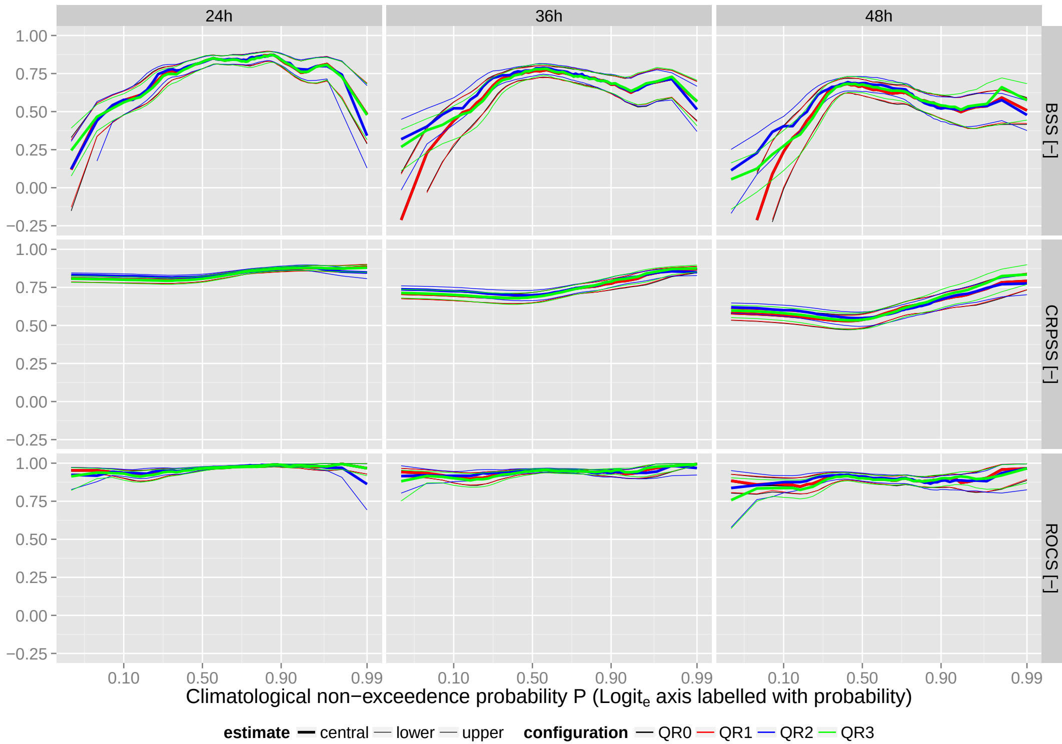

Fig. 13. Verification results for water level forecasts at Welshbridge station (5–9 h lag time). In the rows, three different metrics are shown; from top to bottom these are the Brier Skill Score (BSS), the mean Continuous Ranked Probability Skill Score (CRPSS) and the Relative Operative Characteristic Score (ROCS). Columns show various leadtimes: 24, 36 and 48 h.

HESSD
11, 3811–3855, 2014