Estimation of predictive hydrologic uncertainty using quantile regression and UNEEC methods and their comparison on contrasting catchments

N. Dogulu[1, *], P. López López [1, 2, **], D. P. Solomatine[1, 3], A. H. Weerts[2, 4], D. L. Shrestha[5]

[1]{UNESCO-IHE Institute for Water Education, Delft, the Netherlands}

[2]{Deltares, Delft, the Netherlands}

[3]{Delft University of Technology, the Netherlands}

[4]{Hydrology and Quantitative Water Management Group, Department of Environmental Sciences, Wageningen University, the Netherlands}

[5]{CSIRO Land and Water, Highett, Victoria, Australia}

*currently at: Dept. of Civil Engineering, Middle East Technical University, Ankara, Turkey

**now at: Utrecht University (Utrecht) and Deltares (Delft), the Netherlands

Correspondence to: N. Dogulu (ndogulu@metu.edu.tr)

Abstract

In operational hydrology, estimation of predictive uncertainty of hydrological models used for flood modelling is essential for risk based decision making for flood warning and emergency management. In the literature, there exists a variety of methods analyzing and predicting uncertainty. However, studies devoted to comparing performance of the methods to predict uncertainty are limited. This paper focuses on the methods predicting model residual uncertainty that differ in methodological complexity: quantile regression (QR) and UNcertainty Estimation based on local Errors and Clustering (UNEEC). The comparison of the methods is aimed at investigating how well a simpler method using less input data performs over a more complex method with more predictors. We test these two methods on several catchments from UK that vary in hydrological characteristics and the models used. Special attention is given to the methods' performance at different hydrological conditions. Furthermore, normality of model residuals in data clusters (identified by UNEEC) is analysed. It is found that basin lag time and forecast lead time have large impact on quantification of uncertainty and the presence of normality in model residuals' distribution. In general, it can be said that both methods give similar results. At the same time, it is also shown that UNEEC method provides better performance than QR for small catchments with the changing hydrological dynamics, i.e. rapid response catchments. It is recommended that more case studies of catchments of distinct hydrologic behaviour, with diverse climatic conditions, and having various hydrological features be considered.

1 Introduction

Importance of accounting for uncertainty in hydrological models used in flood early warning systems is widely recognised (e.g. Krzysztofowicz, 2001; Pappenberger and Beven, 2006). Such an uncertainty in the model prediction stems mainly from the four important sources: perceptual model uncertainty, data uncertainty, parameter estimation uncertainty, and model structural uncertainty (e.g. Solomatine and Wagener, 2011). Analysis of *predictive uncertainty* (Todini, 2008) of hydrological models used for flood modelling enable hydrologists and managers to achieve better risk based decision making and thus has the potential to increase the reliability and credibility of flood warning. Therefore, the necessity of estimating predictive uncertainty of rainfall-runoff models is broadly acknowledged in operational hydrology, and the management of uncertainty in hydrologic predictions has emerged as a major focus of interest in both research and operational modelling (Wagener and Gupta, 2005; Liu and Gupta, 2007; Montanari, 2007; Todini, 2008). In this respect comparing different methods, which are often developed and tested in isolation, receives attention of researchers, e.g. as suggested within the HEPEX framework (see van Andel et al., 2013).

While the discussions on the necessity of evaluating the contribution of various sources of errors to the overall model uncertainty are going for a long time (see, e.g. Gupta et al., 2005; Brown and Heuvelink, 2005; Liu and Gupta, 2007), there have been also attempts to estimate the *residual uncertainty*. By residual uncertainty, we understand the remaining model uncertainty assuming that other sources were accounted for (for example by calibrating the parameters), or not considered (all other sources like inaccurate rating curve, inputs, etc.) (Solomatine and Shrestha, 2009). We recognize that there are many sources of uncertainty leading to uncertainty in the model output (their influence is typically explored by running Monte Carlo experiments). However in this paper we consider the uncertainty of model outputs, assuming that parameters, inputs and the data used for model calibration are known (so we don't consider their uncertainty explicitly). Within this context, a (residual) model error is seen as a manifestation of the (residual) model uncertainty.

In this context, two classes of uncertainty analysis methods can be considered. The first one relates to the Bayesian framework with the meta-Gaussian transformation of data as its important part; these methods are based on a rigorous statistical framework. The following techniques and papers can be mentioned: the original Bayesian forecasting system (BFS) and the Hydrological Uncertainty Processor as its part (Krzysztofowicz, 1999; Krzysztofowicz and Kelly, 2000); its implementations and variations described in Montanari and Brath, 2004; Reggiani and Weerts, 2008; Reggiani et al., 2009; Bogner and Pappenberger, 2011; and the Model Conditional Processor (Todini, 2008; Coccia and Todini, 2011).

The other class of methods (of which two are dealt with in this paper) includes more "straightforward" ones which are directly oriented at predicting the properties (quantiles) of the residual error distribution by linear or non-linear regression (machine learning) techniques: quantile regression (QR) (Koenker and Basset, 1978) with its applications in hydrology reported by Solomatine and Shrestha, 2009; Weerts et al., 2011; López López et al., 2013); UNcertainty Estimation based on local Errors and Clustering (UNEEC) that uses machine learning techniques (Shrestha and Solomatine, 2006; Solomatine and Shrestha; 2009); dynamic uncertainty model by regression on absolute error (DUMBRAE) (Pianosi and Raso, 2012). In this paper we consider two methods from this class that differ in their methodological complexity: quantile regression (QR) and UNcertainty Estimation based on local Errors and Clustering (UNEEC).

Quantile regression (Koenker and Basset, 1978; Koenker and Hallock, 2001; Koenker, 2005) is basically a set of linear regression models (typically, two) where predictands (response variables) are the selected quantiles of the conditional distribution of some variables (discharge or water level in the present research study), and predictors are lagged values of the same variable. This methodology allows for examining the entire distribution of the variable of interest rather than a single measure of the central tendency of its distribution (Koenker, 2005). QR models have been used in a broad range of applications, such as economics and financial market analysis (Kudryavtsev, 2009; Taylor, 2007), agriculture (Barnwal and Kotani, 2013), meteorology (Bremnes, 2004; Friederichs and Hense, 2007; Cannon, 2011), wind forecasting (Nielsen et al., 2006; Møller et al., 2008), the prediction of ozone concentrations (Baur et al., 2004; Munir et al., 2012), etc. In hydrological modelling the QR method has been applied as an uncertainty post-processing technique in previous research studies with different configurations.

The configurations of QR differ mainly in two aspects: treatment of quantiles crossing problem (a problem when quantiles of the lower order appear to be larger than those of the higher order) and the quantiles derivation in Normal space using the Normal Quantile Transformation (NQT). Solomatine and Shrestha (2009) make use of the classical QR approach, without considering quantiles crossing and NQT. Weerts et al. (2011), Verkade and

Werner (2011), and Roscoe et al. (2012) apply QR to various deterministic hydrologic forecasts. QR configuration investigated in these studies uses the water level or discharge forecasts as predictors to estimate the distribution quantiles of the model error. It includes a transformation into Normal space using the NQT and the quantile crossing problem is addressed imposing a fixed distribution of the predictand in the crossing domain. Singh et al. (2013) make use of a similar configuration differentiating two cases based on the similarities in information content between calibration and validation data periods. Coccia and Todini (2011) observe that QR's usefulness and performance depend on the assumed patterns in quantiles, e.g. lack of linear variation of the error variance with the magnitude of the forecasts hinders reasonable estimation of the quantiles, especially for high flows/water levels. López López et al. (2014) apply QR to predict the quantiles of the environmental variables itself (water level) rather than the quantiles of the model error, and the four different configurations of QR are compared and extensively verified. It should be noted that by design, the only predictor in QR is the deterministic model output for discharge/water level, and the quantiles of observed discharge/water level are estimated through linear regression.

UNEEC method was introduced 10 years ago (Shrestha and Solomatine, 2006; Shrestha et al., 2006). The method builds a non-linear regression model (machine learning, e.g. an artificial neural network) to estimate the quantiles of the error distribution, and it assumes that residual uncertainty depends on the modelled system state characteristics so that any variable can be used as a predictor. A notable characteristic of UNEEC is special attention to achieve accuracy by local modelling of errors (by clustering and treating clusters separately) so that particularities of different hydrometeorological conditions, i.e. heterogeneities inherent to rainfall-runoff process, are represented through different error *pdf*s. Shrestha and Solomatine (2006) tested the UNEEC method on Sieve catchment in Italy based on the estimates of lower and upper prediction limits corresponding to 90% confidence level (CL). The method was also applied to a different catchment (Brue, in UK; HBV model) and it was compared to the Bayesian meta-Gaussian approach (Montanari and Brath, 2004), as well as the version of Monte Carlo technique GLUE (Beven and Binley, 1992). It was reported that the uncertainty estimates obtained by UNEEC were in fact more acceptable and interpretable than those obtained by the other methods. UNEEC was further extended to estimate several quantiles (thus approximating full *pdf* of the error distribution) and applied to Bagmati catchment in Nepal (Solomatine and Shrestha, 2009), and it was compared to several other methods including QR. It was found that UNEEC method generated consistent and interpretable results which are more accurate and reliable than QR. In the further study (Pianosi et al., 2010) UNEEC was extended so as to include parametric uncertainty (UNEEC-P), however local features of uncertainty were not considered. Nasseri et al. (2013) compared UNEEC with methods which are mainly based on the fuzzy extension principle: IMFEP (Incremental Modified Fuzzy Extension Principle) and MFEP (Modified Fuzzy Extension Principle). It has been shown that the methods provided similar performance on the two monthly water balance models for the two basins in Iran and France.

Solomatine and Shrestha (2009) presented their initial experiments to compare QR and UNEEC on one case study, and Weerts et al. (2011) discussed the experience with QR on another one. In this paper we go further and test the newer variants of these methods on several contrasting catchments that cover a wide range of climatic conditions and hydrological characteristics. The motivation here is to identify possible advantages and disadvantages of using QR and UNEEC methods based on their comparative performance, especially during flooding conditions (i.e. for the data cluster associated with high flow/water level conditions). The knowledge gaps regarding the use of the methods with different parameterizations are addressed. For example, we now incorporate in UNEEC the autoregressive component by considering past error values (in addition to discharge and effective rainfall) in one case study, and model outputs for the state variables soil moisture deficit (*SMD*) and groundwater level (*GW*) are used as predictors (in addition to water level) in another case study. In the QR version implemented, the linear regression model was established to predict the quantiles of observed water levels conditioned on simulated/forecasted water levels. Furthermore, we present results of statistical analysis of error time series to better understand (hydrological) models' quality in relation to its effect on uncertainty analysis results, and to discuss the assumption of normality in the model residuals, particularly in view of the clustering approach employed within the framework of UNEEC method. We apply methods to estimate predictive uncertainty in Brue catchment (southwest UK) and Upper Severn catchments - Yeaton, Llanyblodwel, and Llanerfyl (Midlands, UK).

It should be noticed that by design UNEEC uses a richer set of predictors than QR and a more sophisticated non-linear regression model, so the comparison between simple and complex models may seem unfair. However, more predictors may not bring more information needed for accurate prediction. Only experiments can allow for stating that for each particular case. Our experience with the data-driven models (and both QR and UNEEC are such) showed that adding more predictors does not necessarily mean higher accuracy on unseen data. Parsimony (Box, Jenkins, and Reinsel, 2008) often leads to better generalization. In this study we compare the two uncertainty prediction methods, with the aim of investigating if a simpler method using less input data may possibly perform better than the more complex method with more predictors. Overall, selection of the most appropriate uncertainty processor for a specific catchment is a matter of compromise between its complexity and accuracy in consideration of the data availability and also the characteristics of the catchment, and we believe the findings of such a comparative analysis could be useful for the operational hydrology community.

The remainder of the paper is structured as follows. The next section describes the residual uncertainty analysis methods (QR and UNEEC) and the validation measures used. Section 3 describes the studied catchments and the experimental setup. The results for error and uncertainty analyses are presented and discussed in Section 4. In Section 5 the main conclusions from the study and recommendations for future work are presented.

2 Methodology

2.1 Uncertainty analysis methods

2.1.1 Definitions

As in Solomatine and Shrestha (2009) and Weerts et al. (2011), we consider a deterministic (hydrological) model *M* of a catchment predicting a system output variable $\hat{y}$ given the input data vector x ($x \in X$), and the vector of model parameters θ. There are various sources of error associated with the model output (e.g. discharge), so the system response (i.e. actual discharge) can be expressed as:

$$y_{t+LT} = \hat{y} + e = M(x,\theta) + e \quad (1)$$

where e is the total residual error (in the remainder of the text, the terms "model error" and "model residual" is used interchangeably to refer to e); t is the (discrete) time. The model M can be used in two modes depending on the relation between the lead time (LT: the duration between time of forecast and time for which the forecast is made) of interest and the model time step (Δt):

$$\left\{ \begin{array}{ll} \text{simulation mode,} & LT = 1 \cdot \Delta\text{t} \\ \text{forecasting mode,} & LT > 1 \cdot \Delta\text{t} \end{array} \right\} \quad (2)$$

Given the model structure *M*, and the parameter set θ, the uncertainty analysis methods used in this study, namely QR and UNEEC, estimate the residual uncertainty of a calibrated hydrological model whose parameters and inputs are assumed to be known exactly. In this setup the different sources of uncertainty are not distinguished explicitly. In both methods, the uncertainty model *U* predicts the quantile value q^{τ} and is calibrated for different quantiles (τ), and for various lead times (LT) separately:

$$q^{\tau}_{t+LT} = U(I, \lambda) \quad (3)$$

where I is the input data vector, and λ is the vector of model parameters. In a simplest case when number of quantiles is 2, they form the CL (e.g. 90%) and the corresponding confidence interval, CI. The quantiles computed in this study are τ = 0.05, 0.25, 0.75, and 0.95 allowing for forming the 50% and 90% confidence intervals.

2.1.2 Quantile regression

As mentioned, several QR configurations have been previously investigated for estimating the residual uncertainty. In López López et al. (2014) (in open access) the four alternative configurations of QR for several catchments at the Upper Severn River have been compared and verified. The comparative analysis included different experiments on the derivation of regression quantiles in original and in normal space using NQT, a piecewise linear configuration considering independent predictand domains and avoiding the quantiles crossing problem with a relatively recent technique (Bondell et al., 2010). The intercomparison showed that the reliability and sharpness vary across configurations, but in none of the configurations do these two forecast quality aspects improve simultaneously. Further analysis reveals that skills in terms of the various verification metrics (i.e. Brier skill score, BSS; mean continuous ranked probability skill core, CRPSS; and the relative operating characteristic score, ROCS) are very similar across the four configurations. Therefore, noting also the main idea behind the current study (which is to investigate how well a simpler method using less input data performs over a more complex method with more predictors), the simplest QR configuration (termed there the “QR1: non-crossing Quantile Regresssion”) was applied in this study. QR1 estimates the quantiles of the distribution of water level or discharge in the original domain, without any initial transformation and avoids the quantiles crossing problem. A brief description of the QR configuration used in the present work is given below (for details the reader is referred to López López et al., 2014).

For every quantile τ, we assume a linear relationship between the forecasted (or predicted) value, $\hat{s}$, and the real observed value, s,

$$s = a_\tau \hat{s} + b_\tau \quad (4)$$

where a_τ and b_τ are the parameters of linear regression. By minimising the sum of residuals, one can find the parameters a_τ and b_τ:

$$\min \sum_{j=1}^{J} \rho_\tau (s_j - (a_\tau \hat{s}_j + b_\tau)) \quad (5)$$

where s_j and $\hat{s}_j$ are the j^{th} paired samples from a total of J samples and ρ_τ is the quantile regression function for the quantile τ:

$$\rho_\tau(\varepsilon_j) = \left\{ \begin{array}{cc} (\tau\text{ -}1)\cdot\varepsilon_j, & \varepsilon_j \le 0 \\ \tau\cdot\varepsilon_j, & \varepsilon_j \ge 0 \end{array} \right\} \quad (6)$$

Eqn. 6 is applied for the error (ε_j), which is defined as the difference between the observation (s_j) and the linear QR estimate ($a_\tau \hat{s}_j + b_\tau$) for the selected quantile τ.

Fig. 1 illustrates the estimation of a selection of quantiles, including 0.95, 0.75, 0.25 and 0.05 quantiles. To obtain the QR function for a specific quantile, e.g. $\tau = 0.05$, Eqns.(5) and (6) are applied as follows:

$$\rho_{0.05}(\varepsilon_j) = \left\{ \begin{array}{cc} -0.95\cdot\varepsilon_j, & \varepsilon_j \le 0 \\ 0.05\cdot\varepsilon_j, & \varepsilon_j \ge 0 \end{array} \right\} \quad (7)$$

In case of an ideal model, the 95 % of observed-forecasted pairs would be located above $\tau = 0.05$ quantile linear regression line, and 5 % would remain below it. Considering the two observed-forecasted pairs of the total of J samples, $j = 1$ and $j = 2$, their corresponding errors, ε_1 and ε_2, are:

$$\begin{array}{l} \varepsilon_1 = s_1 - (a_{0.05}\hat{s}_1 + b_{0.05}) < 0 \\ \varepsilon_2 = s_2 - (a_{0.05}\hat{s}_2 + b_{0.05}) > 0 \end{array} \quad (8)$$

Introducing both values in Eqn. (5), QR allows for solving the minimization problem calculating the regression parameters $a_{0.05}$ and $b_{0.05}$ for this particular quantile $\tau = 0.05$:

$$\min(-0.95 \cdot \varepsilon_1 + 0.05 \cdot \varepsilon_2 + ... + \rho_{0.05}(\varepsilon_J)) \quad (9)$$

The procedure explained here can be applied for any quantile, τ .

Figure 1. Quantile regression example scheme considering different quantiles.

2.1.3 UNEEC

In UNEEC, a machine learning model, e.g. an artificial neural network, instance-based learning (e.g. *k*-nearest neighbours) or a M5 model tree, is built to predict uncertainty associated with the model outputs corresponding to the future inputs to a (hydrological) model. The steps involved in UNEEC are summarized below:

- Identify the set of predictor variables (e.g. the lagged rainfall data, soil moisture, flow, etc.) that describe the flow process based on their effect on the model error. These predictors can be selected using Average Mutual Information (AMI) and correlation analysis. Using AMI brings the advantage of detection of nonlinear relationships (Battiti, 1994).
- Identify the fuzzy clusters in the data set in the space of predictor variables (using, e.g. fuzzy c-means method) (Fig. 2). The optimal number of clusters can be determined using the methods described, e.g. in Xie and Benie, 1991; Halkidi et al. 2001; Nasseri and Zahraie, 2011.
- For each cluster *c*, calculate the quantiles, q_c^τ , of the empirical distribution of the model error, taking into account however the membership degree of each data vector to a considered cluster.
- For each data vector, calculate the "global" estimate of the quantile q^τ using the quantiles calculated for each cluster q_c^τ. This is done by weighting them by the corresponding degree of membership of the given data vector to this cluster.

Calculated q^{τ} values for each quantile τ are used as outputs for the uncertainty model *U*.

- Train a machine learning model (*U*) (e.g. ANN or M5 model tree) using the set of predictors as inputs, and the data prepared at the previous step as the output. *U* will be able to predict the quantile value q^{τ} for the new input vectors.

Various machine learning models can be employed; in this study M5 model tree (Quinlan, 1992) has been used for all case studies. A model tree is a tree-like modular model which is in fact equivalent to a piecewise linear function. At non-terminal nodes there are rules that progressively split data into subsets, and finally the linear regression equations in the leaves of the tree built on the data subset that reached this particular leaf. Main reasons for using this technique are its accuracy, transparency (analytical expressions for models are obtained explicitly) and speed in training. Model trees have shown high accuracy in our previous studies (e.g. Solomatine and Dulal, 2003).

Figure 2. An example to fuzzy clustering of input data (the predictors are past rainfall at lag *t-2* and past flow at lag *t-1*) during training of the uncertainty model, U (adapted from *Solomatine*, 2013).

2.2 Validation methods

In this study we use several statistical measures of uncertainty to evaluate and to some extent to compare performances of QR and UNEEC. These are, namely, prediction interval coverage probability (PICP; Shrestha and Solomatine, 2006), mean prediction interval (MPI; Shrestha and Solomatine, 2006), and average relative interval length (ARIL; Jin et al., 2010). PICP has been also used by other authors (e.g. Laio and Tamea, 2007) as an important performance measure to estimate the accuracy of probabilistic forecasts.

PICP should be seen as the most important measure since it shows how many observations fall into the estimated interval. PICP is the probability that the observed values (y_t) lie within the estimated prediction limits computed for a significance level of $1-\alpha$ (e.g. 90%):

$$PICP = \frac{1}{n}\sum_{t=1}^{n} C \qquad \text{where} \qquad C = \begin{cases} 1, & PL_t^{lower} \le y_t \le PL_t^{upper} \\ 0, & otherwise \end{cases} \tag{10}$$

Ideally, PICP value should be equal or close to the specified CL.

MPI computes the average width of uncertainty band (or prediction interval), i.e. the distance between upper and lower prediction limits (PL_t^{upper} and PL_t^{lower}, respectively):

$$MPI = \frac{1}{n}\sum_{t=1}^{n}(PL_t^{upper} - PL_t^{lower}) \tag{11}$$

MPI = 0 means there is no uncertainty at all. MPI is rather simple indicator giving an idea about the distribution sharpness.

ARIL is similar to MPI and considers average width of uncertainty bounds in relation to the observed value:

$$ARIL = \frac{1}{n}\sum_{t=1}^{n}\frac{(PL_t^{upper} - PL_t^{lower})}{y_t} \tag{12}$$

Having the observed value in denominator accounts for the fact that uncertainty (and MPI) is usually higher for higher values of flow and thus has a “normalization” effect. A problem with ARIL is that if the flow is zero or close to zero, ARIL will be infinity or very high. This problem could be helped by removing all observations above a certain threshold from the calculations (a suggestion of one of the reviewers of this paper); we leave this idea for further testing in the future research.

A possibility to combine PICP and ARIL is to use the NUE indicator proposed by Nasseri and Zahraie (2011):

$$NUE = \frac{PICP}{w \times ARIL} \tag{13}$$

Nasseri and Zahraie (2013) recommend that methods with the higher NUE should be preferred over those with the lower NUE, however we do not think this is a universally applicable recommendation: if for two methods PICP is equal and close to the confidence interval (90%) and ARIL for one method is higher (which is not good), then NUE for this method will be actually lower.

There is no single objective measure of the quality of an uncertainty prediction method (since the “actual” uncertainty of the model (error *pdf*) at each time step is not known). Closer PICP is to the CL, higher the trust in a particular uncertainty prediction method should be. In principle, a reliable method should lead to reasonably low values of MPI (and ARIL). However, a wide MPI does not mean that a method estimating prediction interval is inaccurate – it could simply mean that the main model is not very accurate and the high MPI shows that.

PICP indeed evaluates if the expected percentage of observations fall into the predicted interval, and should be seen as an important average indicator of the predictor's performance. However, in case of high noise in the model error (aleatoric uncertainty) the fact that PICP is far from 90% could mean simply that none of the data-driven predictive models can capture the input-output dependencies and to predict quantiles accurately. For comparative studies however, PICP can very well be used: the method with PICP closest to 90% should be seen as the best (with some tolerance). Additional analysis may be carried out to see if the methods developed for the assessment of the probabilistic forecasts quality can be used (Laio and Tamea, 2007) (it is not exactly the same as the residual uncertainty analysed here but the mathematical apparatus seems could be transferrable). In this paper, however, we have not considered these so they can be recommended for exploration and testing in the future studies.

It is also worth mentioning that all considered measures are averages so should be used together with the uncertainty bound plots which visual analysis reveals more information on the capacity of different uncertainty prediction methods during particular periods.

3 Application

3.1 Case studies

3.1.1 Brue catchment

Located in the southwest of England, the Brue River catchment has a history of severe flooding. Draining an area of 135 km^2 to its river gauging station at Lovington (Fig. 3a), the catchment is predominantly rural and of modest relief and gives rise to a responsive flow regime due to its soil properties. The major land use is pasture on clay soil. The mean annual rainfall in the catchment is 867 mm and mean river flow is 1.92 m^3/s (basin average, 1961-1990) (Table 1). This catchment has been extensively used for research on weather radar, quantitative precipitation forecasting and rainfall-runoff modelling, as it has been facilitated with a dense rain gauge network (see, e.g. Moore et al., 2000; Bell & Moore, 2000)

The flow in Brue River was simulated by HBV-96 model (Lindström et al., 1997), which is an update version of the HBV rainfall-runoff model (Bergström, 1976). This lumped conceptual hydrological model consists of subroutines for snow accumulation and melt (excluded for Brue), soil moisture accounting procedure, routines for runoff generation, and a simple routing procedure (Fig. 3b). The input data used are hourly observations of precipitation (basin average), air temperature, and potential evapotranspiration (estimated by modified Penmann method) computed from the 15 minutes data. Model time step is one hour ($\Delta t = 1$ hr). The model is calibrated automatically using adaptive cluster covering algorithm (ACCO) (Solomatine et al., 1999). The data sets used for calibrating and validating the HBV-96 model are based on Shrestha and Solomatine (2008). It should be mentioned that the discharge data on calibration has many peaks which are higher in magnitude compared to those in the validation data.

Figure 3. (a) The Brue catchment showing dense rain gauges network and its river gauging station, Lovington, where the discharge is measured, and (b) Schematic representation of HBV-96 model (Lindström et al., 1997) with routine for snow (upper), soil (middle), and response (bottom) (Shrestha and Solomatine, 2008).

The uncertainty analyses conducted for Brue catchment are based on one-step-ahead flow estimates, i.e. $LT = 1$ hour (simulation mode). Effective rainfall (rainfall minus evapotranspiration) values were used instead of using rainfall data directly.

3.1.2 Upper Severn catchments: Yeaton, Llanyblodwel, and Llanerfyl

Flowing from Cambrian Mountains (610 meters) in Wales, the River Severn is the longest river in Britain (about 354 km). It forms the border between England and Wales and flows into the Bristol Channel. The river drains an area of approximately 10500 km^2 above the monitoring station at Upton on Severn. Mean annual precipitation ranges from approximately 2500 mm in the west to less than 700 mm in the south (EA, 2009). The Upper Severn includes rock formations classified as non-aquifers as well as loamy soils characterised by their high water retention capacity (for more detailed description of the Upper Severn, see Hill and Neal, 1997). Flooding is a major problem at the downstream due to excessive rainfall at the upstream (the Welsh hills), early 2014 floods being the most recent significant floods that occurred.

In this work, the three sub-catchments of Upper Severn River are analysed: Yeaton, Llanyblodwel, and Llanerfyl (Fig. 4). The area, elevation, mean flow, mean annual rainfall and basin lag time (time of concentration) information of the catchments are presented in

Table 1. Yeaton catchment is located at a lower elevation and over a flat area compared to Llanerfyl and Llanyblodwel. This catchment has also the longest basin lag time. The smallest catchment in terms of drainage area is Llanerfyl, which also has the shortest basin lag time (approx. 3-5 hours) leading to flash floods, so that the predictive uncertainty information on flood forecast for this catchment has especially high importance.

Figure 4. The Upper Severn catchments: Yeaton, Llanyblodwel and Llanerfyl.

Table 1. Summary of the main basin characteristics.

In Midlands Flood Forecasting System (MFSS; a Delft-FEWS forecast production system as described in Werner et al., 2013), the Upper Severn catchment is represented by a combination of numerical models for: rainfall-runoff modelling (MCRM; Bailey and Dobson, 1981), hydrological routing (DODO; Wallingford, 1994), hydrodynamic routing (ISIS; Wallingford, 1997), and error correction (ARMA). The input data used within MFSS includes (a) Real Time Spatial data (observed water level and raingauge data as well as air temperature and catchment average rainfall); (b) Radar Actuals, (c) Radar Forecasts, and (d) Numerical Weather Prediction data (all provided by the UK Meteorological Office). The data available was split into two parts for calibration (7 March 2007 08:00 – 7 March 2010 08:00) and validation (7 March 2010 20:00 – 7 March 2013 08:00), preserving similar statistical properties in both data sets.

The forecasting system issues two forecasts per day (08:00 and 20:00 UTC) with a time horizon of two days. First, the estimates of internal states are obtained running the models (which are forced with observed precipitation, evapotranspiration and temperature) in historical mode over the previous period. The state variables for the (hydrological) model are soil moisture deficit (*SMD*, the amount of water required to bring the current soil moisture content to field capacity in the root zone), groundwater level (*GW*), snow water equivalent (*SWE*), and snow density (*SD*). Using a standalone version of MFSS, the system (forced by the forecasted precipitation) is then run forward with a time step of 1 hour.

It is important to note that this case study, unlike Brue catchment, includes errors in the meteorological forecast and the back transformation of discharge to water level – via rating curve – in a lumped manner. Therefore, the effects of *rating curve uncertainty* (Di

Baldassarre and Montanari, 2009; Sikorska et al., 2013; Coxon et al., 2014; Mukolwe et al., 2014) and *precipitation forecast uncertainty* (Kobold and Sušelj, 2005; Shrestha et al., 2013) are accommodated as well.

The uncertainty analysis is aimed at estimating predictive uncertainty for the forecast time series ($\Delta t = 12$ hrs) corresponding to the lead time of interest. In this study, we consider the lead times $LT = 1$, 3, 6, 12, and 24 hours only.

3.2 Experimental setup

In all case studies the QR uncertainty prediction method employs a linear regression model. While in Brue catchment the linear regression model estimates the quantile τ of observed discharge conditioned on simulated discharge, in Upper Severn catchments the linear regression model estimates the quantile τ of observed water level conditioned on forecasted water level. In UNEEC the M5 model tree is employed as the prediction model. Selection of best set of the input variables for UNEEC is based on AMI and correlation analysis, and the number of clusters is identified by the model-based optimization. UNEEC is configured differently for each case, as described below.

3.2.1 Brue catchment

Shrestha and Solomatine (2008) tested UNEEC method on Brue catchment to assess residual uncertainty of the one-step-ahead flow estimates. The predictors of model error identified using AMI and correlation analysis were only lagged discharge (Q_{t-1}, Q_{t-2}, Q_{t-3}) and effective rainfall (RE_{t-8}, RE_{t-9}, RE_{t-10}) values. In this study, however, we try a different set of predictors. In addition to the mentioned variables, we consider also the two most recent past error values (e_{t-1}, e_{t-2}), allowing thus for incorporating the autoregressive features (for this case study it paid off - MPI values decreases (< 5%) during both training and test periods). As in the previous study the number of clusters used was 5.

3.2.2 Upper Severn catchments: Yeaton, Llanyblodwel, and Llanerfyl

In the Upper Severn case studies, a variety of predictors are considered for the model, e.g. observed and modelled water level, forecasted precipitation, and state variables (*GW*, *SMD*, *SWE*, *SD*). Although the benefits of using the soil moisture (observed or modelled) and groundwater level information for modelling rainfall-runoff processes and predicting runoff is well known in the literature (Aubert et al., 2003; Lee and Seo, 2011; Tayfur et al., 2014), we cannot cite any studies exploring the possible advantages of using such information for improving predictive capabilities of uncertainty analysis methods. Therefore, the dependence of model residuals on variables expressing internal state of the catchments is also analysed.

Among the state variables, the most significant correlation with the model error was shown by *GW* and *SMD.* While *GW* was found to be positively correlated with model residuals (i.e. as *GW* increases, error increases too), *SMD* and model error had a negative correlation. The positive correlation between *GW* and model residuals can be explained by the fact that high groundwater levels are associated with excessive precipitation during which model error are higher in magnitude. High soil moisture deficit, on the other hand, indicates that there has been no excessive precipitation and the soil is not filled up with infiltrated water. High evaporation rates (causing soil to dry up) can also result in high soil moisture deficit. It should be noted that the latter is less likely to be valid for the Upper Severn catchments considering the prevailing climate in the region. Accordingly, lower soil moisture deficit is linked with excessive precipitation events such that soil moisture deficit is negatively correlated with the model error.

Eventually, on the basis of studying the correlations and AMI between various candidate predictors and the output, and using expert judgement, the following variables have been chosen to serve as candidate predictors:

- the most recent precipitation ($P_{t\text{-}1}$),
- the observed water level ($H_{obs,\ t\text{-}1}$),
- error ($e_{t\text{-}1}$),
- state variables *GW* and *SMD*.

It should be noted that subscript t-1 denotes the 12 hours delay since the data sets analysed has a time step of 12 hours (see Sect. 3.1.2).

In an attempt of removing least influential inputs, the set of variable above was then subjected to the model-based optimization: the degree of influence of various inputs has been explored by running the UNEEC predictor for different sets of inputs and comparing the resulting PICP and MPI. It was found that there were only negligible changes (and mostly no change) when $P_{t\text{-}1}$ and $e_{t\text{-}1}$ were included or not. Based on this analysis these two variables have been excluded from the further experiments, and only the variables *GW*, *SMD*, $H_{obs,\ t\text{-}1}$ have been used as predictors. Inclusion of *GW* was important since this variable provides more explainable results in terms of PICP and MPI. It should be noted that using *GW* and *SMD* can be considered as a proxy for using the rainfall information.

Fuzzy clustering in UNEEC is carried out by the fuzzy c-means method and employs 6 clusters with the fuzzy exponential coefficient set to 2. The number of clusters was chosen based on computation of Partition Index (SC), Separation Index (S) and Xie and Beni Index (XB) (Bensaid et al., 1996; Xie and Beni, 1991). (It should be mentioned that the sensitivity of PICP and MPI to different number of clusters supports the choice of 6 clusters.)

Within the variables considered in clustering, *GW* is the most influential one. Fig. 5 shows fuzzy clustering of *GW*, *SMD*, and $H_{obs,\,t\text{-}1}$ data for Llanyblodwel catchment (lead time = 6 hrs). This figure contains also the plot of model residuals against *GW* where one can observe heteroscedasticity of model residuals with respect to *GW*. As can be easily seen, while cluster 2 is associated with very high groundwater levels, clusters 4 is associated with the low groundwater level conditions, which might occur due to the low water levels in the river and/or high soil moisture deficit.

Figure 5. Fuzzy clustering of: *GW* (left, top) and its relation with the model residuals (right), *SMD* (left, middle) and $H_{obs,\,t\text{-}}$, (left, bottom) for calibration period (7 March 2007 08:00 – 7 March 2010 08:00) - Llanyblodwel, lead time = 6 hrs.

It must be noted that in this study the hydrological model output is not included as yet another input to UNEEC (along with the observed discharge/water level) in all case studies. However it may be worth exploring this idea in the further studies.

4 Results and discussion

This part focuses on statistical error analysis (Sect. 4.1) and comparison of uncertainty analysis results (Sect. 4.2).

4.1 Statistical error analysis

Understanding the quality of hydrological model quality (e.g. water level forecasts) is important in order to discuss uncertainty analysis results provided by any method. For this purpose we analyse the error time series statistically. We also check the homoscedasticity (the assumption which simplifies the mathematical and computational treatment of random variables) of the model residuals. Furthermore, we investigate the normality of model residuals through probability plots of the *normal* distribution and the *t location-scale* distribution, which *pdf* is given by Eqn. 14 and Eqn. 15, respectively

$$f(x) = (1/\sigma \cdot \sqrt{2\pi}) \cdot e^{-(x-\mu)^2/2\sigma^2} \tag{14}$$

$$f(x) = \frac{\Gamma\left(\nu+1/2\right)}{\sigma \cdot \sqrt{2\pi} \cdot \Gamma \cdot \left(\nu+1/2\right)} \left[\frac{\nu + \left(x-\mu/\sigma\right)^2}{\nu}\right]^{-\left(\nu+1/2\right)} \tag{15}$$

where μ: location parameter (mean), σ: scale parameter (standard deviation), ν: shape parameter (i.e., the number of degrees of freedom), and Γ: gamma function. The *t* location-scale distribution is similar to the normal distribution but has heavier tails making it more prone to outliers. Within this study outliers refer to very high model residuals occurring during extreme precipitation and flow events. In case of normality of data its analysis becomes much simpler, however often this is not the case.

Residual uncertainty varies in time and with the changing hydrometeorological situation, so in this paper we investigate the residuals distribution for different hydrometeorological conditions represented by clusters found within the UNEEC method (on the training dataset).

4.1.1 Brue catchment

The observed discharge plotted against simulated discharge during calibration and validation periods can be seen in Fig. 6a and 6c, respectively. During calibration although the model residuals are lower at flows higher than 35 m^3/s compared to at flows less than 35 m^3/s (in Fig. 6a) it can be seen from Fig. 6c that the HBV-96 model is less accurate in simulating high flows compared to low flows. It is also noteworthy to mention that in calibration (Fig. 6a) there is higher dispersion around the diagonal line than in validation (Fig. 6c.)

Fig. 6b and 6d shows how model residuals change with increasing discharge values during calibration and validation periods, respectively. Clearly, model residuals' of Brue catchment are heteroscedastic, that is to say, the variance of model residuals vary with the effect being modelled, i.e. observed discharge.

Figure 6. Observed discharge, simulated discharge and model residuals during calibration and validation (Brue catchment).

Fig. 7 presents probability plots for model residuals during training. The top left plot compares the two selected distributions (normal distribution and *t* location-scale distribution). The estimated parameters for the best fit to data are μ = 0.0363 m^3/s and σ = 0.7619 m^3/s for normal distribution - same with the empirical parameters. On the other hand, the best fit parameters for *t* location-scale distribution are different: μ = 0.0607 m^3/s, σ = 0.2351 m^3/s and ν = 1.5833. From this figure, one can conclude that the model residuals' distribution is far from being close to normal even though the parameters of the fitted normal distribution are the same with those obtained from the empirical distribution. It is obvious that *t* location-scale distribution provides better fit as it is able to enclose the data at the tails much better compared to fitted normal distribution. Yet, the outliers are still not represented fully.

Figure 7. Probability plots for model residuals (during training) for Brue catchment: comparison of the two fitted distributions: normal vs. *t* location-scale distribution (top left), and the clusters.

Normality of model residuals' distribution is further investigated for different hydrometeorological conditions as identified by clustering in the space of the predictor variables. Analysis of probability plot for each cluster formed indicates that there is no significant departure from normality (with regard to the fitted normal distribution) unlike in the overall model residuals. The most striking result among all clusters is achieved in the one representing very high flow and high rainfall (Cluster 4, 0.95% of total data) (Fig. 7, bottom middle). The distribution of all the other clusters (Cluster 1, 2, 3, and 5) was found to be more or less equally close to normal. When visually compared these distributions were only slightly less close to normal with respect to Cluster 4.

4.1.2 Upper Severn catchments: Yeaton, Llanyblodwel, and Llanerfyl

The quality of (water level) forecasts is assessed based on standard deviation of model error. The results are comparatively presented for different lead times in Fig. 8. It can be clearly seen that during both calibration and validation as lead time increases, the standard deviation of error increases as well. Also, it should be noticed that there is a direct increasing effect of shorter basin lag time on standard deviation. For example, catchment with shortest basin lag time, that is Llanerfyl, has always larger standard deviation for all lead times. On the contrary, the smallest standard deviation always occurs in the catchment having the longest basin lag time, which is Yeaton. This is mainly due to the fact that the basin lag time represents memory of a catchment. Hence, flood forecasting capability of a hydrological model is affected negatively when the basin lag time is short.

Figure 8. Standard deviation of model error during calibration and validation (Upper Severn catchments).

The observed water levels are plotted against forecasted water levels in Llanyblodwel catchment during calibration and validation for lead time = 6 hrs in Fig. 9a and 9c, respectively. Fig. 9b and 9d shows model error plotted against observed water level on the logarithmic scale. Although it is not very clear from Fig. 9a (and Fig. 9c), it is evident from Fig. 9c (and Fig. 9d) that the model error increases with higher water levels, as expected. This confirms the heteroscedasticity of model residuals.

Figure 9. Observed water level, forecasted water level and model residuals during calibration and validation (Llanyblodwel, lead time = 6 hrs).

Normality of model residuals for Llanyblodwel catchment for all lead times was investigated (see Fig. 10, top left). Visual inspection of probability plots, superimposed on which the line joining the 25^{th} and 75^{th} percentiles of the fitted normal distributions, reveals that errors are not normally distributed, i.e. the data does not fall on the straight line as it is especially the case for the tails. It should be realized that the departure from normality increases with longer lead times. The top right plot in Fig. 10 compares the two selected distributions (normal distribution and *t* location-scale distribution) for model residuals during training. It can be concluded that neither the normal distribution nor the *t* location-scale distribution provides a good fit to the data.

Figure 10. Probability plots for model residuals (during training) for Llanyblodwel catchment: comparison of fitted normal distributions for all the lead times (top left), comparison of the two fitted distributions: normal vs. *t* location-scale distribution (top right) and the clusters (lead time = 6 hrs).

Furthermore, a normality check for model residuals' distribution is made individually for the data clusters corresponding to particular hydrometeorological conditions. The variables used for clustering are groundwater level (*GW*), soil moisture deficit (*SMD*), and observed water level ($H_{obs,\ t-1}$). It is seen that the level of achieving normality in model residuals' distribution for each cluster is substantially poorer if compared to the Brue catchment. This can be explained by the fact that the error time series data being analysed has a time step of 12 hrs which is long enough to hinder effects of varying water levels on error. Another reason can be related to the nature of model residuals, e.g. forecasted precipitation is used to predict water levels. This brings a great amount of uncertainty and a higher difference between the actual and the predicted water levels (i.e. higher model residuals). It is also worth mentioning that the distribution closest to normal is found in the data cluster representing high groundwater levels, high water levels, and low soil moisture deficit (Cluster 2, comprising 4.6% of the total data set) (Fig. 10, middle). Distributions of Clusters 1, 3, 4, 5, and 6 are far from normal.

Both Brue and Llanyblodwel case studies indicate that it is not possible to understand the origin of the model error in uncertainty assessment looking at the probability plots of model residuals for each cluster. However, what is worthwhile to mention that it is mostly the extreme events making the overall distribution non-Gaussian. Classifying data so that different hydrometeorological conditions (most importantly, the extreme events), are separated helps to achieve homogeneity, and thus normality in model residuals' distribution. Therefore clustering can be suggested as an alternative to transformation of model residuals before applying any statistical methods on them.

4.2 Uncertainty prediction by QR and UNEEC

Uncertainty analysis results from both methods are evaluated and compared employing the validation measures explained in Section 2.2.

4.2.1 Brue catchment

Validation measures PICP, MPI, and ARIL are provided in Table 2. In terms of PICP, even though QR provides PICP values slightly closer to 90% and 50% during training, UNEEC was found to be more reliable in validation especially for the 90% CL. While the narrowest prediction interval on average is given by UNEEC during training for both 90% and 50% CL, comparable MPI values are obtained during validation. QR has smaller ARIL values particularly for the 90% CL. However on aggregate UNEEC yields better results over QR, especially in validation.

Table 2. Uncertainty analysis results for 90% and 50% confidence levels (Brue catchment).

Looking at Fig. 11a, visual analysis of 90% prediction intervals for the highest flow period in validation reveals that neither UNEEC nor QR is perfectly able to enclose the observations of high flows. Overall, in validation the analysis results from UNEEC and QR are comparable for the highest peak event (Table 2). For medium peaks in validation, however, QR produces wider uncertainty bounds in comparison to UNEEC. This is illustrated in Fig. 11b. For this medium peak event it should be noted that the higher MPI (and ARIL) value by QR is not manifested in PICP – both methods have very close PICP values (Table 2). One of the reasons for this may relate to the fact that by design UNEEC uses more predictors that explain the (past) catchment behaviour and hence is able to "memorize" catchment behaviour better, and this is especially pronounced during the longer periods of medium flows rather than during high flows having shorter duration.

Figure 11. Comparison of prediction limits for 90% confidence level during validation: (a) for the highest peak event (16 December 1995 04:00 – 28 December 1995 16:00), and (b) for a medium peak event (6 January 1996 00:00 – 18 January 1996 12:00).

We have also compared performance of QR and UNEEC for each cluster found by UNEEC during training. Unlike for the whole data set (which is highly heterogeneous due to extremes in rainfall-runoff process) analysis for each individual cluster focuses on more homogeneous data sets. Table 3 shows the corresponding PICP, MPI and ARIL. In general, it is difficult to decide which method is better – results are mixed. However there is one observation that can be made. For most clusters there is a dependency between PICP and MPI: typically the higher MPI corresponds to PICP being closer to the CL (90%). This may be explained by the fact that for narrow MPIs PICP would be under "pressure" and be lower (however it would be difficult to generalize). For example, for the high flow cluster (Cluster 4) QR appears to be better in terms of PICP, whereas UNEEC ends up with very narrow MPI and this is probably the reason why its PICP could not reach 90% CL.

The reported comparison was done for the clusters found by UNEEC during training. In principle a similar comparison can be also made for the homogeneous groups of data in the validation set, however this may not have much sense since this set imitates the model in operation, and in operation all models are run for individual input vectors at each time step of the model run, and not for the whole set of data (so the “validation set” in operation will never exist).

Table 3. PICP, MPI, and ARIL values for each cluster (training, 90% confidence level, Brue): UNEEC vs. QR.

4.2.2 Upper Severn catchments: Yeaton, Llanyblodwel, and Llanerfyl

For these catchments, in order to reflect performance for different lead times better, we are using the graphical representation of results.

Fig. 12 shows the PICP values plotted against the MPI for the calibration and validation periods. The most important general conclusion is that both methods show excellent results in terms of PICP for 90% CL. For the 50% CL the results seem to be worse, especially for UNEEC – but the reader should take into account that for the low lead times the hydrological models are very accurate, hence MPI is extremely narrow (especially for 50% CL) and it is no surprise PICP cannot be accurately calculated. Further, for the 90% CL, the following can be said: for Yeaton QR does slightly better than UNEEC; for Llanyblodwel both methods are equally good; for Llanerfyl: UNEEC method is a bit better than QR.

Figure 12. Comparison of UNEEC and QR based on both PICP and MPI during calibration period (7 March 2007 08:00 – 7 March 2010 08:00) and validation period (7 March 2010 20:00 – 7 March 2013 08:00) for 90% and 50% confidence level (The size of the marker represents the lead time, i.e. bigger the marker, longer the lead time).

For the further analysis, Fig. 13 presents MPI and ARIL values for the 90% CL on calibration and validation data sets. It can be seen that with the increase of the lead time, the forecast error obviously increases, and the values of both indicators follow. In view of the (high) model accuracy, the relatively low MPI values in Yeaton catchment are not surprising for both methods. Overall, the results are mixed: for some catchments QR is marginally better, for other catchments UNEEC has higher performance.

Figure 13. MPI (left) and ARIL (right) values obtained during calibration period (7 March 2007 08:00 – 7 March 2010 08:00) and validation period (7 March 2010 20:00 – 7 March 2013 08:00) for 90% confidence level.

For the further comparison of estimated prediction limits through uncertainty plots, three cases are selected based on the relationship between basin lag time and lead time. These cases are (1) Yeaton, lead time = 3 hrs (lead time < basin lag time), (2) Llanyblodwel, lead time = 6 hrs (lead time ≈ basin lag time), and (3) Llanerfyl, lead time = 12 hrs (lead time > basin lag time). The fundamental idea here is to understand how well the residual uncertainty is assessed with regard to forecast lead time and its relation to basin lag time. The catchment with the longest basin lag time (Yeaton) is considered for Case 1, where the effect of a very short lead time is to be investigated. Here on this decision, there is the deliberate intention to combine the condition of having more accurate model outputs (i.e. extremely small residuals) as well. Case 3, on the other hand, is important to understand lead time-basin lag time relationship for the worst situation: relatively poor quality of forecasting model and the longest lead time. This is the most critical case since the performance of predictive uncertainty method's performance has a bigger role in operational decision making process. Apart from these two extreme cases, Case 2 represents a balanced situation where the lead time of interest and basin lag time are approximately equal. Llanyblodwel catchment is chosen for this case as its model has a moderate predictive accuracy. Fig. 14 compares the computed prediction limits by QR and UNEEC for these cases during the latest 11 months period of validation (April 2012 - February 2013). It was during late 2012 that Upper Severn catchment suffered from serious flooding and this period corresponds to the right half of the plots. The most salient observations from Fig. 14 are as follows:

- In Llanerfyl, one can notice a strange behaviour of the model causing sharp changes in forecasted water levels (unstable model outputs), and thus in prediction limits. Considering that Llanerfyl catchment has a basin lag time of ~3-5 hrs, hydrological conditions in the catchment, e.g. water levels, can change significantly in 12 hrs (Δt, time step of the data set). Therefore, it is not surprising that the sharpest changes occur in this catchment's hydrograph as compared to Yeaton and Llanyblodwel. One can observe even more significant changes in the second half period of the hydrograph. It is necessary to mention that these oscillating changes appear as a consequence of the forecasting model's extremely poor performance.

- For the low water levels in Yeaton and Llanyblodwel, UNEEC gives wider prediction intervals as compared to QR. A possible explanation for this can be encapsulation of groundwater level information in UNEEC. Groundwater levels remains at higher levels for longer periods than water levels in the river (i.e. due to slow and long response time of groundwater levels to changing hydrometeorological conditions). Thus, using *GW* as an input variable in its nonlinear model, UNEEC has the potential to provide uncertainty band of larger widths for water levels when the groundwater level is high.
- For the medium water levels in Yeaton and Llanybldowel, QR gives wider prediction intervals as compared to UNEEC, which is confirmed by the higher MPI and ARIL (without any significant improvements in PICP) values for QR (Table 4) obtained for medium water levels. This is particularly true on falling limb part of the hydrographs as exemplified in Fig. 15a and Fig. 15b (for Yeaton and Llanyblodwel, respectively). The average of the MPI values corresponding to three examples shown from Yeaton and Llanyblodwel, respectively, are 0.0204 and 0.0201 meters for UNEEC whereas for QR it is 0.0418 and 0.0295 meters.
- For peak water levels in Yeaton and Llanyblodwel catchments, it is mostly QR that produces higher upper prediction limit than UNEEC. Yet, this does not contribute to the overall performance of the method significantly. On the contrary, it is seen in some cases that such high upper prediction limits makes the uncertainty band unnecessarily too wide.
- Continuous peaks prevail in Llanerfyl catchment (as its basin lag time is way shorter than the forecast lead time of interest). Such continuous peaks occur during certain periods in Llanyblodwel catchment too. In most of these cases, UNEEC gives narrower uncertainty band, and wider prediction interval computed by QR is redundant. That is to say, it doesn't contribute QR method's performance (as measured by PICP) at all in terms of its ability to enclose more observations within the band. For peak water levels, however, QR is slightly more informative than UNEEC.
- Noticeably, upper prediction limits obtained by QR in Llanerfyl catchment for the long-lasting falling limb part of the hydrograph (indicated by arrows in Fig. 14c) are too high, e.g. even greater than those provided by UNEEC. QR (in this study, by design) is a method building simple linear regression models considering only

observed water levels on forecasted water levels. Having rather simple mathematical formulation, it might be that sensitivity of the computed upper prediction limit to the magnitude of water level increases, and shows an amplifying effect on uncertainty band width.

Figure 14. Comparison of prediction limits for 90% confidence level during validation (1 April 2012 – 7 March 2013): (a) Yeaton, lead time = 3 hrs, (b) Llanyblodwel, lead time = 6 hrs, (c) Llanerfyl, lead time = 12 hrs.

Figure 15. Comparison of prediction limits for falling limb part of the hydrographs (medium water levels) for 90% confidence level during validation: (a) Yeaton, lead time = 3 hrs, (b) Llanyblodwel, lead time = 6 hrs, (c) Llanerfyl, lead time = 12 hrs.

Table 4. PICP, MPI, and ARIL values for MEDIUM water levels (validation, 90% confidence level): UNEEC vs. QR.

Table 5 shows the values of validation measures (PICP, MPI, and ARIL) for each cluster (obtained during training) for Llanyblodwel catchment (lead time = 6 hrs). For flood management the cluster 2 (4.6% of all data) – with the high groundwater levels, and hence potentially corresponding to flood conditions – could be the most interesting one. In UNEEC, the highest MPI value was obtained for this cluster with a relatively bad PICP value compared to the other clusters. Similar to UNEEC, the largest MPI was obtained for this cluster with QR method also. Both methods provide equally bad PICP values. Giving a wider uncertainty band than UNEEC on average, QR is less capable of estimating reasonable prediction limits for very high groundwater levels. This is also supported by its greater (12%) ARIL value compared to UNEEC.

PICP and MPI values for the cluster 4 should be mentioned as well. This cluster represents the situations with the very low water levels, very low groundwater levels, and very high soil moisture deficit, and constitutes 16.6% of the whole data. In comparison to UNEEC, QR provides a PICP value very close to 90% CL despite its slightly lower MPI. Thus, one can say that UNEEC fails in providing reliable uncertainty estimates for the extreme condition associated to very low water and groundwater levels. This can be due to the effect of using state variables as predictors. All in all, the state variables are calculated by the model and they cannot reflect real catchment conditions accurately, especially when the (hydrological) model is not very accurate. That is particularly true for the extreme events considering that models mostly fail in simulating such events.

Overall, UNEEC is worse than QR on for one cluster but better or equal on all other clusters, however, in general, both methods in terms of PICP show reasonably good results.

Table 5. PICP, MPI, and ARIL values for each cluster (training, 90% confidence level, Llanyblodwel, lead time = 6 hrs): UNEEC vs. QR.

5 Conclusions and recommendations

This study should be seen as accompanying the study by López López et al. (2014) (and earlier work on UNEEC and QR) and presents a comparative evaluation of uncertainty analysis and prediction results from QR and UNEEC methods on the four catchments that vary in hydrological characteristics and the models used: Brue catchment (simulation mode) and Upper Severn catchments - Yeaton, Llanyblodwel, and Llanerfyl (forecasting mode). The latter set of case studies is important from a practical perspective in that the effect of lead time on uncertainty analysis results and its relation with basin lag time is demonstrated. For both QR and UNEEC different model configurations than their previous applications are considered. One of reasons to compare these two methods was to understand if a simpler linear method (QR) using less input data performs well compared to a more complex (non-linear) method (UNEEC) with more predictors. The following conclusions can be drawn from the results of this study:

- In terms of easiness of setup (data preparation and calibration), preference should be given to QR simply because it is a simpler linear method with one input variable (in this study), whereas UNEEC has more steps and requires more data analysis. However, the model setup is carried out only once, and in operation both methods can be easily used and both have very low running times (a fraction of a second on a standard PC) since they are based on algebraic calculations.
- In almost all case studies both methods adequately represent residual uncertainty and provide similar results consistent with understanding of the hydrological picture of the catchment and the accuracy of the (hydrological) models used. We can recommend both methods for the use in flood forecasting.

- In one case study, Llanerfyl, we found that UNEEC was giving more adequate estimates than QR. This catchment has a shorter basin lag time and the model outputs for this catchment were characterized by a relatively high error, so our conclusion was that probably in such a rapid response catchment the UNEEC's more sophisticated non-linear models were able to capture relationships between the hydrometeorological and state variables, and the quantiles better than the QR's linear model.
- A useful finding is that inclusion of a variable representing groundwater level (*GW*) as a predictor in UNEEC improves its performance for the Upper Severn catchments. This can be explained by the fact that this variable has a high level of information content about the state of a catchment. However, it should be noted that in other catchments using such information can be misleading due to slow (and long) response time of groundwater levels to changing hydrometeorological conditions. Yet, overall, it can be advised to make use of variables which can be representative of the hydrological response behaviour of a catchment for improving the predictive capacity of data-driven methods.

There are limitations of the presented research (aspects that have not been taken into account in this paper due to the time and project settings constraints), which can be also seen as recommendations for the future research.

We recommend comparing the two presented methods (QR and UNEEC) with more predictive uncertainty methods which use different methodologies, such as HUP (Krzysztofowicz, 1999), the more recent MCP (Todini 2008) and DUMBRAE (Pianosi and Raso, 2012). Yet another recommendation (induced by the referee's and the Editor's suggestions) is to extend the list of the possible performance measures and to test the applicability of the methods developed for the assessment of the probabilistic forecasts quality (Laio and Tamea, 2007) which mathematical apparatus is transferrable to the problem of residual uncertainty prediction.

It can be also recommended to test capabilities of different predictive uncertainty methods on theoretical cases with the known distributions, as well as on the catchments of distinct hydrologic behaviour, with diverse climatic conditions, and having various hydrological features. In this study, we found that the basin lag time is a notable characteristic of a catchment having great influence on uncertainty analysis results (as measured by PICP and

MPI). When the lag time is longer, the catchment memorizes more information regarding its hydrological response characteristics.

On the other hand, exploring the performance of different methods on *similar catchments* (Huang et al., 2012; Toth, 2013; Gregor et al., 2013; Laaha et al., 2013) and finding bases for generalized guidelines on the selection of most appropriate predictive uncertainty method in operational flood forecasting practices is also important and could be considered in the further studies as well.

When different predictive uncertainty methods are evaluated based on their comparative performance, it is more important to have validation measures incorporating certain aspects of rainfall-runoff process, i.e. varying flow conditions. For example, the accuracy of the hydrological model decreases during high flow events, and thus the amount of residual uncertainty increases. This necessitates exploring validation measures linking the prediction interval to the (hydrological) model quality, e.g. by employing the weighted mean prediction interval (Dogulu et al., 2014).

There are other possibilities for further improvements in the both presented methods. For example, the different configurations of QR, the alternative clustering techniques for UNEEC, as well as using in it instance-based learning (e.g. locally weighted regression) as the predicting model can be explored further.

Acknowledgements

The results presented in this paper were obtained during the thesis research of the European Erasmus Mundus Flood Risk Management Master students Nilay Dogulu and Patricia López López. The Environment Agency is gratefully acknowledged for providing the data for Upper Severn catchments in the UK. We are very grateful to the two anonymous reviewers and the HESS Editor, Professor E. Todini, for very useful comments and suggestions.

References

Aubert, D., Loumagne, C., and Oudin, L.: Sequential assimilation of soil moisture and streamflow data in a conceptual rainfallrunoff model, J. Hydrol., 280, 145–161, doi: 10.1016/S0022-1694(03)00229-4a, 2003.

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

Barnwal, P. and Kotani, K.: Climatic impacts across agricultural crop yield distributions: An application of quantile regression on rice crops in Andhra Pradesh, India, Ecol. Econ., 87, 95–109, 2013.

Battiti, R.: Using mutual information for selecting features in supervised neural net learning, IEEE T. Neural Networ., 5, 537–550, 1994.

Baur, D., Saisana, M., and Schulze, N.: Modelling the effects of meteorological variables on ozone concentration—a quantile regression approach, Atmos. Environ., 38, 4689–4699, 2004.

Bell, V. A. & Moore, R. J.: The sensitivity of catchment runoff models to rainfall data at different spatial scales, Hydrol. Earth Syst. Sci., 4, 653–667, 2000.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchmentsRep. RHO7, 118Swed. Meteorol. and Hydrol. Inst., Norrköping, Sweden, 1976.

Bensaid, A. M., Hall, L. O., Bezdek, J. C., Clarke, L. P., Silbiger, M. L., Arrington, J. A., and Murtagh, R. F.: Validity-guided (re) clustering with applications to image segmentation, IEEE T. Fuzzy Syst., 4, 112–123, 1996.

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction. Hydrol. Process., 6, 279–298, doi 10.1002/hyp.3360060305, 1992.

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.

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

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C.: Time series analysis: forecasting and control (4th ed.), p.16, John Wiley & Sons, Inc., Hoboken, New Jersey, 2008.

Bremnes, J. B.: Probabilistic wind power forecasts using local quantile regression, Wind Energy, 7, 47–54, doi: 10.1002/we.107, 2004.

Brown, J. D. and Heuvelink, G. B. M.: Assessing uncertainty propagation through physically based models of soil water flow and solute transport, in: Encyclopedia of Hydrological Sciences, edited by: Anderson, M. G., John Wiley, New York, 2005.

Cannon, A. J.: Quantile regression neural networks: Implementation in R and application to precipitation downscaling, Comput. Geosci., 37, 1277–1284, 2011.

Coccia, G. and Todini, E.: Recent developments in predictive uncertainty assessment based on the model conditional processor approach. Hydrol. Earth Syst. Sci., 15, 3253–3274, 2011.

Coxon, G., Freer, J., Westerberg, I., Woods, R., Smith, P. and Wagener, T.: A generalised framework for large-scale evaluation of discharge uncertainties across England and Wales, EGU General Assembly, Vienna, Austria, 27 April–2 May 2014, EGU2014-10157, 2014.

Di Baldassarre, G. and Montanari, A.: Uncertainty in river discharge observations: a quantitative analysis, Hydrol. Earth Syst. Sci., 13, 913–921, doi:10.5194/hess-13-913-2009, 2009.

Dogulu, N., Solomatine, D. P., and Shrestha, D. L.: Applying clustering approach in predictive uncertainty estimation: a case study with the UNEEC method, EGU General Assembly, Vienna, Austria, 27 April–2 May 2014, EGU2014-5992, 2014.

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

Friederichs, P. and Hense, A.: Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression, Mon. Weather Rev., 135, 2365–2378, doi:10.1175/MWR3403.1, 2007.

Gupta, H. V., Beven, K. J., and Wagener, T.: Model calibration and uncertainty estimation, in: Encyclopedia of Hydrological Sciences, edited by: Anderson, M. G., John Wiley, New York, 2005.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M.: On clustering validation techniques, J. Intell. Inf. Syst., 17, 107–145, doi:10.1023/A:1012801612483, 2001.

Hill, T., & Neal, C.: Spatial and temporal variation in pH, alkalinity and conductivity in surface runoff and groundwater for the Upper River Severn catchment. Hydrol. Earth Syst. Sci., 1, 697–715, 1997.

Huang, Y., Bardossy, A., and Wagener, T.: Identification of Functional Catchment Similarity, AGU Fall Meeting, 3–7 December, 2012, Abstract #H13A-1305, 2012.

Jin, X., Xu, C. Y., Zhang, Q., and Singh, V. P.: Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model, J. Hydrol., 383, 147–155, doi: 10.1016/j.jhydrol.2009.12.028, 2010.

Kobold, M. and Sušelj, K.: Precipitation forecasts and their uncertainty as input into hydrological models, Hydrol. Earth Syst. Sci., 9, 322–332, 2005.

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

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

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

Krzysztofowicz, R.: Bayesian theory of probabilistic forecasting via deterministic hydrologic model, Water Resour. Res., 35, 2739-2750, doi:10.1029/1999WR900099, 1999.

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

Krzysztofowicz, R. and Kelly, K. S.: Hydrologic uncertainty processor for probabilistic river stage forecasting. Water Resour. Res., 36, 3265-3277, doi:10.1029/2000WR900108, 2000.

Kudryavtsev, A. A.: Using quantile regression for rate-making, Insur. Math. Econ., 45, 296–304, doi: 10.1016/j.insmatheco.2009.07.010, 2009.

Laaha, G., Salinas, J. L., and Blöschl, G.: Exploring the link between spatial hydrologic variability and catchment similarity for the purpose of regionalisation, EGU General Assembly, Vienna, Austria, 7–12 April 2013, EGU2013-5433, 2013.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sci., 11, 1267–1277, doi:10.5194/hess-11-1267-2007, 2007.

Lee, H., Seo, D.-J., and Koren, V.: Assimilation of streamflow and in-situ soil moisture data into operational distributed hydrologic models: Effects of uncertainties in the data and initial model soil moisture states, Adv. Water Resour., 34, 1597–1615, doi: 10.1016/j.advwatres.2011.08.012, 2011.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272–288, doi: 10.1016/S0022-1694(97)00041-3, 1997.

Liu, Y. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water Resour. Res., 43, W07401, doi:10.1029/2006WR005756, 2007.

López López, P., Verkade, J. S., Weerts, A. H., and Solomatine, D. P.: Alternative configurations of Quantile Regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River: a comparison, Hydrol. Earth Syst. Sci., 18, 3411–3428, doi:10.5194/hess-18-3411-2014, 2014.

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

Møller, J. K., Nielsen, H. A., and Madsen, H.: Time-adaptive quantile regression, Comput. Stat. Data An., 52, 1292–1303, 2008.

Montanari, A.: What do we mean by ’uncertainty’? The need for a consistent wording about uncertainty assessment in hydrology, Hydrol. Proc., 21, 841–845, 2007.

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.

Munir, S., Chen, H., and Ropkins, K.: Modelling the impact of road traffic on ground level ozone concentration using a quantile regression approach, Atmos. Environ., 60, 283–291, 2012.

Moore, R. J., Jones, D. A., Cox, D. R. and Isham, V. S.: Design of the HYREX raingauge network, Hydrol. Earth Syst. Sci., 4, 523–530, 2000.

Mukolwe, M. M., Di Baldassarre, G., Werner, M. and Solomatine, D. P.: Flood modelling: parameterisation and inflow uncertainty, P. I. Civil Eng.-Wat. M., 167, 51–60, 2014.

Nasseri, M. and Zahraie, B.: Application of simple clustering on space-time mapping of mean monthly rainfall pattern, Int. J. Climatol., 31, 732–741, doi: 10.1002/joc.2109, 2011.

Nasseri, M., Zahraie, B., Ansari, A., and Solomatine, D. P.: Uncertainty assessment of monthly water balance models based on Incremental Modified Fuzzy Extension Principle method, J. Hydroinform., 15, 1340–1360, doi:10.2166/hydro.2013.159, 2013.

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.

Pappenberger, F. and Beven, K. J.: Ignorance is bliss: Or seven reasons not to use uncertainty analysis, Water Resour. Res., 42, W05302, doi:10.1029/2005WR004820, 2006.

Pianosi, F. and Raso, L.: Dynamic modeling of predictive uncertainty by regression on absolute errors, Water Resour. Res., 48, W03516, doi:10.1029/2011WR010603, 2012.

Pianosi, F., Shrestha, D. L. and Solomatine, D. P.: ANN-based representation of parametric and residual uncertainty of models, IEEE IJCNN, 1–6, doi:10.1109/IJCNN.2010.5596852, 2010.

Quinlan, J. R.: Learning with continuous classes, in: Proceedings of the 5th Australian joint Conference on Artificial Intelligence, Hobart, Tasmania, 16–18 November 1992, 343–348, 1992.

Reggiani, P. and Weerts, A.: A Bayesian approach to decision-making under uncertainty: An application to real-time forecasting in the river Rhine, J. Hydrol., 356, 56–59, doi: 10.1016/j.jhydrol.2008.03.027, 2008.

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.

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.

Shrestha, D. L., Robertson, D. E., Wang, Q. J., Pagano, T. C., and Hapuarachchi, H. A. P.: Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose, Hydrol. Earth Syst. Sci., 17, 1913-1931, doi:10.5194/hess-17-1913-2013, 2013.

Shrestha, D. L. and Solomatine, D. P.: Machine learning approaches for estimation of prediction interval for the model output, Neural Networks, 19, 225–235, doi: 10.1016/j.neunet.2006.01.012, 2006.

Shrestha, D. L., Rodriguez, J., Price, R. K., Solomatine, D. P.: Assessing model prediction limits using fuzzy clustering and machine learning. Proc. 7th Int. Conf. On Hydroinformatics, Nice, France, 2006.

Shrestha, D. L. and Solomatine, D. P.: Data-driven approaches for estimating uncertainty in rainfall-runoff modelling, Int. J. River Basin Management, 6, 109–122, 2008.

Sikorska, A. E., Scheidegger, A., Banasik, K., and Rieckermann, J.: Considering rating curve uncertainty in water level predictions, Hydrol. Earth Syst. Sci., 17, 4415-4427, doi:10.5194/hess-17-4415-2013, 2013.

Singh, S.K., McMillan, H. and Bárdossy, A.: Use of the data depth function to differentiate between case of interpolation and extrapolation in hydrological model prediction, J. Hydrol., 477, 213–228, doi: 10.1016/j.jhydrol.2012.11.034, 2013.

Solomatine, D.P. and Dulal, K. N.: Model trees as an alternative to neural networks in rainfall–runoff modelling, Hydrol. Sci. J., 48, 399–411, 2003.

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.

Solomatine, D. P. and Wagener, T.: Hydrological Modeling, in: Treatise on Water Science, edited by: Peter A. Wilderer, Elsevier, 2011.

Solomatine, D. P., Dibike, Y. B., and Kukuric, N.: Automatic calibration of groundwater models using global optimization techniques, Hydrol. Sci. J., 44, 879–894, 1999.

Solomatine, D. P.: Modelling Theory and Uncertainty (An Introduction) - Part 2: Analysis of model uncertainty, Lecture Notes. UNESCO-IHE Institute for Water Education, Delft, the Netherlands, 2013.

Tayfur, G., Zucco, G., Brocca, L. and Moramarco, T.: Coupling soil moisture and precipitation observations for predicting hourly runoff at small catchment scale, J. Hydrol., 510, 363–371, doi:10.1016/j.jhydrol.2013.12.045, 2014.

Taylor, J. W.: Forecasting daily supermarket sales using exponentially weighted quantile regression, Eur. J. Oper. Res., 178, 154–167, 2007.

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

Toth, E.: Catchment classification based on characterisation of streamflow and precipitation time-series, Part I: Classification of streamflow signatures, EGU General Assembly, Vienna, Austria, 7–12 April 2013, EGU2013-14007, 2013.

van Andel, S. J., Weerts, A., Schaake, J. and Bogner, K.: Post-processing hydrological ensemble predictions intercomparison experiment, Hydrol. Process., 27, 158–161, doi:10.1002/hyp.9595, 2013.

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.

Wagener, T. and Gupta, H. V.: Model identification for hydrological forecasting under uncertainty, Stoch. Env. Res. Risk A., 19, 378–387, doi:10.1007/s00477-005-0006-5, 2005.

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

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

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.

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, 30 doi:10.1016/j.envsoft.2012.07.010, 2013.

Xie, X. L. and Beni, G.: A validity measure for fuzzy clustering, IEEE T. Pattern. Anal., 13, 841–847, 1991.

FIGURES AND TABLES

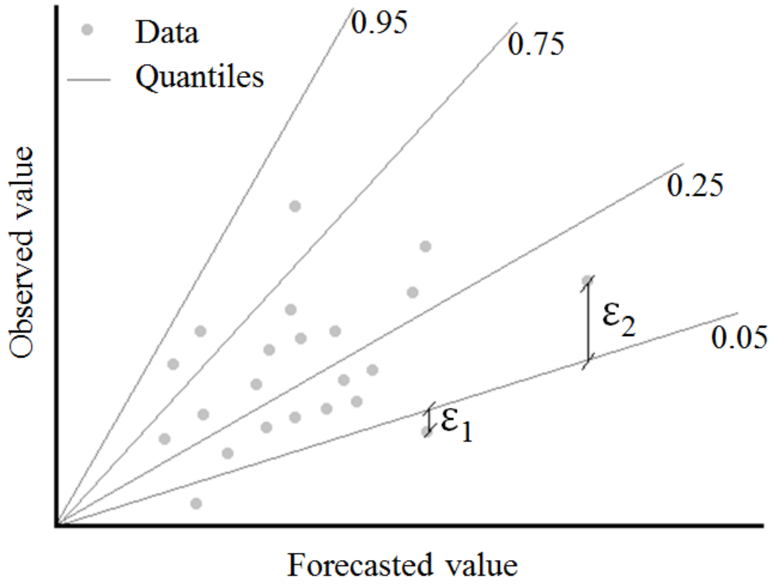

Figure 1. Quantile regression example scheme considering different quantiles.

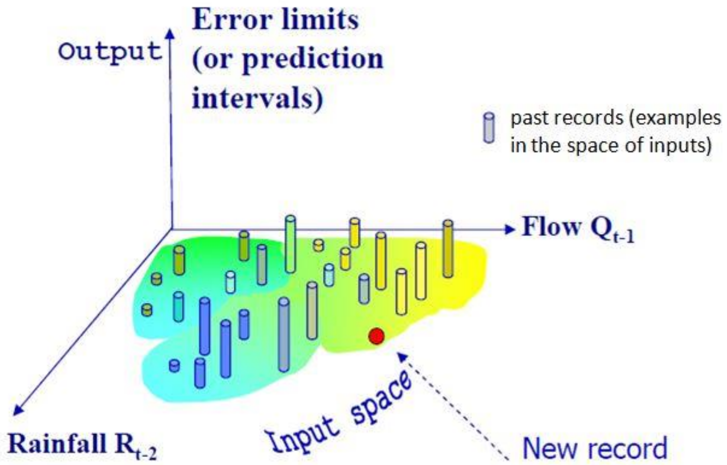

Figure 2. An example to fuzzy clustering of input data (the predictors are past rainfall at lag *t-2* and past flow at lag *t-1*) during training of the uncertainty model, U (adapted from Solomatine, 2013).

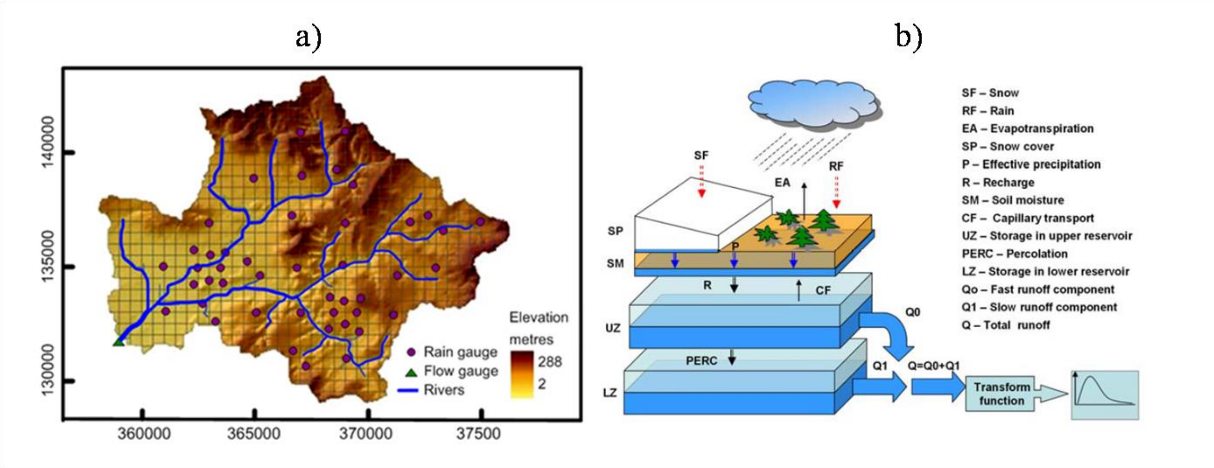

Figure 3. (a) The Brue catchment showing dense rain gauges network and its river gauging station, Lovington, where the discharge is measured, and (b) Schematic representation of HBV-96 model (Lindström et al., 1997) with routine for snow (upper), soil (middle), and response (bottom) (Shrestha and Solomatine, 2008).

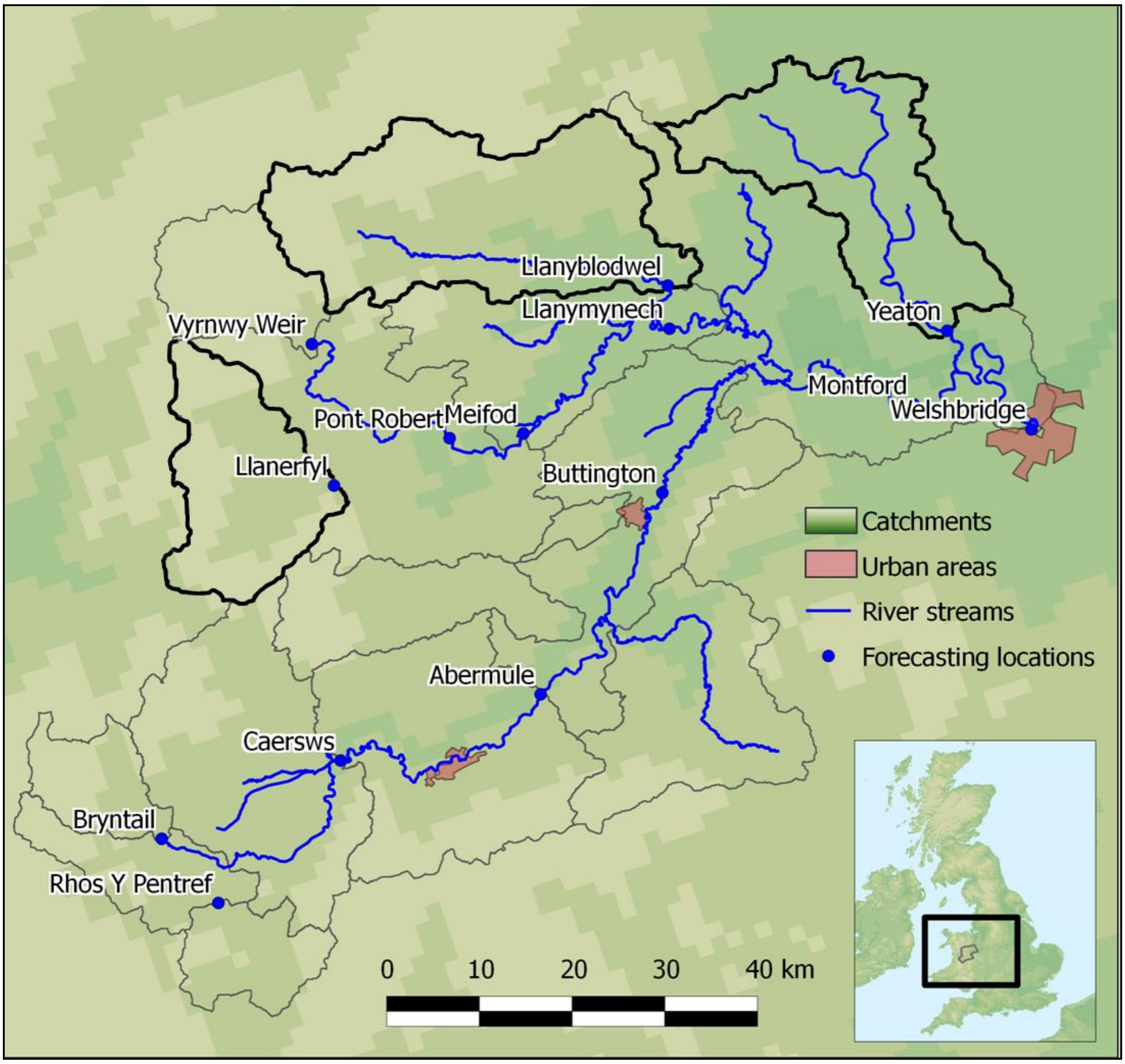

Figure 4. The Upper Severn catchments: Yeaton, Llanyblodwel and Llanerfyl.

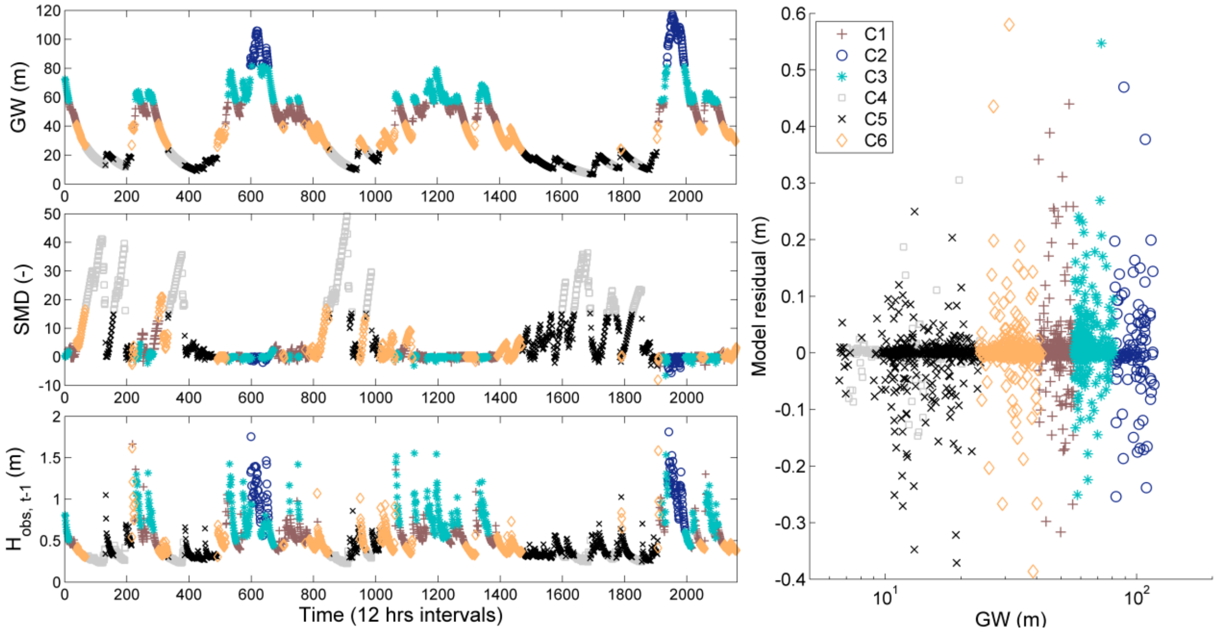

Figure 5. Fuzzy clustering of: *GW* (left, top) and its relation with the model residuals (right), *SMD* (left, middle) and $H_{obs,\ t\text{-}1}$ (left, bottom) for calibration period (7 March 2007 08:00 – 7 March 2010 08:00) - Llanyblodwel, lead time = 6 hrs.

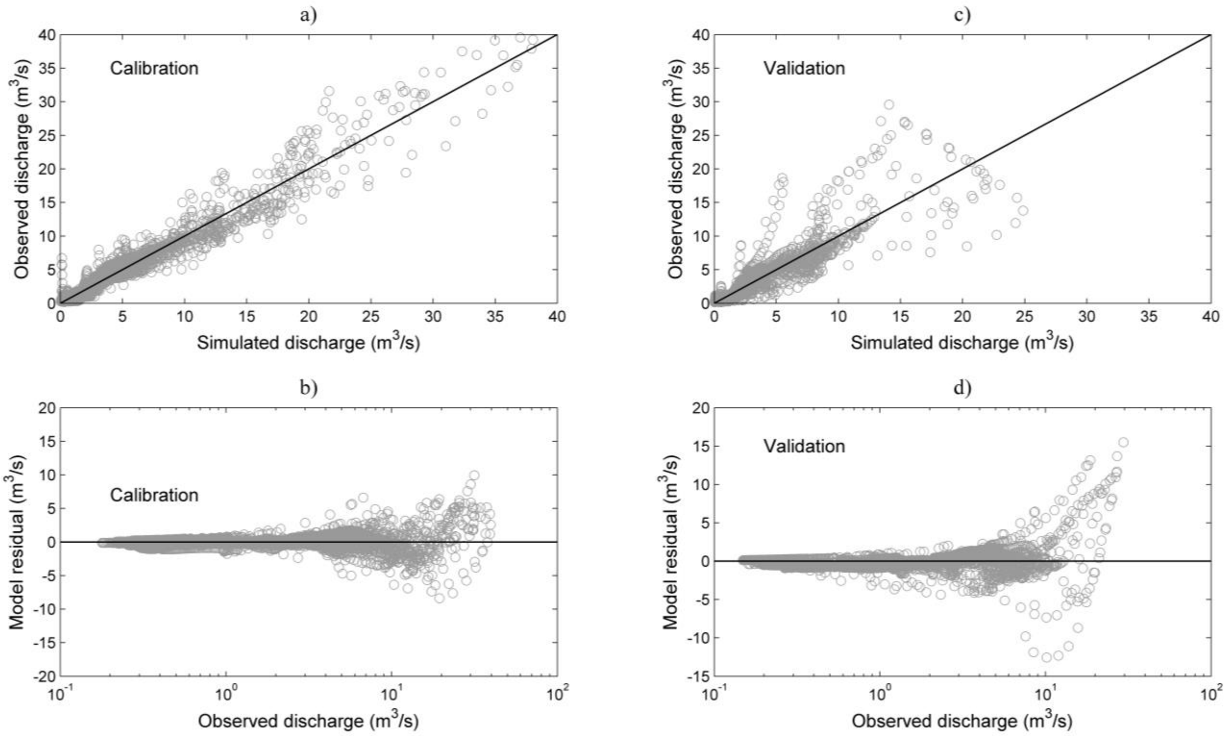

Figure 6. Observed discharge, simulated discharge and model residuals during calibration and validation (Brue catchment).

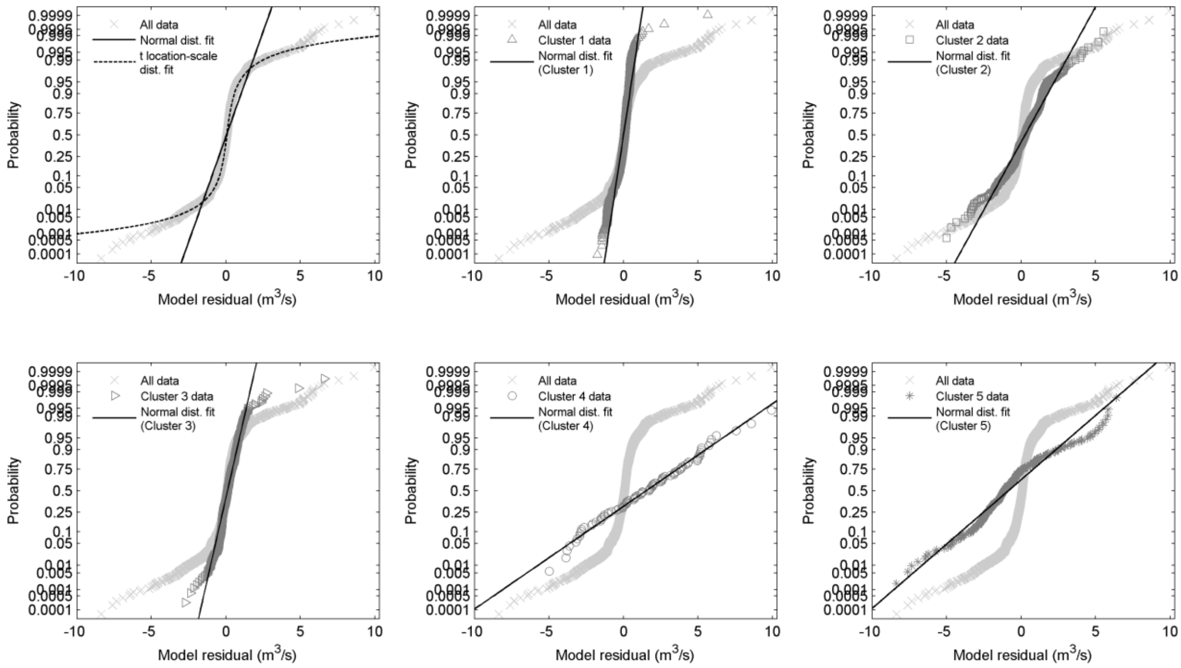

Figure 7. Probability plots for model residuals (during training) for Brue catchment: comparison of the two fitted distributions: normal vs. *t* location-scale distribution (top left), and the clusters.

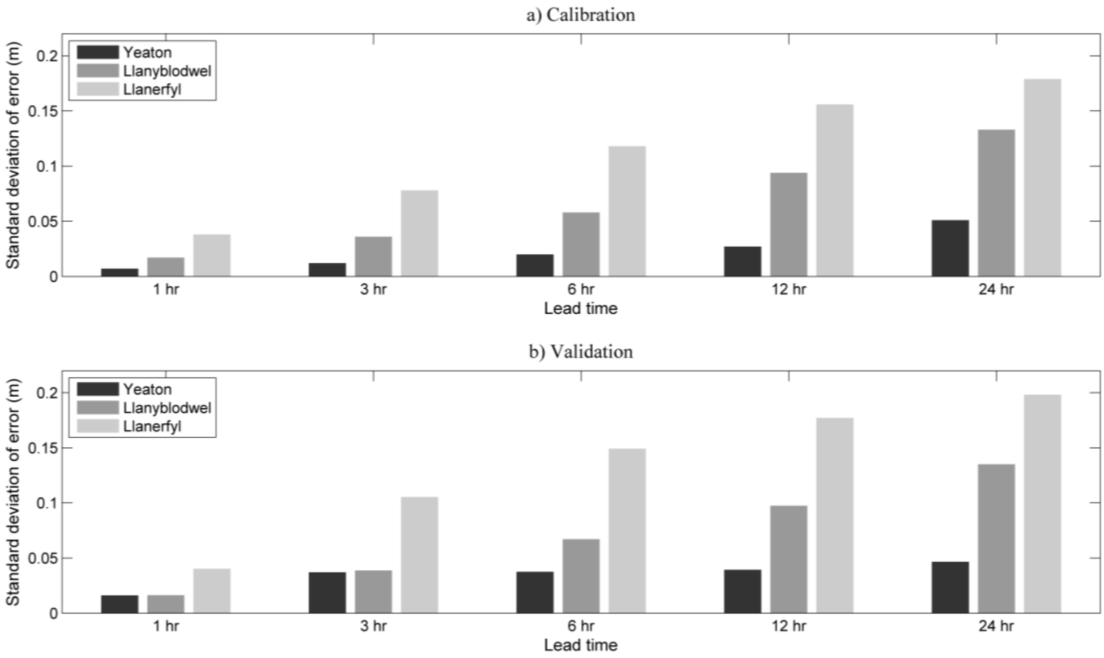

Figure 8. Standard deviation of model error during calibration and validation (Upper Severn catchments).

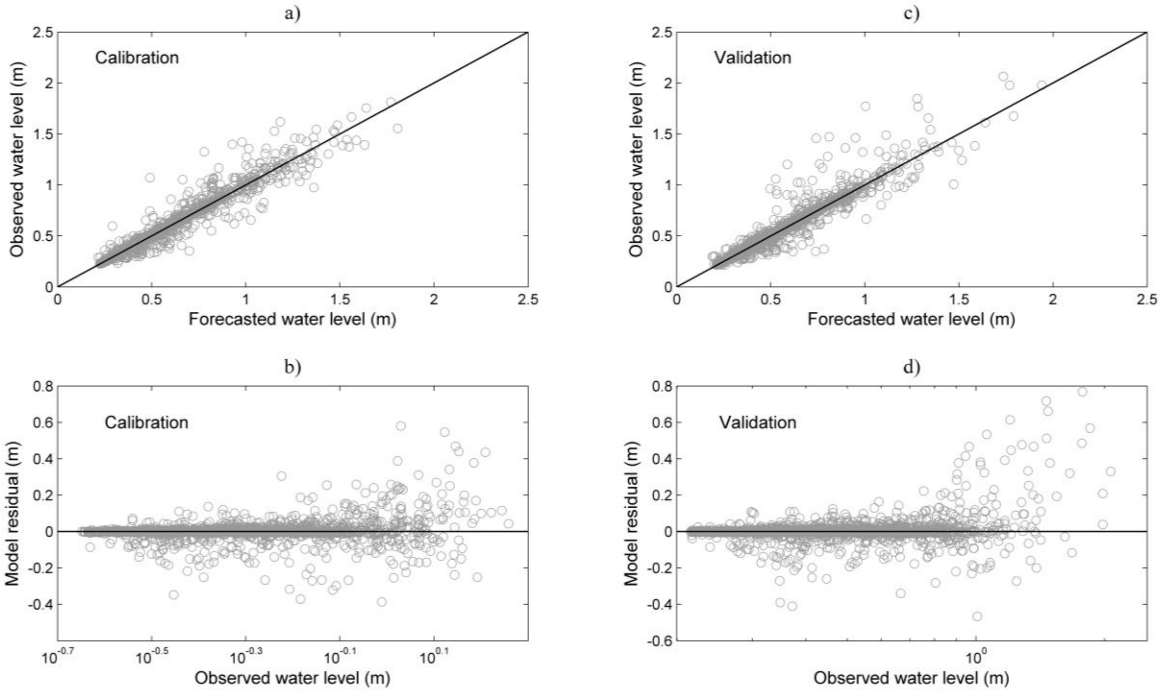

Figure 9. Observed water level, forecasted water level and model residuals during calibration and validation (Llanyblodwel, lead time = 6 hrs).

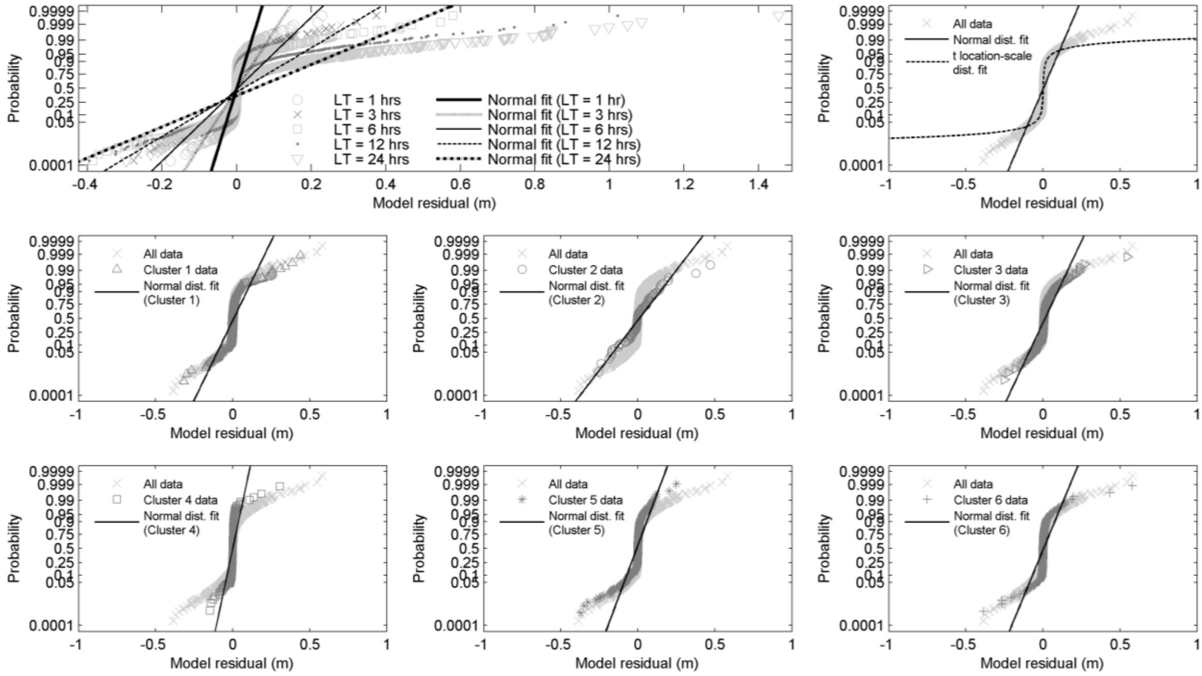

Figure 10. Probability plots for model residuals (during training) for Llanyblodwel catchment: comparison of fitted normal distributions for all the lead times (top left), comparison of the two fitted distributions: normal vs. *t* location-scale distribution (top right) and the clusters (lead time = 6 hrs).

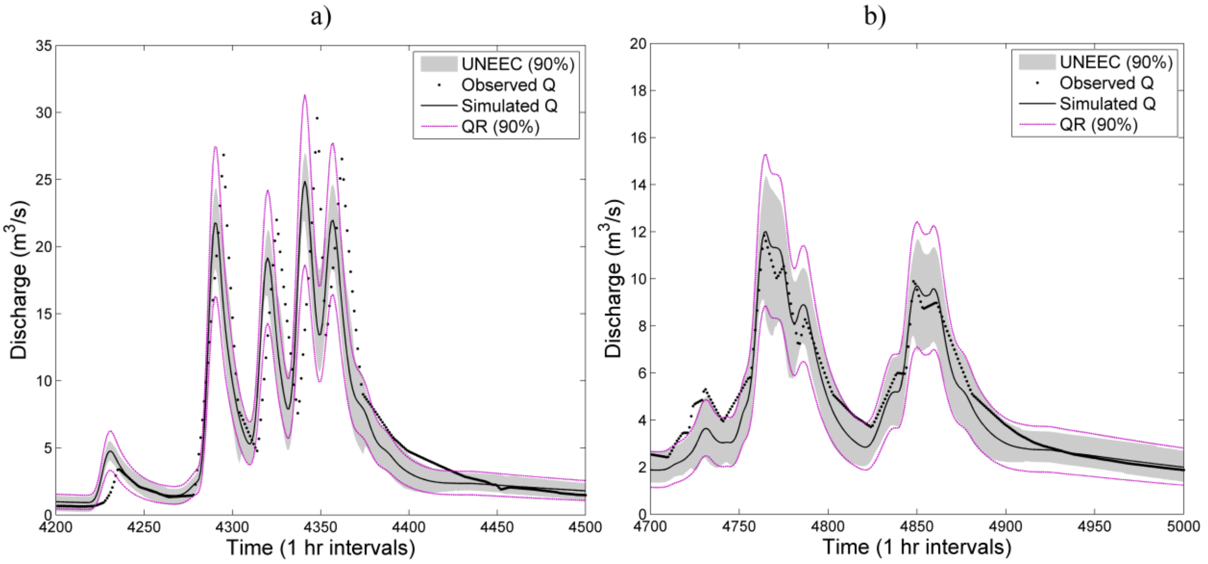

Figure 11. Comparison of prediction limits for 90% confidence level during validation: (a) for the highest peak event (16 December 1995 04:00 – 28 December 1995 16:00), and (b) for a medium peak event (6 January 1996 00:00 – 18 January 1996 12:00).

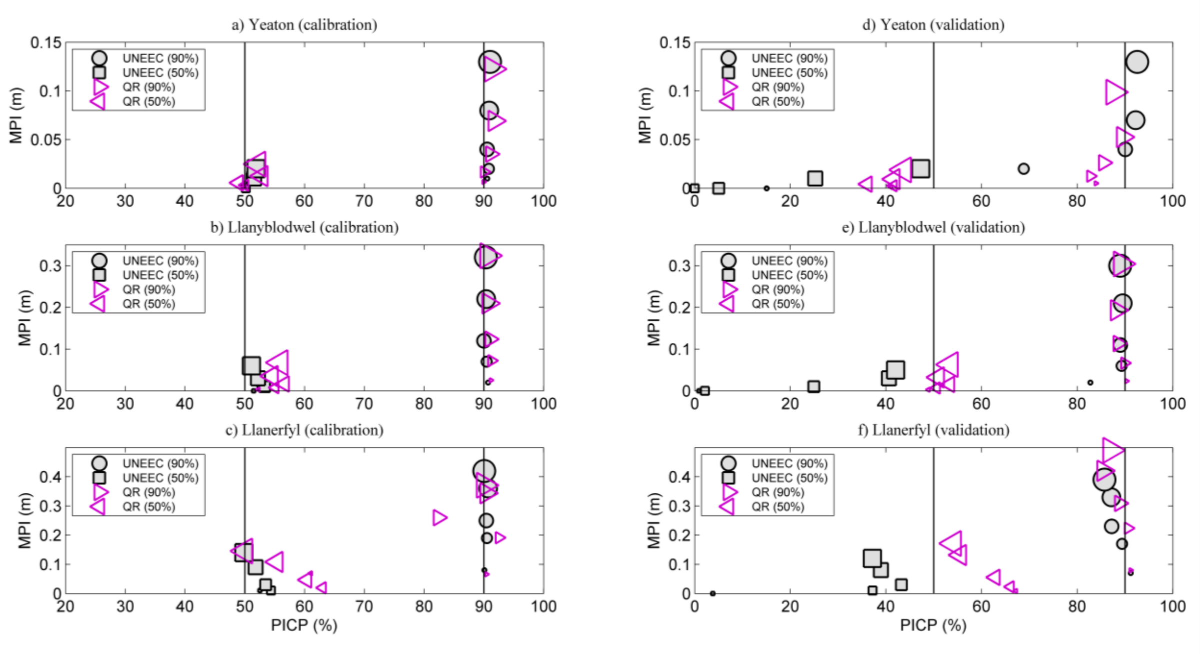

Figure 12. Comparison of UNEEC and QR based on both PICP and MPI during calibration period (7 March 2007 08:00 – 7 March 2010 08:00) and validation period (7 March 2010 20:00 – 7 March 2013 08:00) for 90% and 50% confidence level (The size of the marker represents the lead time, i.e. bigger the marker, longer the lead time).

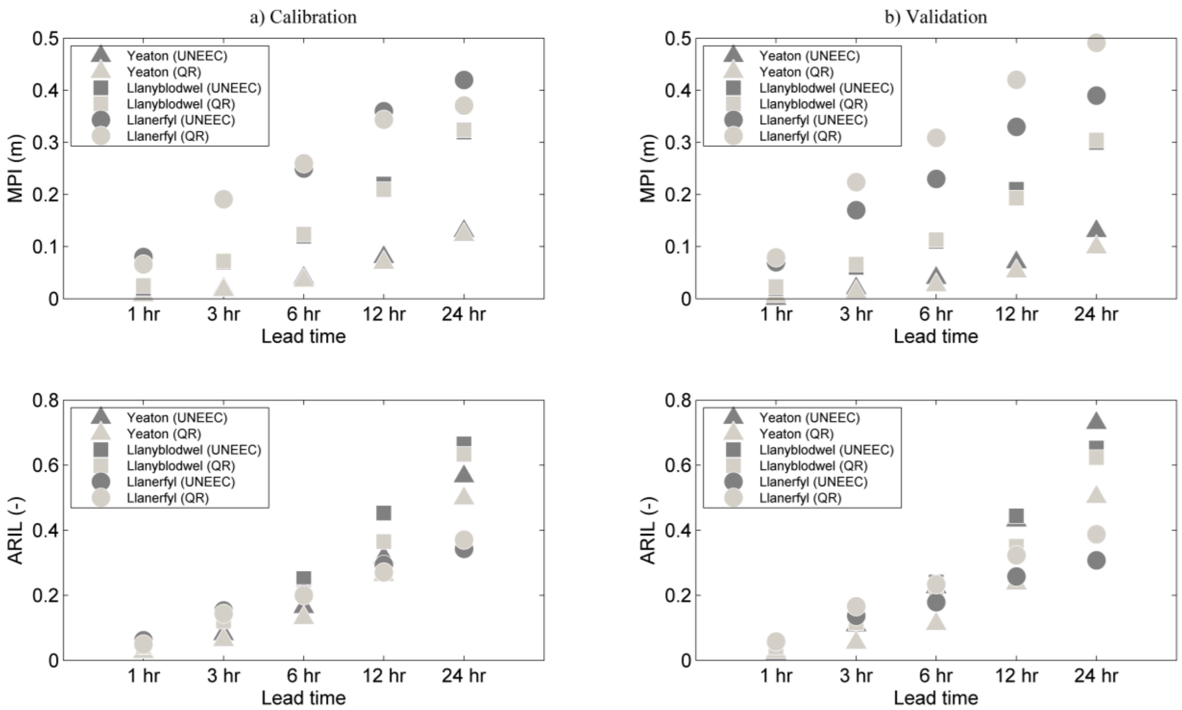

Figure 13. MPI (left) and ARIL (right) values obtained during calibration period (7 March 2007 08:00 – 7 March 2010 08:00) and validation period (7 March 2010 20:00 – 7 March 2013 08:00) for 90% confidence level.

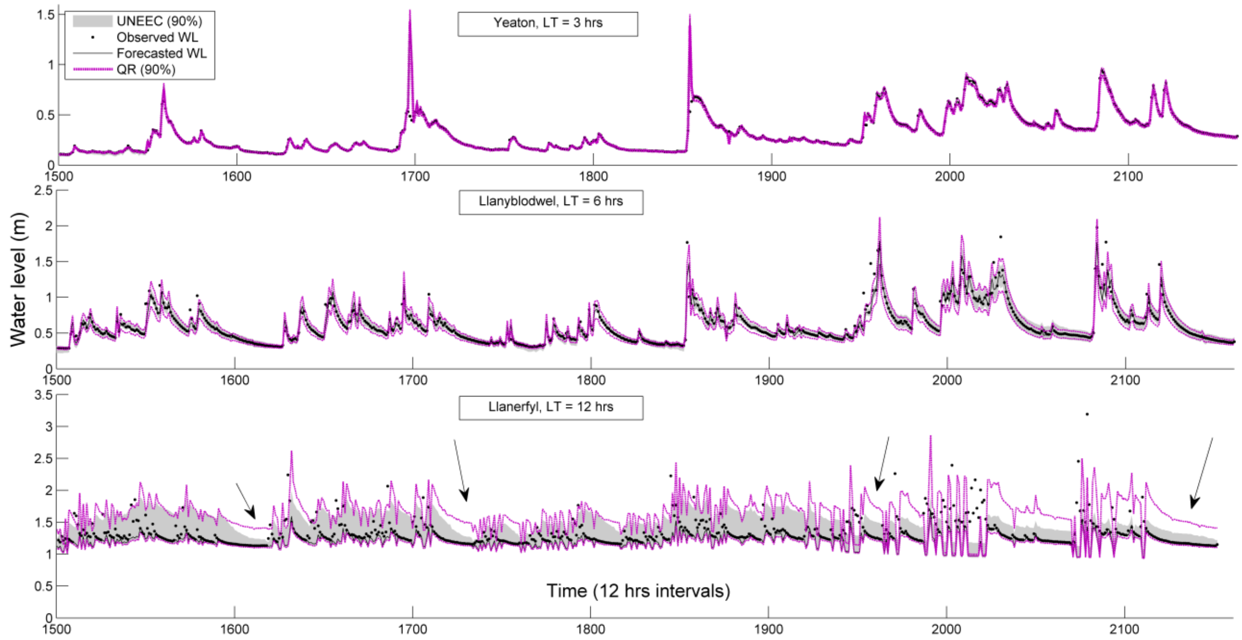

Figure 14. Comparison of prediction limits for 90% confidence level during validation (1 April 2012 – 7 March 2013): (a) Yeaton, lead time = 3 hrs, (b) Llanyblodwel, lead time = 6 hrs, (c) Llanerfyl, lead time = 12 hrs.

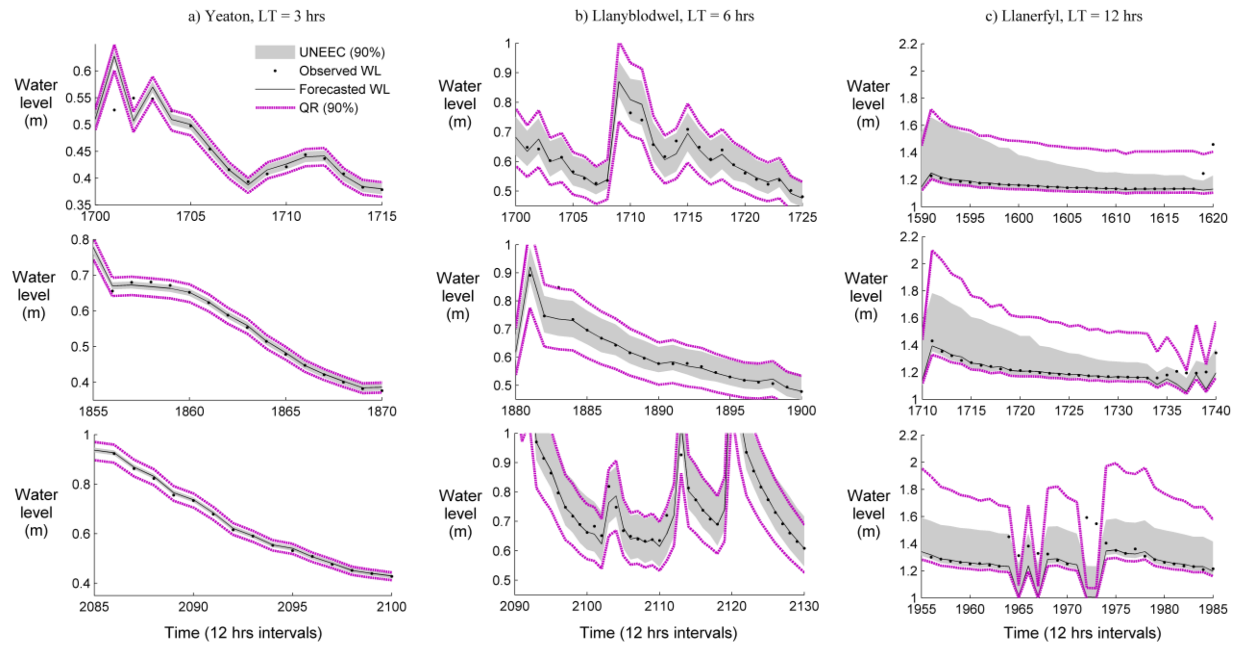

Figure 15. Comparison of prediction limits for falling limb part of the hydrographs (medium water levels) for 90% confidence level during validation: (a) Yeaton, lead time = 3 hrs, (b) Llanyblodwel, lead time = 6 hrs, (c) Llanerfyl, lead time = 12 hrs.

Table 1. Summary of the main basin characteristics for the catchments selected.

Catchment name	Drainage area (km^2)	Elevation (m)	Mean flow (m^3/s)	Mean annual rainfall (mm)	Highest river level recorded (m)	Basin lag time (hr)
Brue	135	≈ 20	1.92*	867*	4.45***	8 - 9
Yeaton	180.8	61.18	†1.60	†767	1.13***	15 - 20
Llanyblodwel	229	77.28	†6.58	†1267	2.68***	7 - 10
Llanerfyl	≈ 100	151	> 10**	> 1300**	3.59***	3 - 5

* Basin average for the period 1961-1990.
** Rough estimates based on the data available for 2006-2013.
*** http://apps.environment-agency.gov.uk/river-and-sea-levels/
† Computed for the periods 1963-2005 and 1973-2005 for Yeaton and Llanyblodwel, respectively and taken from UK Hydrometric Register (Marsh and Hannaford, 2008).

Table 2. Uncertainty analysis results for 90% and 50% confidence levels (Brue catchment).

Confidence level		PICP (%)		MPI (m^3/s)		ARIL (-)	
		UNEEC	*QR*	*UNEEC*	*QR*	*UNEEC*	*QR*
TR	90%	91.19	90.00	1.58	1.69	1.86	1.47
	50%	51.28	50.01	0.54	0.58	0.55	0.46
VD	90%	88.29	82.33	1.37	1.39	2.35	1.83
	50%	30.29	32.75	0.45	0.47	0.67	0.57
VD *(highest peak event)*	90%	57.14	62.79	2.86	3.47	0.66	0.78
	50%	27.91	30.90	1.06	1.28	0.24	0.27
VD *(medium peak event)*	90%	88.04	87.04	2.36	2.75	0.51	0.61
	50%	55.81	50.50	0.90	1.00	0.20	0.22

TR: *Training*, *VD*: *Validation*

Table 3. PICP, MPI, and ARIL values for each cluster (training, 90% confidence level, Brue): UNEEC vs. QR.

Cluster No	Number of data	UNEEC			QR		
		PICP (%)	*MPI (m^3/s)*	*ARIL (-)*	*PICP (%)*	*MPI (m^3/s)*	*ARIL (-)*
1[a]	5447 (62.3%)	92.12	1.14	2.67	88.16	0.88	1.96
2	787 (9.0%)	82.08	2.98	0.50	84.5	3.51	0.57
3	2167 (24.7%)	94.46	1.44	0.53	96.72	1.94	0.71
4[b]	83 (0.95%)	74.70	7.55	0.33	90.36	12.00	0.49
5	266 (3.05%)	77.44	5.96	0.48	89.47	7.58	0.58

[a] Low flows, low rainfall.
[b] High Flows, high rainfall.

Table 4. PICP, MPI, and ARIL values for MEDIUM water levels (validation, 90% confidence level): UNEEC vs. QR.

Catchment	Water level range (medium)	Number of data	UNEEC			QR		
			PICP (%)	MPI (m)	ARIL (-)	PICP (%)	MPI (m)	ARIL (-)
Yeaton	0.3 - 0.6 m	281 (13%)	82.56	0.0212	0.054	86.48	0.0299	0.074
Llanyblodwel	0.5 - 0.8 m	540 (25%)	89.63	0.1377	0.223	93.52	0.1680	0.269
Llanerfyl	1.3 - 1.6	570 (26.5%)	84.91	0.4156	0.297	85.09	0.5572	0.398

Table 5. PICP, MPI, and ARIL values for each cluster (training, 90% confidence level, Llanyblodwel, lead time = 6 hrs): UNEEC vs. QR.

Cluster No	Number of data	UNEEC			QR		
		PICP (%)	MPI (m)	ARIL (-)	PICP (%)	MPI (m)	ARIL (-)
1	413 (19.1%)	88.62	0.1492	0.271	93.95	0.1506	0.250
2[a]	100 (4.6%)	85.00	0.2964	0.288	95.00	0.3538	0.326
3	336 (15.5%)	90.18	0.1798	0.249	94.94	0.2283	0.287
4[b]	359 (16.6%)	93.04	0.0518	0.182	89.14	0.0305	0.100
5	535 (24.8%)	89.53	0.1128	0.308	85.79	0.0742	0.179
6	416 (19.2%)	90.38	0.0920	0.212	92.31	0.1021	0.208

[a] High groundwater levels
[b] Low groundwater levels