QUANTIFYING UNCERTAINTY OF FLOOD FORECASTING USING DATA DRIVEN MODELS

DURGA LAL SHRESTHA\(^1\) and DIMITRI P. SOLOMATINE\(^2\)

\(^1\) Research Fellow
\(^2\) Associate Professor
UNESCO-IHE Institute for Water Education
P.O. Box 3015, 2601 DA Delft, The Netherlands
(Tel: 31-15-2151-764, Fax: 31-15-2122-921, e-mail: d.shrestha@unesco-ihe.org)
(Tel: 31-15-2151-715, Fax: 31-15-2122-921, e-mail: d.solomatine@unesco-ihe.org)

Abstract

Flooding is a complex and inherently uncertain phenomenon. Consequently forecasts of it are inherently uncertain in nature due to various sources of uncertainty including model uncertainty, input uncertainty and parameter uncertainty. Several approaches have been reported to quantify and propagate uncertainty through flood forecasting models using probabilistic and fuzzy set theory based methods. In this paper, a method for quantifying uncertainty in flood forecasting using data driven modeling techniques is presented. Uncertainty of the model outputs is estimated in terms of prediction intervals. First, prediction intervals for training (calibration) data using clustering techniques to identify the distinguishable regions in input space with similar distributions of model errors were constructed. Secondly, a data driven model was built using the computed prediction intervals as targets. Thirdly, the constructed model was validated by estimating the prediction intervals for unseen (test) data. The method was tested on Sieve and Bagmati river basins using various data driven models. Preliminary results show that the method is superior to another method (uniform interval method) estimating prediction intervals.

Keywords: Flood forecasting, Uncertainty; Prediction intervals; Data driven models; Models trees; Neural networks; Lumped conceptual rainfall-runoff model

1. INTRODUCTION

Since flood is a complex and uncertain phenomenon, forecast of it is also inherently uncertain in nature. Incorporating the uncertainty estimation in the forecast can help the decision maker within the flood forecasting and warning system and thus enhance the reliability and credibility of both the forecasting and the warning system in real time flood management. Quantifying uncertainty within the flood forecasting enables an authority to set risk-based criteria for flood warning, provides information necessary for making rational decisions and offers potential for additional economic benefits of forecasts to every rational decision maker (Krzysztofowicz, 2001).

The uncertainties in flood forecasting have four important sources (e.g. Refsgaard and Storm, 1996): a) input data (e.g. precipitation, temperature, etc.) b) data used for calibration (e.g. runoff) c) model parameters d) imperfection of model structures. The error sources a) and b) depend on the quality of data, whereas c) and d) are more model specific.

Uncertainty in flood forecasting can be estimated using several approaches. The first approach is to forecast the model outputs probabilistically (Krzysztofowicz, 2001) which is
based on Bayesian framework. The second approach is to estimate uncertainty by analyzing the statistical properties of the model errors that occurred in reproducing the observed historical data. This approach has been widely used in statistical (Harnett and Murphy, 1980; Wonnacott and Wonnacott, 1996) and machine learning communities (Nix and Weigend, 1994; Heskes, 1997) for time series forecasting. In this approach uncertainty is estimated in terms of confidence intervals or prediction intervals. Recently Montanari and Brath (2004) proposed meta-Gaussian (Kelly and Krzysztofowicz, 1997) approach for the estimation of rainfall-runoff simulation uncertainty. It estimates the probability distribution of the model error conditioned by the value of the contemporary simulated river flows. The third approach is to use simulation and re-sampling based techniques such as Monte Carlo (Kuczera and Parent, 1998) and Generalised Likelihood Uncertainty Estimation (GLUE) (Beven and Binley, 1992) methods. The fourth approach is based on fuzzy theory based method (Maskey et al., 2004). This approach provides a non-probabilistic approach for modelling the kind of uncertainty associated with vagueness and imprecision.

The first and the third approaches mentioned require the prior distributions of the uncertain input parameters or data to be propagated through the model to the outputs. The second approach requires certain assumptions about the data and the errors, and, obviously, the relevancy and accuracy of such approach depends on the validity of these assumptions. The last approach requires knowledge of the membership function of the quantity subject to the uncertainty.

In this paper, we propose a new method for quantifying uncertainty in flood forecasting. Ideologically, it is close to the second approach i.e. uses the historical model residuals. Since the model residuals result from various sources of uncertainties as mentioned above, the method aims at estimating their aggregate effect without attempting to separate each individual contribution. In the situations when distributions of parameters or distributions of model errors are not known, we propose to use data driven modelling techniques to quantify uncertainty of flood forecasting in the form of prediction intervals (PIs). Data driven modelling induce a causal relationship between sets of input and output data in the form of a mathematical device, which in general is not related to the physics of the real world situation (Solomatine and Price, 2004). Various data driven techniques, such as linear regression, locally weighted regression (LWR), M5 model trees (MT) (Quinlan, 1992) and neural networks (ANN) can be used to predict the PIs using all or part of input vectors that are used to train (calibrate) the forecasting models. Upper and lower PIs are computed separately on the basis of model errors observed during training (calibration) of the forecasting models. Fuzzy C-means clustering (Bezdek, 1981) technique is used to determine zones of the input space having similar distributions of model errors. The PIs of each cluster are determined on the basis of empirical distributions of the errors associated with instances belonging to this cluster. The PIs of each instance are then determined according to grade of their membership in each cluster.

After training of the model that is used to predict the PIs, they are applied to estimate the PIs for the forecasts made by the four data driven models (linear regression, LWR, MT, and ANN) in Sieve (Italy) river basin and Simulation made by the two-tank lumped conceptual rainfall-runoff model (Sugawara, 1995) in Bagmati (Nepal) river basin (DHM, 2000). Their outputs are compared to uniform interval method (UIM) as benchmark method using prediction intervals coverage probability (PICP) and mean prediction interval (MPI).

The presented method does not need any information about the distribution of parameters and is not based on any assumptions about the distribution of errors. Most of the existing methods estimate upper and lower PIs, which are symmetrical about the point estimate. The presented method, however, computes upper and lower PIs independently and they are not
necessarily symmetrical. The method is robust and flexible and it can be applied to the outputs of any models regardless of its class or structure.

2. PREDICTION INTERVALS

Prediction intervals (PIs) are intervals within which a future unknown value (e.g., point forecast) is expected to lie with a prescribed probability. The upper and lower limits are called prediction limits or bounds. The prescribed probability is called confidence level. PIs should be distinguished from the Confidence Interval (CI). The CI deals with the accuracy of our estimate of the true regression (statisticians formulate this differently - as estimation of the mean value of the outputs). In contrast, the PI deals with the accuracy of our estimate with respect to the observed value. It should be noted that the PI is wider than the CI. The PIs are widely used and present a comprehensible form of uncertainty measure. The PIs can be easily derived from probability density function and even from variance of the quantity subject to uncertainty.

Most of the methods to construct a $100(1-\alpha)\%$ prediction limits for model output $y$ used in practice are essentially of the following form:

$$PL_U = y + z_{\alpha/2}\sigma_e$$
$$PL_L = y - z_{\alpha/2}\sigma_e$$

where $PL_U$ and $PL_L$ are the upper and lower prediction limits respectively, $z_{\alpha/2}$ is the value of the standard normal variate with cumulative probability level of $\alpha/2$, and $\sigma_e$ is the standard deviation of the errors. Computation of $\sigma_e^2$ is not so easy in most of practical situation especially for multivariate models containing many equations or depending on non-linear relationships. Nevertheless, bootstrap (Efron, and Tibshirani, 1993) techniques have been used in estimating uncertainty of neural networks models (e.g. Nix and Weigend, 1994; Heskes, 1997) to estimate $\sigma_e^2$. These techniques are generally computationally time consuming.

3. METHODOLOGY

Due to the various sources of uncertainty as mentioned in the section 1, it is not surprising that model outputs do not match the observed values well. The essence of the proposed method is that historical residuals between the model outputs and the corresponding observed data are the best available quantitative indicators of the discrepancy between the real-world process and its model. These residuals between the model outputs and observed values give the valuable information that can be used to assess the model uncertainty. These residuals are often function of the model input’s values and can be modelled.

The method identifies $n$ zones (regions or clusters) in input space reproducing different distributions of historical residuals. It is assumed that a region in input space that is associated with any particular cluster has similar residuals or residuals with similar distributions. Having identified the clusters, the PIs for the particular cluster are computed from empirical distributions of the historical residuals that belong to the cluster under consideration. For instance, in order to construct $100(1-\alpha)\%$ PIs, $\alpha/2*100$ and $(1-\alpha/2)*100$ percentile values are taken from empirical distributions of residuals for lower and upper PIs respectively. Typical value for $\alpha$ is 0.05, so this corresponds to 95% interval. If input space is divided into crisp clusters, e.g., by K-means clustering method, and each instance belongs to only one cluster, this computation is straightforward. However, in case of fuzzy clustering...
e.g., by C-means method where each instance belongs to more than one cluster and is associated with membership grades that indicates the degree with which the instance belong to a given cluster, the computation of the above percentiles should take this into account. The following expression gives the lower PI, denoted by $PIC_L$ for cluster $j$:

$$PIC_L(j) = e(i)$$

$$i: \sum_{k=1}^{i} \mu_j(k) < \alpha / 2 \sum_{k=1}^{N} \mu_j(k)$$

(2)

where $i$ is the maximum value of it that satisfies the above inequality, $e(i)$ is the error (residuals) associated with the instance $i$ (instances are sorted in ascending order), $\mu_j(i)$ is the membership grade of $i^{th}$ instance to cluster $j$. Similar type of expression can be obtained for the upper PI ($PIC_U$). This is illustrated in Fig. 1.

Fig. 1 Computation of prediction intervals in case of using fuzzy C-means clustering.

Once the PIs are computed for each cluster, the PIs for each instance in input space can be computed. Computation of the PIs for each instance also depends upon the clustering techniques employed. For example, if K-means clustering is employed, then the PI for each instance in the particular cluster is the same as that of the cluster. However, if fuzzy C-means clustering is employed, PIs are computed using the weighted mean of PIs of each cluster as:

$$PI_L(i) = \frac{c}{\sum_{j=1}^{c} \mu_j(i)} PIC_L(j)$$

(3)

where $PI_L(i)$ is the lower PI for $i^{th}$ instance, $c$ is number of clusters. Similar type of expression can be obtained for the upper PI.

Once PIs for each input instance is computed, it is possible to construct mapping function $f_{PI}$ that estimate underlying functional relationship between input $x$ and target or desired PIs as:

$$PI = f_{PI}(x)$$

(4)

where $PI$ without subscripts denotes the PI in general. Mapping function $f$ may be of any form from linear regression to non-linear functions such as ANN. In other words, given a set of $N$ data pairs \{x(n), PI(n)\}, $n = 1, \ldots, N$, we can train data driven model such as ANN to estimate underlying function relating $x$ to $PI$. It is to be noted that the target or desired variable might be either intervals or limits. Mapping function $f_{PI}$ will be referred to as the prediction intervals model (PIM) in this paper.
The performance of the PIM is evaluated using the prediction interval coverage probability (PICP). Even though target values of the PIs are not available in the validation data set, it is interesting to know whether the observed values (e.g. runoff) from the forecasting model are inside the estimated prediction bounds. By definition, the prediction bounds enclose the true but unknown value \((1 - \alpha)\%\) of times on average (typically 95%). The PICP is the probability that the target of an input vector lies within the computed prediction bounds and is estimated by the corresponding frequency as follows:

\[
\text{PICP} = \frac{\text{count}(t)}{V}
\]

where \(V\) is number of data in the validation set, \(t\) is target value for which prediction bounds are constructed. If the clustering techniques and the PIM are optimal, then the PICP value will be consistently close to the \((1 - \alpha)\%\).

Yet another performance measure for PIs can be introduced here as well. This is the mean prediction interval (MPI), which is an average prediction interval on validation data set and it measures the ability to enclose target values inside prediction bounds on cost of its width. MPI can be estimated by

\[
\text{MPI} = \frac{1}{V} \sum_{i=1}^{V} \left[ |P_{L}(i)| + |P_{U}(i)| \right]
\]

4. EXPERIMENTS AND RESULTS

4.1. STUDY AREA

The method was applied in the two-study areas: Sieve river basin and Bagmati river basin. The Sieve river basin is a tributary of the Arno river basin having length of 56 km and located in the Tuscany region of Italy. The Arno basin that includes the Sieve catchment was extensively studied and used as a case study for various physically-based hydrologic modelling exercises (e.g. Todini, 1996). Data-driven methods were used by Solomatine and Dulal (2003). It has a drainage area of 822 km\(^2\). The basin covers mostly hills and mountainous areas. The climate of the basin is temperate and humid. Three months of hourly discharge (Q) and precipitation (R) data were available (Dec. 1959 to Feb. 1960, 2160 data points) which represent various types of hydrological conditions.

Bagmati river basin is a medium sized river basin of Nepal with a catchment area of about 3700 km\(^2\) up to the Nepal-India border. It originates from the southern slope of Shivapuri lek (Mahabharat within Kathmandu valley) passes through the inner Mahabharat range and stretches to the plains of Terai (ending at Nepal-India border). The available data consists of 3 rainfall stations, 1 discharge station and 1 climatic station for air temperature. The daily time series data for rainfall, air temperature and discharge for 8 years (from 1988 to 1995, 2920 data points) were collected and analysed. The mean daily arial precipitation was calculated using a Thiessen polygon method and the daily evapotranspiration was computed using the modified Penman method recommended by FAO.

4.2. PROCEDURE

Whole data set was split into training (calibration) and validation parts. Same training data set used in forecasting or simulation model was also used in training PIM. Similarly, same validation data set used in forecasting model was again used in validating PIM. Furthermore clustering was employed to the same training data to construct PIs. For Sieve river basin validation data constitutes first 300 points starting 01/12/59, 07:00 to 13/12/59,18:00 while remaining 1854 data points starting from 13/12/59, 19:00 to 28/02/60, 00:00 constitutes the
training dataset. For Bagmati river basin, 1998 training data is from period 03/01/88 to 22/06/93, and 922 validation data - from period 23/06/93 to 31/12/95.

For Sieve river basin, various data driven models from on simple linear regression to non-linear regression such as ANN were fitted on training data set to forecast river flows $Q_{t+i}$ several hours ahead ($i=1, 3$ or $6$). The input variables to these forecasting models are based on the previous values of flow ($Q_{t-q}$) and previous values of effective rainfall ($Re_{t-Vr}$), where $\tau_{q}$ being between 0 and 2 hours and $\tau_{r}$ being between 0 and 5 hours depending upon lead time of forecasts. Selection of input variable to the forecasting models are based on the linear correlation coefficient and understanding of physical insight between the input variables and sought target or desired variable (see Table 1). Then the trained model was used to forecast on unseen data set i.e. the validation data set.

For Bagmati river basin, 2 tank rainfall-runoff conceptual model is calibrated using training data. The model has a total of 8 parameters, which were estimated by automatically optimizing using GLOBE (Solomatine, 2005).

Having model outputs for the training and validation data sets, the PIM was constructed to estimate the PIs on validation data set as follows. Fuzzy C-means clustering technique was first employed to construct the PIs for each cluster in the training data set and then to construct the PIs for each instance in the training data set. Note that the input to the PIM may constitute all or only some of the variables used in the forecasting model. Table 1 shows the description of input variables to the forecasting or simulation model, clustering, and the PIM. The target variables to the PIM are the upper and the lower prediction bounds. Alternately, the target variable may be the upper and the lower PIs, but in this case, predicted values from the forecasting model have to be added to obtain the prediction bounds. The PIs were constructed for 95% confidence level unless specified. The performance of the PIM is assessed by the PICM and the MPI introduced in Section 3.

The optimal number of clusters in Fuzzy C-means clustering was identified using the performance index $P$ (Amiri, 2003) and the Xie-Beni index $S$ (Xie, and Beni, 1991). Sometimes the minimal values of these indices are not so pronounced, so the experiments were repeated with different numbers of clusters.

Table 1: Input data to the models.

<table>
<thead>
<tr>
<th>Case study</th>
<th>Designation</th>
<th>Target Variables</th>
<th>Forecasting/Simulation Model</th>
<th>Clustering</th>
<th>Prediction Intervals Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sieve river basin</td>
<td>SieveQ1</td>
<td>$Q_{t+1}$</td>
<td>$Re_{t}$, $Re_{t-1}$, $Re_{t-2}$, $Re_{t-3}$, $Re_{t-4}$, $Re_{t-5}$, $Q_{t}$, $Q_{t-1}$, $Q_{t-2}$</td>
<td>Same as forecasting</td>
<td>Same as forecasting</td>
</tr>
<tr>
<td></td>
<td>SieveQ3</td>
<td>$Q_{t+3}$</td>
<td>$Re_{t}$, $Re_{t-1}$, $Re_{t-2}$, $Re_{t-3}$, $Q_{t}$, $Q_{t-1}$</td>
<td>Same as forecasting</td>
<td>Same as forecasting</td>
</tr>
<tr>
<td></td>
<td>SieveQ6</td>
<td>$Q_{t+6}$</td>
<td>$Re_{t}$, $Q_{t}$</td>
<td>Same as forecasting</td>
<td>Same as forecasting</td>
</tr>
<tr>
<td>Bagmati river basin</td>
<td>BagmatiQ</td>
<td>$Q_{t}$</td>
<td>$R_{t}$, $E_{t}$</td>
<td>$R_{t}$, $E_{t}$, $Q_{t}$</td>
<td>$R_{t}$, $R_{t-1}$, $E_{t}$, $E_{t-1}$, $Q_{t-1}$</td>
</tr>
</tbody>
</table>

4.3. RESULTS AND DISCUSSION

Fig. 2 shows clustering of input examples in Sieve river basin for 1 hour ahead prediction. Cluster 1 (C1) contains input examples with very high runoff, whereas cluster 5 (C5) is associated with very low values of runoff. Fuzzy C-means clustering is able to identify
clusters corresponding to various mechanism of runoff generation process such high flow, base flow etc.

Fig. 2 Clustering of input examples in training data set in SieveQ1 using fuzzy C-means clustering.

The comparison of performance (in terms of the PICP and the MPI) for SieveQ1 using various data driven models shows that performance of MT is better than that of the other models. The PICP for MT is 96.67% and MPI is 15.25 m$^3$/s. We also compared the results with uniform interval method (UIM). The UIM constructs single PI from empirical distributions of errors on the whole training data and applied uniformly to the validation data set. The PICP using the UIM for MT is 91.13% and MPI is 11.18 m$^3$/s. Simple model like linear regression also contains PICP close to 95%, but with very wide prediction bound.

Fig. 3 Computed prediction bounds for SieveQ1 using MT model. Black line corresponds to predicted runoff, while grey shading represents 95% prediction bounds. Small circles represent observed value out of computed prediction bounds.
Fig. 3 shows the computed prediction bounds for SieveQ1 on validation data set using MT. It is to be noticed that only few examples are outside the computed prediction bounds with reasonable width of prediction bound.

We also computed the PICP and the MPI for SieveQ3 and SieveQ6. Although PICP are very close to 95%, MPIs are very wide as compared to SieveQ1 which is not surprising. Fig. 4 shows the deviation of the PICPs from the desired confidence level for Sieve river basin using MT. The PIs were constructed for various confidence levels ranging from 10% to 99%. It is to be noticed from Fig. 4 that the PICPs are very close to desired confidence levels at values higher than 80%, and in practice the PIs are constructed around this value. Furthermore, it can be noted that the PIs are too narrow in most of cases as the PICPs are below the straight line. Such evidence was also reported by Chatfield (2000). In these cases the PIM underestimates uncertainty of the model outputs.

![PICP vs Confidence Level](image)

**Fig. 4** The PICP for different values of confidence level using MT model in Sieve river basin.

Fig. 5 presents fan chart showing MPIs for Sieve river basin in different forecast lead times for different degree of confidence levels. From figure it is evident that the width of the PIs increases with the increase of the confidence level. Moreover it is also illustrated that the width of PIs increases as forecast lead time increases. Thus uncertainty of model forecast increases as lead time increases.

![Fan Chart](image)

**Fig. 5** A fan chart showing mean prediction intervals for flow prediction in Sieve river basin upto 6 hours ahead. The darkest strip covers 10% probability and the lightest coves 99%.
Table 2 shows the cluster centers and computed PIs in Bagmati river basin. It is to be noticed that cluster C3 corresponds to very high values of rainfall, evapotranspiration and runoff, while C5 is to with very low values of these. Computed PIs are consistent with values of inputs as C3 has high value of PIs, while C5 has low value of PIs. This fact is also supported by Table 3, which shows very high degree of correlation of computed PIs with rainfall and runoff.

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Cluster centre</th>
<th>Lower prediction interval</th>
<th>Upper prediction interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( R_t ) (mm/day)</td>
<td>( E_t ) (mm/day)</td>
<td>( Q_t ) (m(^3)/s)</td>
</tr>
<tr>
<td>C1</td>
<td>0.96</td>
<td>0.78</td>
<td>57.98</td>
</tr>
<tr>
<td>C2</td>
<td>2.92</td>
<td>1.16</td>
<td>116.62</td>
</tr>
<tr>
<td>C3</td>
<td>28.72</td>
<td>1.30</td>
<td>515.51</td>
</tr>
<tr>
<td>C4</td>
<td>4.37</td>
<td>1.38</td>
<td>153.95</td>
</tr>
<tr>
<td>C5</td>
<td>0.38</td>
<td>0.51</td>
<td>27.96</td>
</tr>
</tbody>
</table>

Table 3. Correlation coefficient of computed prediction intervals of cluster with value of cluster centers

<table>
<thead>
<tr>
<th>Inputs</th>
<th>( R_t )</th>
<th>( E_t )</th>
<th>( Q_t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower prediction interval</td>
<td>-0.965</td>
<td>-0.736</td>
<td>-0.990</td>
</tr>
<tr>
<td>Upper prediction interval</td>
<td>0.927</td>
<td>0.799</td>
<td>0.964</td>
</tr>
</tbody>
</table>

Using 5 numbers of clusters, experiments were repeated to compute PIs for output of simulation model using different data driven models as conducted in Sieve river basin. The results show that MT is superior as compared to other data driven models. Only 5.1% of the validation data fall outside the computed 95% PIs (1.4% and 3.7% below lower and above the upper PI respectively) with MPI of 332.77 m\(^3\)/s. Furthermore, these points are quite uniformly distributed along the range of the observed runoff. The UIM gives PICP of 92.6% with MPI of 389.25. Due to lack of space, supporting figures are not presented here.

**5. CONCLUSIONS**

This paper presents a novel method for quantifying uncertainty of flood forecasting models using data driven modelling techniques. Uncertainty of flood forecasting is estimated in terms of the PIs. Computed PIs explicitly take into account all sources of uncertainty of the model outputs without attempting to separate the contribution given by the different sources of uncertainty. The methodology is independent of the structure or class of the forecasting model as it requires only the model outputs. Unlike existing techniques the methodology does not require the knowledge of prior distribution of parameters and does not rely on any assumptions about the distribution of errors. Most of the known techniques compute upper and lower prediction intervals symmetrically. The presented methodology computes upper and lower intervals independently.

The methodology has been applied in order to estimate the PIs of the outputs of flood forecasting models in two case studies: Sieve and Bagmati river basins. The result shows that the computed prediction bounds are satisfactory as the PICP are very close to desired degree of confidence level with reasonable width of the prediction bounds. We also compare our results with the UIM and the results demonstrate that the new method performs consistently better than the UIM.
ACKNOWLEDGMENT

The authors acknowledge the valuable input of Mr. Javier Rodriguez during his M.Sc. study at UNESCO-IHE Institute for Water Education in 2003. This work is partly supported by the EU project Integrated Flood Risk Analysis and Management Methodologies (FLOODsite), contract GOCE-CT-2004-505420.

REFERENCES


