QUANTIFYING UNCERTAINTY OF FLOOD FORECASTING USING DATA DRIVEN MODELS

DURGA LAL SHRESTHA[1] and DIMITRI P. SOLOMATINE[2]

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

Flooding is a complex and inherently uncertain phenomenon. Consequently flood forecasts are inherently uncertain in nature due to various sources of uncertainty including model uncertainty, input uncertainty and parameter uncertainty. 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. Several approaches have been reported to quantify and propagate uncertainty through flood forecasting models using probabilistic (e.g. Krzysztofowicz, 2001; Beven & Binely, 1992) and fuzzy set theory based methods (e.g. Maskey et al., 2004). The former approach requires the prior distributions of the uncertain input parameters or data to be propagated through the model to the outputs. While the later approach requires membership function of the quantity subject to the uncertainty.

In this paper we propose a new method for quantifying uncertainty in flood forecasting based on analyzing the statistical properties of the model errors. 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 from 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) and neural networks (ANN) can be used to predict the PIs using all or part of input instances 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 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 in Bagmati (Nepal) river basin. Their outputs are compared to *uniform interval method* (UIM) using *prediction intervals coverage probability* (PICP) and *mean prediction interval* (MPI). The UIM constructs single PI from empirical distributions of errors on the whole training data and applied uniformly to the validation data set.

The results show that 96.67% (PICP) of the validation data are inside the computed prediction bounds for 95% confidence level for 1 hour ahead prediction in Sieve river basin with average prediction bound of 15.25 m^3/s (MPI) using data driven model: MT. We also repeated experiments with 3 hour and 6 hour ahead prediction in Sieve catchment and obtained PICPs very close to 95%, but with very wide MPIs.

We also conducted experiments for the Bagmati river basin. The results show that 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 m^3/s.

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. The method is robust and flexible and it can be applied to the outputs of any models regardless of its class or structure.

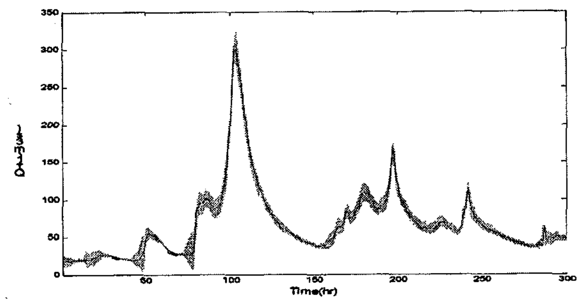

Fig. 1 Computed prediction bounds for SieveQ1 using MT model. Black line corresponds to predicted runoff, while grey shading represents 95% prediction bounds. Small circles represent observed value out of computed prediction bounds.

REFERENCES

Beven, K. J. & Binely, J., 1992. The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, 6, 279-298.

Krzysztofowicz, R., 2001. The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, 249, 2-9.

Maskey, S., Guinot, V. & Price R. K., 2004. Treatment of precipitation uncertainty in rainfall-runoff modelling, *Advances in Water Resources*, 27, 889-898.

Solomatine, D. P. & Price, R. K., 2004. Innovative Approached to Flood Forecasting Using Data Driven and Hybrid Modelling. In *Proc. of the 6th Int. Conference on Hydroinformatics*, Liong et al. (eds.), World Scientific Publishing Company, Singapore.