WATER RESOURCES RESEARCH, VOL. 45, W00B11, doi:10.1029/2008WR006839, 2009



# A novel method to estimate model uncertainty using machine learning techniques

Dimitri P. Solomatine[1,2] and Durga Lal Shrestha[1]

Received 15 January 2008; revised 14 October 2008; accepted 31 October 2008; published 20 January 2009.

[1] A novel method is presented for model uncertainty estimation using machine learning techniques and its application in rainfall runoff modeling. In this method, first, the probability distribution of the model error is estimated separately for different hydrological situations and second, the parameters characterizing this distribution are aggregated and used as output target values for building the training sets for the machine learning model. This latter model, being trained, encapsulates the information about the model error localized for different hydrological conditions in the past and is used to estimate the probability distribution of the model error for the new hydrological model runs. The M5 model tree is used as a machine learning model. The method is tested to estimate uncertainty of a conceptual rainfall runoff model of the Bagmati catchment in Nepal. In this paper the method is extended further to enable it to predict an approximation of the whole error distribution, and also the new results of comparing this method to other uncertainty estimation approaches are reported. It can be concluded that the method generates consistent, interpretable and improved model uncertainty estimates.

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

## 1. Introduction

[2] Uncertainty has always been inherent in water resources engineering and management. For example, in coastal and river flood defenses it was treated implicitly through conservative design rules, or explicitly by probabilistic characterization of meteorological events leading to extreme floods. Along with the recognition of the uncertainty of physical processes, the uncertainty analysis of models of these processes has become a popular research topic over the last decade. Rapid growth in computational power, the increased availability of distributed hydrological observations and an improved understanding of the physics and dynamics of water systems permit more complex and sophisticated models to be built. While these advances in principle lead to more accurate (less uncertain) models, at the same time if such complex (distributed) models with many parameters and data inputs are not parameterized properly or lack input data, they could be an inaccurate representation of reality. This prompts more studies into the model uncertainty of various types.

[3] The model errors are typically seen as the mismatch between the observed and the simulated system behavior. In the context of hydrological modeling they are unavoidable owing to the inherent uncertainties in the process. These uncertainties stem mainly from the four important sources [see, e.g., *Melching*, 1995; *Refsgaard and Storm*, 1996; *Gupta et al.*, 2005] and relate our understanding and measurement capabilities regarding the real-world system under study: (1) uncertainties in input data (e.g., precipitation and temperature); (2) uncertainties in data used for calibration, (e.g., output data such as streamflow); (3) uncertainties in model parameters; and (4) uncertainties due to imperfect model structure.

[4] Explicit recognition of uncertainty is not enough; in order to have this notion adopted by decision makers in water resources management, uncertainty should be properly estimated and communicated [*Pappenberger and Beven*, 2006]. The research community, however, has done quite a lot in moving toward the recognition of the necessity of complementing point forecasts of decision variables by the uncertainty estimates. Hence, there is a widening recognition of the necessity to (1) understand and identify of the sources of uncertainty; (2) quantify uncertainty; (3) evaluate the propagation of uncertainty through the models; and (4) find means to reduce uncertainty. A number of methods have been proposed in the literature to estimate model uncertainty. In general, these methods fall into six categories [see, e.g., *Montanari*, 2007; *Shrestha and Solomatine*, 2008]: (1) analytical methods [see, e.g., *Tung*, 1996], (2) approximation methods (e.g., first-order second moment method [*Melching*, 1992]), (3) simulation and sampling-based methods [e.g., *Kuczera and Parent*, 1998], (4) Bayesian methods (e.g., ''generalized likelihood uncertainty estimation'' (GLUE) by *Beven and Binley* [1992]), (5) methods based on the analysis of model errors [e.g., *Montanari and Brath*, 2004] and (6) methods based on fuzzy set theory [see, e.g., *Maskey et al.*, 2004].

[5] Most of the existing methods (e.g., categories 3 and 4) analyze the uncertainty of the uncertain input variables by propagating it through the deterministic model to the outputs, and hence requires the assumption of their distributions. Most of the approaches based on the analysis of the

[1]UNESCO-IHE Institute for Water Education, Delft, Netherlands.
[2]Water Resources Section, Delft University of Technology, Netherlands.


0043-1397/09/2008WR006839$09.00

model errors require certain assumptions regarding the residuals (e.g., normality and homoscedasticity). Obviously, the relevancy and accuracy of such approaches depend on the validity of these assumptions. The fuzzy theory-based approach requires knowledge of the membership function of the quantity subject to the uncertainty which could be very subjective. Furthermore, the majority of the uncertainty methods deal only with a single source of uncertainty. For instance, Monte Carlo-based methods analyze the propagation of uncertainty of parameters (measured by the probability distribution function, pdf) to the pdf of the output. Similar types of analysis are performed for the input or structural uncertainty independently. Note that the methods based on the analysis of the model errors typically compute the uncertainty of the ''optimal model'' (i.e., the model with the calibrated parameters and the fixed structure), and not of the ''class of models'' (i.e., a group of models with the same structure but parameterized differently) as, for example, Monte Carlo methods do.

[6] The contribution of various sources of errors to the model error is typically not known and, as pointed out by *Gupta et al.* [2005], disaggregation of errors into their source components is often difficult, particularly in hydrology where models are nonlinear and different sources of errors may interact to produce the measured deviation. Nevertheless, evaluating the contribution of different sources of uncertainty to the overall uncertainties in model prediction is important, for instance, for understanding where the greatest sources of uncertainties reside, and, therefore directing efforts toward these sources [*Brown and Heuvelink*, 2005]. In general, relatively few studies have been conducted to investigate the interaction between different sources of uncertainty and their contributions to the total model uncertainty [*Engeland et al.*, 2005; *Gupta et al.*, 2005]. For the decision-making process, especially in water resources management, it is more important to know the total model uncertainty accounting for all sources of uncertainty than the uncertainty resulting from individual sources. Recently *Shrestha and Solomatine* [2006, 2008] presented the basis of a novel method to estimate the uncertainty of the optimal model that takes into account all sources of errors without attempting to disaggregate the contribution given by their individual sources. The approach is referred to as an ''uncertainty estimation based on local errors and clustering'' (UNEEC). The method uses clustering and machine learning techniques to estimate the uncertainty of a process model by analyzing its residuals (errors). The distribution of model error is conditioned on the input and possible state variables of the model including the lagged variable of the observed response variable. Since the pdf of the model error is estimated through empirical distribution, it is not necessary to make any assumption about residuals. The method is computationally efficient, and therefore can be easily applied to computationally demanding process models. The method described here is based on the concept of optimality instead of equifinality as it analyzes the historical model residuals resulting from the optimal model (both in structure and parameters). If compared to earlier publications, in this paper the UNEEC method is extended further by introducing several quantiles of the error distribution, another case study is considered, and the results are also compared to those produced by several methods of uncertainty estimation.

## 2. Brief Overview of Machine Learning Techniques

[7] A machine learning technique is an algorithm that estimates (or induces) a hitherto unknown mapping (or dependency) between the inputs (predictors) and outputs (predictands) of a physical system from the available data [*Mitchell*, 1997]. As such a dependency (model) is discovered, it can be used to predict the future outputs of the system from the known input values. Machine learning techniques, based on observed data $\mathbf{D} = (\mathbf{X}, \mathbf{y}) = \{\mathbf{x}_t, y_t\}$, $t = 1, 2, \ldots, n$, try to identify (learn) the target function $f(\mathbf{x}_t, \mathbf{w})$ describing how the real system behaves, where $\mathbf{X}$ is the matrix ($\mathbf{x}$, vector) of the input data, $\mathbf{y}$ is the vector ($y$, scalar) of systems' response, $n$ is the number of data, $\mathbf{w}$ is the parameter vector of the function. Learning (or ''training'') here is the process of minimizing the difference between observed response $y$ and model response $\hat{y}$ through an optimization procedure. Model $f$ is often called a ''data-driven model.'' For a recent overview of data-driven modeling in water-related issues see, e.g., *Solomatine and Ostfeld* [2008].

[8] A review of the application of machine learning techniques to estimate the uncertainty of either process or machine learning-based rainfall runoff modeling can be found in the work by *Shrestha and Solomatine* [2008]. Sections 2.1 and 2.2 present a brief overview of the machine learning technique used in this study.

### 2.1. Piecewise Linear Regression Models: Model Trees

[9] A model tree (MT) is a hierarchical (or tree-like) modular model which has splitting rules in nonterminal nodes and linear regression functions at the leaves of the tree. The M5 algorithm constructs progressively hierarchical linear models that relate the input data to the corresponding values of output by dividing the input space. The input data is either associated with a leaf or is split into subsets and the same process is applied recursively to the subsets. The splitting criterion for the M5 algorithm is based on treating the standard deviation of the output values that reach a node as a measure of the error at that node, and calculating the expected reduction in error as a result of testing each attribute at that node. After examining all possible splits, M5 chooses the one that maximizes the expected error reduction. The MT is analogous to a piecewise linear function, learns efficiently, and can tackle tasks with very high dimensionality: up to hundreds of variables. As compared to other machine learning techniques, MT learning is fast and the results are interpretable. Details of MT and its M5 algorithm can be found elsewhere [see, e.g., *Witten and Frank*, 2000; *Solomatine and Dulal*, 2003].

### 2.2. Cluster Analysis

[10] The objective of cluster analysis is to partition a data set into subsets (clusters), so that the data in each subset share some common trait: often proximity according to some defined similarly measure. There are basically three types of clustering algorithms: exclusive, overlapping, and hierarchical. In the first case, the data are grouped in such a way that each data point belongs to a definite cluster and it

cannot be included in another cluster. The example of this type of clustering is "K-means" clustering. On the contrary, the second type, the overlapping clustering uses fuzzy sets [*Zadeh*, 1965] to cluster data, so that each data point belongs to several clusters with some degree of so-called fuzzy membership in the range [0, 1]. The best-known method of fuzzy clustering is the "fuzzy C-means" [*Bezdek*, 1981]. There are also nonfuzzy clustering methods leading to overlapping clusters, such as those based on the mixture of Gaussians. A hierarchical clustering algorithm finds successive clusters splitting the previously established clusters. Detailed description of the clustering algorithms can be found elsewhere [see, e.g., *Mitchell*, 1997].

## 3. Methodology

[11] The historical model residuals (errors) between the model prediction $\hat{y}$ and the observed data $y$ are the best available quantitative indicators of the discrepancy between the model and the real-world system or process, and they provide valuable information that can be used to assess the predictive uncertainty. The residuals and their distribution are often the functions of the model input variables and can be predicted by building separate model mapping of the input space to the model residuals or even the pdf of error. In other words, the idea here is to learn the relationship between the distribution of the model errors and the input variables and to use this information to predict the distribution of the model error when it predicts the output variable (e.g., runoff) in the future. It is assumed that the process model error is a proper indicator of model uncertainty and explained as follows.

[12] A deterministic model $M$ of a real-world system predicting a system output variable $y^*$ given input vector $\mathbf{x}(\mathbf{x} \in \mathbf{X})$ is considered. Let $y$ be the measurement of an unknown true value $y^*$, made with error $\varepsilon_y$. Various types of errors propagate through the model $M$ while predicting the observed output $y$ and have the following form:

$$y = y^* + \varepsilon_y = M(\mathbf{x}, \theta) + \varepsilon_s + \varepsilon_\theta + \varepsilon_x + \varepsilon_y \quad (1)$$

where $\theta$ is a vector of the model parameters values, $\varepsilon_s$, $\varepsilon_\theta$, and $\varepsilon_x$, are the errors associated with the model structure $M$, parameter $\theta$ and input vector $\mathbf{x}$, respectively. In most practical cases, it is difficult to estimate the error components of equation (1) unless some important assumptions are made. Thus, the different components that contribute to the total model error are generally treated as a single lumped variable and equation (1) can be reformulated as

$$y = \hat{y} + \varepsilon \quad (2)$$

where $\hat{y}$ is the model output and $\varepsilon$ is the total remaining (or residual) error. The UNEEC method estimates the uncertainty associated with the given model structure $M$, and parameter set $\theta$ by analyzing historical model residuals $\varepsilon$ which is an aggregate effect of all sources of error. Thus, the uncertainty estimated with the UNEEC method is consistent only for the given model structure and the parameter set $\theta$. It does not mean that the model structure and parameter uncertainty are ignored, but it is assumed that the uncertainty associated with the wrong model structure, inaccurate parameter values, and observational errors (if any) are manifested implicitly in the model residuals. This type of uncertainty analysis based on the model residuals is different from the classical uncertainty analysis methods where uncertainty of parameters, input data (presented by pdf) or plausible model structures are propagated to the pdf of the output.

[13] The UNEEC method starts by selecting the single best model structure from the plausible model structures in reproducing the observed behavior of the system. This ensures that the uncertainty associated with the wrong choice of the model structure is reduced as much as possible. Then it requires the prior identification of an optimal model parameter set, which can be achieved by calibration procedure aimed at minimizing some error function. This ensures minimizing the uncertainty associated with inaccurate estimate of parameter values. Observational errors can be reduced by the improved observational techniques and understanding of the characteristics of such errors. Of course, the model error $\varepsilon$ cannot be eliminated completely. The aim here is to build a model to estimate the pdf of the model error $\varepsilon$ conditioned to the input and/or state variables of the process model. Since the predictive uncertainty of the model output is more important than the pdf of the model error, the latter is then transferred to the predictive uncertainty by using information on the model predictions (described later in the section). Note that even when the optimal process model is used to produce a deterministic prediction, it does not, however, exclude the possibility of using some combination (ensemble) of "good" (but not optimal) models having the same structure but different in the values of the parameters, which could result from a Monte Carlo exercise.

[14] The UNEEC method consists of three main parts: (1) clustering; (2) estimation of the error probability distribution for clusters; and (3) building the overall model of the error probability distribution. These parts are described in this section (see Figure 1).

### 3.1. Clustering the Data

[15] Clustering of data is an important step of the UNEEC method. Its goal is to partition the data into several natural groups that can be interpreted. By data we understand here the vectors of some variable (input) space, and the input space here means not only input variables of the process model, but also all the relevant state variables which characterizes different mechanism of the modeled process, e.g., runoff generation process. The input data which belong to the same cluster will have similar characteristics and correspond to similar real-life situations. Furthermore, the distributions of the model errors within different clusters have different characteristics. This seems to be a strong assumption of the UNEEC method, which would be reasonable to test before applying it. In hydrological modeling this assumption seems to be quite natural: a hydrological model is often inaccurate in simulating extreme events (high consecutive rainfalls) which can be identified in one group by the process of clustering: resulting in high model residuals (wide error distribution). When data in each cluster belong to a certain "class" (in this case, a hydrological situation), local error models can be built: they will be more robust and accurate than the global model which is fitted on the whole data.

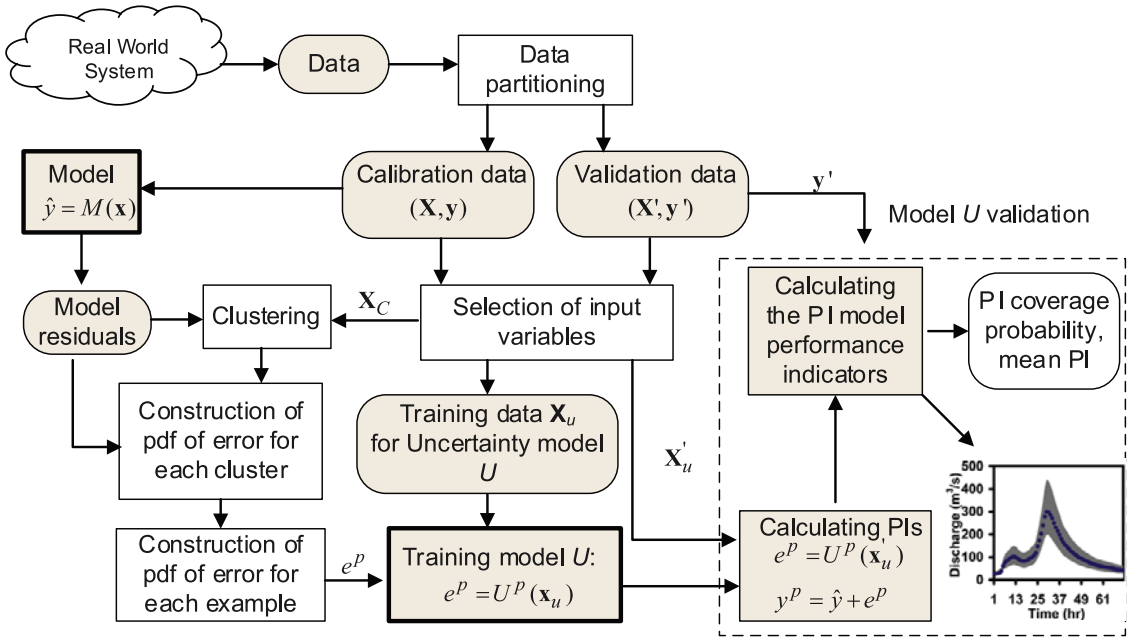


**Figure 1.** The generalized framework of the uncertainty estimation based on local errors and clustering (UNEEC) method. The UNEEC method has three steps: clustering, estimation of the probability distribution of the model error, and building model $U$ for probability distribution of the error. Once the model $U$ is trained in the calibration data set ($\mathbf{X}_u$), the model can be used to predict the probability distribution of the model error for the new data input.

[16] Before clustering the input space, the most relevant input variables should be selected from the data **D**. Since clustering is unsupervised learning where desired output of the system being studied is not known in advance, the selection of the variables in application to the process model are done by incorporating domain (hydrological) knowledge. The data set $\mathbf{X}_c$ is the matrix of input data constructed from the data **D** for partitioning into several natural groups (clusters). Additional analysis between model residuals and variables constituting **D** may be needed to choose the adequate variables for building $\mathbf{X}_c$ from **D**. Often $\mathbf{X}_c$ encompasses the variables of **D** and additional lagged vectors of some variables of **D** based on correlation and/or average mutual information (AMI) analysis.

### 3.2. Estimating Probability Distribution of the Process Model Error

[17] Typically the process model is nonlinear and contains many parameters. This will hinder the analytical estimation of the pdf of the model error. Thus the empirical pdf of the model error for each cluster is independently estimated by analyzing historical model residuals on the calibration data. In order to avoid a biased estimate of pdf or its quantiles of the model error, it is important to check if there is any overfitting by the process model on the calibration data. It is also possible to use leave-one-out cross validation [see, e.g., *Cawley et al.*, 2004] to overcome the bias estimate of the quantiles if the computational burden of running the process model is not prohibitive. However, such a cross-validation technique may be impractical in hydrological modeling because of computational load resulting from training multiple models. Another solution is to use a separate calibration sample data set that has not been used to calibrate the model, provided enough data are available. Note that when dealing with limited calibration data, the empirical distribution might be a very poor approximation of the theoretical distribution, so the reliability of such a method depends on the availability of data.

[18] Since the pdf of the model error is estimated for each cluster, it depends on the clustering method used. For example, in the case of K-means clustering where each instance of data belongs to only one cluster, the quantiles are taken from the empirical error distribution for each cluster independently. However, in the case of fuzzy clustering method (FCM) where each instance belongs to more than one cluster, and is associated with several membership functions, the computation of the quantiles should take this into account. The following expression gives the $p$th [0, 1] quantile of the model error for cluster $i$:

$$ec_i^p = \varepsilon_t \qquad t : \sum_{k=1}^{t} \mu_{i,k} < p \sum_{t=1}^{n} \mu_{i,t} \tag{3}$$

where $t$ is the maximum integer value running from unity that satisfies the above inequality, $\varepsilon_t$ is the residual associated with the $t$th data (data are sorted with respect to the associated residual), and $\mu_{i,t}$ is the membership function of the $t$th data to cluster $i$. This is not the only way of calculating quantiles for fuzzy clusters, and we tested several of them before choosing the one presented; unfortunately the space available does not allow for providing the details. An alternative would be to use the threshold of the membership degree in selecting the points to be included in sorting for each cluster.

### 3.3. Building a Model for Probability Distribution of the Process Model Error

[19] In order to estimate or predict the pdf of the process model error (or some quantiles of it) for the unseen input vectors, a machine learning model was built which will have predictive power after being trained using the calibration data. This model is referred to as an "uncertainty model" $U$ and can be built using several approaches [see *Shrestha and Solomatine*, 2008]. In this paper, the nonlinear regression method (e.g., model tree) is used.

[20] In order to train the model $U$, the pdf (or its quantiles) of the model error has to be estimated for the individual input data vector where the information about the model residuals is known. Since the empirical pdf of the model error for the clusters are already computed following the method described in section 3.2, the input data being the member of the clusters share this information of distribution. However, it is worth mentioning that the estimation of quantiles for the individual input data vector depends on the types of clustering techniques employed. For example, in K-means clustering the input data share the same information of pdf of the error for a particular cluster, thus ignoring the variation of the error distribution inside the cluster. However, there are other possibilities such as using distance function (distance between the centers of the cluster to the input vector) as a weight to vary error distribution.

[21] In the case of fuzzy clustering an approach that can be termed "fuzzy committee" is used to compute the quantiles for each individual input data vector and given by

$$e_t^p = \sum_{i=1}^{c} \mu_{i,t}^{2/m} ec_i^p \Big/ \sum_{i=1}^{c} \mu_{i,t}^{2/m} \tag{4}$$

where $e_t^p$ is the $p$th quantile of the pdf of the error for $t$th input data, $ec_i^p$ is the $p$th quantile of the pdf of the error for cluster $i$, and $m$ is the smoothing exponential coefficient. It should be noted that equation (4) is indeed a soft combination of the quantiles of each cluster depending upon the membership function values. This formulation has an additional advantage of smoothing of the quantiles across the input data. The smoothing can be increased with a higher value of $m$.

[22] Once the quantiles of the pdf of the model error for each example in the training data are obtained, machine learning model $U$ (that estimates the underlying functional relationships between the input vector $\mathbf{x}_u$ and the computed quantiles) is constructed:

$$e^p = U^p(\mathbf{x}_u;\, \theta^p) \tag{5}$$

where $\theta^p$ is the parameters vector of the model $U^p$ for the $p$th quantile. Please note that the calibration data set for model $U$ is ($\mathbf{X}_u$, $\mathbf{e}^p$), where $\mathbf{X}_u$ is input data constructed from $\mathbf{X}$ described below, $\mathbf{e}^p$ is a vector of $p$th quantiles. Thus model $U$, after being trained on input data $\mathbf{X}_u$, encapsulates the pdf of the model error and maps the input $\mathbf{x}_u$ to the pdf or quantiles of the process model error. It is worthwhile noting that the model $U$ can take any form, from linear to nonlinear regression function such as an artificial neural network (ANN). The choice of the model depends on the complexity of the problem to be handled and the availability of data. Once the model $U$ is trained on the calibration data $\mathbf{X}_u$, it can be employed to estimate the quantiles or the pdf of the model error for the new data input.

[23] As previously mentioned, the predicted quantiles of the pdf of the model error should be transferred to a more meaningful and understandable entity: predictive uncertainty of the model output. The quantile of the predictive uncertainty of the model output can be estimated as

$$y^p = \hat{y} + e^p \tag{6}$$

where $y^p$ is the $p$th quantile of the model output. One can see that equation (6) is the reformulation of equation (2). In order to estimate, for example, 90% prediction interval, it is necessary to build two models, $U^5$ and $U^{95}$, that will predict 5% and 95% quantiles, respectively.

[24] The issue here is to construct the calibration data $\mathbf{X}_u$ to build the regression model $U$. In most practical cases, data set $\mathbf{X}_u$ can be constructed from the set $\mathbf{D}$. Since the nature of models $M$ and $U$ is very different, additional analysis such as correlation or AMI analysis between the quantiles of the error pdf and the variables constituting $\mathbf{D}$ is needed to choose the adequate variables with proper lags. For example, if model $M$ is a conceptual hydrological model, it would typically use rainfall ($R_t$) and evapotranspiration ($E_t$) as input variables to simulate the output variable runoff ($Q_t$). However, the uncertainty model $U$, whose aim is to predict the pdf of the error of the simulated runoff, may be trained with the possible combination of rainfall and evapotranspiration (or effective rainfall), their past (lagged) values, the lagged values of runoff, and, possibly, their combinations.

### 3.4. Validation of the UNEEC Method

[25] The UNEEC method is validated by (1) measuring predictive capability of uncertainty model $U$ (e.g., using root-mean-squared error); (2) measuring the statistics of the uncertainty estimation; and (3) visualizing plots of prediction intervals with the observed hydrograph.

[26] Validation of model $U$, however, is not straightforward owing to the following. A validation procedure typically assumes that for a new situation (in our case, new values of rainfall, runoff, etc.) it is possible to compare the output variable value calculated by the model (say, 5% quantile calculated by the model $U^5$) to the corresponding "measured" value. However, it is impossible to measure the value of a quantile (in this case a 5% quantile) (whereas it is possible to validate the performance of model $M$ since it is possible to measure runoff). One may argue that if there is a large validation data set available, it is possible to generate the quantiles in the validation data set either (1) directly from the whole set or (2) from the clusters found in the validation set using the same procedure used when building model $U$. However, in option 1 the comparison would not be fair since model $U$ uses different (local, cluster-based) models of uncertainty, and in option 2 the clusters generated will be different from those generated during building of $U$ and hence comparison would not be proper either. We use an alternative method to validate the model $U$ proposed by *Shrestha et al.* [2006].

[27] Calibration data $\mathbf{X}_u$ was divided into two parts: training data, which constitutes major part of the calibration data (in this study, 67%), were selected randomly from the

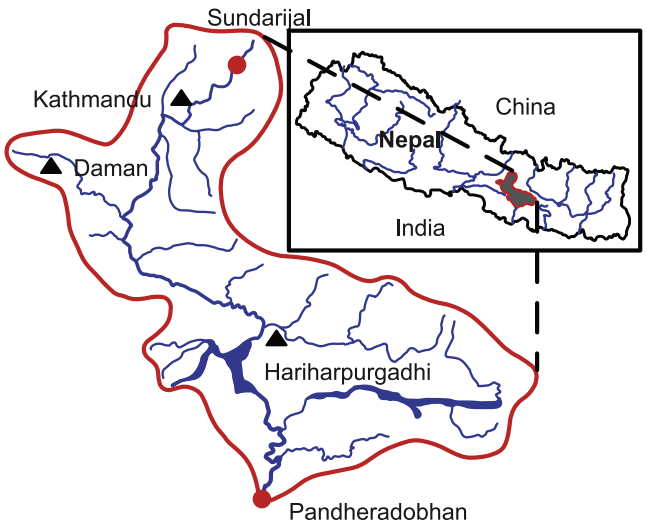


**Figure 2.** Location map of the Bagmati catchment considered in this study. Triangles denote the rainfall stations, and circles denote the discharge gauging stations. Discharge measured at Pandheradobhan is used for the analysis (adopted from *Solomatine et al.* [2008] with permission from Blackwell).

calibration pool without replacement. The training data are used to train the model *U*. The remaining data from the pool is the cross-validation (or test) data set which is used to perform ''intermediate'' tests of model *U* with the purpose to select its structure, parameters and the input variables. Random (or close to random) selection ensures statistical similarity of these data sets. Once the model *U* is tested on the cross-validation (test) data set, the model *U* with the best structure can be retrained on the full calibration data so as to increase its accuracy.

[28] Even though the quantiles of the model error pdf and, consequently, the quantiles of the model output are not observable in the validation data set, it is interesting to know if the distribution of the observed data fits the distribution predicted by model *U*. When predicting only two quantiles (or prediction interval, PI), this problem is equivalent to counting how many of the observed values are inside the prediction interval. In this case, validation of the UNEEC method can be done by evaluating two statistics: prediction interval coverage probability (PICP) and mean prediction interval (MPI). The former statistic measures the probability that the observed values lie within the estimated PIs. The latter statistic computes the average width of the PIs and gives an indication of the model uncertainty. In principle, these measures can also be computed individually for each cluster by training a classifier that would be able to attribute the new input vectors to one of the clusters or by computing the distance between the new input vectors and the cluster centroid vectors.

## 4. Application

### 4.1. Study Area

[29] The area selected for this study is the Bagmati catchment in Nepal. It lies in the central mountainous region of country and encompasses nearly 3700 km$^2$ within Nepal and reaches the Ganges River in India. The catchment area draining to the gauging station at Pandheradobhan is about 2900 km$^2$ (see Figure 2) and it covers the Kathmandu valley including the source of the Bagmati River at Shivapuri and surrounding Mahabharat mountain ranges. The catchment covers eight districts of Nepal and is a perennial water body of Kathmandu. The length of the main channel is about 195 km within Nepal and 134 km above the gauging station.

[30] Time series data of rainfall of three stations (Kathmandu, Hariharpurgadhi, and Daman) within the basin with a daily resolution for 8 years (1988 to 1995) was collected. The mean areal rainfall was calculated using Thiessen polygons. Although this method is not recommended for mountainous regions, the mean rainfall is consistent with the long-term average annual rainfall computed with the isohyetal method [*Chalise et al.*, 1996]. The long-term mean annual rainfall of the catchment is about 1500 mm with 90% of the rainfall occurring during the four months of the monsoon season (June to September). Daily flows were recorded from only one station at Pandheradobhan. Long-term mean annual discharge of the river at the station is 151 m$^3$/s but the annual discharge varied from 96.8 m$^3$/s in 1977 to 252.3 m$^3$/s in 1987 [*Department of Hydrology and Meteorology*, 1998]. The daily potential evapotranspiration was computed using the modified Penman method recommended by FAO [*Allen et al.*, 1998].

[31] Two thousand daily records from 1 January 1988 to 22 June 1993 were selected for calibration of the process model (in this study the HBV hydrological model, see section 4.2) and data from 23 June 1993 to 31 December 1995 was used for the validation (verification) of the process model. The first two months of calibration data were used as a warming-up period and hence excluded in the study. The separation of the 8 years of data into calibration and validation was done on the basis of previous studies. These data sets were used with all uncertainty analysis methods employed in this study.

### 4.2. Process Model: Conceptual Hydrological Model HBV

[32] A simplified version of the HBV-96 model (see Figure 3) was used as the process model to simulate river flows for the case study. The HBV model [*Bergström*, 1976] is a rainfall runoff model, which includes conceptual numerical descriptions of hydrological processes at the catchment scale. The model consists of subroutines for the meteorological interpolation, snow accumulation and melt, evapotranspiration estimation, a soil moisture accounting procedure, routines for runoff generation, and finally, a simple routing procedure between the subbasins and in lakes. It is possible to run the model separately for several subbasins and then add the contributions from all the subbasins. For the basins of considerable elevation range, subdivision into elevation zones can also be made. This subdivision is made for the snow and soil moisture routines only. Each elevation zone can further be divided into different vegetation zones (e.g., forested and nonforested areas).

[33] Input data are observations of precipitation, air temperature and estimates of potential evapotranspiration. The time step is usually one day, but it is possible to use shorter time steps. The evaporation values used are normally monthly averages although it is possible to use daily values. Air temperature data are used for calculations of snow

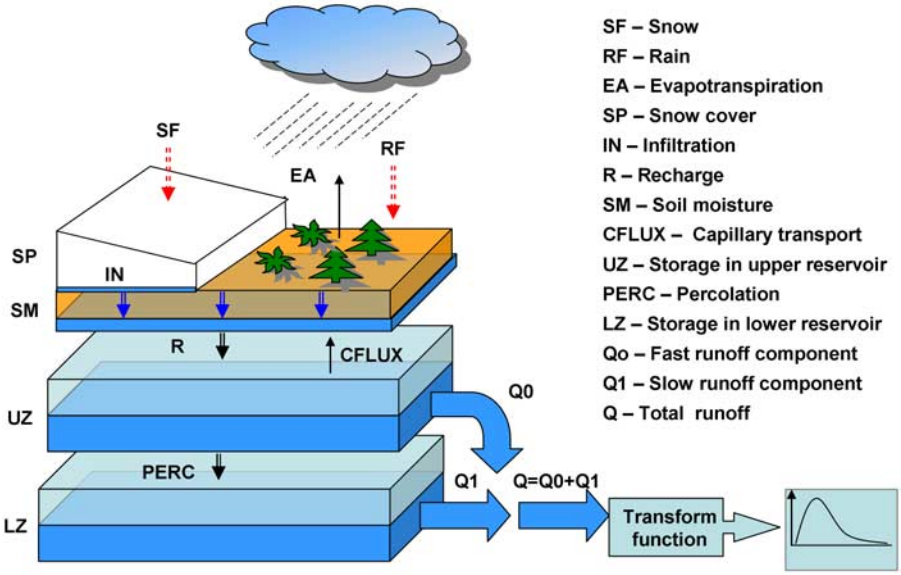


**Figure 3.** Schematic representation of the HBV-96 model (after *Lindström et al.* [1997]) with routines for snow, soil, and runoff response (reproduced from *Shrestha and Solomatine* [2008] with permission from the International Association for Hydraulic Research).

accumulation and melt. It can also be used to adjust potential evaporation when the temperature deviates from normal values, or to calculate the potential evaporation. The detailed description of the model can be found, e.g., in the work by *Lindström et al.* [1997].

## 5. Results and Discussions

### 5.1. Analysis of the Simulation Results

[34] A version of the HBV model with 13 parameters (4 parameters for snow, 4 for soil, and 5 for the response routine) was used. Since there is no snowfall in the catchment, the snow routine was excluded leaving only 9 parameters (see Table 1). The model was first calibrated using the global optimization routine, adaptive cluster covering algorithm, ACCO [*Solomatine et al.*, 1999] to find the best set of parameters, and subsequently the manual adjustments of the parameters was made. ACCO is a random search global optimization algorithm which was implemented in the global optimization tool, GLOBE (available at http://www.data-machine.com).

[35] The ranges of parameters values for automatic calibration were set on the basis of the ranges of calibrated values from the other model applications [e.g., *Braun and Renner*, 1992] and the hydrologic knowledge of the catchment. The ranges were extended when the solutions were found near the border of the parameter ranges and recalibration of the model was done with the extended ranges of the parameters. The model was calibrated using the *Nash and Sutcliffe* [1970] efficiency ($R_{eff}$) as the objective function. Finally, manual fine tuning of the parameters followed the automatic procedure by visual comparison of the observed and simulated hydrographs.

[36] The $R_{eff}$ value of 0.83 was obtained for the calibration period; this value corresponds to the root-mean-squared error (RMSE) value of 92.31 $m^3/s$. The model was subsequently validated by simulating the flows for the independent validation data set. The $R_{eff}$ was 0.87 for this period with the RMSE value of 127.6 $m^3/s$. Please note that the standard deviation of the observed discharge in the validation period is 54% higher than that in the calibration period and this apparently affects the increased performance in the validation period with respect to $R_{eff}$. The simulated and observed hydrographs along with rainfall and simulation error are shown in Figure 4.

**Table 1.** Ranges and Optimal Values of the HBV Model Parameters[a]

| Parameter | Description and Unit | Ranges | | Calibrated Value |
|---|---|---|---|---|
| | | Minimum | Maximum | |
| FC | Maximum soil moisture content (mm) | 50 | 550 | 450 |
| LP | Limit for potential evapotranspiration | 0.3 | 1 | 0.90 |
| ALFA | Response box parameter | 0 | 4 | 0.1339 |
| BETA | Exponential parameter in soil routine | 1 | 6 | 1.0604 |
| K | Recession coefficient for upper tank (day) | 0.05 | 0.5 | 0.3 |
| K4 | Recession coefficient for lower tank (day) | 0.01 | 0.3 | 0.04664 |
| PERC | Maximum flow from upper to lower tank (mm/day) | 0 | 8 | 7.5 |
| CFLUX | Maximum value of capillary flow (mm/day) | 0 | 1 | 0.0004 |
| MAXBAS | Transfer function parameter (day) | 1 | 3 | 2.02 |

[a]The uniform ranges of parameters are used for calibration of the Hydrologiska Byråns Vattenbalansmodell (HBV) model using the adaptive cluster covering algorithm and for analysis of the parameter uncertainty of the HBV model by the generalized likelihood uncertainty estimation method.

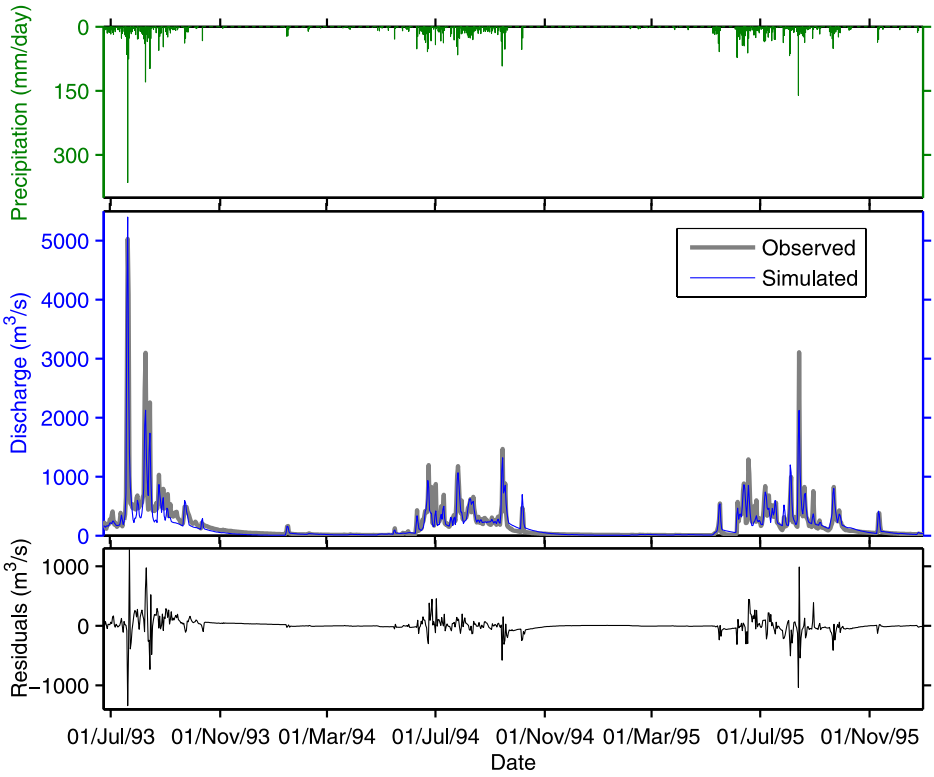


**Figure 4.** Simulated discharge for the Bagmati catchment for a validation period. The top and bottom show the precipitation and model errors, respectively, during the same period.

### 5.2. Analysis of the Model Residuals

[37] The analysis of the model residuals in the calibration period shows that the model residuals are highly correlated with the observed flows. Most of the high flows have relatively high residuals whereas the low flows have small residuals. The presence of heteroscedasticity in the residuals is observed as well. The normal probability plots of the residuals in the calibration and in the validation periods show that the residuals are not normally distributed. Kolmogorov-Smirnov and Lilliefors [*Lilliefors*, 1967] tests support this hypothesis. These tests of normality and homoscedasticity suggest that in order to provide a reliable estimation of the model uncertainty, the transformations of model residuals will be required if statistical methods are to be applied.

### 5.3. Clustering

[38] The clustering is performed using Fuzzy C-means algorithm based on the previous experience [*Shrestha and Solomatine*, 2006, 2008]. Selection of the input variables and the optimal number of clusters is discussed in sections 5.3.1 and 5.3.2.

#### 5.3.1. Selection of the Input Variables

[39] Several approaches have been reported in the literature [e.g., *Guyon and Elisseeff*, 2003; *Bowden et al.*, 2005] to select the model input variables; we follow the similar approaches. The input variables $\mathbf{X}_c$ used in clustering are constructed from the rainfall, the potential evapotranspiration, and the observed discharge. Several structures of the input data including the lagged variables were considered following the analysis of the correlation and AMI between the rainfall, runoff, and evapotranspiration with the model residuals. It appeared that the inclusion of the potential evapotranspiration does not improve the results obtained for the cross-validation data set, and it can be said that its inclusion would introduce "confusion" into the model. For example, during the low flow season (i.e., the dry season of April and May) there is very high potential evapotranspiration due to the hot weather in this period. The calibration of the hydrological model shows that the model captures the low flow reasonably well. However, this hydrological condition (low flow, negligible or zero rainfall, and very high potential evapotranspiration) is not identified as low flow season in clustering. So it was decided not to include the potential evapotranspiration as a separate variable, but rather to use the effective rainfall (rainfall minus evapotranspiration for rainfall greater than evapotranspiration, and zero otherwise). The following conventions are used throughout this manuscript while defining the input variables: $RE_{t-\tau}$, effective rainfall at time $t-\tau$; $Q_{t-\tau}$, discharge at time $t-\tau$; $\varepsilon_t$, model residuals at time $t$; $\tau$, lag time (0, 1, 2, ..., $\tau_{\max}$).

[40] After some trials with the different input combinations the following variables were selected for clustering: $RE_t$ and $Q_t$. The principle of parsimony was followed by avoiding the use of a large number of inputs. In addition to this the absolute values of model residuals were used, which explicitly forced the input data having similar values of the model residuals to be in one group. Since the rainfall and discharge have different units with different orders of magnitude and their values do not represent the same quantities, all input variables were normalized to the interval [0, 1]. This can prevent the model from being dominated

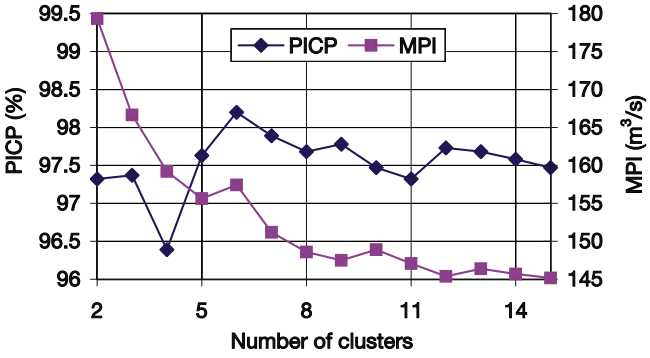


**Figure 5.** Sensitivity of the statistics of uncertainty measures with the number of clusters for the calibration period. Prediction interval coverage probability (PICP) measures the probability that the observed discharge values lie within the estimated prediction intervals. Mean prediction interval (MPI) is the average width of the prediction intervals.

by variables with large values, and is commonly used in machine learning techniques.

#### 5.3.2. Selection of the Number of Clusters

[41] An important issue in clustering is to identify the optimal number of clusters. Three validation indices are used to select the optimal number of clusters: partition index (SC), separation index (S), and Xie-Beni index (XB) [*Bensaid et al.*, 1996; *Xie and Beni*, 1991]. The comparison of the indices shows that the optimal number of clusters is 5.

[42] Figure 5 depicts the sensitivity of uncertainty measures to the number of clusters $c$. It is observed that MPI decreases with the increase of $c$. However, in the case of PICP there is no obvious pattern. The MPI fluctuates around the value 97.5% after $c = 5$. At $c = 5$, MPI and deviation of PICP from the desired confidence level (i.e., 90%) is smaller compared to those with $c = 6$. This value is also consistent with the previous research by *Shrestha and Solomatine* [2006] which has shown that this value is reasonable to represent the different situations related to the runoff generation process.

#### 5.3.3. Analysis of Clusters

[43] Figure 6 shows fuzzy clustering of the input examples. The input variables (effective rainfall $RE_t$ (Figure 6a) and discharge $Q_t$ (Figure 6b)) are on the abscissa, and the model residuals are on the ordinate. Please note that each input data belongs to all 5 clusters with different degrees of membership. However, on the plot the cluster which has the maximum membership function is shown. It is observed that there is a well-defined pattern of model residuals with the input variables such as $RE_t$ and $Q_t$. One can see that the high flows and the high (effective) rainfall generally have higher values of model residuals, and this is identified well by the clustering process (cluster C3). On the other hand, the conditions characterized by the low flows are also separated into one cluster (cluster C1) which has a very low value of the model residuals.

[44] Figure 7 presents the separation of the hydrograph with respect to the different clusters. One may notice that the majority of the flows have the highest membership in cluster C1 which can be interpreted as the base flow. The peaks and most of the high flows are attributed to cluster C3. Some high flows, especially in the recession part of the hydrograph, are attributed to cluster C5, because these examples have less or no rainfall and cannot be grouped into C3 which has high values of both flow and rainfall. It can be said that Fuzzy C-means clustering was able to identify the clusters corresponding to the various mechanisms of the runoff generation process, such as peak flow with high rainfall, high flow with no or less rainfall (recession part of the hydrograph) and base flow, etc.

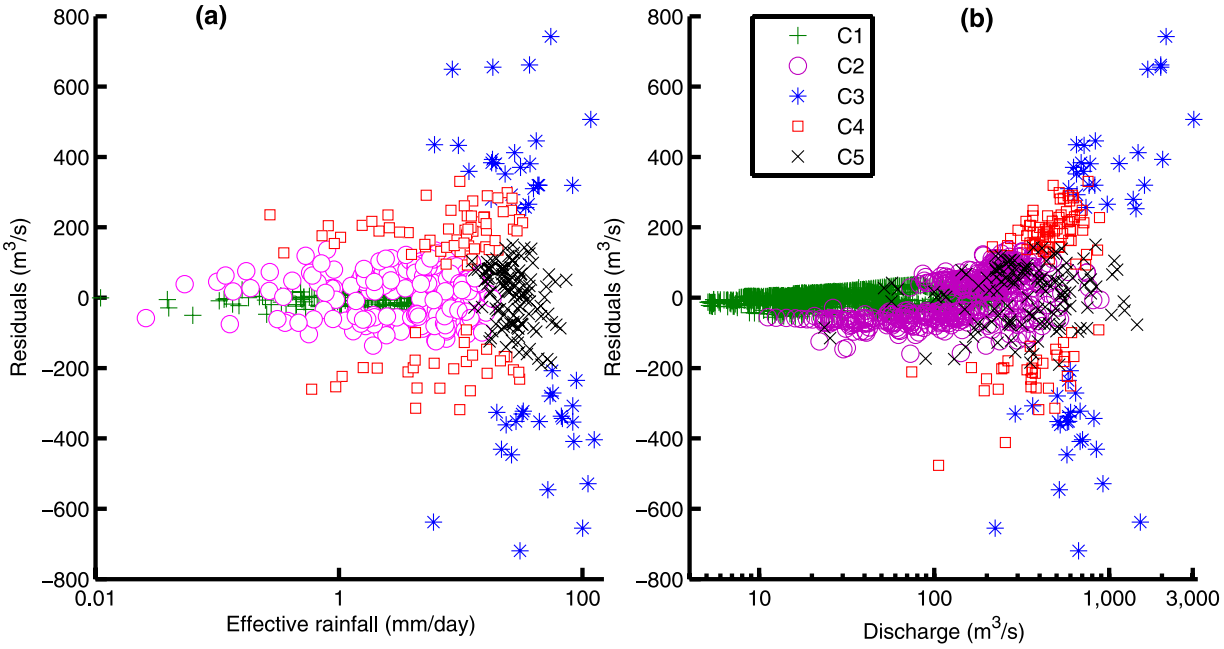


**Figure 6.** Fuzzy clustering of the input data in the calibration period (from 1 January 1988 to 22 June 1993) showing (a) effective rainfall and (b) discharge. The labels C1 through C5 indicate the cluster ID.

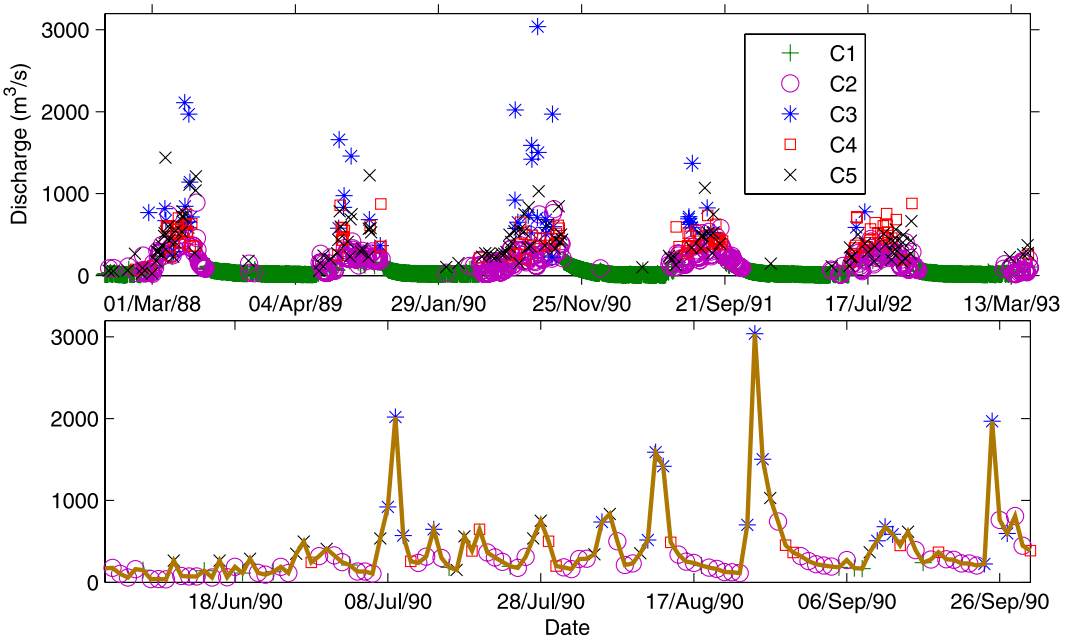


**Figure 7.** Fuzzy clustering of the input data in the calibration period (from 1 January 1988 to 22 June 1993). The bottom shows the enlarged view of the monsoon event of 1990.

### 5.4. Selection of the Input Variables for the Uncertainty Model U

[45] In order to select the most important influencing variables for the uncertainty model $U$, an approach similar to the one used for clustering was followed. Correlation analysis and AMI analysis between input variables $RE_t$ and $Q_t$ (including lags) and the quantiles of the model error were conduced. Figure 8a shows the correlation coefficient and AMI of $RE_t$ and its lagged variables up to 7 days, i.e., $RE_{t-1}$, $RE_{t-2}$, . . ., $RE_{t-7}$ with the 5% and 95% quantiles. It is observed that the variables $RE_t$ and $RE_{t-1}$ are strongly correlated with both the quantiles, so these two variables are included in the input vector $\mathbf{x}_u$.

[46] The correlation and AMI analyses between $Q_t$ and the quantiles of the model error are presented in Figure 8b. It is also observed that $Q_t$ and $Q_{t-1}$ are strongly correlated with the quantiles. Although the lag 2 variable $Q_{t-2}$ also has high correlation, only $Q_t$ and $Q_{t-1}$ were included in the input vector. The reason is that the flow $Q_t$ is highly autocorrelated and inclusion of too many lagged variables of $Q_t$ may lead to the redundancy of the model structure. Note that during the model application, $Q_t$ is not available and we use its approximation made by model $M$. Although the simulated $Q_t$ may bring additional uncertainty to model $U$, our experiments have shown that this approach resulted in the more accurate model $U$ (in terms of PICP and MPI).

### 5.5. Selection and Validation of the Uncertainty Model U

[47] M5 model tree (MT) was used as an uncertainty prediction model $U$. There are certain advantages of using MT if compared to other machine learning methods; it is

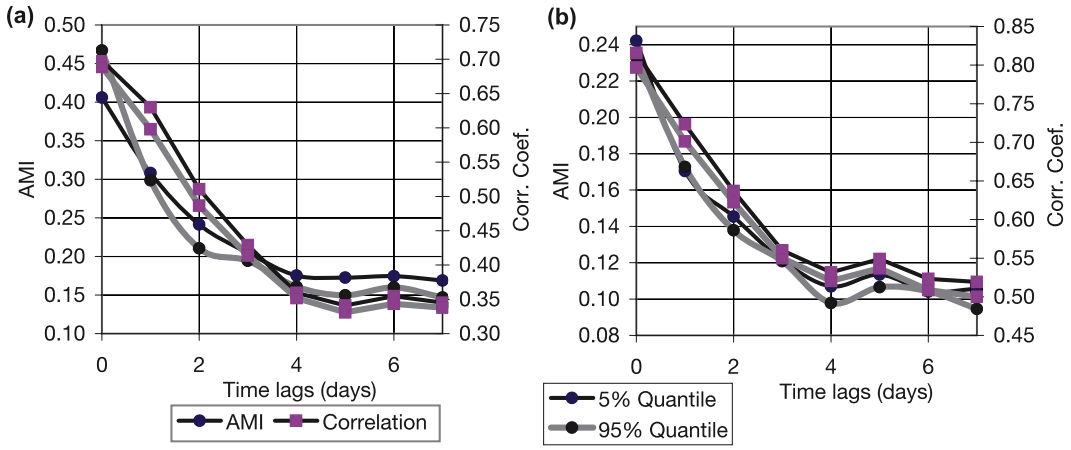


**Figure 8.** Average mutual information and correlation coefficient of 5% and 95% quantiles of the model errors with (a) effective rainfall and (b) discharge. The thin dark line shows the 5% quantile, and the thick gray line shows the 95% quantile.

**Table 2.** Mean and Standard Deviation Statistics of the 5% and 95% Quantiles of the Model Error, and the Performances of Uncertainty Prediction Models $U$[a]

| Data Set | Mean | SD | RMSE | $R_{eff}$ |
|---|---|---|---|---|
| Training | 73.60 (86.29) | 77.54 (73.60) | 28.44 (27.25) | 0.87 (0.86) |
| Cross validation | 67.36 (79.58) | 71.75 (67.70) | 26.69 (27.41) | 0.86 (0.84) |
| Calibration | 71.55 (84.07) | 75.72 (71.76) | 27.74 (26.70) | 0.87 (0.86) |

[a]Performances are RMSE and $R_{eff}$. The training and cross-validation data constitute 67% and 33%, respectively, of the calibration data (data used to calibrate hydrological model). Values in parentheses correspond to statistics of 95% quantiles of the model error and the performance of uncertainty prediction models $U^{95}$. SD, RMSE, and $R_{eff}$ are standard deviation of model errors, root-mean-squared error, and Nash-Sutcliffe efficiency between predicted and target values of the quantiles, respectively.

simple, easy, and fast to train. The results are interpretable, understandable, and reproducible. *Solomatine and Dulal* [2003] have shown that MT can be used as an alternative to ANN in rainfall runoff modeling. There is only one parameter in MT, the pruning factor (or, alternatively, the minimum number of data allowed in each linear model component), which controls the complexity of the model. The following shows the structure of the input data for the models $U$ to predict 5% and 95% quantiles:

$$e^5 = U^5(RE_t, RE_{t-1}, Q_t, Q_{t-1}; pf)$$
$$e^{95} = U^{95}(RE_t, RE_{t-1}, Q_t, Q_{t-1}; pf) \quad (7)$$

where $e^5$ and $e^{95}$ are the 5% and 95% quantiles of the model error, respectively, and $pf$ is the pruning factor. The following is the example of generated model tree by $U^5$ for $e^5$ with $pf = 4$:

$Q_t$ <= 73: LM1 (773)
$Q_t$ > 73:
| $Q_t$ <= 337:
| | $RE_t$ <= 8.73:
| | | $Q_{t-1}$ <= 255: LM2 (203)
| | | $Q_{t-1}$ > 255: LM3 (74)
| | $RE_t$ > 8.73: LM4 (77)
| $Q_t$ > 337: LM5 (173)

LM1: $e^5 = 11.7 + 1.73RE_t + 0.909RE_{t-1} + 0.293Q_{t-1} + 0.302Q_t$

LM2: $e^5 = 44.7 + 0.282RE_t + 0.15RE_{t-1} + 0.0114Q_{t-1} + 0.198Q_t$

LM3: $e^5 = 80.4 + 0.282RE_t + 1.6RE_{t-1} + 0.0114Q_{t-1} + 0.0405Q_t$

LM4: $e^5 = 44.3 + 3.22RE_t + 0.167RE_{t-1} + 0.153Q_{t-1} + 0.0242Q_t$

LM5: $e^5 = 146 + 1.7RE_t + 0.0706Q_{t-1} + 0.00895Q_t$

There are five linear models (namely LM1, LM2, LM3, LM4, and LM5) generated for various intervals of $Q_t$, $RE_t$, and $Q_{t-1}$. Note that numbers inside the parenthesis are the numbers of the data vectors in the subsets. Similar structures of the linear models are obtained for $e^{95}$. Table 2 shows the mean and standard deviation statistics of the generated quantiles of the model error and performance of $U^5$ and $U^{95}$ in training, cross-validation and calibration data sets.

[48] The mean and the standard deviation of the quantiles on the training and cross-validation data sets are very consistent with high variability of the original data. The performances of the uncertainty models $U$ as measured by RMSE and $R_{eff}$ for 5% and 95% quantiles are quite reasonable in both training and cross-validation data sets in spite of the high variability of the quantiles. The RMSE and $R_{eff}$ values of models $U$ on the calibration data set are quite consistent with those for the training and cross-validation sets and this ensures the predictability of the models $U$. Figure 9 depicts the scatterplot for the generated quantiles and predicted quantiles in the cross-validation data. It is observed that the both models are quite good for approximating the relationship between the input space variables and the quantiles of the model error.

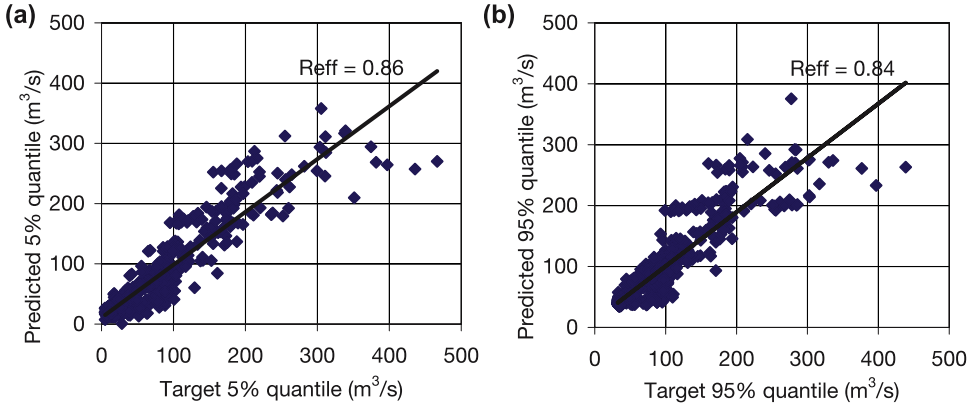


**Figure 9.** Scatterplot of the predicted and the target quantiles of the model errors for (a) 5% and (b) 95% quantiles in the cross-validation data (part of the calibration data).

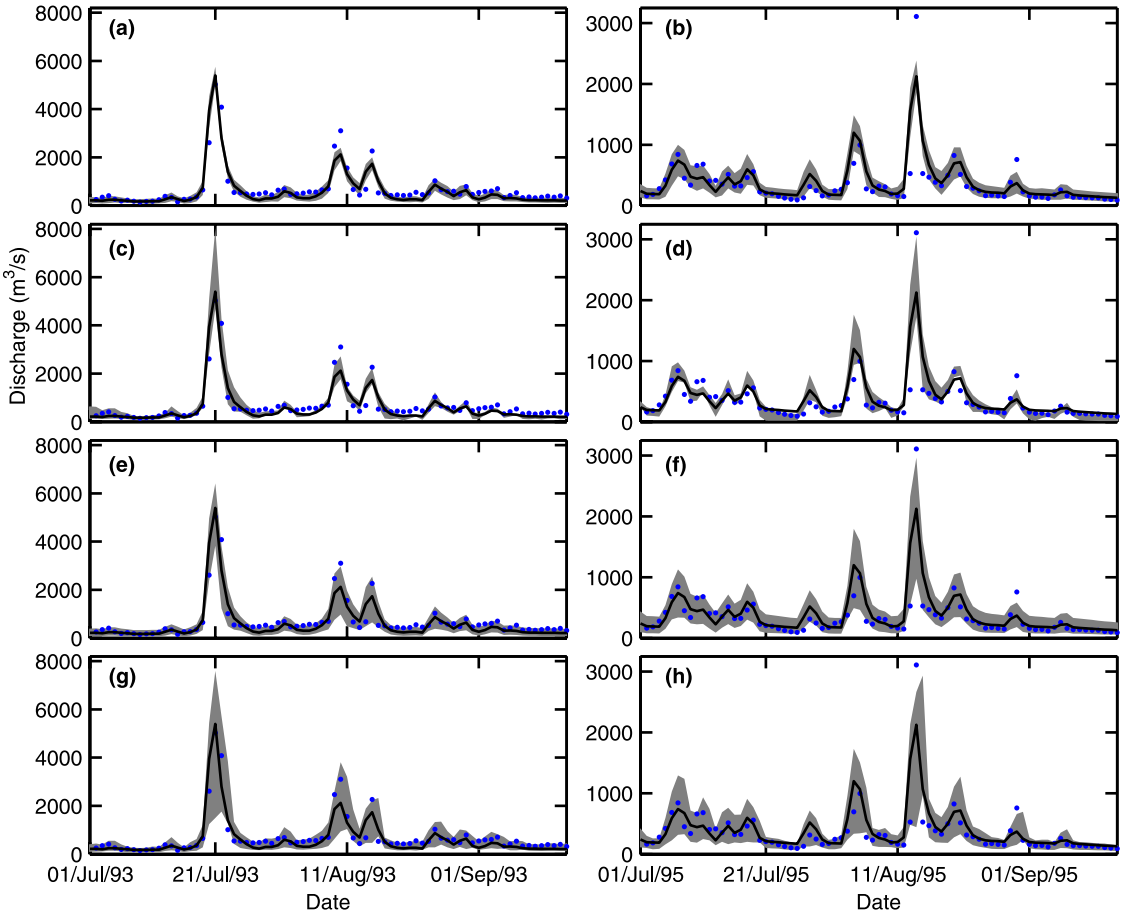


**Figure 10.** A comparison of 90% prediction bounds (darker shaded region) estimated with (a and b) the UNEEC method, (c and d) the generalized likelihood uncertainty estimation (GLUE) method, (e and f) the meta-Gaussian method, and (g and h) the quantile regression method in the validation period. Figures 10a, 10c, 10e, and 10g show the monsoon period of 1993, and Figures 10b, 10d, 10f, and 10h show the monsoon of 1995. The dots show the observed discharge; the line shows the simulated discharge.

### 5.6. Analysis of the Process Model Uncertainty

[49] 5% and 95% quantiles of the model residuals were computed for each of the clusters. Investigating the clusters with their centers and quantiles reveals that the clusters with high values of rainfall and runoff have the high values of the quantiles, while the clusters with low values of rainfall and runoff have the low values of the quantiles. Figure 10 shows the observed discharge, the 90% hydrograph prediction uncertainty and comparison to the other three methods. The details of the comparison follow later. Figure 10a highlights the flood event that occurred during the monsoon of 1993. Interestingly enough the HBV model captures the highest peak flow very well and consequently the peak flow is bracketed by the predicted PIs. Figure 10b focuses on another monsoon event during 1995. This event was underestimated by the HBV model. One can see that the estimated uncertainty bound fails to enclose the highest peak discharge of the 1995 monsoon.

[50] It was observed that 88.07% of the observed data points are enclosed within the computed PIs. 6.4% of the validation data points fall below the lower PI, whereas 5.53% data points fall above the upper PI. The average width of the uncertainty bounds; that is, MPI is 165 $m^3/s$. This value is reasonable if compared with the order of magnitude of the model error in the validation data. The further analysis reveals that the distribution of the observed discharge below the lower PI is relatively consistent with the observed discharge. As far as the upper PI is concerned, less data are outside in the low flow (range of 0–250 $m^3/s$). This means that the upper PIs are unnecessarily overestimated. However, the width of the upper PI in the intermediate flows (range of 250–750 $m^3/s$) is considerably narrower.

## 6. Comparison With the Other Methods

[51] In this section the results are compared with the widely used Monte Carlo method GLUE, the meta-Gaussian, and the QR method. Note that the comparison with GLUE is performed purely for illustration since the latter analyzes the parametric uncertainty and other methods: the uncertainty based on the analysis of the optimal (calibrated) model residuals.

### 6.1. GLUE Method

[52] Experiment for the GLUE method [*Beven and Binley*, 1992] is setup as follows: (1) prior feasible ranges of parameter values are set to be the same as those used in automatic calibration of the HBV model (see Table 1);

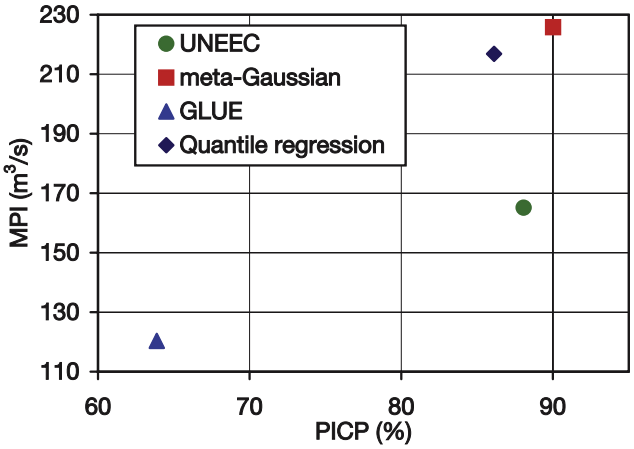


**Figure 11.** A comparison of the statistics of uncertainty estimated with the UNEEC, meta-Gaussian, GLUE, and quantile regression methods in the validation period. The statistics of the uncertainty are measured with PICP and MPI.

(2) likelihood measure was based on the Nash-Sutcliffe efficiency criterion used also by *Beven and Freer* [2001]; (3) rejection threshold values was set to 0.7; (4) the number of behavioral parameter sets was set to 25,000.

[53] The comparison results are reported in Figures 10c and 10d. One may notice the differences among the prediction bounds estimated by UNEEC and GLUE, but, again, these two techniques are different in nature. The key difference is that GLUE, as implemented here, accounts only for parameter uncertainty whereas the UNEEC method assumes the use of the optimal model and treats all other sources of uncertainty in an aggregated form. Note that the width of the prediction bounds obtained by the GLUE method varies with the rejection threshold and the likelihood measure to a great extent. For example, the lower rejection threshold produces relatively wider uncertainty bounds. (It is well known that the chosen value of the rejection threshold considerably influences the GLUE results).

[54] It can be noticed that only 63.9% of the observed discharge values in the validation data fall inside the 90% prediction bounds estimated by the GLUE method. As expected, the width of the prediction bounds is smaller than of those obtained with the other methods. The average value of the prediction bound width is 120.35 $m^3/s$ (see Figure 11). The further detailed analysis reveals that only 6.51% of the observed discharges are below lower PIs. The majority of the observed flows (29.61%) fall above the upper PIs.

### 6.2. Meta-Gaussian Method

[55] The meta-Gaussian approach [see, e.g., *Kelly and Krzysztofowicz*, 1997; *Montanari and Brath*, 2004] computes model uncertainty by estimating the pdf of the model error conditioned by the contemporary value of the simulated river flow. In this approach, both model residuals and simulated model outputs are transformed into the Gaussian domain by normal quantile transform (NQT). The meta-Gaussian approach is based on certain assumptions about the model residuals: they should be Gaussian and homoscedastic. However, in practical application, some of these basic assumptions are not satisfied (see Figure 6). Thus the model residuals were transformed according to the outline presented by *Montanari and Brath* [2004] to stabilize the variance of the model residuals. The description of the meta-Gaussian approach can be found in the work by *Montanari and Brath* [2004].

[56] The 90% prediction bounds estimated with the meta-Gaussian approach are shown in Figures 10e and 10f. Quite interestingly, it is found that 90% (more accurately, 90.02%) of the observed discharge values in validation data fall inside the estimated 90% prediction bounds (see also Figure 11). Further analysis of the results reveals that 3.7% of the observed data are below the lower PI whereas 6.3% of data are above the upper PI. However, one can see that the bounds width is consistently larger and the average width of the prediction bounds is 225.79 $m^3/s$ which is about 35% larger than that estimated by the UNEEC method.

### 6.3. Quantile Regression Techniques

[57] Quantile regression (QR) introduced by *Koenker and Bassett* [1978] is a statistical technique intended to estimate the conditional quantile functions. Just as the classical linear regression methods based on minimizing sums of squared residuals enable one to estimate the models for conditional mean of the response variable, given certain values of the predictor variables, QR methods offer a mechanism for building the models for the conditional median function, and the full range of other conditional quantile functions. Note that in opposition to UNEEC, QR is a linear regression method and is based on the whole data set, and does not include the local specialized nonlinear models as is done in UNEEC.

[58] QR method was used to compute the 5% and 95% error quantiles. The results are reported in Figures 10g and 10h. It is observed that 86.12% of the observed discharge values in validation data fall inside the estimated 90% prediction bounds (see Figure 11). Further analysis of the results reveals that 4.01% of the observed data are below the lower PI whereas 9.87% of data are above the upper PI. The average width of the prediction bounds is 216.86 $m^3/s$.

## 7. More Accurate Estimation of the Probability Distribution

[59] In the previous sections, only 90% prediction intervals, i.e., 5% and 95% quantiles, were estimated. In this section, the applicability of the method to derive the more accurate estimation of the pdf is demonstrated. Several quantiles (such as 2.5, 5, 10:10:90, 95, 97.5%) were computed for the calibration data after clustering of the input space. Then regression models are trained for each quantile independently:

$$e^p = U^p(RE_t, RE_{t-1}, Q_t, Q_{t-1}; pf) \tag{8}$$

where $p$ = 2.5, 5, 10:10:90, 95, 97.5%. In this experiment, a total of 13 regression models have been trained. The trained models $U^p$ were used to predict the quantiles on the validation data set. Note that since MTs are used as regression models, it takes only a couple of seconds to train a single model, so the computational cost to estimate the full distribution is not a major concern. However, this could be an issue for computationally extensive algorithms, such as

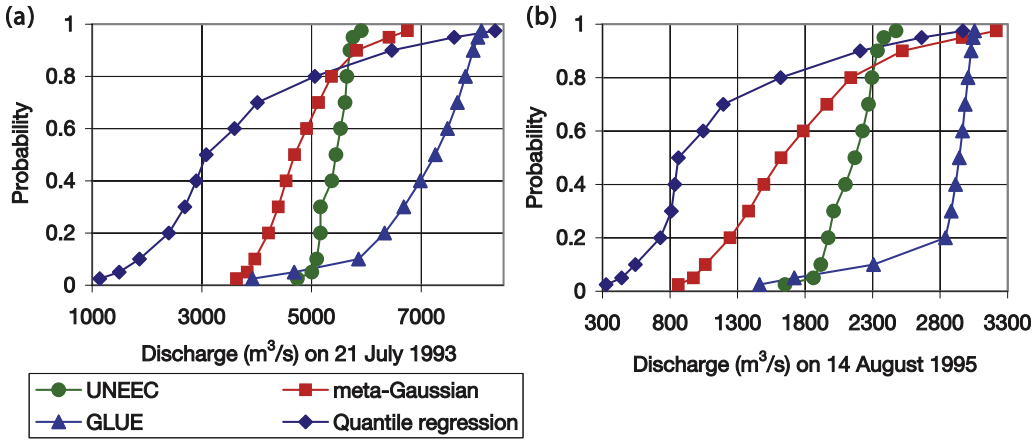


**Figure 12.** A comparison of estimation of cumulative probability distribution for peak discharges of the monsoon period of (a) 1993 and (b) 1995.

support vector machine or ANN with long records of input data.

[60] The cumulative probability distribution (cdf) for the peak discharge of the flood dated 21 July 1993 (the highest peak event of the 1993 monsoon shown in Figure 10) is shown in Figure 12a. The cdf computed from GLUE, meta-Gaussian, and QR methods are also presented for comparison. One can notice that the cdf computed from the UNEEC method is relatively steep. The GLUE and QR methods produce a comparatively flat cdf. For this particular flood event this means that the uncertainty is very high as estimated by GLUE and QR and it is lower as predicted by UNEEC. The meta-Gaussian method gives the intermediate results. Additional analysis of the cdf for the flood event of 14 August 1995 supports the finding that the uncertainty estimated with UNEEC is consistently lower for the flood events (Figure 12b).

[61] Further analysis was performed in order to compute PICPs and MPIs for various confidence levels ranging from 20% to 95%. The results are presented in Figure 13. As far as the PICP is concerned, the ideal would be to follow the thick gray line (Figure 13a). Points below this line indicate that less data are bracketed by the uncertainty bounds. On the other hand, if more data were enclosed in the uncertainty bounds, the PICP line would be above the ideal line. It can be seen that the PICPs computed with the meta-Gaussian and QR methods are very close to the desired confidence levels. GLUE produces consistently lower values of PICPs. In UNEEC more data are enclosed at lower values of the confidence levels.

[62] As far as the MPI is concerned, the GLUE method generates relatively narrower uncertainty bounds (Figure 13b). MPI estimated by UNEEC and meta-Gaussian methods are comparable at the lower values of the confidence levels. However, uncertainty bounds estimated by the meta-Gaussian method increase faster as compared to those obtained with UNEEC after 60% confidence levels. The MPI estimated with the QR method is similar to that obtained with the meta-Gaussian method.

## 8. Limitations and the Possible Extensions of the Method

[63] This section discusses some issues concerning the limitations and possible extensions of the UNEEC method. Since the machine learning technique is the core of the method, the problem of extrapolation, which is a well known problem of machine learning techniques, is also

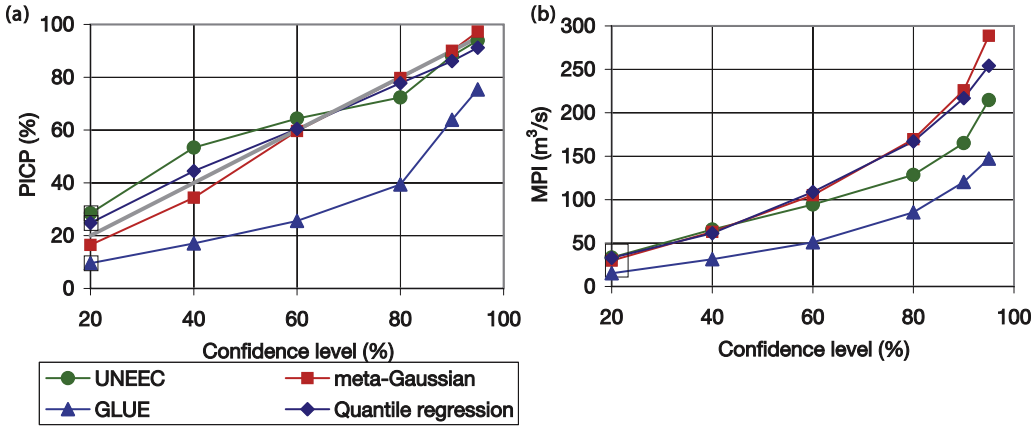


**Figure 13.** A comparison of the statistics of uncertainty measures. (a) PICP. In an ideal case, the plot between PICP and confidence level follows the thick gray line. (b) MPI for different values of the confidence level.

present. This means the results are reliable only within the boundaries of the domain where the training data are given. In order to avoid the problem of extrapolation, an attempt should be made to ensure that the training data includes all possible combinations of the events including the extremes, and this is not always possible since the extremes tend to be rather rare events.

[64] Another issue is that the reliability and accuracy of the uncertainty analysis depend on the accuracy of the regression models used, so attention should be given to this aspect. For example, one could use the cross-validation data set to check the accuracy of the model. Note, however, that in the case study considered the regression models were quite accurate.

[65] As mentioned before, this method relies on the concept of optimality instead of equifinality. If the assumption of the existence of a single ''best'' model is not valid, then all of the models that are considered ''good'' should be considered, as is done when the concept of equifinality is adopted. This can be achieved by combining such models in an ensemble, or by generating the metamodels of uncertainty for each possible combination of the model structure and parameter set, or even involving the uncertainty associated with the input data. Consequently, instead of having a single set of uncertainty bounds for each forecast, there will be a set of such bounds generated. The authors are currently exploring the applicability of the concept of equifinality and the results will be reported in due course.

## 9. Conclusion

[66] The assessment of the total (overall) model uncertainty of the optimal (calibrated) rainfall runoff models has received relatively little attention in the research literature. Most research typically focuses on one single source of uncertainty and the majority of the studies are oriented toward parametric uncertainty. There are many situations, however, when the contribution of the parameter uncertainty to the total uncertainty is smaller compared to the other types, for instance input (rainfall) uncertainty or structure uncertainty. The consequence of considering only parametric uncertainty is that the predictive uncertainty bounds estimated are too narrow and thus considerable part of the observed data fall outside these bounds. Furthermore, the disaggregation of the total model uncertainty into its source components is difficult, particularly in cases common to hydrology where the model is nonlinear and complex and different sources of uncertainty may interact.

[67] This paper presents an extension of the method (termed UNEEC) developed earlier [*Shrestha and Solomatine*, 2006, 2008] by making it possible to obtain full pdf of the model output. The method is tested on a new case study and compared to more uncertainty analysis techniques. Our approach assumes the model error (mismatch between the observed and modeled value) to be an indication of model uncertainty. The novelty of the approach is in the following: (1) no assumptions are made about the pdf of residuals; (2) building the uncertainty model specialized for a particular area of the state space (hydrometeorological condition) which is identified by fuzzy clustering; and (3) building the uncertainty model using machine learning techniques.

[68] The presented method was used to estimate the uncertainty (expressed as pdf) of a conceptual hydrological HBV model applied to the Bagmati catchment in Nepal. The comparisons with other uncertainty estimation methods (GLUE, meta-Gaussian, and quantile regression) are also reported. It is shown that the UNEEC method generates the consistent and interpretable uncertainty estimates, and this is an indicator that it can be a valuable tool for assessing uncertainty of various predictive models.

[69] A number of possible extensions of the method are suggested, and their feasibility and effectiveness are currently being explored.

[70] **Acknowledgment.** The work described in this manuscript was partly supported by the European Community's Sixth Framework Program through the grant to the budget of the Integrated Project FLOODsite, contract GOCE-CT-2004-505420.

## References

Allen, R. G., L. S. Pereira, D. Raes, and M. Smith (1998), *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*, FAO, Rome.

Bensaid, A. M., L. O. Hall, J. C. Bezdek, L. P. Clarke, M. L. Silbiger, J. A. Arrington, and R. F. Murtagh (1996), Validity-guided (re) clustering with applications to image segmentation, *IEEE Trans. Fuzzy Syst.*, *4*(2), 112–123, doi:10.1109/91.493905.

Bergström, S. (1976), Development and application of a conceptual runoff model for Scandinavian catchments, *Rep. RHO 7*, Swedish Meteorol. and Hydrol. Inst., Norrköping, Sweden.

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

Beven, K., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *J. Hydrol. Amsterdam*, *249*, 11–29, doi:10.1016/S0022-1694(01)00421-8.

Bezdek, J. C. (1981), *Pattern Recognition With Fuzzy Objective Function Algorithms*, Kluwer Acad., Norwell, Mass.

Bowden, G. J., G. C. Dandy, and H. R. Maier (2005), Input determination for neural network models in water resources applications. Part 1—Background and methodology, *J. Hydrol. Amsterdam*, *301*, 75–92, doi:10.1016/j.jhydrol.2004.06.021.

Braun, L. N., and C. B. Renner (1992), Application of a conceptual runoff model in different physiographic regions of Switzerland, *Hydrol. Sci. J.*, *37*(3), 217–232.

Brown, J. D., and G. B. M. Heuvelink (2005), Assessing uncertainty propagation through physically based models of soil water flow and solute transport, in *Encyclopedia of Hydrological Sciences*, edited by M. G. Anderson, pp. 1181–1195, John Wiley, New York.

Cawley, G. C., N. L. C. Talbot, R. J. Foxall, S. R. Dorling, and D. P. Mandic (2004), Heteroscedastic kernel ridge regression, *Neurocomputing*, *57*, 105–124, doi:10.1016/j.neucom.2004.01.005.

Chalise, S. R., M. L. Shrestha, K. B. Thapa, B. R. Shrestha, and B. Bajracharya (1996), *Climatic and Hydrological Atlas of Nepal*, Int. Cent. for Integrated Mt. Dev., Kathmandu, Nepal.

Department of Hydrology and Meteorology (1998), *Hydrological Records of Nepal, Stream Flow Summary*, Kathmandu, Nepal.

Engeland, K., C.-Y. Xu, and L. Gottschalk (2005), Assessing uncertainties in a conceptual water balance model using Bayesian methodology, *Hydrol. Sci. J.*, *50*(1), 45–63, doi:10.1623/hysj.50.1.45.56334.

Gupta, H. V., K. J. Beven, and T. Wagener (2005), Model calibration and uncertainty estimation, in *Encyclopedia of Hydrological Sciences*, edited by M. G. Anderson, pp. 2015–2031, John Wiley, New York.

Guyon, I., and A. Elisseeff (2003), An introduction to variable and feature selection, *J. Mach. Learning Res.*, *3*(7–8), 1157–1182, doi:10.1162/153244303322753616.

Kelly, K. S., and R. Krzysztofowicz (1997), A bivariate meta-Gaussian density for use in hydrology, *Stochastic Hydrol. Hydraul.*, *11*(1), 17–31, doi:10.1007/BF02428423.

Koenker, R., and G. Bassett (1978), Regression quantiles, *Econometrica*, *46*(1), 33–50, doi:10.2307/1913643.

Kuczera, G., and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The metropolis algorithm, *J. Hydrol. Amsterdam*, *211*, 69–85, doi:10.1016/S0022-1694(98)00198-X.

Lilliefors, H. W. (1967), On the Kolmogorov-Smirnov test for normality with mean and variance unknown, *J. Am. Stat. Assoc.*, *62*(318), 399–402, doi:10.2307/2283970.

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

Maskey, S., V. Guinot, and R. K. Price (2004), Treatment of precipitation uncertainty in rainfall-runoff modelling: A fuzzy set approach, *Adv. Water Resour.*, *27*, 889–898, doi:10.1016/j.advwatres.2004.07.001.

Melching, C. S. (1992), An improved first-order reliability approach for assessing uncertainties in hydrological modelling, *J. Hydrol. Amsterdam*, *132*, 157–177, doi:10.1016/0022-1694(92)90177-W.

Melching, C. S. (1995), Reliability estimation, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 69–118, Water Resour. Publ., Highlands Ranch, Colo.

Mitchell, T. M. (1997), Machine Learning, 414 pp., McGraw-Hill, Singapore.

Montanari, A. (2007), do we mean by uncertainty? The need for a consistent wording about uncertainty assessment in hydrology, *Hydrol. Processes*, *21*, 841–845, doi:10.1002/hyp.6623.

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

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part 1: A discussion of principles, *J. Hydrol. Amsterdam*, *10*, 282–290, doi:10.1016/0022-1694(70)90255-6.

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

Refsgaard, J. C., and B. Storm (1996), Construction, calibration and validation of hydrological models, in *Distributed Hydrological Modelling*, *Water Sci. Technol. Libr.*, vol. 22, edited by M. B. Abbott and J. C. Refsgaard, pp. 41–54, Kluwer Acad., Dordrecht, Netherlands.

Shrestha, D. L., and D. P. Solomatine (2006), Machine learning approaches for estimation of prediction interval for the model output, *Neural Networks*, *19*(2), 225–235, doi:10.1016/j.neunet.2006.01.012.

Shrestha, D. L., and D. P. Solomatine (2008), Data-driven approaches for estimating uncertainty in rainfall runoff modelling, *J. River Basin Manage.*, *6*(2), 109–122.

Shrestha, D. L., J. Rodriguez, R. K. Price, and D. P. Solomatine (2006), Assessing model prediction limits using fuzzy clustering and machine learning, in *Proceedings of the 7th International Conference on Hydroinformatics*, Res. Publ., Chennai, India.

Solomatine, D. P., and K. N. Dulal (2003), Model trees as an alternative to neural networks in rainfall—Runoff modelling, *Hydrol. Sci. J.*, *48*(3), 399–411, doi:10.1623/hysj.48.3.399.45291.

Solomatine, D. P., and A. Ostfeld (2008), Data-driven modelling: Some past experiences and new approaches, *J. Hydroinformatics*, *10*(1), 3–22, doi:10.2166/hydro.2008.015.

Solomatine, D. P., Y. Dibike, and N. Kukuric (1999), Automatic calibration of groundwater models using global optimization techniques, *Hydrol. Sci. J.*, *44*(6), 879–894.

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

Tung, Y.-K. (1996), Uncertainty and reliability analysis, in *Water Resources Handbook*, edited by L. W. Mays, pp. 7.1–7.65, McGraw-Hill, New York.

Witten, I. H., and E. Frank (2000), *Data Mining: Practical Machine Learning Tools With Java Implementations*, Morgan Kaufmann, San Francisco, Calif.

Xie, X. L., and G. Beni (1991), A validity measure for fuzzy clustering, *IEEE Trans. Pattern Anal. Mach. Intel.*, *13*(8), 841–847, doi:10.1109/34.85677.

Zadeh, L. A. (1965), Fuzzy sets, *Inf. Control*, *8*(3), 338–353, doi:10.1016/S0019-9958(65)90241-X.

D. L. Shrestha and D. P. Solomatine, UNESCO-IHE Institute for Water Education, P.O. Box 3015, NL-2601 DA Delft, Netherlands. (d.solomatine@unesco-ihe.org)