Residual uncertainty estimation using instance-based learning with applications to hydrologic forecasting

Wani, Omar; Beckers, Joost V.L.; Weerts, Albrecht H.; Solomatine, Dimitri P.

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Hydrology and
Earth System
Sciences

Residual uncertainty estimation using instance-based learning with applications to hydrologic forecasting

Omar Wani[1,2,a,b], Joost V. L. Beckers[2], Albrecht H. Weerts[2,3], and Dimitri P. Solomatine[1,4,5]

[1]IHE Delft Institute for Water Education, Delft, the Netherlands
[2]Deltares, Delft, the Netherlands
[3]Hydrology and Quantitative Water Management Group, Department of Environmental Sciences, Wageningen University, Wageningen, the Netherlands
[4]Water Resources Section, Delft University of Technology, Delft, the Netherlands
[5]Water Problems Institute of RAS, Moscow, Russia
[a]currently at: Institute of Environmental Engineering, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
[b]currently at: Swiss Federal Institute of Aquatic Science and Technology (Eawag), Dübendorf, Switzerland

Correspondence to: Omar Wani (owani@student.ethz.ch)

Abstract. A non-parametric method is applied to quantify residual uncertainty in hydrologic streamflow forecasting. This method acts as a post-processor on deterministic model forecasts and generates a residual uncertainty distribution. Based on instance-based learning, it uses a k nearest-neighbour search for similar historical hydrometeorological conditions to determine uncertainty intervals from a set of historical errors, i.e. discrepancies between past forecast and observation. The performance of this method is assessed using test cases of hydrologic forecasting in two UK rivers: the Severn and Brue. Forecasts in retrospect were made and their uncertainties were estimated using kNN resampling and two alternative uncertainty estimators: quantile regression (QR) and uncertainty estimation based on local errors and clustering (UNEEC). Results show that kNN uncertainty estimation produces accurate and narrow uncertainty intervals with good probability coverage. Analysis also shows that the performance of this technique depends on the choice of search space. Nevertheless, the accuracy and reliability of uncertainty intervals generated using kNN resampling are at least comparable to those produced by QR and UNEEC. It is concluded that kNN uncertainty estimation is an interesting alternative to other post-processors, like QR and UNEEC, for estimating forecast uncertainty. Apart from its concept being simple and well understood, an advantage of this method is that it is relatively easy to implement.

1 Introduction

Hydrologic forecasts for real-life systems are inevitably uncertain (Beven and Binley, 1992; Gupta et al., 1998; Refsgaard et al., 2007). This, among other things, is due to the uncertainties in the meteorological forcing, in the modelling of the hydrologic system response and in the initial state of the system at the time of forecast. It is well accepted that, compared to a simple deterministic forecast, additional information about the expected degree of accuracy of that forecast is valuable and generally leads to better decision making (Krzysztofowicz, 2001). Various techniques have therefore been developed to quantify uncertainties associated with the meteorological model input (van Andel et al., 2013), the initial state of the model (Li et al., 2009) and the hydrologic models themselves (Deletic et al., 2012; Coccia and Todini, 2011). Frameworks and guidelines have been developed to incorporate uncertainty analysis of environmental models effectively in decision making (Arnal et al., 2016; Reichert et al., 2007; Refsgaard et al., 2007). Broadly, there are three basic approaches to uncertainty estimation: (i) explicitly defining a probability model for the system response, e.g. Todini (2008), (ii) estimation of statistical properties of the error time series in the post-processing phase of model forecast, e.g. Dogulu et al. (2015), and (iii) methods using Monte Carlo sampling of inputs and/or parameters, aimed

at getting a range of model outputs, e.g. Beven and Binley (1992) and Freer et al. (1996). Other uncertainty estimation techniques may employ a combination of these approaches (Del Giudice et al., 2013). Some techniques focus on one source of uncertainty, such as the model parameter uncertainty (Benke et al., 2008) or the model structure uncertainty (Butts et al., 2004), while others focus on combined uncertainties stemming from model parameters, model structure deficits and inputs (Schoups and Vrugt, 2010; Evin et al., 2013; Del Giudice et al., 2013). In this context, it is important to note that apart from estimating uncertainty of model parameters during calibration, uncertainty estimation for hydrologic forecasting requires quantification of predictive uncertainty, which includes uncertain system response in addition to different combinations of model parameters (Renard et al., 2010; Coccia and Todini, 2011; Dotto et al., 2012).

In this paper, we will restrict ourselves to the class of uncertainty estimators called post-processors. These methods usually do not discriminate between different sources of uncertainty. They "aggregate" all sources into a so-called residual uncertainty. Post-processing methods assume the existence of a single calibrated model with an optimal set of model parameters, and build a statistical or machine learning model of the residual uncertainty. Typically, these techniques relate a combination of model inputs and/or outputs to the model error distribution. Various post-processors have been developed and applied to hydrologic modelling, such as a meta-Gaussian error model (Montanari and Brath, 2004), UNEEC (Solomatine and Shrestha, 2009), quantile regression (Weerts et al., 2011), and DUMBRAE (Pianosi and Raso, 2012). Quantile regression (QR) is a relatively straightforward post-processing technique that relates the probability of residual errors to the model forecast (the predictand) by a regression model that is derived from historical forecasts and observations. QR has been successfully applied for uncertainty quantification in hydrologic forecasts with various modifications (Weerts et al., 2011; Verkade et al., 2013; Roscoe et al., 2012; López López et al., 2014; Hoss and Fischbeck, 2015), whereas UNEEC involves a machine learning technique for building a non-linear regression model of error quantiles (Solomatine and Shrestha, 2009). UNEEC includes three steps: (1) fuzzy clustering of input data in the space of "relevant" variables; (2) estimating the probability distribution function of residual errors for each cluster and (3) building a machine learning model (e.g. an artificial neural network) of the prediction interval for a given probability (Dogulu et al., 2015). Many other uncertainty estimation techniques, such as DUMBRAE (Pianosi and Raso, 2012), HUP (Krzysztofowicz, 1999), model conditional processor (Coccia and Todini, 2011), Bayesian revision (Reggiani et al., 2009) and Bayesian model averaging (Raftery et al., 2005), make explicit assumptions about the nature of the probability distribution function of error. This is not necessary for QR and UNEEC (López López et al., 2014; Dogulu

et al., 2015). Nevertheless, in QR and UNEEC assumptions need to be made about the form of the regression function that is used to calculate the quantiles.

In an attempt to explore the utility of easier-to-implement post-processing techniques, we employ a simple non-parametric forecast method for residual uncertainty quantification. This method uses kNN search to learn about the past residual errors, which avoids having to make explicit assumptions about the nature of the error distribution and tuning of distribution parameters. Instance-based learning has been used in meteorology and hydrology before for resampling of precipitation and streamflows, most notably by Lall and Sharma (1996), who used the k nearest-neighbour (kNN) method for resampling of monthly streamflow sequences. kNN search has also been used in a non-parametric simulation method to generate random sequences of daily weather variables (Rajagopalan and Lall, 1999). They defined a weighting function for probability where the predictand is resampled from k values. Jules and Buishand (2003) used nearest-neighbour resampling to generate multi-site sequences of daily precipitation and temperature in the Rhine basin. Also, instance-based learning has been used as a data-driven model for hydrologic forecasting (Solomatine et al., 2008; Solomatine and Ostfeld, 2008). Beckers et al. (2016) use nearest-neighbour resampling to generate monthly sequences of climate indices and related precipitation and temperature series for the Columbia River basin. Specifically in the context of error modelling, a version of UNEEC that uses kNN instance-based learning as its basic machine learning technique to predict the residual error quantiles was compared to the original ANN-based UNEEC in Shrestha and Solomatine (2008). However, kNN can also be used without the complicated UNEEC procedure that includes fuzzy clustering. The application of kNN has recently been tested for forecast updating by constructing a deterministic error prediction model (Akbari and Afshar, 2014). Similarly, it has been shown that model errors can be resampled using kNN, after explicitly accounting for input and parameter uncertainty, to generate uncertainty intervals (Sikorska et al., 2015). In this paper we extend the simplification of kNN resampling for uncertainty estimation. We present an application of the kNN method to generate residual uncertainty estimates for a predictand, using a fixed time series of input and fixed model parameters, and explore whether this approach, being simpler than many other uncertainty quantification approaches mentioned above, is a useful or even a better alternative.

To demonstrate its use, we employ a relatively simple configuration of kNN resampling to produce uncertainty intervals for hydrologic forecasting. The next section explains the method in more detail and describes the validation procedure, i.e. the performance indicators. In Sect. 3, the method is applied to two case studies, each with a different system response (discharge and water level). The performance of kNN uncertainty estimation as a function of forecast lead time is

analysed in the first case study. The second case study is used to further validate the performance of kNN uncertainty estimation and analyse its sensitivity to the choice of search space and the value of k. Also, the influence of systematic bias in the hydrologic model on the uncertainty intervals generated by kNN search is explored in the second case study. For both case studies, performance indices of kNN resampling are compared to those of QR and UNEEC. And finally in Sect. 4, we discuss the usability of kNN search as a post-processor uncertainty estimator in hydrologic forecasting.

2 Method

2.1 kNN error model

The kNN residual uncertainty estimator can be seen as a zero-order local error quantile model built from a kNN search. Let us define a vector $\boldsymbol{v}$ in n-dimensional space of variables (the search space) on which the residual uncertainty is assumed to be statistically dependent.

$$\boldsymbol{v} = \left[v^1, \ldots, v^n \right] \tag{1}$$

The cumulative probability distribution function C of residual errors at prediction time step t conditioned on $\boldsymbol{v} = \boldsymbol{v}_t$ is defined as

$$C_t \left(e | \boldsymbol{v} = \boldsymbol{v}_t \right) = P_t \left(E \leq e | \boldsymbol{v} = \boldsymbol{v}_t \right), \tag{2}$$

where P is the probability function and E denotes the random variable for residual errors. Residual error is defined throughout this paper as the difference between the simulated values and the observed values for a hydrologic system response f, like discharge or water level.

$$e = f_{\text{simluated}} - f_{\text{observed}} \tag{3}$$

We are making the assumption of stationarity in time so that past error distributions are representative of the future:

$$C_t \left(e | \boldsymbol{v} = \boldsymbol{v}_t \right) = C_{\text{p}} \left(e | \boldsymbol{v} = \boldsymbol{v}_t \right). \tag{4}$$

The subscript p denotes historical time series. Therefore C_{p} is the cumulative distribution function of residual errors from the past. In Eq. (4), C_{p} is being conditioned to the input variable vector at time t. Nevertheless, as we only have single realizations of the error variable E for each historical point, we relax the constraint of $\boldsymbol{v} = \boldsymbol{v}_t$. Instead, we assume that the nearby neighbours of $\boldsymbol{v}_t$ in n-dimensional space will have a similar probability distribution of errors to $\boldsymbol{v}_t$ and that these historical errors are samples from $C_{\text{p}} \left(e | \boldsymbol{v} = \boldsymbol{v}_t \right)$. An empirical probability distribution can thus be constructed using the kNN historical errors:

$$C_t \left(e_t | \boldsymbol{v} = \boldsymbol{v}_t \right) \approx C_{\text{p}} \left(e | r_{\text{p}} \leq r_k \right), \tag{5}$$
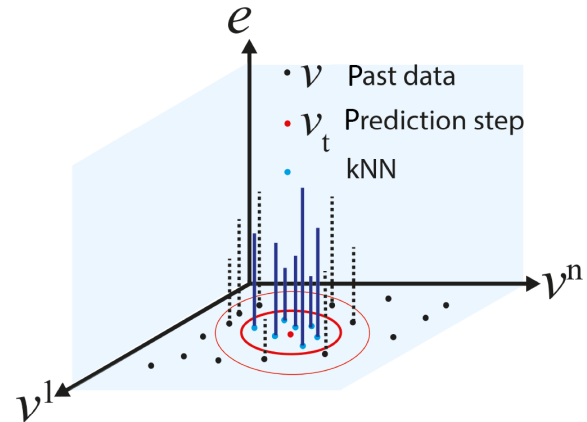

Figure 1. Dependence of error samples on the value of k. For larger values of k, points are at a greater distance from $\boldsymbol{v}_t$ (the prediction step), thus compromising the conditioning of the residual error probability distribution on $\boldsymbol{v}_t$ (Eq. 5).

where r_{p} is the Euclidean distance in n-dimensional space of input variables.

$$r_{\text{p}} = |\boldsymbol{v}_{\text{p}} - \boldsymbol{v}_t| = \sqrt{\left[\sum_{i=1}^{n} \left(v_{\text{p}}^i - v_t^i \right)^2 \right]} \tag{6}$$

$\boldsymbol{v}_{\text{p}}$ is the input variable vector of the past data point in the cloud of such past data points $\boldsymbol{v}$ (Fig. 1) and r_k is the distance to the kth nearest neighbour of $\boldsymbol{v}_t$. The choice of the input variable vector is a problem in itself since it should include only the most relevant variables that determine the forecast uncertainty. In this study, the input variable vector is chosen based on correlation between the candidate variables and the past errors. If the correlation between the error time series and a particular candidate variable is relatively high, then it can be included in the input variable vector space. Other, more sophisticated methods involving the mutual information can be used as well (Fernando et al., 2009). This will be exemplified in the case studies described in the next section. To represent the relative importance of input variables used in the search, dimensions of the input variable vector space can be suitably weighted in. Also, the model-based methods can be used where models are built for each considered candidate input variable set, and the choice is made based on their relative performance. These, however, were not explored in this study; it rather focused on the usability of the kNN search in its most basic implementation for uncertainty quantification. Nevertheless, we do demonstrate the sensitivity of the uncertainty intervals to the choice of input variable vector.

In order to level variables with different magnitudes, they are normalized. If σ_i represents the standard deviation of in-

put variable i calculated using the past data, then

$$r_{\mathrm{p}} = \sqrt{\left[\sum_{i=1}^{n} \frac{\left(v_{\mathrm{p}}^{i} - v_{t}^{i}\right)^2}{\sigma_i^2}\right]}. \tag{7}$$

Once the input variable vector space is decided, the probability of non-exceedance of a forecast error is calculated empirically by sampling from the conditional error distribution:

$$C_t\left(e_t | \boldsymbol{v} = \boldsymbol{v}_t\right) \approx C_{\mathrm{p}}\left(e | r_{\mathrm{p}} \leq r_k\right) = j / k, \tag{8}$$

where j is the rank of value e (for which the probability of non-exceedance is being computed) in the ascending array of k error values. The kNN search is thus employed to generate a sample and to build an empirical error distribution for this predictive uncertainty quantification. Such a mathematical description does not employ explicit regression models for predicting quantiles, which can be seen as a disadvantage in extrapolating outside available data. Also, as this configuration of kNN used in this research generates residual error quantiles, which capture the mismatch between measurement values and simulated values, the uncertainty in observational data is not considered. The generated quantiles are aimed at capturing the measured system response and do not attempt to capture the true response of the hydrologic system.

As one would expect, due to the nature of our sampling approximation (Eq. 8), the number of nearest neighbours, k, will affect the empirical conditional probability distribution of errors. If k is very large, many data points that are quite distant from $\boldsymbol{v}_t$ (Fig. 1) will be selected and the conditioning on the current forecast situation will not be valid. Large values of k will thus yield error distributions with larger uncertainty intervals – resembling the marginal error distribution. If k is small, the set of k errors will be small and subject to sampling error, so this set will not adequately represent the uncertainty distribution at $\boldsymbol{v}_t$. The tail of a distribution is more prone to sampling errors compared to its mean. Thus, to attain an acceptable degree of convergence, many more samples are required for quantiles corresponding to bigger prediction intervals (van der Vaart, 1998). For improved performance, the value of k can be subject to optimization of some cost function: the optimal value of k could be the one that enables a reasonable estimate of the uncertainty quantiles and additionally we may require that the sensitivity of the error distribution to k is small. In this study, we carry out such optimization using quite a simple heuristic guideline – the value of k is varied until the probability distribution of errors stabilizes and becomes less sensitive to the value of k for a few model predictions. We also demonstrate the sensitivity of uncertainty intervals to the value of k in one of the case studies. The choice of this relatively simple procedure for error quantile generation using kNN resampling is a reasonable starting point to assess its potential for residual uncertainty. This study explores the potential of uncertainty estimation using kNN in as simple a way as possible, and then compares its performance to two other residual uncertainty estimators. More advanced application of kNN, for example using fuzzy weights and kNN sampling to assign prediction intervals (Shrestha and Solomatine, 2008) or through explicit consideration of uncertainty in parameter and input by sampling them from their distributions, has been successfully shown (Sikorska et al., 2015).

To summarize, the steps for uncertainty quantification using kNN resampling are as follows.

1. Compose the input variable vector space ($\boldsymbol{v}$) on which uncertainty will be conditioned. Correlation analysis can help find the most relevant variables.

2. Set the number of neighbours k.

3. For a forecast at prediction time step t, identify the set of k nearest neighbours to the input vector $\boldsymbol{v}_t$. This set represents the hindcasts (forecasts in retrospect) most similar to $\boldsymbol{v}_t$.

4. Use the residual errors from these k points to build an empirical error distribution for the forecast at time step t.

5. Finally, identify the errors corresponding to the required quantiles (probabilities of non-exceedance) from this empirical distribution (in this paper, we use the 5–95 and 25–75 % quantiles).

2.2 Validation methods

Three statistical measures have been employed in this study to check the effectiveness of uncertainty estimation techniques, namely prediction interval coverage probability ($\mathrm{PICP}_{\mathrm{PI}}$), the mean prediction interval ($\mathrm{MPI}_{\mathrm{PI}}$) (see e.g. Shrestha and Solomatine, 2008; Dogulu et al., 2015) and the Alpha Index (Renard et al., 2010). $\mathrm{PICP}_{\mathrm{PI}}$ represents the percentage of observations (C) covered by a prediction interval (PI) corresponding to a certain probability of occurrence (in our case 90 and 50 %).

$$\mathrm{PICP}_{\mathrm{PI}} = \frac{N_{\mathrm{in}}}{N_{\mathrm{obs}}} \cdot 100\,\%, \tag{9}$$

where N_{in} is the number of observations located within the PI and N_{obs} is the total number of observations. These metrics are calculated using the following equations:

$$\mathrm{PICP}_{90} = \frac{1}{n}\sum_{i=1}^{n} C_{90} \cdot 100\,\%, \quad \mathrm{PICP}_{50} = \frac{1}{n}\sum_{i=1}^{n} C_{50} \cdot 100\,\%, \tag{10}$$

$$C_{90} = \left\{ \begin{array}{ll} 1, & \text{if } q_{i,\,0.05} \leq q_i \leq q_{i,\,0.95} \\ 0, & \text{else} \end{array} \right\},$$

$$C_{50} = \left\{ \begin{array}{ll} 1, & \text{if } q_{i,\,0.25} \leq q_i \leq q_{i,\,0.75} \\ 0, & \text{else} \end{array} \right\}, \tag{11}$$

where $q_{i,0.95}$ and $q_{i,0.05}$ are values with 95 and 5 % probabilities of non-exceedance at time i. Thus the region bound within these two values will have a confidence interval of 90 %. Similarly, $q_{i,0.75}$ and $q_{i,0.25}$ represent the boundaries for 50 % C. The MPI is the average width of the confidence intervals corresponding to a particular probability. It is a measure of the magnitude of the uncertainty.

$$\mathrm{MPI}_{90} = \frac{1}{n}\sum_{i=1}^{n}\left(q_{i,0.95} - q_{i,0.05}\right)$$

$$\mathrm{MPI}_{50} = \frac{1}{n}\sum_{i=1}^{n}\left(q_{i,0.75} - q_{i,0.25}\right) \qquad (12)$$

We also quantify the reliability of the predicted error quantiles by comparing it to the observed error quantiles. The mismatch between the observed ($q_{\mathrm{obs},j}$) and predicted ($j/100$) error quantiles can be summarized by the Alpha Index (α).

$$\alpha' = \frac{1}{100}\sum_{j=1}^{100}\left|q_{\mathrm{obs},j} - j/100\right| \qquad (13)$$

$$\alpha = 1 - 2\alpha' \qquad (14)$$

There have been discussions whether an isolated verification index can capture all the aspects that make a probabilistic forecast good or bad (Laio and Tamea, 2007). The choice of a verification index for an uncertainty estimation technique should also be dependent on the purpose of hydrologic forecast. For example, Coccia and Todini (2011) evaluate the performance of model conditional processors for flood forecasting using the predicted and observed probabilities of exceedance over a threshold. Also, in their study predicted error quantiles are compared to observed error quantiles. López López et al. (2014) and Dogulu et al. (2015) use PICP and MPI, among other verification measures, to access the performance of QR and UNEEC. This study will limit the comparison of kNN resampling with other techniques to PICP and MPI only, which give a reasonable assessment of performance. Nevertheless, it does not preclude the possibility that the uncertainty estimation techniques perform differently if evaluated using other indices.

3 Case studies

The performance of kNN resampling was evaluated by applying the technique to hydrological forecasting for several catchments in two different parts of England. The two case studies provide two different hydrologic conditions for testing and include different models for prediction. Also, different kinds of system responses are being predicted in the two case studies – water level and discharge. The accuracy of the quantified prediction intervals was deduced by using validation data sets. Also, the first case study was used to evaluate the impact of changing lead time on uncertainty of hydrologic models and its quantification using kNN resampling.
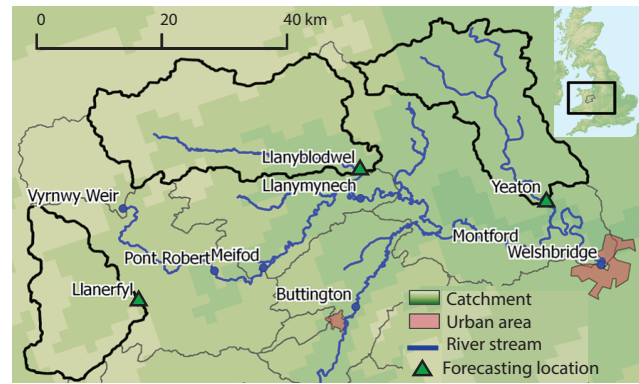

Figure 2. Upper Severn subcatchments with gauging stations (from López López et al., 2014).

3.1 Upper Severn catchment

3.1.1 Catchment description

The Upper Severn region is located in the Midlands, UK (Fig. 2). The River Severn, with a total length of 354 km, is the longest river in the UK. Its course acts as a geographic delineation between England and Wales, finally draining into the Bristol Channel. The overall River Severn catchment area is 10 459 km^2. Around 2.3 million people live in this region. The area is predominantly rural, but there are also a number of highly urbanized parts. The area covering the upper reaches of the River Severn, from its source on Plynlimon to its confluence with the River Perry upstream of Shrewsbury in Shropshire, is called the Upper Severn catchment. The Upper Severn catchment is predominantly hilly. It is dominated on the western edge by the Cambrian Mountains and a section of Snowdonia National Park (River Severn CFMP; EA, 2009).

The Severn catchment has a diverse geology. The headwaters of the river rise on Silurian mudstones, siltstones and grits and flow eastwards over these same rock formations. These rock formations do not allow water to flow easily through them. Therefore they are classified as non-aquifers with only limited potential for groundwater abstraction. Further west, in the Middle Severn section, the River Severn encounters sandstones, which are classified as a major aquifer and are highly permeable, highly productive and able to support large groundwater abstractions (River Severn CFMP; EA, 2009). The climate of the Severn catchment is generally temperate, experiencing modest to high precipitation depending on topography. Welsh mountains can receive over 2500 mm of precipitation per annum, whereas the rest of the catchment receives rainfall similar to the UK average – less than 700 mm per annum. The test forecast locations used in this study are Llanerfyl, Llanyblodwel and Yeaton. Table 1 lists the basin and hydrological information for these subcatchments (López López et al., 2014).

Catchment	Area (km^2)	Mean annual rain (mm)	Mean flow ($m^3 s^{-1}$)	Max water level (m)
Llanerfyl	125	1077	> 10	3.59
Llanyblodwel	229	1267	6.58	2.68
Yeaton	180.8	767	1.6	1.13

3.1.2 Experimental set-up

The flood forecasting system for the River Severn is organized in a sequential manner, being composed of a number of separate systems that are effectively linked. This forecasting system works with a high degree of automation and efforts have been made to involve a minimum amount of human intervention. The UK Environment Agency uses the Midlands Flood Forecasting System (MFFS) to do flood forecasting and to help in warning operation. The MFFS in turn is based on the Delft-FEWS (Flood Early Warning System) platform (Werner et al., 2013). Within the MFFS, there are lumped numerical models for rainfall–runoff (MCRM; Bailey and Dobson, 1981) and models for hydrologic (DODO; Wallingford, 1994) and hydrodynamic routing (ISIS; Wallingford, 1997). The rainfall input for the MFFS is acquired from ground measurements via rain gauges, from radar measurements or from numerical weather prediction data. The MFFS predicts ahead in time the response of the Upper Severn subcatchments but, as expected, the quality of forecast deteriorates with increasing lead time.

To do uncertainty analysis for the MFFS, hindcasting or reforecasting is done and then results are compared to the observed data. All the input time series used for hindcasting are taken from measured data. In this study, the reforecasting period was kept equal to the one employed in the studies of López López et al. (2014) and Dogulu et al. (2015). The chosen period is from 1 January 2006 to 7 March 2013. Data in the period till 6 March 2007 are used for the model spin-up. The remaining period is used for the calibration and validation of the uncertainty estimation techniques. Forecasts are made on a 12-hourly basis – at 08:00 and 20:00 daily, up to a lead time of 48 h. kNN resampling was applied for forecasts at 10 different lead times: 1, 3, 6, 9, 12, 18, 30, 36, 42, and 48 h. To choose an input variable vector for kNN resampling, correlation analysis was done between residual error and contenders for input variable vector space, namely simulated water level (H^{sim}), measured water level (H^{obs}) and residual error (E^{obs}) from various time steps t. The analysis was done to assist in a manual selection of input variable vectors. The correlation between residual error and water level reduces fast with time lag between the two time series.

Therefore it is enough to choose relatively simple and small-dimensional input variable vector spaces. For lead times, l, up to 6 h we chose

$$\boldsymbol{v} = \left[H_t^{sim}, H_{t-l}^{obs}, e_{t-l}^{obs} \right]. \tag{15}$$

For higher lead times, uncertainty has only been conditioned on H_t^{sim} as the residual error becomes less and less correlated with variable values as measured several hours behind the prediction time.

$$\boldsymbol{v} = \left[H_t^{sim} \right] \tag{16}$$

Ninety-nine values of residual errors were sampled from the nearest neighbourhood to generate an empirical distribution at each prediction step. This allowed us to get the "resolution" of 1 percentile in the generated empirical distribution. To develop confidence in the chosen value of k, we checked for a few prediction steps how sensitive the generated empirical distribution is to the value of k. Four different instances of $\boldsymbol{v}_t$ were chosen. Each instance represents a prediction step in the input variable vector space (the red circle in Fig. 1), with different hydrologic conditions. The plots of the cumulative mean square difference between probability density functions (pdfs) of varying k were generated. Cumulative mean square difference (Eq. 17) serves as an index to show how much the empirical pdfs change with changing k. We get a decreasing slope with increasing k. It shows that the pdfs become almost identical for values of k around 100. If $P_{k_i}(e)$ is the probability density for a residual error e calculated through k_i nearest neighbours using kNN resampling, for probability functions corresponding to discrete bin size Δe, the cumulative difference is defined as

cumulative difference

$$= \sum_{k_i=10}^{k_i=k} \sum_{\text{first } e \text{ bin}}^{\text{last } e \text{ bin}} \left[\Delta e \cdot P_{k_i}(e) - \Delta e \cdot P_{k_{i-1}}(e) \right]^2. \tag{17}$$

The various values of k_i that were tested are −10, 30, 50, 70, 90, 100, 110, 130 and 150. Using the information from Fig. 3, a value of $k = 99$ does not seem to be heavily affected by sampling errors. Nevertheless, it is not a mathematically calibrated value of k and therefore is likely to be sub-optimal. However, it should still be able to provide reasonably representative samples from the error distribution, as is suggested by Fig. 3.

3.1.3 Results

Figure 4 shows two hydrographs for the same event, where model predictions were made at different lead times. From the graph of lead time 48 h it is evident that the error quantiles that kNN resampling produces are not forced to have zero mean. Therefore the model prediction can sometimes lie
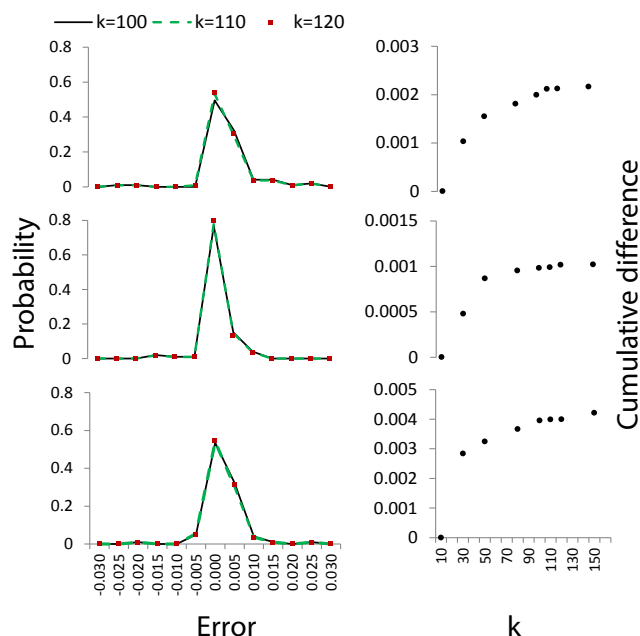
Figure 3. Dependence of the residual error probability function on the value of k for three didactic values of v_t (in each row in this plot). The probability is computed for error bins of size 0.005 units each. The graphs show that for k from around 90 to 120, the corresponding empirical error distributions become almost identical.

outside the predicted quantiles. This is because kNN resampling learns from past instances where the model has consistently underpredicted or overpredicted the flow, so it corrects for this bias. The hydrographs capture the low flows and the peaks well. It can also be seen that for high flows the errors are usually higher than for medium and low flows. The residual error distribution is thus heteroscedastic, i.e. the variance depends on the magnitude of the predicted flow. The autocorrelation can be checked by plotting errors versus time, whereas performance of an error model with regard to heteroscedasticity can be estimated by plotting reliability diagrams for different magnitudes of flow, which would mean different water levels in this case.

The plotting of error time series (Fig. 5) for various lead times shows some recurring trends across all three subcatchments. The errors are small for small lead time forecasts and the spread of error time series increases with increasing lead time. Moreover, the errors do not look autocorrelated for smaller lead times, whereas for the higher lead times autocorrelation becomes more prominent. This can be ascribed to the memory of the hydrologic system. If the system response is higher than what the model simulates for a particular lead time, then the system response is likely to be higher for the next time step as well. As the errors become larger, they tend to lose their independence property. This is captured by the error samples generated by kNN resampling as well. The rate at which autocorrelation deteriorates for ob-

Table 2. Alpha Index (α) for high flows corresponding to different lead times of Upper Severn subcatchments.

Lead time (h)	1	12	24	48
Llanerfyl	0.92	0.87	0.79	0.64
Llanyblodwel	0.93	0.95	0.93	0.90
Yeaton	0.97	0.94	0.94	0.75

served residual errors corresponds well to the kNN resampling error samples' autocorrelation. It can be seen that kNN resampling preserves the autocorrelation in the error time series without using an autoregressive model.

To check the performance of kNN resampling for various flow magnitudes, the simulation values were divided into low and high flows – the lowest and highest 10 % of water levels simulated in the validation phase respectively. The reliability diagrams (Fig. 6) show that the overall performance of error quantiles for all water levels is good for low and medium lead times. The reliability decreases with high lead times (24 h and above). The reliability plots show that kNN resampling performs better for high flows compared to low flows, even for higher lead times. For low flows and high lead times, the forecast probability of non-exceedance is higher than the observed relative frequency. Nevertheless, from 0.90 probability of non-exceedance and above, the reliability curve comes back to the desired 45° line. For flood forecasting it is important to model the high and medium flows well. kNN resampling delivers quite reliable quantiles for such flow regimes. The deteriorating model performance with higher lead times gets reflected in the performance of kNN resampling quintiles as well.

To assess the performance of kNN resampling relative to other established post-processor uncertainty estimation techniques, comparisons with QR and UNEEC have been carried out. The results for QR have been taken from López López et al. (2014) and the results for UNEEC from Dogulu et al. (2015). QR results for uncertainty estimation were available for all the lead times as done using kNN resampling and, from UNEEC, only for lead times of 1, 3, 6, 9, 12, and 24 h. Values of PICP and MPI are shown in Fig. 7, together with results from UNEEC and QR. The Alpha Index (α) is reported for several lead times in Table 2. As expected, the MPI of all the uncertainty estimation techniques increases with increasing lead time. Comparison between kNN resampling and QR has been made for three locations and 10 lead times in the validation period. Model simulations were run two times each day. Verification indices for uncertainty analysis were calculated separately for each lead time and each location. Considering the 90 and 50 % quantiles as two prediction intervals, this allowed for the evaluation of PICP and MPI 60 times (Fig. 7). kNN resampling has a higher PICP in 67 % of the cases and a smaller MPI for 73 % of the cases. A comparison between kNN resampling and UNEEC was
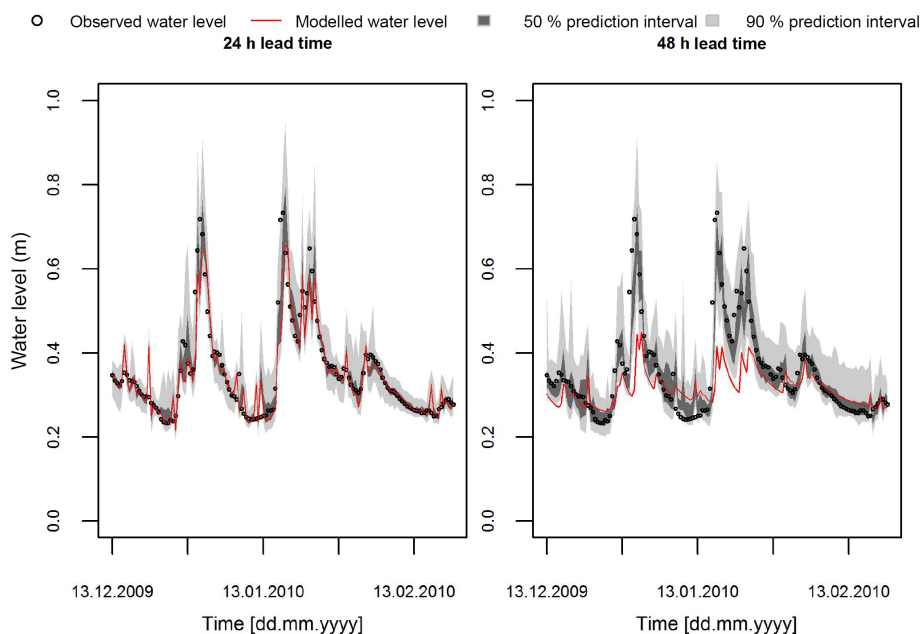
Figure 4. Prediction intervals for the Yeaton catchment using kNN resampling. The hydrographs are shown for the two different lead times. The 50 % prediction interval is the interval between the 25 and 75 % quantiles of residual error, and the 90 % quantile is the interval between the 5 and 95 % quantiles. The reporting time interval is 12 h.

made for three locations and five lead times for the validation. For each location and each lead time, the 90 and 50 % quantiles were generated, which allowed for the evaluation of PICP and MPI 30 times (Fig. 7). The PICP of kNN resampling is higher in 60 % of the cases and the MPI is smaller in 36 % of the cases. Based on these results we concluded that, for this case study, kNN resampling generally produces narrower confidence bands and provides a better coverage of the probability distribution than the other methods in the majority of forecasts, especially showing better performance for the larger lead times.

3.2   River Brue

3.2.1   Catchment description

The River Brue, located in the south-west of England, has a history of severe flooding. The test forecast location used in this study is Lovington, where the upstream catchment area is 135 km^2 (Fig. 8). The catchment is predominantly rural and the soil consists of clay and sand. This kind of soil and the modest relief give rise to a slowly responsive flow regime. The mean annual rainfall in the catchment is 867 mm; the mean river flow is 1.92 m^3 s^{-1} and has a maximum flow of 39.58 m^3 s^{-1}. This catchment has been extensively used for research on weather radar, quantitative precipitation forecasting and hydrologic modelling.

3.2.2   Experimental set-up

For the Brue catchment the simplified version of the HBV rainfall–runoff model has been used (Bergström, 1976). The HBV-96 model is a lumped conceptual model (Lindström et al., 1997). Like most other conceptual models, HBV consists of subroutines for snow accumulation and melt, soil moisture accounting and surface runoff, and employs a simple routing scheme. The input for the HBV model consists of precipitation (basin average), air temperature and potential evapotranspiration (estimated by the modified Penman method using automatic weather data available). Historical input data are available for a period of 1994–1996. Predictions are only made for 1 h lead time. Uncertainty analysis is done for a chosen period from 24 June 1994 to 31 May 1996. Hindcasts were made on a daily basis, using a warm state from a historical run. The hindcasts were split into calibration and validation set at 24 June 1995 for the uncertainty estimation techniques. The calibration data set was used to calibrate (train) UNEEC and QR, and for the resampling of errors using the kNN algorithm. The resampled errors were used to estimate prediction intervals for the predictions from the validation data set. Each of the two data sets represents almost a full year of observations. Three input variable vectors were chosen based on the results of correlation analysis, from simple to complex. This allows us to study the dependence on the choice of search space. Input variable vector (ivv) 3 for kNN resampling and UNEEC is the same for this comparison, whereas QR only uses Q_{sim} (Dogulu et al., 2015). The
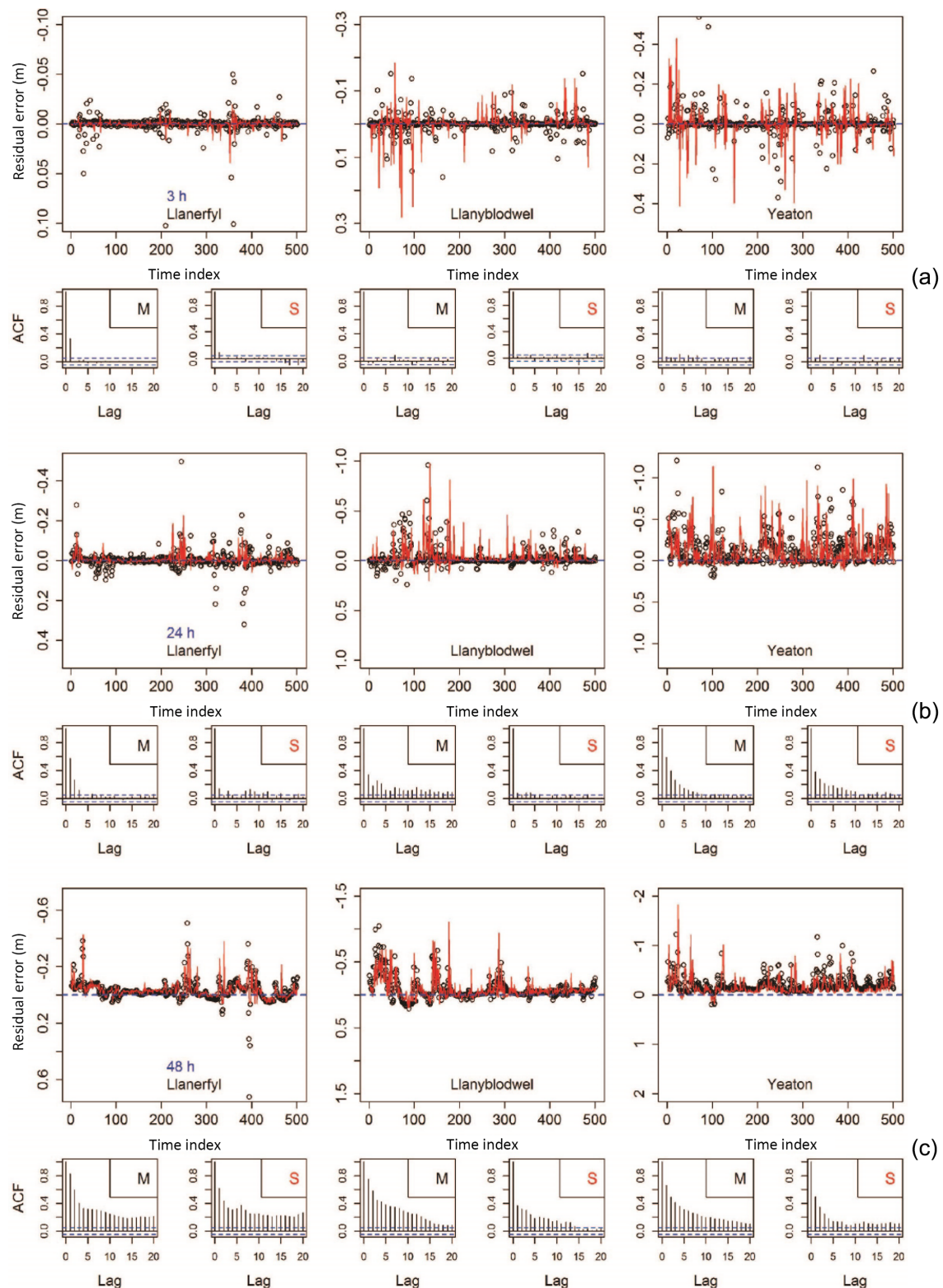
Figure 5. Plots of error samples and their autocorrelation (ACF). The error time series generated using kNN resampling are in red. Black circles represent the observed errors, i.e. obtained after measuring water level and comparing it to the simulated water level. M stands for measured and S for simulated. The lead times for each row are **(a)** 3 h, **(b)** 24 h and **(c)** 48 h.
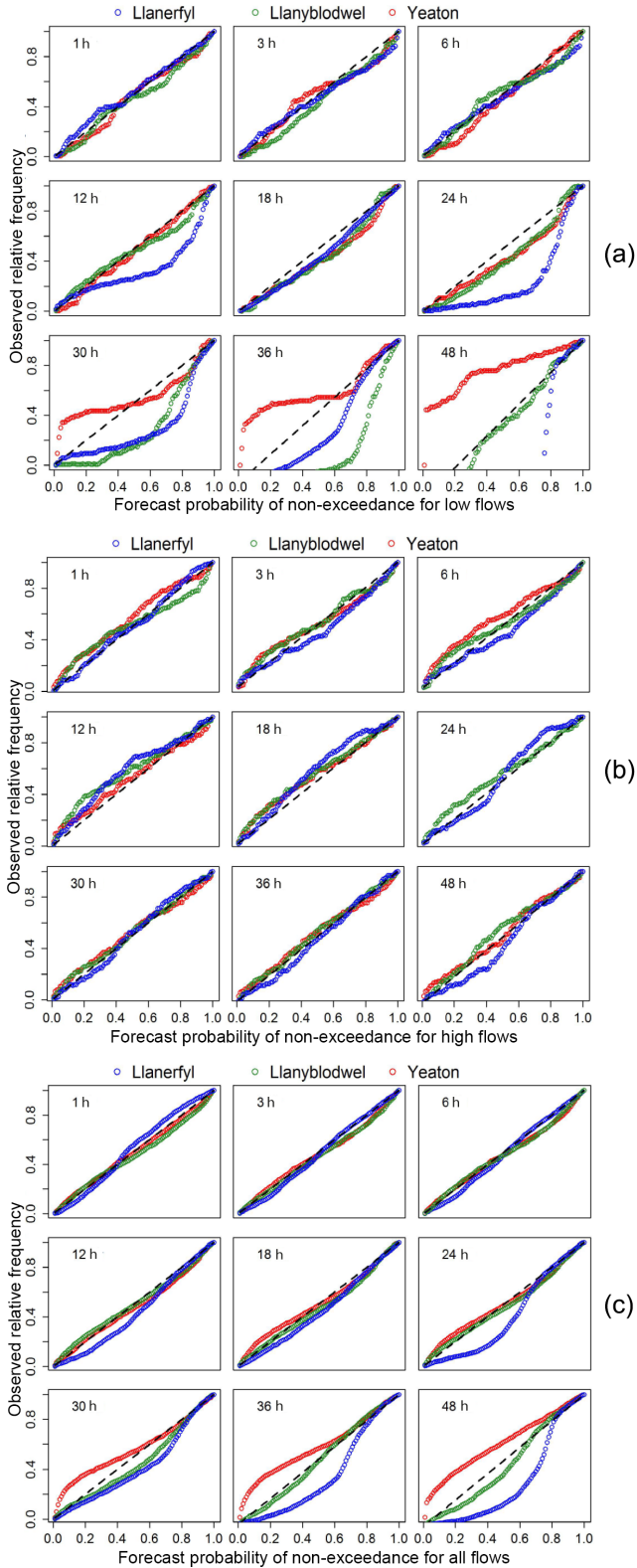
Figure 6. Reliability diagram from Upper Severn subcatchments for **(a)** low, **(b)** high and **(c)** all flows (Llanerfyl: blue; Llanyblodwel: green; Yeaton: red).
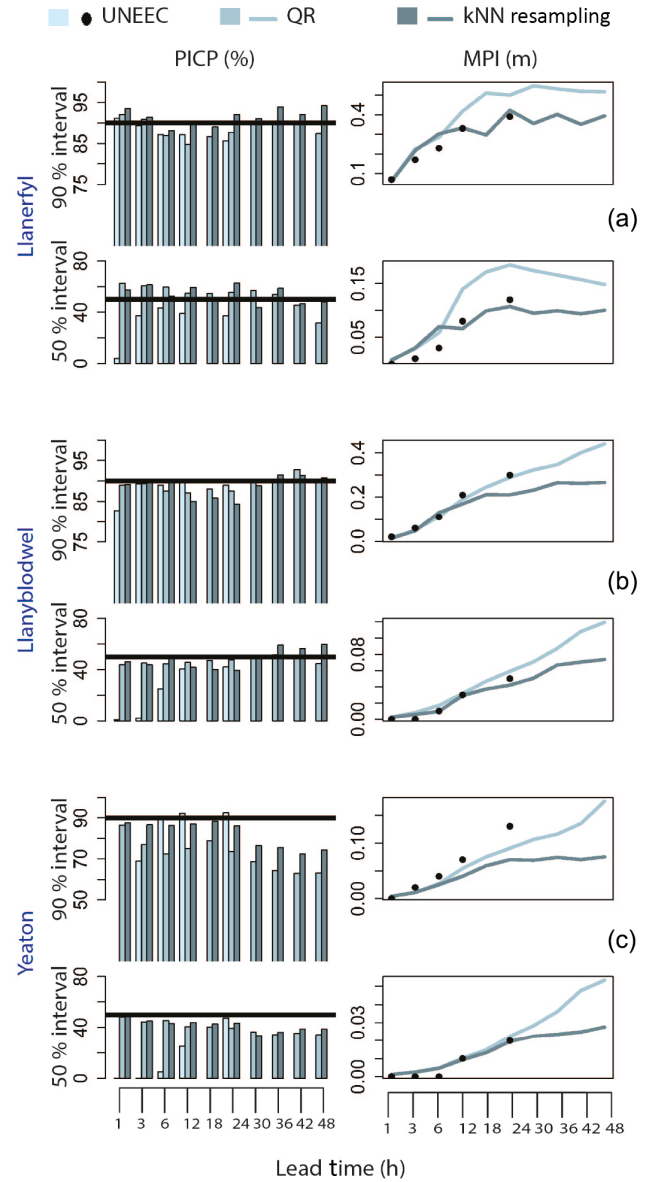
Figure 7. PICP and MPI comparison for Upper Severn subcatchments – **(a)** Llanerfyl, **(b)** Llanyblodwel and (c) Yeaton.

three input variable vectors used are

$$\boldsymbol{v}_{(ivv1)} = \left[Q_t^{sim} \right], \tag{18}$$

$$\boldsymbol{v}_{(ivv2)} = \left[Q_t^{sim}, e_{t-1}^{obs} \right], \tag{19}$$

$$\boldsymbol{v}_{(ivv3)} = \left[R_{t-8}^{obs}, R_{t-9}^{obs}, R_{t-10}^{obs}, Q_{t-1}^{obs}, Q_{t-2}^{obs}, Q_{t-3}^{obs}, \right.$$
$$\left. e_{t-1}^{obs}, e_{t-2}^{obs} \right], \tag{20}$$

where R is the effective rainfall, Q is the discharge and e is the residual error. Considering t as the prediction time, then the subscripts of the various input variables represent the time and the superscripts sim and obs mean they are sim-
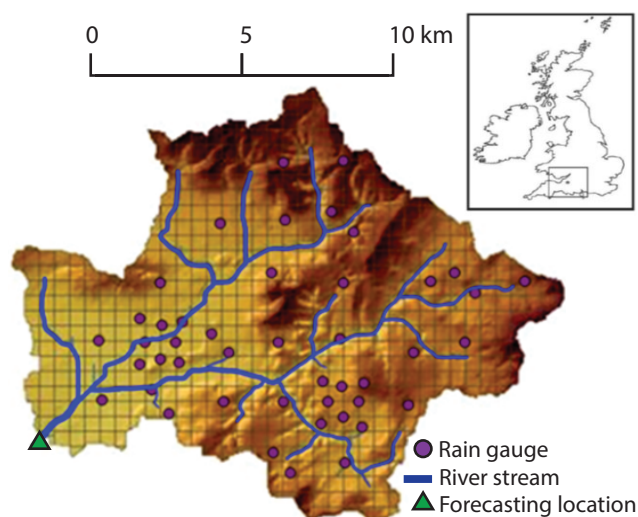
Figure 8. Brue catchment (from Shrestha and Solomatine, 2008).

ulated and observed values respectively. The number of nearest neighbours was chosen to be 99 and 199, to analyse its influence on uncertainty quantification. Uncertainty analysis was done for a calibrated HBV model as well as a model with a unit systematic bias. The bias was introduced to the simulation results of the calibrated model by simple addition. The aim of a biased model for uncertainty quantification using kNN resampling is to assess the performance of kNN resampling when the residuals are not zero mean.

3.2.3 Results

kNN resampling was applied to a single historical simulation and compared to observations. The simulated hydrographs for the highest discharge event with 50 and 90 % prediction intervals are shown in Fig. 9. The residual distribution of kNN resampling is generally a non-zero mean. Therefore we see that the prediction intervals may sometimes deviate from the deterministic model prediction quite significantly. The ability of kNN resampling to search for similar hydrologic conditions, like rainfall, and discharge in the past, and to learn from the residuals, allows it to make more representative error distributions. For example, in Fig. 9, the falling limb of the hydrograph shows that the prediction band generated by kNN resampling captures the observed flow for input variable vectors 2 and 3, even though the model shows a noticeable mismatch with the measurements. This can be explained by considering the history of errors that the model made during such hydrologic conditions in the past. And as kNN resampling learns that the model consistently underestimates in such cases, the corresponding error distribution corrects for this bias. The results of the PICP and MPI are shown in Table 3 together with results from UNEEC and QR (Dogulu et al., 2015). As can be seen from the table, kNN resampling's performance is comparable to that of UNEEC and

QR for this case study. The prediction intervals generated by kNN resampling are smaller, compared to the other two uncertainty estimation techniques, while the coverage probability is similar. It indicates that kNN resampling is able to learn well from past data and condition the probability of residual errors well. The Alpha Index for the validation phase is also high (0.96). We also notice that three different input variable vectors show different degrees of performance (Fig. 9). The past errors, e_{t-1}^{obs}, seem to be informative in this case, providing very narrow conditional error probabilities.

Apart from evaluating the usability of kNN resampling for calibrated models, the performance of kNN resampling quantiles generated by kNN resampling for a model with systematic bias was also checked. Figure 10 shows that the performance of kNN resampling does not diminish under systematic bias. The reliability of the generated quantiles remains almost unfazed. As a systematic bias will not affect the autocorrelation structure of the residual errors, the autocorrelation of error samples generated through kNN resampling also remains unchanged. Nevertheless, we see a shift in the mean of the sample time series, which is roughly equal to unity. The reliability of quantiles generated using kNN resampling for high flows (highest 10 % in the validation period) is poorer than for all flows. The invariance of kNN resampling performance to model bias makes it a robust postprocessor technique; however, unlike in the case of Upper Severn subcatchments, the technique's performance diminishes for high flows.

4 Discussion and conclusions

The application of kNN resampling to two case studies shows that the forecast uncertainty intervals are relatively narrow and still capture the observations well. The expected increase in uncertainty for longer lead times is also reproduced well and the probability coverage of kNN resampling remains good, as verified from historical observations. This is in accordance with previous research (Sikorska et al., 2015). The error samples generated by kNN resampling reproduce two important characteristics of residual errors in hydrologic models, namely autocorrelation and heteroscedasticity. Also, for applications to flood modelling, the high flows are most important and the uncertainty quantification by kNN resampling for the Upper Severn shows reasonable reliability for this high-flow regime. For the Brue, the performance is poorer. This can be attributed to the inadequacy of representative high flows in the calibration phase in combination with the choice of the input variable vector. The highest flow in calibration time series is $15.4 \, \mathrm{m^3 \, s^{-1}}$, whereas in validation time series it is $29.9 \, \mathrm{m^3 \, s^{-1}}$. It is also shown that the technique is generally robust to the performance of the underlying deterministic model. If the model has systematic biases, kNN resampling learns from the past errors of the model and recreates the systematic bias in the empirical error distribu-
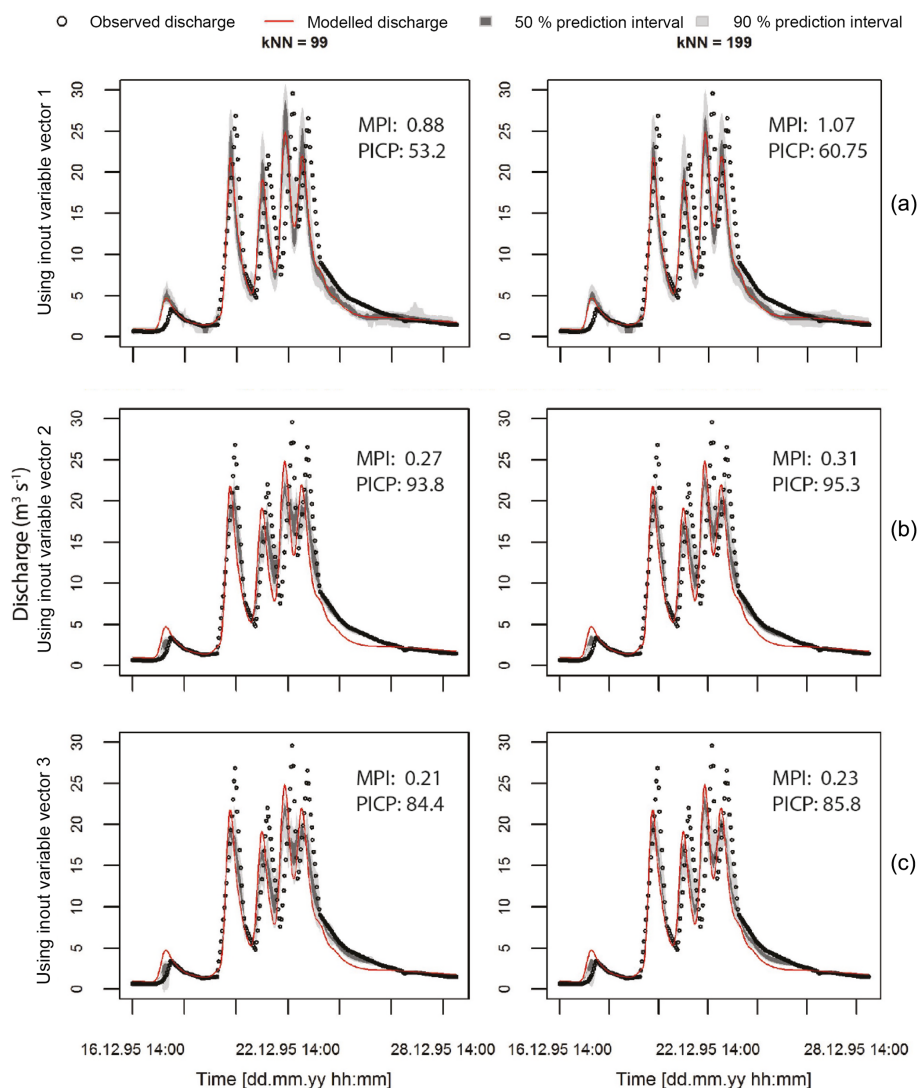
Figure 9. The 50 and 90 % prediction intervals for the Brue catchment using kNN resampling. The hydrographs are shown for two different k values (99, 199) and three different input variable vectors given by **(a)** Eq. (18), **(b)** Eq. (19) and **(c)** Eq. (20). This is the largest event in the validation time series. (The 50 % prediction interval is the interval between the 25 and 75 % quantiles of residual error, and the 90 % quantile is the interval between the 5 and 95 % quantiles. MPI and PICP correspond to the whole validation time series.)

Table 3. Performance of various uncertainty estimation techniques for the Brue catchment. For kNN resampling and UNEEC the same input variable vector is used (Eq. 20). For QR only Q_{sim} is used.

	PICP (expected 90 %)			MPI ($m^3\,s^{-1}$)		
	UNEEC	QR	kNN	UNEEC	QR	kNN
Calibration	91.19	90.00	86.3	1.58	1.69	0.51
Validation	88.29	82.33	84.42	1.37	1.39	0.21

tion mean, thus maintaining the performance of prediction intervals. Our results on systematic error correction by kNN resampling substantiate the findings from previous research on forecast updating using kNN (Akbari and Afshar, 2014).

These findings from this study are confirmed by three quantitative indicators of forecast reliability. The comparison of kNN resampling uncertainty estimates to those generated by QR and UNEEC shows that the mean prediction intervals
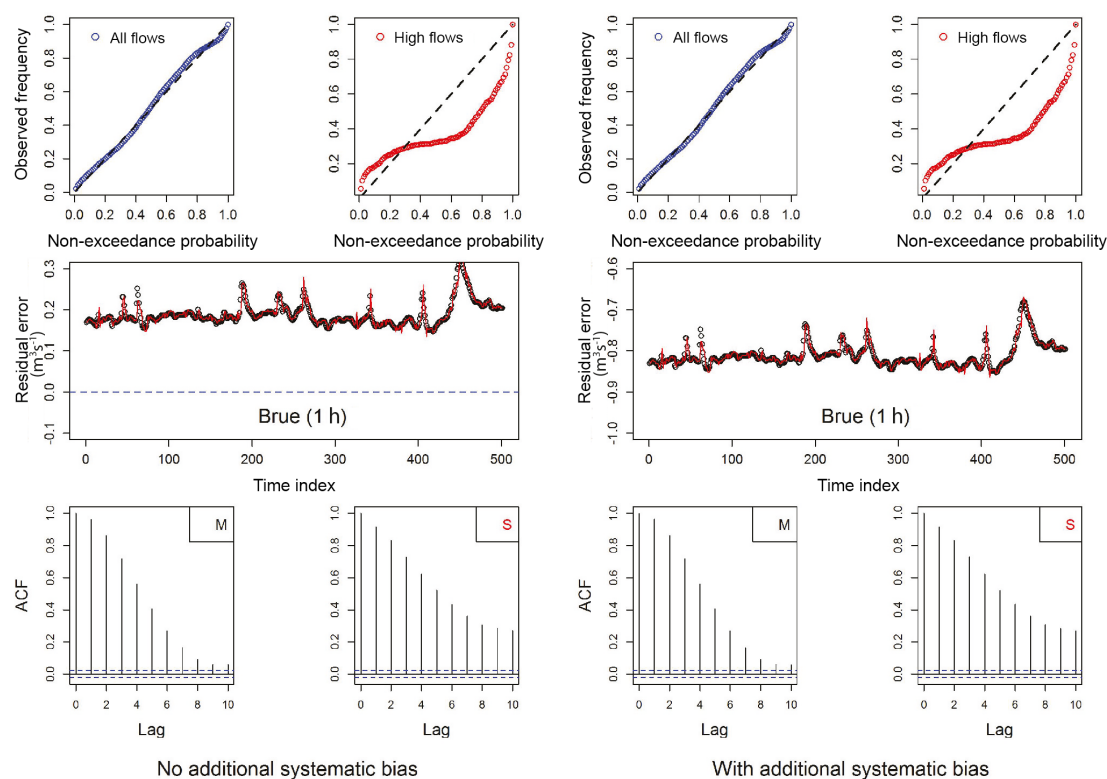
Figure 10. Effect on reliability of quantiles and autocorrelation of error samples on adding a systematic bias to the model artificially. kNN samples, generated using input variable vector 3 (Eq. 20), are plotted in red, and observed errors in black circles. M stands for measured and S for simulated.

(MPIs) generated by kNN resampling are generally smaller. A significantly smaller MPI using kNN resampling, as in the case of the Brue, is in part due to the conditioning on input variable vectors as compared to UNEEC and QR. As the values of k in this study have been restricted to 99 and 199, the error distribution tends to be much narrower than the marginal error distribution. The conditional distribution will turn into a marginal distribution when the number of k is equal to the time steps in the calibration time series. A more quantitative dependence on the k value and MPI will need further research. Apart from a narrow MPI, we also find that kNN resampling is generally able to capture the expected ratio of observations within its intervals (PICP) most of the time, or at least be close to the expected value.

As in the case of all other data-driven methods, the applicability of kNN resampling depends on the availability of sufficiently long and representative historical forecasts and observations. The historical series should include several occurrences of forecasting situations that are similar to the current situation. In extreme cases, the kind of kNN search proposed here will select the most similar historical situations which may or may not be representative of the current situation. In contrast to the methods like QR and UNEEC that build explicit predictive regression models which are able to extrapolate for the data which are beyond the limits of the calibration (training set), kNN resampling does not extrapolate. This could be seen as a disadvantage. On the other hand, however, the extrapolation that is done by regression techniques could also be seen as doubtful. It is not a given that the most extreme historical situations are less representative of the uncertainty of an extremely high flow than an extrapolated result. The results in this paper show that kNN resampling has a good or poor reliability for the highest values in the validation set, depending on the case study and the choice of input variable vector. Due to the non-parametric nature of kNN resampling, the increasing variance of residual errors for higher values of predictands is generally adequately taken into account.

As kNN resampling, like other post-processors, learns about the residual error process from the past, the historical records should be representative of the current forecast conditions. In changing conditions, this may not be true. Changing conditions may be caused, for example, by climate change or more local changes in the catchment like deforestation and dam building. This is a common problem for all data-driven statistical estimators and is not unique to kNN resampling. Care needs to be taken to use data time series which do not outright violate the assumptions regarding the invariance of catchment and climate.

One of the few calibration parameters of kNN resampling is the number of nearest neighbours k. In this study, k has been chosen by a simple heuristic technique. For optimal performance, it would be advisable to calibrate k for each application in a more systematic way. We do show for Brue that the sensitivity of the uncertainty intervals to the value of k is not significant, when changing it from 99 to 199. However, we also expect that the optimal value of k will depend on the length of the historical data series and on the uncertainty quantiles of interest. In the context of search space, in this research, the input variable vector has been chosen by correlation analysis. It can be recommended to use more sophisticated procedures for real-life applications, which can capture the non-linear dependence between the error process and input variable vector candidates. Improvements in performance can possibly be achieved by seeking a better set of input variables for each forecast location and lead time of interest.

In conclusion, kNN resampling can be considered a relatively simple machine learning technique to predict hydrologic residual uncertainty. The errors from the similar hydrologic conditions in the past are used as samples for the residual error probability distribution and the samples are collected by a k nearest-neighbour search. The application of this technique to case studies Brue and Upper Severn sub-catchments has shown promising results. In comparison to many other data-driven techniques, kNN resampling has the advantage of avoiding assumptions about the nature of the residual error distribution: the instance-based learning approach is non-parametric and non-regressive and requires little calibration. The method was shown to be able to quantify hydrologic uncertainty to an accuracy that is comparable to other techniques like QR and UNEEC. Given the relatively small effort in setting up the method, the performance of kNN resampling in uncertainty quantification is more than acceptable when compared to other post-processor error models.

5   User interface

A website has been developed as part of this research to help generate uncertainty intervals using kNN resampling for a given time series of predictions. Address: www.modeluncertainty.com.

Data availability.   Data are available upon request.

Competing interests.   The authors declare that they have no conflict of interest.

References

Akbari, M. and Afshar, A.: Similarity-based error prediction approach for real-time inflow forecasting, Hydrol. Res., 45, 589–602, https://doi.org/10.2166/nh.2013.098, 2014.

Arnal, L., Ramos, M.-H., Coughlan de Perez, E., Cloke, H. L., Stephens, E., Wetterhall, F., van Andel, S. J., and Pappenberger, F.: Willingness-to-pay for a probabilistic flood forecast: a risk-based decision-making game, Hydrol. Earth Syst. Sci., 20, 3109–3128, https://doi.org/10.5194/hess-20-3109-2016, 2016.

Bailey, R. A. and Dobson, C.: Forecasting for floods in the Severn catchment, Journal of the Institution of Water Engineers and Scientists, 35, 168–178, 1981.

Beckers, J. V. L., Weerts, A. H., Tijdeman, E., and Welles, E.: ENSO-conditioned weather resampling method for seasonal ensemble streamflow prediction, Hydrol. Earth Syst. Sci., 20, 3277–3287, https://doi.org/10.5194/hess-20-3277-2016, 2016.

Benke, K. K., Lowell, K. E., and Hamilton, A. J.: Parameter uncertainty, sensitivity analysis and prediction error in a water-balance hydrological model, Math. Comput. Model., 47, 1134–1149, https://doi.org/10.1016/j.mcm.2007.05.017, 2008.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, Swedish Meteorological and Hydrological Institute, Norrköping, Sweden, SMHI Rep. RHO 7, 134 pp., 1976.

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

Butts, M. B., Payne, J. T., Kristensen, M., and Madsen, H.: An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, J. Hydrol., 298, 242–266, https://doi.org/10.1016/j.jhydrol.2004.03.042, 2004.

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, https://doi.org/10.5194/hess-15-3253-2011, 2011.

Deletic, A., Dotto, C. B. S., McCarthy, D. T., Kleidorfer, M., Freni, G., Mannina, G., Uhl, M., Henrichs, M., Fletcher, T. D., Rauch, W., Bertrand-Krajewski, J. L., and Tait, S.: Assessing uncertain-

ties in urban drainage models, Phys. Chem. Earth Pt. A/B/C, 42–44, 3–10, https://doi.org/10.1016/j.pce.2011.04.007, 2012.

Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., and Rieckermann, J.: Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, Hydrol. Earth Syst. Sci., 17, 4209–4225, https://doi.org/10.5194/hess-17-4209-2013, 2013.

Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H., and Shrestha, D. L.: Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments, Hydrol. Earth Syst. Sci., 19, 3181–3201, https://doi.org/10.5194/hess-19-3181-2015, 2015.

Dotto, C. B. S., Mannina, G., Kleidorfer, M., Vezzaro, L., Henrichs, M., McCarthy, D. T., Freni, G., Rauch, W., and Deletic, A.: Comparison of different uncertainty techniques in urban stormwater quantity and quality modelling, Water Res., 46, 2545–2558, https://doi.org/10.1016/j.watres.2012.02.009, 2012.

EA: Environment Agency: River levels: Midlands, available at: http://www.environment-agency.gov.uk/homeandleisure/floods/riverlevels/ (last access: 1 October 2013), 2009.

Evin, G., Kavetski, D., Thyer, M., and Kuczera, G.: Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, Water Resour. Res., 49, 4518–4524, https://doi.org/10.1002/wrcr.20284, 2013.

Fernando, T. M. K. G., Maier, H. R., and Dandy, G. C.: Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach, J. Hydrol., 367, 165–176, https://doi.org/10.1016/j.jhydrol.2008.10.019, 2009.

Freer, J., Beven, K., and Ambroise, B.: Bayesian Estimation of Uncertainty in Runoff Prediction and the Value of Data: An Application of the GLUE Approach, Water Resour. Res., 32, 2161–2173, https://doi.org/10.1029/95WR03723, 1996.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resour. Res., 34, 751–763, https://doi.org/10.1029/97wr03495, 1998.

Hoss, F. and Fischbeck, P. S.: Performance and robustness of probabilistic river forecasts computed with quantile regression based on multiple independent variables, Hydrol. Earth Syst. Sci., 19, 3969–3990, https://doi.org/10.5194/hess-19-3969-2015, 2015.

Jules, J. B. and Buishand, T. A.: Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation, Clim. Res., 25, 121–133, 2003.

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

Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, J. Hydrol., 249, 2–9, https://doi.org/10.1016/S0022-1694(01)00420-6, 2001.

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

Lall, U. and Sharma, A.: A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series, Water Resour. Res., 32, 679–693, https://doi.org/10.1029/95wr02966, 1996.

Li, H., Luo, L., Wood, E. F., and Schaake, J.: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, J. Geophys. Res., 114, D04114, https://doi.org/10.1029/2008jd010969, 2009.

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, https://doi.org/10.1016/S0022-1694(97)00041-3, 1997.

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, https://doi.org/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.

Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, Water Resour. Res., 40, W01106, https://doi.org/10.1029/2003WR002540, 2004.

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

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Mon. Weather Rev., 133, 1155–1174, https://doi.org/10.1175/mwr2906.1, 2005.

Rajagopalan, B. and Lall, U.: A k-nearest-neighbor simulator for daily precipitation and other weather variables, Water Resour. Res., 35, 3089–3101, https://doi.org/10.1029/1999wr900028, 1999.

Refsgaard, J. C., van der Sluijs, J. P., Hojberg, A. L., and Vanrolleghem, P. A.: Uncertainty in the environmental modelling process – A framework and guidance, Environ. Modell. Softw., 22, 1543–1556, https://doi.org/10.1016/j.envsoft.2007.02.004, 2007.

Reggiani, P., Renner, M., Weerts, A. H., and van Gelder, P. A. H. J. M.: Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system, Water Resour. Res., 45, W02428, https://doi.org/10.1029/2007WR006758, 2009.

Reichert, P., Borsuk, M., Hostmann, M., Schweizer, S., Spörri, C., Tockner, K., and Truffer, B.: Concepts of decision support for river rehabilitation, Environ. Modell. Softw., 22, 188–201, https://doi.org/10.1016/j.envsoft.2005.07.017, 2007.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resour. Res., 46, W05521, https://doi.org/10.1029/2009WR008328, 2010.

Roscoe, K. L., Weerts, A. H., and Schroevers, M.: Estimation of the uncertainty in water level forecasts at ungauged river locations using quantile regression, International Journal of River Basin Management, 10, 383–394, https://doi.org/10.1080/15715124.2012.740483, 2012.

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resour. Res., 46, W10531, https://doi.org/10.1029/2009WR008933, 2010.

Shrestha, D. L. and Solomatine, D. P.: Data-driven approaches for estimating uncertainty in rainfall-runoff modelling, International Journal of River Basin Management, 6, 109–122, https://doi.org/10.1080/15715124.2008.9635341, 2008.

Sikorska, A. E., Alberto, M., and Demetris, K.: Estimating the Uncertainty of Hydrological Predictions through Data-Driven Resampling Techniques, J. Hydrol. Eng., 20, A4014009, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000926, 2015.

Solomatine, D. P. and Ostfeld, A.: Data-driven modelling: some past experiences and new approaches, J. Hydroinform., 10, 3–22, https://doi.org/10.2166/hydro.2008.015, 2008.

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

Solomatine, D. P., Maskey, M., and Shrestha, D. L.: Instance-based learning compared to other data-driven methods in hydrological forecasting, Hydrol. Process., 22, 275–287, https://doi.org/10.1002/hyp.6592, 2008.

Todini, E.: A model conditional processor to assess predictive uncertainty in flood forecasting, International Journal of River Basin Management, 6, 123–137, https://doi.org/10.1080/15715124.2008.9635342, 2008.

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

van der Vaart, A. W.: Asymptotic Statistics, Asymptotic Statistics, 3, 443 pp., https://doi.org/10.2307/2530729, 1998.

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

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

Wallingford: HR Wallingford, ISIS software, HR Wallingford, Hydraluic Unit, Wallingford, UK, available at: http://www.isisuser.com/isis/ (last access: 1 October 2013), 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, https://doi.org/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, https://doi.org/10.1016/j.envsoft.2012.07.010, 2013.