Journal of Hydrology 382 (2010) 20–33

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Uncertainty assessment and ensemble flood forecasting using bootstrap based artificial neural networks (BANNs)

Mukesh Kumar Tiwari, Chandranath Chatterjee *

Agricultural and Food Engineering Department, Indian Institute of Technology, Kharagpur, West Bengal 721 302, India

ARTICLE INFO

Article history:
Received 9 May 2009
Received in revised form 18 August 2009
Accepted 9 December 2009

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Taha Ouarda, Associate Editor

Keywords:
Uncertainty
Flood forecasting
Bootstrap
ANNs
Ensemble

SUMMARY

A reliable hydrologic prediction is essential for planning, designing and management activities of water resources. This study quantifies the parametric uncertainty involved in flood forecasting using artificial neural network (ANN) models. Hourly water level forecasting models are developed and uncertainty assessment is carried out for hourly water level forecasting. Hourly water level data of five upstream gauging stations of a large river basin are considered for hourly water level forecasting. The uncertainty associated with hourly flood forecast is investigated using the bootstrap based artificial neural networks (BANNs). Ensemble prediction is made by averaging the output of member bootstrapped neural networks. Results obtained indicate that BANN-hydrologic forecasting models with confidence bounds can improve their reliability for flood forecasts. It is illustrated that the confidence intervals based on BANNs are capable of quantifying the parametric uncertainty for short as well as for long lead time forecasts. Study shows the ensemble prediction is more consistent and reproducible. The study also analyzes the effect of length of training datasets and performance of split sample validation in BANNs modeling. The results illustrate that short length of training datasets with appropriate representation can perform similar to models with long length training datasets. The results also illustrate that the bootstrap technique is capable of solving the problems of over-fitting and underfitting during training of BANN models as the results without cross validation show similar performance compared to results obtained using cross validation technique.

Introduction

Flood forecasting is desirable to have sufficient lead time for taking appropriate flood prevention measures as well as evacuation actions. Hydrological models for flood forecasting are simplification of a more complex system and the natural processes are described with mathematical equations and the corresponding parameters are derived from observations and experience leading to uncertainty. The reliability of the model estimated discharge/water level is affected by three major sources of uncertainties (Bates and Townley, 1988): data uncertainty (quality and representativeness of data), model structure uncertainty (ability of the model to describe the catchment's response), and parameter uncertainty (adequate values of model parameters). The quantification of these uncertainties is important for practical decision making. It is difficult to assess the data uncertainty because the magnitude of data errors is often unknown and any attempt to model these deviations is ultimately based on a guess.

Black box models in the form of artificial neural networks (ANNs) have gained momentum in last few decades for flood forecasting. The ability of ANN in mapping complex nonlinear rainfall runoff relationship has increased the number of application in rainfall runoff modeling and river discharge forecasting (Agarwal and Singh, 2004; Campolo et al., 2003; Jain and Srinivasulu, 2004). Substantial literature on ANN have been reported in ASCE (2000a,b), Dawson and Wilby (2001) and Maier and Dandy (2000). The quantification of the uncertainty associated with the results provided by ANN models is essential for their confident and reliable use in practice. To quantify the parametric uncertainty in the ANN computation bootstrap procedure is employed which is based on resampling technique (Efron and Tibshirani, 1993; Tibshirani, 1996; Twomey and Smith, 1998; Zio, 2006).

The bootstrap (Efron, 1979) is a computational procedure that uses intensive resampling with replacement, in order to reduce uncertainty (Efron and Tibshirani, 1993). In addition, it is the simplest approach since it does not require complex computations of derivatives and Hessian-matrix inversion involved in linear methods or the Monte Carlo solutions of the integrals involved in the Bayesian approach (Dybowski and Roberts, 2000). Bootstrap technique has been used successfully in hydrological modeling and it is

* Corresponding author. Tel.: +91 3222 283158.
E-mail addresses: mukesh_k_tiwari@yahoo.com (M.K. Tiwari), cchatterjee@agfe.iitkgp.ernet.in (C. Chatterjee).

0022-1694/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2009.12.013

the topic of current research. Documented applications of bootstrap ranges from estimating means, confidence intervals, parameter uncertainties and network design techniques (e.g. Cover and Unny, 1986; Woo, 1989; Lall and Sharma, 1996; Sharma et al., 1997; Tasker and Dunne, 1997). Bootstrap technique has also been used in artificial neural network model development. Abrahart (2003) employed bootstrap technique to continuously sample the input space in the context of rainfall runoff modeling and reported that it offered marginal improvement in terms of greater accuracies and better global generalizations. He suggested further involving bootstrap technique for estimating confidence interval of the outputs. Jeong and Kim (2005) used ensemble neural network (ENN) using bootstrap technique to simulate monthly rainfall–runoff. They concluded that ENN is less sensitive to the input variable selection and the number of hidden nodes than the single neural network (SNN). Jia and Culver (2006) used the bootstrap technique to estimate the generalization errors of neural networks with different structures and to construct the confidence intervals for synthetic flow prediction with a small data sample. Han et al. (2007) studied the uncertainties involved in real-time prediction in using an ANN model. They proposed a method to understand the uncertainty in ANN hydrologic models with the heuristic that the distance between the input vector at prediction and all the training data provide a valuable indication on how well the prediction would be. They concluded that for long term predictions, the ANN showed superior performance but that was only probabilistic depending on how the calibration and test events were arranged. However their method did not quantify the uncertainty of the model parameters or the predictions. Srivastav et al. (2007) proposed a method of uncertainty analysis for ANN hydrological models which was based on bootstrap technique. They developed an ANN model for forecasting the river flow at 1 h lead time and the results revealed that the proposed method of uncertainty analysis is very efficient and can be applied to an ANN based hydrological model.

To the best of our knowledge no studies have been reported in the hydrologic literature that have used bootstrap based ANNs for uncertainty assessment and for making ensemble probabilistic forecast with confidence interval for multi step ahead (different lead times) forecast. Earlier neural bootstrap studies have also revealed that variation in forecasts due to changes in structure or architecture are small in comparison to those that arise from sample splitting (LeBaron and Weigend, 1998). In this study the uncertainty in neural network based forecast and the estimation of confidence intervals that arise from parametric uncertainty have been investigated. An attempt is made to quantify the uncertainty involved in flood forecasting and to make ensemble forecasts for 1–10 h lead time using bootstrap based ANNs (BANNs) in Mahanadi River basin where flood forecasting is a critical issue. Further, an attempt is also made to compare the performance of BANN models for different length of training datasets and to check the utility of cross validation technique with BANNs.

Theory

Artificial neural networks

Artificial neural networks are information processing systems composed of simple processing elements (nodes) linked by weighted synaptic connections (Muller and Reinhardt, 1991; Rumelhart and McClelland, 1986). They reconstruct the complex nonlinear input/output relations by combining multiple simple functions, by analogy with the functioning of the human brain.

In all generality, let us consider an ANN to be trained for performing the task of nonlinear regression, i.e. estimating the underlying nonlinear relationship existing between a vector of input variables x and an output target y, assumed mono-dimensional for simplicity of illustration, on the basis of a finite set of input/output data examples:

$$T_n \equiv \{(x_n, y_n), \quad n = 1, 2, \ldots, n_p\} \tag{1}$$

It can be assumed that the target y is related to the input vector x by an unknown nonlinear deterministic function $\mu_y(x)$ corrupted by a Gaussian white noise $\varepsilon(x)$, viz.

$$y = \mu_y(x) + \varepsilon(x) \quad \varepsilon(x) = N(0, \sigma_\varepsilon^2(x)) \tag{2}$$

The objective of the regression task is to estimate $\mu_y(x)$ by means of a regression function $f(x_n; \widehat{w})$, dependent on the set of synaptic weights $\widehat{w}$ to be properly determined on the basis of the available set T. The parameters are usually determined by a training procedure which aims at minimizing the quadratic error function:

$$E = \frac{1}{2n_p} \sum_{n=1}^{n_p} (\widehat{y}_n - y_n)^2 \tag{3}$$

where $\widehat{y}_n = f(x_n; \widehat{w})$ is the network output corresponding to input x_n. If the network architecture and training parameters are suitably chosen and the minimization done to determine the weights values is successful, the obtained function $f(x_n; \widehat{w})$ gives a good estimate of the unknown, true function $\mu_y(x)$. Indeed, it is possible to show that in the ideal case of an infinite training data set and perfect minimization algorithm, a neural network trained to minimize the error function in Eq. (3) provides a function f which performs a mapping from the input x into the expected value of the target y, i.e. the true deterministic function $E[y|x] = \mu_y(x)$ (Bishop, 1995). In other words, the network averages over the noise on the data and discovers the underlying deterministic generator. Unfortunately, all the training datasets are finite and there is no guarantee that the selected minimization algorithm can achieve the global minimum.

Bootstrapped artificial neural networks (BANNs)

In practical regression problems, it is crucial to associate the corresponding measures of confidence to the estimates obtained. In the case of ANN estimation, this requires that the various sources of uncertainty affecting the determination of the weights $\widehat{w}$ be properly accounted for (Tibshirani, 1996; Twomey and Smith, 1998; Dybowski and Roberts, 2000). In this respect, for what concerns the estimate $f(x_n; \widehat{w})$ of $\mu_y(x)$, it must be considered that, from a probabilistic point of view, the data set $T \equiv \{(x_n, y_n), n = 1, 2, \ldots, n_p\}$ used for training the network is only one of an infinite number of possible data sets which may be drawn within the given input volume V_x and from the underlying statistical error distribution. In other words, this variability in the training data set is due to the variability in the sampling of the input vectors x_n, $n = 1, 2, \ldots, n_p$ and in the random fluctuation of the corresponding target output y_n. Each possible training set T can give rise to a different set of network weights $\widehat{w}$. Correspondingly, there is a distribution of regression functions $f(x_n; \widehat{w})$ with variance (with respect to the training set T):

$$\sigma_f^2(x) = E\{[f(x_n; \hat{w}) - E[f(x_n; \hat{w})]]^2\} \tag{4}$$

Since in practice a neural network structure is not a perfect algorithm, it systematically under/over estimates the correct result, i.e. the expected value $E[f(x_n; \widehat{w})]$ is not equal to the true underlying deterministic function, $\mu_y(x)$, their difference being the so-called *bias*. Of course, the bias would be zero in the case of a perfect neural network. The variance with respect to all possible training data sets of the regression error $f(x; \widehat{w}) - \mu_y(x)$ is:

$$\begin{aligned} E\{[f(x;\widehat{w}) - \mu_y(x)]^2\} &= E\{[f(x;\widehat{w}) - E[f(x;\widehat{w})] + E[f(x;\widehat{w})] - \mu_y(x)]^2\} \\ &= E\{[f(x;\widehat{w}) - E[f(x;\widehat{w})]]^2 + [E[f(x;\widehat{w})] - \mu_y(x)]^2 \\ &\quad + 2[f(x;\widehat{w}) - E[f(x;\widehat{w})]] \cdot [E[f(x;\widehat{w})] - \mu_y(x)]\} \\ &= E\{[f(x;\widehat{w}) - E[f(x;\widehat{w})]]^2\} + \{E[f(x;\widehat{w})] - \mu_y(x)\}^2 \end{aligned} \tag{5}$$

where the first term is the variance (Eq. (4)) of the distribution of the regression function $f(x;\widehat{w})$, whereas the second term is the square of the bias. The term corresponding to the double product vanishes since $E\{f(x;\widehat{w}) - E[f(x;\widehat{w})]\} = 0$. Another source of uncertainty in the estimate of $\mu_y(x)$ comes from an inappropriate choice of the network architecture. Indeed, in case of a network with too few nodes, i.e. too few parameters, a large bias occurs since the regression function $f(x;\widehat{w})$ has insufficient flexibility to model the data adequately, which results in poor generalization properties of the network when fed with new input patterns. On the other side, excessively increasing the flexibility of the model by introducing too many parameters, e.g. by adding nodes, increases the variance term because the network regression function tends to over-fit the training data. Thus, in both cases, the network performs poorly in the generalization phase. Additional sources of uncertainty in the network performance arise from the minimization algorithm itself which may get stuck in a local minimum of the error function and from the fact that the training may be stopped prematurely, before reaching the minimum. The quantification of the accuracy of the estimate $f(x;\widehat{w})$ of the true deterministic function, $\mu_y(x)$, in terms of confidence intervals entails the assumption of a distribution for the regression error $[f(x;\widehat{w}) - \mu_y(x)]$ in Eq. (5) and the estimation of its variance. In practice, it is common to assume that the bias, i.e. the second term in Eq. (5), is negligible with respect to the first term, i.e. the variance of the distribution of regression function values $f(x;\widehat{w})$. Actually, neural networks trained on finite data sets are biased estimators (for example, they will always tend to over smooth a sharp peak, as almost any other model (Heskes, 1997)). However, in many applications the variance term indeed dominates the bias term (Stuart et al., 1992) and, furthermore, if it were possible to compute the bias component its value should be used a priori as a correction to obtain a more accurate regression function $f(x;\widehat{w})$. For these reasons, we concentrate on the problem of estimating the variance term $\sigma_f^2(x)$ using the bootstrap technique.

The bootstrap technique is based on resampling with replacement of the available data set and training an individual network on each resampled instance of the original data set. The bootstrap method can be used to expand upon a single realization of a distribution or process to create a set of bootstrap samples that can provide a better understanding of the average and variability of the original unknown distribution or process.

Assume that the data consists of a random sample $T_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ of size n drawn from population of unknown probability distribution F, where $t_i = (x_i, y_i)$ is a realization drawn independently and identically distributed (i.i.d.) from F and consists of a predictor vector x_i, and the corresponding output variable y_i. Let $\widehat{F}$ be the empirical distribution function for T_n with mass $1/n$ on $t_1, t_2, \ldots, t_n$; and let T^* be a random sample of size n taken from i.i.d. with replacement from $\widehat{F}$, where t_i is a single random observation $t_i = (x_i, y_i)$. The set of B bootstrap samples can be represented as $T^1, T^2, \ldots, T^b, \ldots, T^B$, in which B is the total number of bootstrap samples and ranges usually from 50 to 200 (Efron and Tibshirani, 1993). For each T^b, a ANN prediction model is constructed and the output is represented as $f_{\text{ANN}}(x_i, w_b/T^b)$, built using all n observations. The performance of the trained ANN model is evaluated using the observation pairs that are not included in a bootstrap sample and the average performance of these ANNs on their corresponding testing sets is used as an estimate of the generalization error of the ANN model developed on T_n. The generalization error

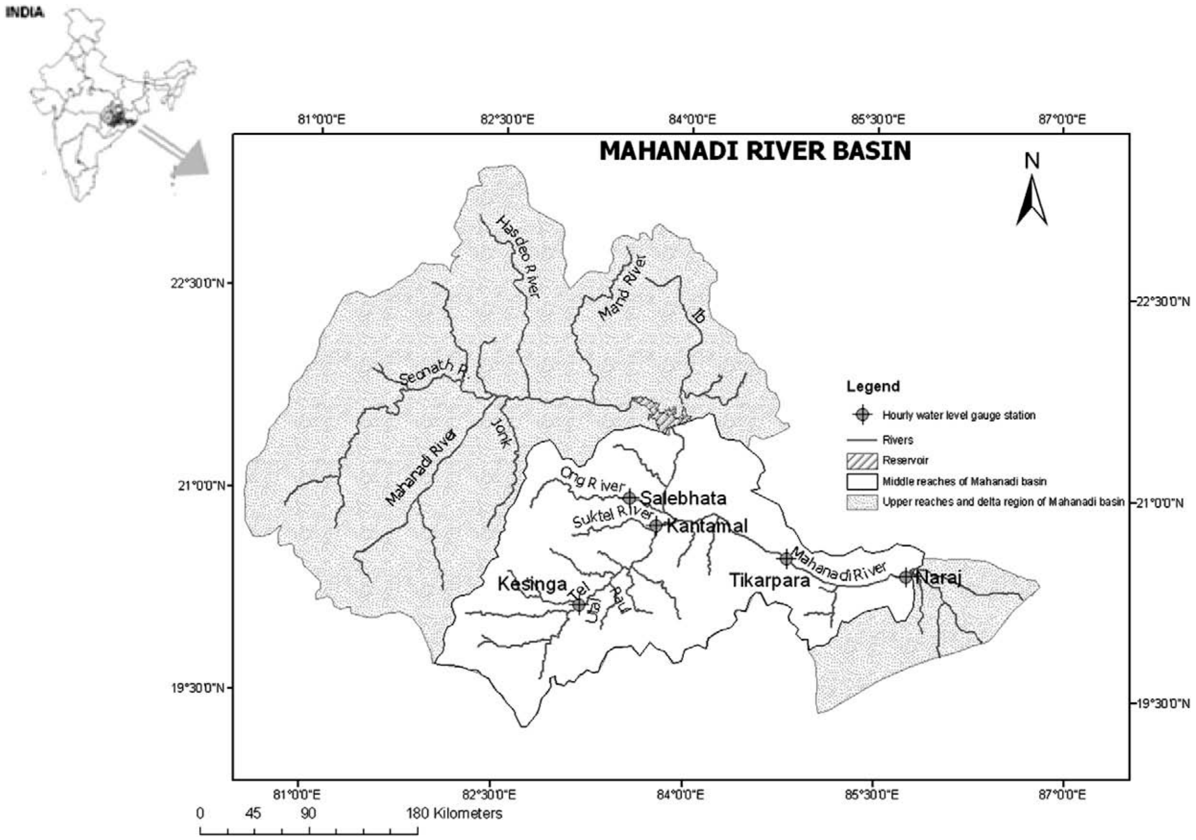


Fig. 1. Index map of the middle reaches of Mahanadi River basin showing location of different gauging stations.

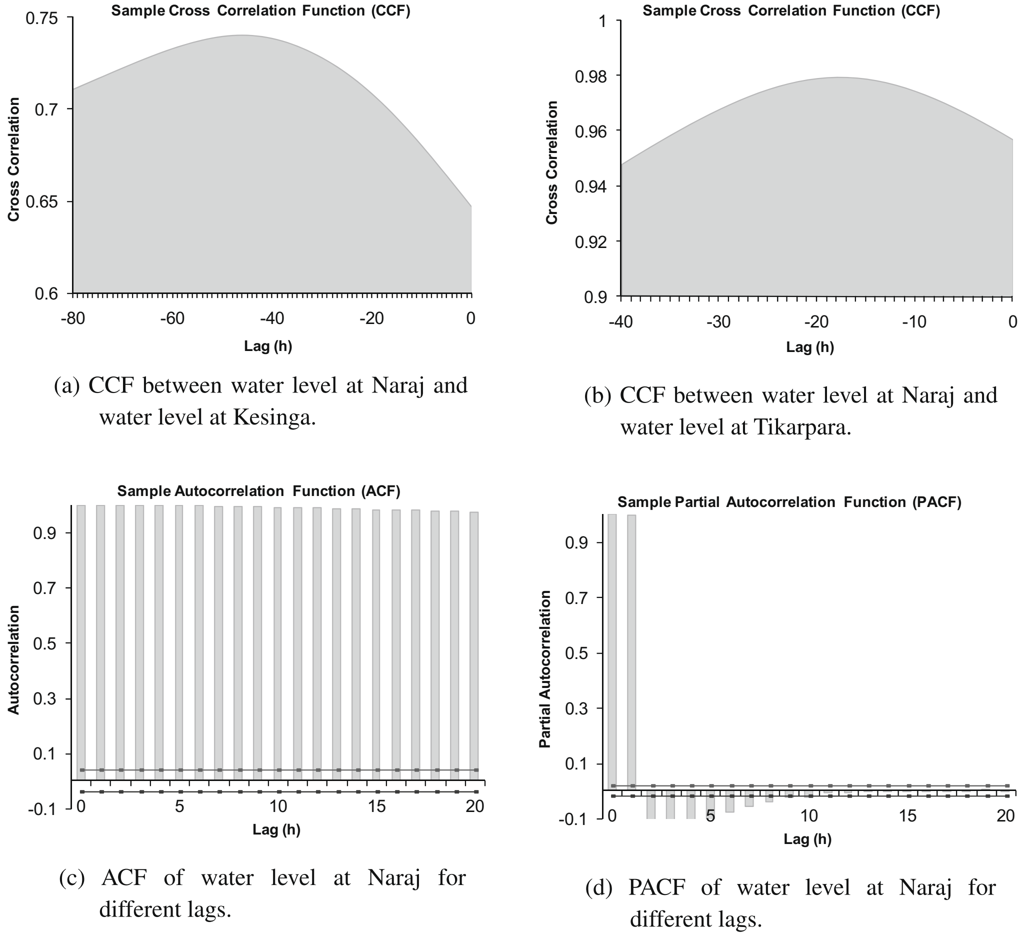
of an ANN model can be estimated by its "E_0" estimate (Twomey and Smith, 1998).

$$\widehat{E}_0 = \frac{\sum_{b=1}^{B}\sum_{i\in A_b}(y_i - f_{\mathrm{ANN}}(x_i, w_b/T^b))^2}{\sum_{b=1}^{B}\#(A_b)} \tag{6}$$

The output of the ANN developed based on the bootstrap sample T^b is represented as $f_{\mathrm{ANN}}(x, w_b/T^b)$, where x is a particular input vector; w_b is the weight vector, A_b is the set of indices for the observation pairs not included in the bootstrap sample T^b, $\#(A_b)$ is the number of observation pair indices in A_b.

Table 1
Statistics of the data set for hourly water level forecasting.

Year	Statistics (m)	Kesinga	Salebhata	Kantamal	Tikarpara	Naraj
2001	Mean	170.71	131.76	123.31	60.71	23.88
	Standard deviation	1.38	0.83	1.86	3.99	1.36
	Maximum	176.70	135.45	129.95	72.25	27.22
	Minimum	169.18	130.71	120.93	55.63	22.14
2002	Mean	169.30	131.15	121.21	56.79	22.31
	Standard deviation	0.53	0.61	0.85	2.11	0.84
	Maximum	171.53	133.54	124.07	65.16	25.26
	Minimum	168.46	130.40	120.27	54.45	20.89
2003	Mean	170.48	131.84	122.93	60.29	23.54
	Standard deviation	1.13	1.32	1.89	4.58	1.59
	Maximum	176.05	139.90	130.37	73.20	27.05
	Minimum	168.84	130.12	120.56	55.21	21.16
2004	Mean	170.15	131.24	122.13	58.14	22.88
	Standard deviation	0.81	0.63	1.34	2.71	0.98
	Maximum	173.37	134.41	129.34	67.29	25.93
	Minimum	169.16	130.43	120.27	55.21	21.61
2005	Mean	169.95	131.04	122.04	58.99	23.14
	Standard deviation	1.12	0.68	1.57	3.52	1.42
	Maximum	177.41	135.04	130.10	69.12	26.14
	Minimum	168.62	130.14	120.21	54.23	20.57



(a) CCF between water level at Naraj and water level at Kesinga.

(b) CCF between water level at Naraj and water level at Tikarpara.

(c) ACF of water level at Naraj for different lags.

(d) PACF of water level at Naraj for different lags.

Fig. 2. Correlation statistics for input structure identification.

For a new input x, the bootstrapped neural network estimate $\widehat{\theta}(x)$ is given by the average of the B bootstrapped estimates

$$\widehat{\theta}(x) = \frac{1}{B}\sum_{b=1}^{B} f_{\mathrm{ANN}}(x, w_b/T^b) \tag{7}$$

and the estimate $\widehat{\sigma}^2_{boot}(x)$ of the variance $\sigma^2_f(x)$ in (4) is given by:

$$\widehat{\sigma}^2_{boot}(x) = \frac{\sum_{b=1}^{B}\sum_{i=A_b}[y_i - f_{\mathrm{ANN}}(x_i, W_b/T^b)]^2}{B-1} \tag{8}$$

The confidence interval at the α% significance level indicates that in repeated application of the technique, the frequency with which the confidence interval would contain the true value is 100∗(1 − α)%. A typical value of α is 0.05 which corresponds to (1 − 0.05)∗100% = 95% confidence limits. A 100∗(1 − α)% confidence interval covering the true flow $\mu_y(x)$ can be estimated in the following equation (Efron and Tibshirani, 1993):

$$\widehat{\theta}(x) \pm t^{\alpha/2}_{n-p}\sigma(x) \tag{9}$$

$\sigma(x)$ is the standard deviation of B bootstrapped estimates, $t^{\alpha/2}_{n-p}$ is the $\alpha/2$ percentile for the Student t distribution with n–p degrees of freedom; n is the total number of flow observations; and p is the total number of parameters in the ANN model.

Comparison of model performance

The Nash–Sutcliffe efficiency (E), root mean square error (RMSE), mean absolute error (MAE), and percentage peak deviation (P_{dv}) performance measures are used to evaluate the accuracy of the bootstrap based ANN. The Nash–Sutcliffe efficiency (E) introduced by Nash and Shutcliff (1970) is still one of the most widely used criteria for the assessment of model performance. The E provides a measure of the ability of a model to predict values that are different from the mean. RMSE and MAE provide different types of information about the predictive capabilities of the model. The RMSE measures the goodness-of-fit relevant to high flow values whereas the MAE yields a more balanced perspective of the goodness-of-fit at moderate flows (Karunanithi et al., 1994).

(i) Nash–Sutcliffe coefficient (E): it is expressed as:

$$E = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O}_i)^2} \tag{10}$$

where O_i and P_i are the observed and predicted flow, $\overline{O}_i$ is the mean of the observed flow; n is the number of data points. The value of Nash–Sutcliffe coefficient varies between $-\infty$ and 1. The closer the value to 1, the better is the model performance.

(ii) Root mean square error (RMSE): it is expressed as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \tag{11}$$

(iii) Mean absolute error (MAE): it is expressed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \tag{12}$$

(iv) Percentage deviation in peak (P_{dv}): it is defined as

$$P_{dv} = \frac{P_p - O_p}{O_p}100 \tag{13}$$

where O_p and P_p are the peaks of observed and predicted flow.

Table 2
Most significant input vectors selected using cross correlation statistics (CCF, ACF, PACF).

Stations	Input variables (lags) (h)
Naraj	1–8
Tikarpara	15–20
Kantamal	37–40
Salebhata	31–36
Kesinga	43–49

Table 3
Division of the datasets for four BANN models.

Models	Datasets (year and number of records)		
	Training	Cross validation	Testing
(a) BANN-CV-3	2001,02,03 (7101)	2004 (2367)	2005 (2367)
(b) BANN-WCV-3	2001,02,03,04 (9468)	–	2005 (2367)
(c) BANN-CV-2	2001,02 (4734)	2004 (2367)	2005 (2367)
(d) BANN-WCV-2	2001,02,04 (7101)	–	2005 (2367)

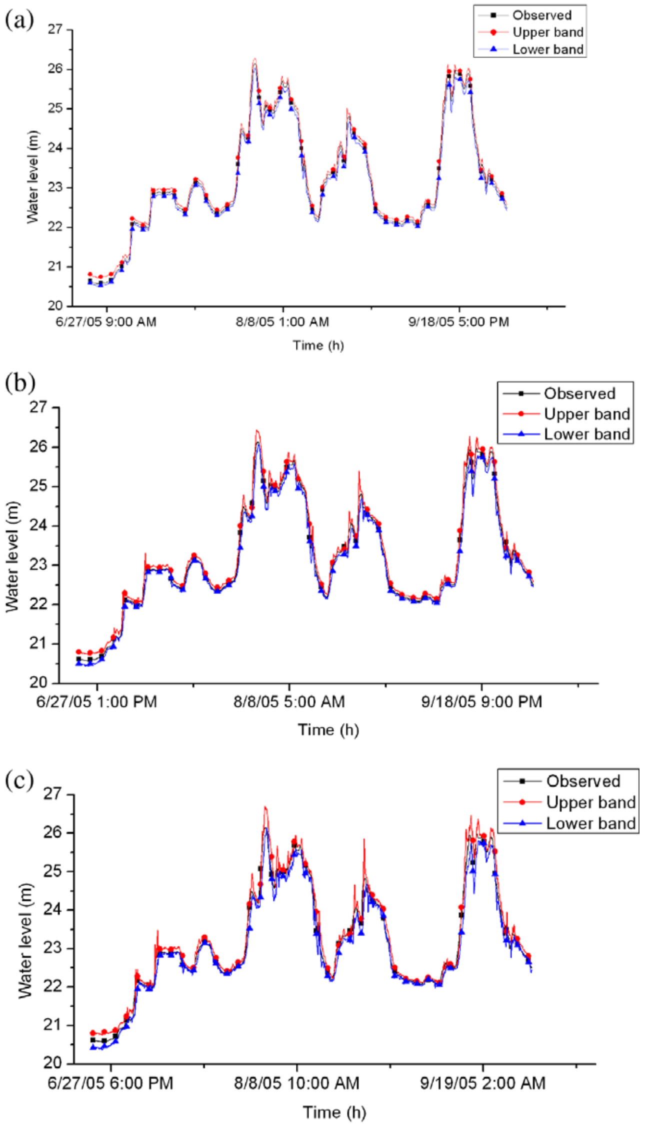


Fig. 3. Observed water levels with predicted 95% confidence bands using BANN-CV-3 for (a) 1 h, (b) 5 h and (c) 10 h lead forecasts.

Study area and data used

Mahanadi River basin which is the fourth largest river basin in India is selected for this study. The Mahanadi River flows to the Bay of Bengal in east-central India draining an area of 141,589 km^2 and has a length of 851 km. It lies between east longitudes 80°30′–86°50′ and north latitudes 19°21′–23°35′. About 53% of the basin is in the state of Chhatisgarh, 46% is in the coastal state of Orissa, and the remainder of the basin is in the states of Jharkhand and Maharashtra. Numerous dams, irrigation projects, and barrages are present in the Mahanadi River basin, the most prominent of which is Hirakud dam. The middle reaches of lower Mahanadi River basin located in Orissa between 82°E19°N and 86°E22°N and encompassing a geographical area of 47558.6 km^2 forms the study area (Fig. 1). The main river reach extends from Hirakud dam to Naraj having a total length of 358.4 km. The main soil types found in the study area are red and yellow soils. The normal annual rainfall is 1458 mm and temperature in this region varies from 14 °C to 40 °C. The average monthly pan evaporation of the area varies from 2.4 mm to 14.6 mm. Most of the rainfall and river flow occur during the monsoon season, from June to September. In the Delta region of the Mahanadi River basin flooding is a serious problem during monsoon season. Naraj, which is situated at the mouth of the Delta is selected for hourly water level forecasting.

The data used for the study consists of hourly water level of five gauging stations (Kesinga, Salebhata, Kantamal, Tikarpara, and Naraj) during the monsoon period starting from 23 June at 9 am and ending on 29 September at 11 pm for years 2001–2005. Thus, the number of hourly water level data per gauging site for 1 year is 2367 and that for 5 years is 11,385 (2367 × 5 = 11,835). Some of the statistical properties of the water level data are presented in Table 1. The location of different gauging stations is shown in Fig. 1.

Methodology

Selection of inputs to the model

One of the most important steps in the ANN hydrologic model development process is the determination of significant input variables (Bowden et al., 2005a,b). The current study used a statistical approach suggested by Sudheer et al. (2002) to identify the appropriate input vectors. The method is based on the heuristic that the

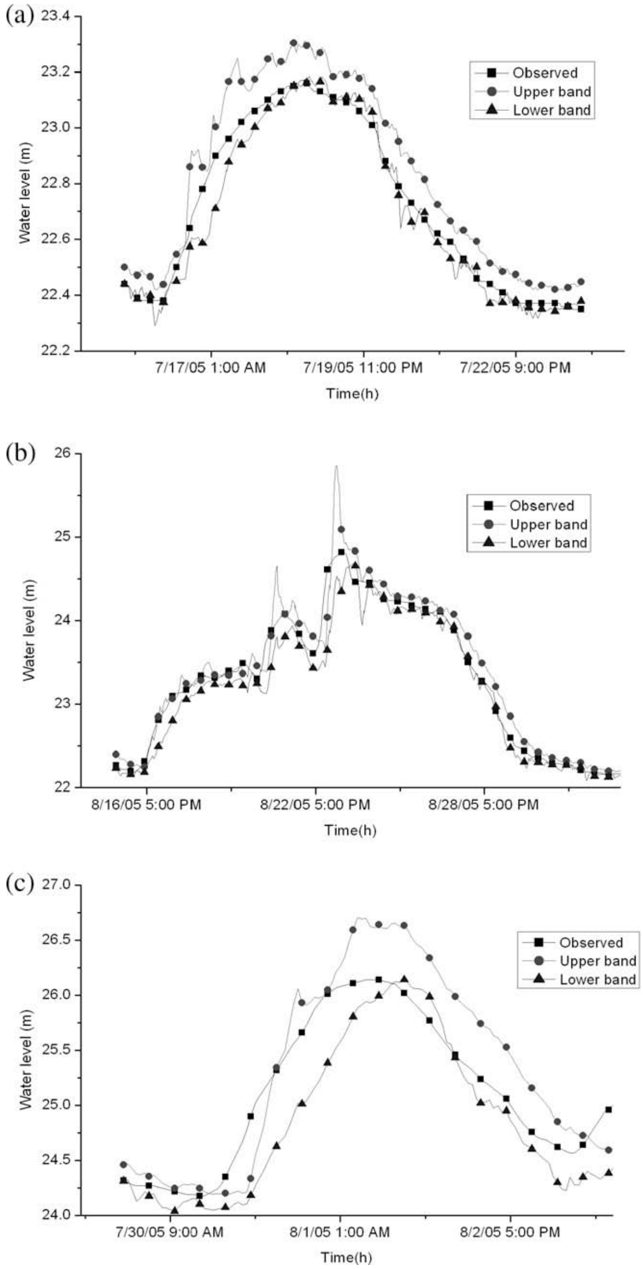


Fig. 4. Observed water levels with predicted 95% confidence bands using BANN-CV-3 model for (a) low, (b) medium and (c) high water level profiles for 10 h lead time forecast.

Table 4
Number of data points in low-, medium-, and high- water level categories.

Category	Number of data points		Percentage of total data points	
	Testing	Training	Testing	Training
Low ($x < \mu$)	1429	3922	60.37	55.23
Medium ($\mu \leqslant x \leqslant \mu + 2\sigma$)	914	2652	38.61	37.35
High ($x > \mu + 2\sigma$)	24	527	1.01	7.42
Total	2367	7101	100	100

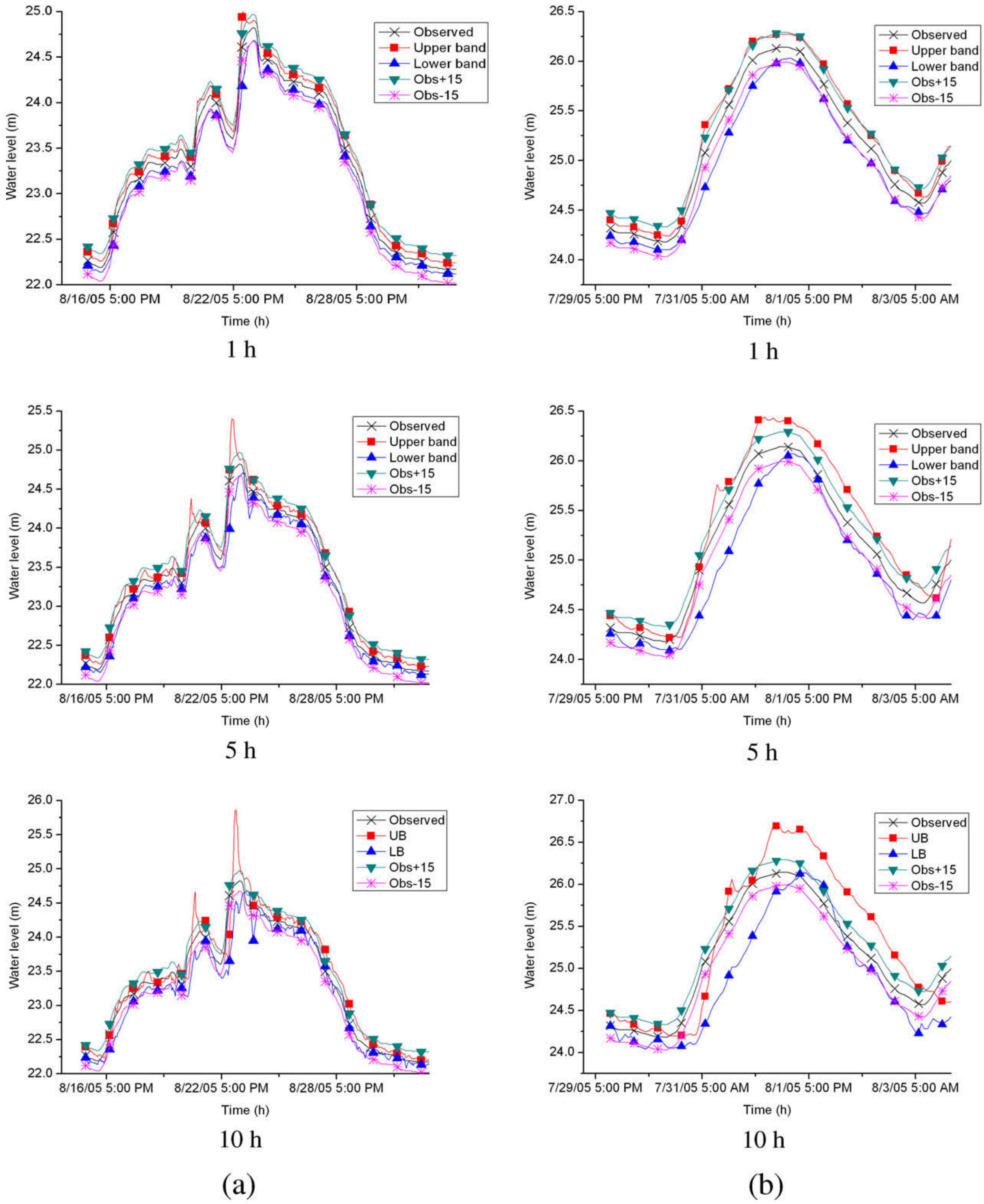
μ = 23.14.
$\mu + 2\sigma$ = 25.98.
μ and σ are mean and standard deviation, respectively for testing dataset.

Table 5
Statistics for low, medium, and high peak water level forecast results.

Water level (m)	Lead time (h)	Observed water level (m)	Mean water level forecasted (m)	95% upper bound (m)	95% lower bound (m)	Standard deviation (m)	Skewness
Low	1	23.17	23.18	23.25	23.10	0.04	−0.07
Low	5	23.17	23.19	23.25	23.13	0.03	−1.23
Low	10	23.17	23.22	23.29	23.15	0.03	−2.12
Medium	1	24.82	24.79	24.90	24.68	0.06	−0.31
Medium	5	24.82	24.77	24.87	24.67	0.05	−1.32
Medium	10	24.82	24.74	24.90	24.68	0.15	−3.08
High	1	26.14	26.14	26.27	26.00	0.07	0.34
High	5	26.14	26.19	26.39	25.99	0.10	−1.43
High	10	26.14	26.32	26.70	25.94	0.19	−3.44

potential influencing variables corresponding to different time lags can be identified through statistical analysis of the data series that uses cross correlation function (CCF), autocorrelation function (ACF), and partial autocorrelation function (PACF) between the variables.

The CCF between the water level at Kesinga and water level at Naraj show a significant correlation from 43 to 49 h lags (Fig. 2a). The CCF between the water level at Tikarpara and water level at Naraj show a significant correlation for 15–20 h lags (Fig. 2b). The ACF and the corresponding 95% confidence bands from lag 0



1 h

1 h

5 h

5 h

10 h

10 h

(a)

(b)

Fig. 5. Observed values of flood hydrographs along with ±15 cm band and predicted 95% confidence bands using BANN-CV-3 model for 1 h, 5 h and 10 h lead time for (a) medium water level and (b) high water level profiles.

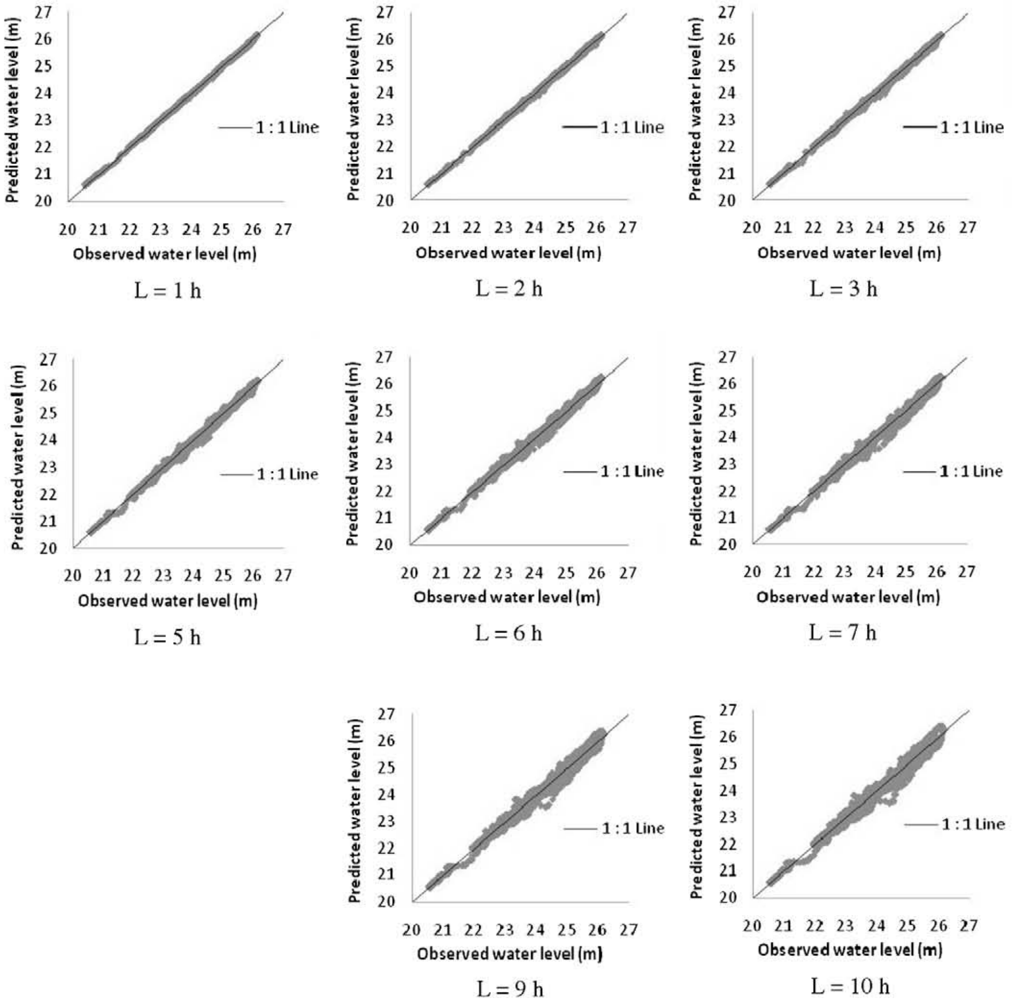


Fig. 6. Scatter plots of observed and predicted water level of testing datasets for different lead time *L* from 1 to 10 h.

Table 6
Performance measures for hourly water level forecasting during validation for different lead times with BANN-CV-3 model.

Lead time (h)	*E* (%)	RMSE (m)	MAE (m)	Peak deviation (%)
1	99.97	0.02	0.02	0.03
2	99.95	0.03	0.02	0.07
3	99.90	0.04	0.03	0.20
4	99.86	0.05	0.03	0.27
5	99.78	0.06	0.04	0.36
6	99.69	0.08	0.05	0.51
7	99.58	0.09	0.06	0.63
8	99.45	0.10	0.07	0.70
9	99.31	0.12	0.08	0.80
10	99.12	0.13	0.08	0.95

to lag 20 h are estimated for the standardized water level series of Naraj (Fig. 2c). The ACF showed a significant correlation, at the 95% confidence level from 0 to more than 20 h. Similarly, the PACF and corresponding 95% confidence limits are estimated for lags 0–20 h (Fig. 2d). The PACF showed significant correlation up to lag 8 (8 h) and, thereafter, fell within the confidence band. The decaying pattern of the PACF confirms the dominance of the autoregressive process, relative to the moving-average process. The above analysis of auto- and partial autocorrelation coefficients suggested incorporating water level values of up to 8 h lag in the input vector to the network. The same process is applied to select significant inputs from the hourly water level series of five gauge stations that is used for hourly flood forecasting. The total input vectors identified are presented in Table 2.

Model structure identification

The identification of the optimal network geometry is one of the major tasks in developing an ANN model. While the number of input output nodes are problem dependant, there is no direct and precise way of determining the optimal number of hidden nodes. The model architecture of an ANN is generally selected through a trial-and-error procedure (Sudheer and Jain, 2004), as currently there is no established methodology for selecting the appropriate network architecture prior to training (Coulibaly et al., 2001). Most ANN models use split sample validation approach to check the ANN models generalization. To apply the split sample validation approach, the available data set is split into a training set, a testing set, and a validation set. The training dataset is used to train the model to estimate the model parameters; the testing dataset is used to test the generalization capability of the model, whereas the validation dataset is used to stop the training process at the time when the error increases for the validation set. This method of stopping the training process before the minimum error based on the training set is reached is often called the early stopping method of training (Haykin, 1998). It prevents a complex network from over-fitting the available data set. One problem with the cross validation approach is that the early stopping method of ANN training is heuristic and lacks theoretical reasoning (Anders and Korn, 1999). Another limitation is that model performance of an ANN model could heavily depend on the partitioning of the available data, especially when the available data set is small (Maier and Dandy, 2000). To overcome these limitations, the optimal model structure is selected using the generalization error (Eq. (6)) based on bootstrap resampling technique (Twomey and Smith, 1998; Jia and Culver, 2006).

The optimum number of hidden neurons is calculated using the generalization error. The generalization error is calculated for 1 h lead time forecast using the mean/ensemble prediction for 50 bootstrapped model and the observed water levels. The number of hidden neurons for all other lead times is taken to be same as for 1 h lead time to maintain consistency. The ANN structures are tested for 1–15 hidden neurons and 0.1–0.9 learning coefficient and momentum and the structure with seven hidden neurons with learning coefficient equal to 0.3 and momentum equal to 0.8 for which the generalization error is minimum is chosen as the optimal structure.

Bootstrapped artificial neural networks (BANNs) development

The ANN models are developed using the most significant inputs which are first log-transformed and then linearly scaled to the range (0, 1) for ANN modeling (e.g. Campolo et al., 1999). The activation function is usually selected to be a continuous and bounded nonlinear transfer function, for which the sigmoid and hyperbolic tangent functions are commonly used (e.g. Haykin, 1998; Govindaraju and Rao, 2000). In this study, a tangent sigmoid function is used as the neuron transfer function for ANN. The computational efficiency of the training process is an important consideration for BANN modeling. A computationally efficient

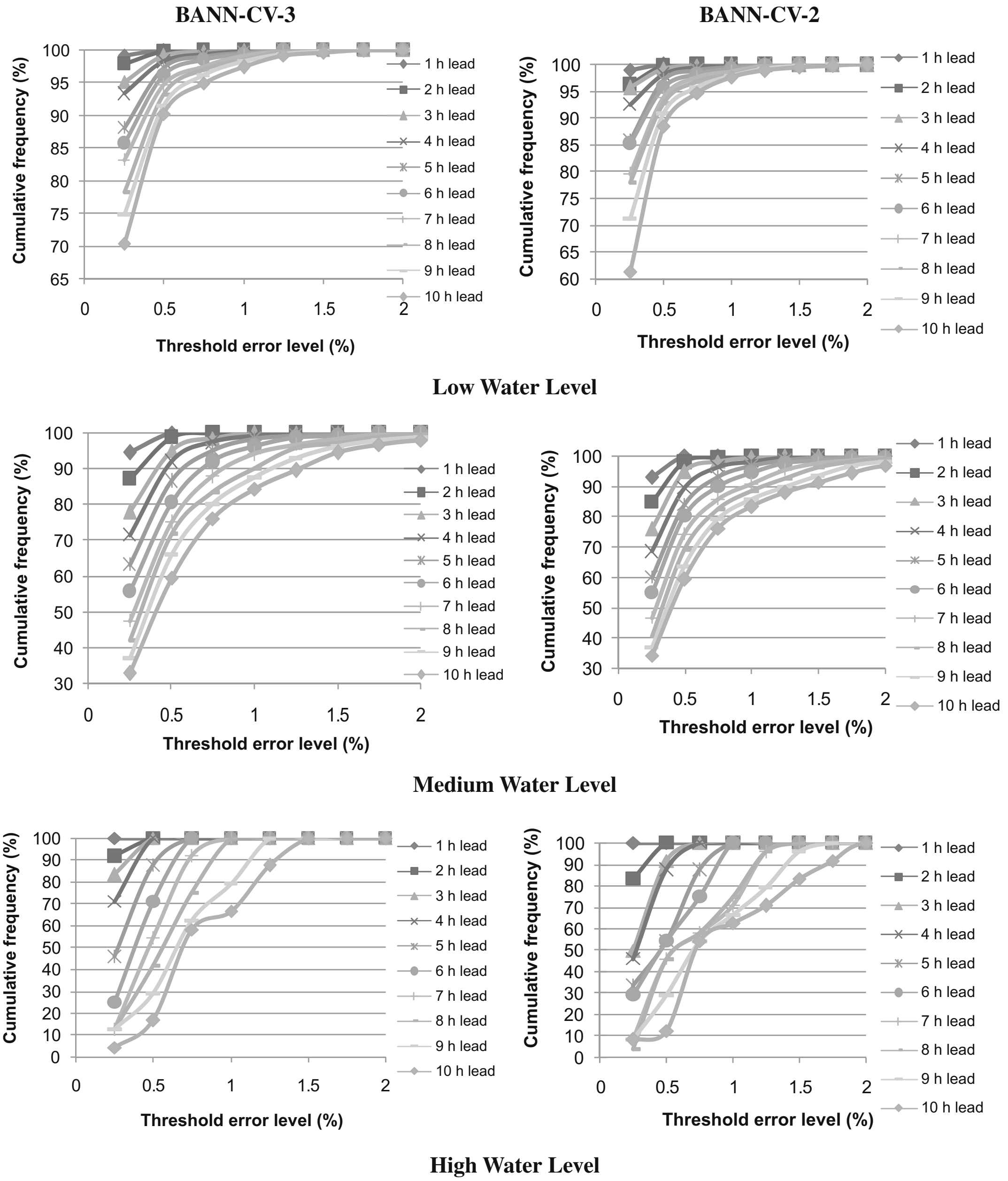


Fig. 7. Distribution of forecast error across different error thresholds for 1–10 h lead time forecasts for low, medium and high water level profiles.

second-order training method, the Levenberg–Marquardt method (Hagan and Menhaj, 1994), is used to minimize the mean squared error between the forecast and the observed gauges. BANN models are developed to illustrate the parametric uncertainty and for ensemble flood forecasting. The BANN models are applied for 1–10 h lead time forecast. In a BANN model, each bootstrapped sample is used to develop an ANN, and many ANNs are developed and then combined to approximate the relationship between model inputs and outputs. The confidence band represents the uncertainty in the neural network predictions. Bootstrapped neural networks use a range of weight sets instead of a single set. The confidence band show the limits to which the predictions could have varied based on the weight set used. The final water level forecasts are the average of these simulations. Bootstrap.xla an Excel Add–In (Barreto and Howland, 2006) is used to develop 50 bootstrap resamples for 1–10 h lead time gauge forecasts. Hourly data of 5 years for monsoon period from years 2001 to 2005 (11,835 datasets) are used for this study.

In this study four BANN models are developed. They are: (a) BANN with cross validation (BANN-CV-3), (b) BANN without cross validation (BANN-WCV-3), (c) BANN with cross validation (BANN-CV-2), (d) BANN without cross validation (BANN-WCV-2). The details of the datasets involved in the models are shown in Table 3. Model (a) is used to assess the uncertainty and for ensemble flood forecasting. Combination of models (a), (b) and (c), (d) are used to assess the performance of split sample validation techniques while the combination of models (a), (c) and (b), (d) are used to compare the performance of models with different length of training datasets.

Results and discussion

Uncertainty in water level forecasts

Fig. 3 shows the observed water level with the 95% confidence band predicted by BANN-CV-3 model for 1 h, 5 h and 10 h lead forecasts for all the testing datasets of 2005. The upper band (UB) and lower band (LB) are the 95% confidence bands which show the uncertainty in neural network predictions. It is obvious that for short lead times the confidence band is narrow and widens as the lead time increases. The width of the confidence bands is more for higher water levels compared to lower water levels.

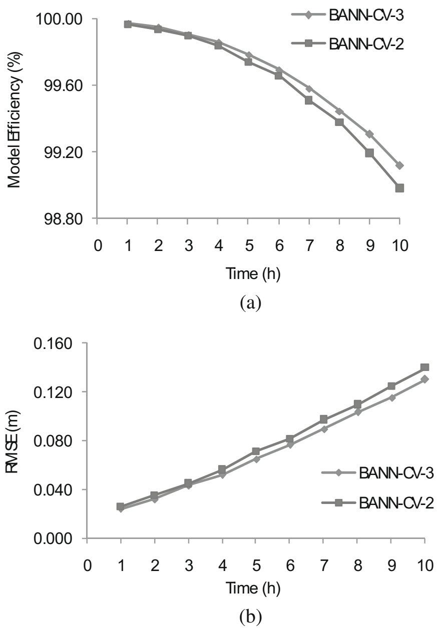


Fig. 8. Performance comparison of BANNs with 2 year and 3 year training dataset: (a) model efficiency (%) and (b) RMSE (m).

In order to assess the uncertainty assessment for different lead time forecasts predicting different magnitudes of flows, a "partitioning analysis" (Jain and Srinivasulu, 2004) is carried out by dividing the total gauge range into low-, medium-, and high-magnitude gauges. Table 4 presents the partitioning of water level for testing period based on the relative spread of the gauges from the mean. Model BANN-CV-3 is used to illustrate the uncertainty in water level forecasts. Peak water level values of hydrographs which fall in low, medium, and high water levels are selected. Table 5 depicts some of the statistics for empirical distribution of predicted water levels of 50 bootstrapped models for 1 h, 5 h and 10 h lead times for actual water level 23.17 m, 24.82 m and 26.14 m as peak values of low, medium and high water level hydrographs, respectively. Skewness values in Table 5 show that the empirical prediction distributions from the 50 bootstrapped models do not deviate excessively from the normal distribution for low, medium and high water levels for 1 h lead time. During

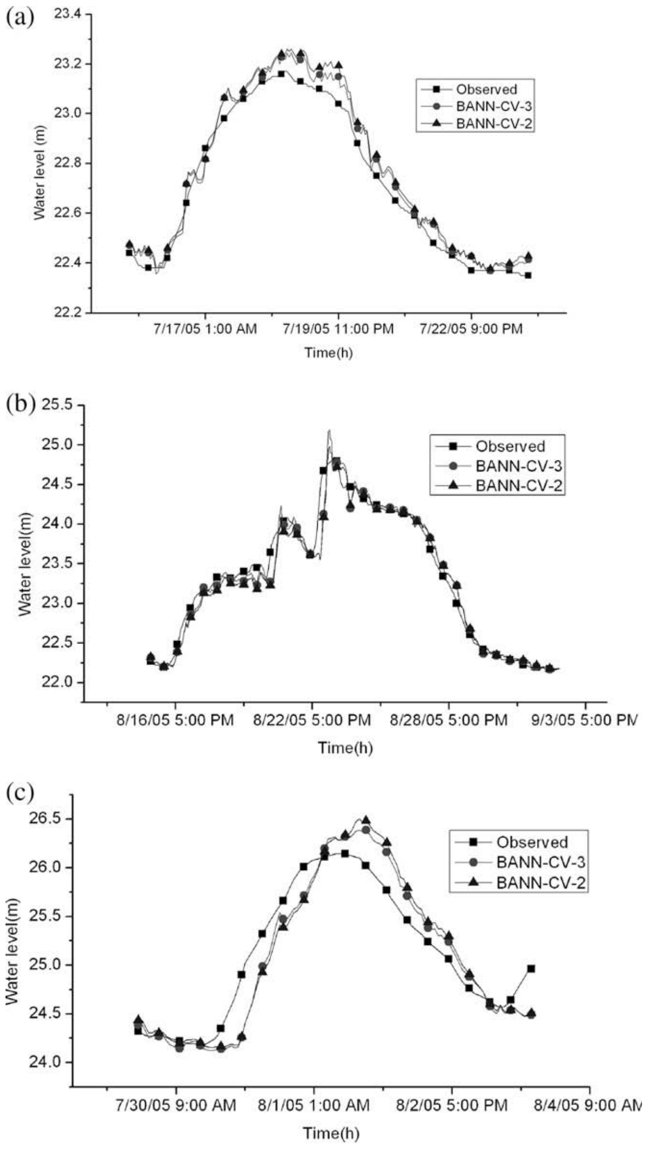


Fig. 9. Comparison of ensemble prediction of BANN-CV-3 and BANN-CV-2 for 10 h lead time forecast for (a) low water levels, (b) medium water levels and (c) high water levels.

higher lead time predictions, the distributions are obviously biased and skewed for all the three water level profile which indicates internal weaknesses in the process description of the model. Table 5 depicts that the width of the confidence band widens for medium and higher water level forecasts but for a particular water level profile (low, medium or high) the width is almost consistent for 1–10 h lead time forecasts. The results show that the peaks of the hydrographs are overestimated for almost all the lower and higher water level and underestimated for medium water level. This can be attributed not only to the weakness in the model structure but also to the complex physical process of the hydrological cycle within the system. Inspite of this all the peaks of selected observed hydrographs from low, medium and high water levels lie within the 95% confidence bands. Fig. 4 shows the confidence band for low, medium and high water level forecasts for 10 h lead time. This analysis is carried out for hydrographs from low, medium and high water level profiles. It is observed that for low water level forecast the confidence bands are not smooth, which may be due to noisy datasets and unsaturated condition of the basin. It is also observed that the rising limb of the water level hydrographs is biased towards upper confidence band whereas recession limb is biased towards lower confidence band and this phenomenon is more prominent for higher water level values. It is observed from Table 5 that the values of low and high water level are over predicted and the peak value of medium water level profile is under estimated. Many researchers have reported that ANN models fail to capture and underestimate the peak flows in a hydrograph (Sudheer et al., 2002; Srivastav et al., 2007). This study reveals that almost half of the values of flood hydrographs are underestimated and another are over estimated and a large amount of values in low-, medium-, and high water level profiles fall within the confidence band (Fig. 4) and show the strength of bootstrapping in uncertainty assessment. Inspite of having a low representation of high water levels in the training set which is 7.42%, 20 (83% i.e. 20 of 24) values lie within the confidence bands. Overall most of the observed values fall in the confidence band and shows the capability of BANN model in quantifying the parametric uncertainty for all the water level profiles.

In the existing flood forecasting technique in India which use gauge to gauge correlation, a simple and common criteria used to evaluate the forecast performance, considers ± 15 cm variation between predicted and observed water level as satisfactory (Central Water Commission, 1989). Fig. 5 shows the observed values along with ± 15 cm band and the predicted 95% confidence band using BANN-CV-3 model for hydrographs in medium and high water level profiles for 1 h, 5 h, and 10 h lead time forecasts. It is observed that for hydrographs in medium water level profile most of the predicted 95% confidence band lie within ±15 cm band for all the lead time forecasts. In higher water level profile the predicted 95% confidence band lie within the ±15 cm band only for 1 h lead time forecast. This shows the higher uncertainties in forecasting the high water levels.

Ensemble water level forecasting

BANN-CV-3 model is used to make ensemble predictions by averaging the 50 bootstrapped ANN models predictions. Fig. 6 shows the scatter plots of observed versus BANN-CV-3 predicted water levels for 1–10 h lead time forecast. Fig. shows that model predictions are in good agreement with the observed water levels. Table 6 shows the performance of the models for 1–10 h lead times forecast in terms of E, RMSE, MAE and P_{dv}. It is observed that the model performs very well eventhough the model performance deteriorates with increase in lead times. This is a general tendency for longer lead times as the previous water level values contain less information than for the shorter lead times for mapping the water level values at longer time horizon. The consistency in terms of all the performance indices for 1–10 h lead time shows the strength of bootstrapping aggregation technique.

It has been reported that the coefficient of efficiency or model efficiency or E can be high (80% or 90%) even for poor models, and the best models do not produce values which, on first examination, are impressively higher (Legates and McCabe, 1999). The RMSE statistic indicates only the model's ability to predict a value away from mean (Hsu et al., 1995).

Therefore, in order to test the effectiveness of the model developed, it is important to test the model using some other performance evaluation criteria such as absolute relative error (ARE) and threshold statistics (Jain et al., 2001; Jain and Ormsbee, 2002; Nayak et al., 2005). The ARE and threshold statistics (TS) not only give the performance index in terms of predicting water levels but also the distribution of the prediction errors.

$$\mathrm{RE}_t = \frac{y_t^o - y_t^c}{y_t^o} \times 100 \tag{14}$$

where RE_t is the relative error in forecast at time t expressed as a percentage, y_t^o is the observed water level at time t, y_t^c is the computed water level at time t.

The threshold statistic for a level of x% is a measure of consistency in forecasting errors from a particular model. The threshold statistics is represented as TS_x and expressed as a percentage. This criterion can be expressed for different levels of absolute relative error from the model. It is computed for x% level as

$$TS_x = \frac{Y_x}{n} \times 100 \tag{15}$$

where Y_x is the number of computed water level values (out of n total computed) for which the absolute relative error is less than x% from the model.

It is clear from the Fig. 7 that the BANN-CV-3 model can forecast 97.4%, 84.3% and 66.7% of the total number of low, medium and high water levels, respectively with less than 1% relative error for 1–10 h lead time forecasts. It is also observed from Fig. 7 that in the entire water level profile of the testing period BANN-CV-3 model predicts 90% of low water level values up to 10 h lead time,

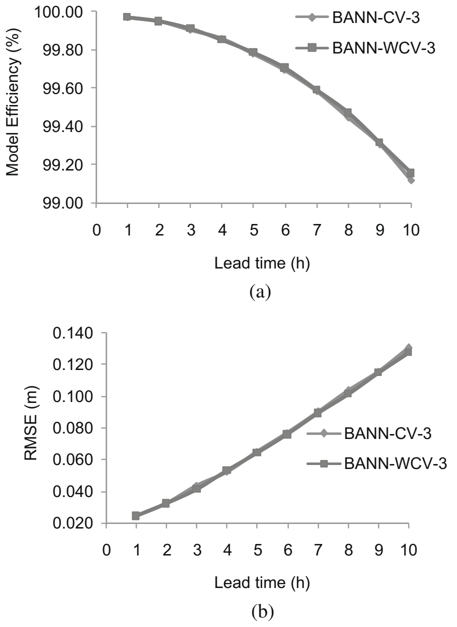
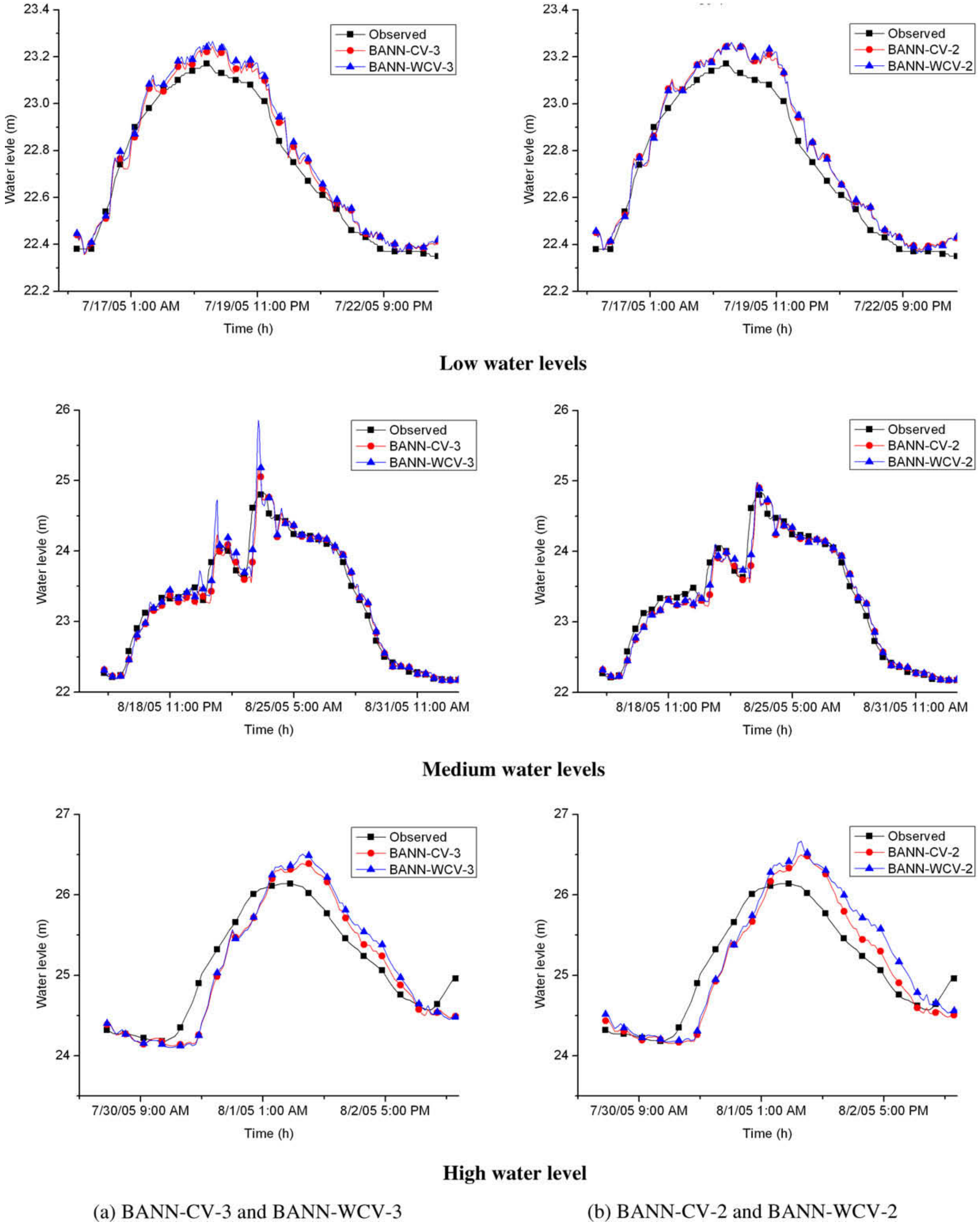


Fig. 10. Performance comparison between models with cross validation technique and without cross validation techniques: (a) model efficiency (%) and (b) RMSE (m).

80% of medium water level values up to 6 h lead time and 70% of high water level values up to 6 h lead time with less than 0.5% of relative error. This 0.5% relative error can be considered as equivalent to the practiced criteria of ±15 cm variation.

Effect of length of training datasets

To illustrate the performance of BANN models for different length of training datasets another model BANN-CV-2 is developed and its ensemble predictions compared with BANN-CV-3 model. The comparison is carried out for entire testing datasets for 1–10 h lead time forecasts (Fig. 8) and for selected hydrographs in low, medium, and high water level profiles for 10 h lead time forecast (Fig. 9). Fig. 8 shows the effect of length of training datasets with two performance indices, E and RMSE for 1–10 h lead time forecasts. It is observed that for 1 h lead time forecast there is no effect of the length of training datasets. However for longer lead times the BANN-CV-3 model performance is marginally better than



Low water levels

Medium water levels

High water level

(a) BANN-CV-3 and BANN-WCV-3

(b) BANN-CV-2 and BANN-WCV-2

Fig. 11. Comparison of ensemble prediction of (a) BANN-CV-3 and BANN-WCV-3 and (b) BANN-CV-2 and BANN-WCV-2 for 10 h lead time for low-, medium-, and high-water levels profile.

BANN-CV-2 model. It is observed from Fig. 9 that both the models underestimate the rising limb and over estimate the recession limb values and this phenomenon is more prominent for high water level values. It is also seen that for low and high water level there is no significant difference between the mean water level forecasted by BANN-CV-2 and BANN-CV-3 models. However, for medium water level forecasts, model BANN-CV-3 over predicts peak values (this is a double-peaked hydrograph) whereas BANN-CV-2 model simulates the observed peak values more satisfactorily. The results obtained for medium water level indicates that with proper representation, short length training datasets can perform similar to the long length training datasets. It is observed from Fig. 7 that BANN-CV-2 model can forecast 90% of the total number of low, medium and high gauge values for 10 h, 4 h and 3 h lead time, respectively whereas model BANN-CV-3 can forecast 90% of the total number of low, medium and high gauge values for 10 h, 4 h and 4 h lead time, respectively. BANN-CV-3 model forecasts all the high water level values below 1.5% relative error whereas BANN-CV-2 model forecasts all the high water level values within 2% relative error. This difference in forecast error for higher water level may be due to the low representation of higher water levels in the training data set of the BANN-CV-2 model as this model does not include the datasets of 2003 which is a flood year. This implies that models with proper representation of datasets can perform similar and inclusion of new datasets that are beyond the calibration datasets can marginally improve the performance of BANN model predictions.

Effect of split sample validation

In this study the models BANN-CV-3 and BANN-WCV-3 and models BANN-CV-2 and BANN-WCV-2 are used to compare the performance of model based on cross validation and without cross validation technique for 1–10 h lead time forecasts. It is observed from Fig. 10 that the BANN model based on cross validation performs almost similar to the model without cross validation for 1–10 h lead times.

Performance of the BANN-CV-3 and BANN-WCV-3 models is also assessed for low, medium and high water level profiles. Fig. 11 shows that hydrographs for low and high water level under estimate the rising limb and overestimate the recession limb, whereas this phenomena is not prominent in medium gauge forecasts. This may be due to the reason that higher water levels lack proper representation whereas low water levels represent unsaturated conditions of the whole or a part of the basin. Whereas the medium water levels have a good representation in the training datasets and have less effect of unsaturated conditions of the basin. The figures for medium water level especially for peak values show that BANN-WCV-3 model performs marginally better than BANN-CV-3 model whereas the performance of model BANN-CV-2 is very similar to BANN-WCV-2 model. This shows that it is not necessary to use early stopping criteria when bootstrap technique is employed for ensemble flood forecasting.

Summary and conclusions

Assessment of uncertainty involved in forecasts of different lead time from artificial neural networks is necessary to assure reliable use of the results. The study evaluates the parametric uncertainty and importance of ensemble flood forecasting in ANN modeling. Five year water level data of five water level gauging stations are used to assess the uncertainty involved in hourly water level forecasting for 1–10 h lead times. The uncertainty associated with hourly flood forecast is investigated using the bootstrap resampling technique based artificial neural networks (BANNs). Four neural network models (BANN-CV-2, BANN-WCV-2, BANN-CV-3, BANN-WCV-3) are developed in this study. BANN-CV-3 model performs very well up to 1 h lead forecast. The results obtained using BANN-CV-3 indicate that the uncertainty increases for longer lead time but reliability of forecast can be increased by making confidence interval and ensemble predictions. The BANN-CV-3 model is used to perform bootstrap aggregation of multi-model ensembles which produced averaged outputs and yielded more stable solution compared to the traditional ANN methodologies. The effect of the length of the training data set and the performance of the early stopping criterion are also investigated using BANN models. It is found that there is no effect of length of training datasets for 1 h lead time forecasts whereas for longer lead time longer length of training datasets can improve the model performance but it is dependent on data representation. It is also found that the early stopping criterion does not improve the model performance when combined with the bootstrap technique. Based on past literature, in this study the architecture is optimised only for 1 h lead forecast and this architecture is used for all lead times up to 10 h forecast, eventhough the complexity of the model for different lead times could be different. The number of input vectors may also play a role for different lead time forecasts. These issues need to be addressed properly in future studies. One of the findings of this study is that short length data with proper representation can produce better results as compared to long length datasets, however, further studies are necessary for selection of optimum length of datasets that produces minimum error. It is also observed that the rising limb of the water level hydrographs is biased towards upper confidence band whereas recession limb is biased towards lower confidence band. Further, ensemble prediction shows that the rising limb of the flow hydrograph is underestimated whereas the falling limb is overestimated by the models particularly for higher water level values. This can be interpretated as a systematic shift of predicted values. Studies may be conducted to correct this shift using bias correction techniques.

Acknowledgements

The authors would like to thank the Mahanadi and Eastern River Organization, Central Water Commission (CWC), Bhubaneswar for providing the necessary hydrological data. The authors would also like to thank the anonymous reviewers for their constructive comments which was of immense help in improving the manuscript.

References

Abrahart, R.J., 2003. Neural network rainfall–runoff forecasting based on continuous resampling. Journal of Hydroinformatics 5 (1), 51–61.

Agarwal, A., Singh, R.D., 2004. Runoff modelling through back propagation artificial neural network with variable rainfall–runoff data. Water Resources Management 18 (3), 285–300.

Anders, U., Korn, O., 1999. Model selection in neural networks. Neural Networks 12, 309–323.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology., 2000a. Artificial neural networks in hydrology I: preliminary concepts. Journal of Hydrological Engineering, ASCE 5 (2), 115–123.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology., 2000b. Artificial neural networks in hydrology II: hydrologic applications. Journal of Hydrological Engineering, ASCE 5 (2), 124–137.

Barreto, H., Howland, F.M., 2006. Introductory Econometrics: Using Monte Carlo Simulation with Microsoft Excel. Cambridge University Press, New York.

Bates, B.C., Townley, L.R., 1988. Nonlinear, discrete flood event models: 3. Analysis of prediction uncertainty. Journal of Hydrology 99, 91–101.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press.

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005a. Input determination for neural network models in water resources applications. Part 2 – forecasting salinity in a river. Journal of Hydrology 301, 93–107.

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005b. Input determination for neural network models in water resources applications. Part 1 – background and methodology. Journal of Hydrology 301, 75–92.

Campolo, M., Andreussi, P., Soldati, A., 1999. River flood forecasting with a neural network model. Water Resources Research 35 (4), 1191–1197.

Campolo, M., Soldati, A., Andreussi, P., 2003. Artificial neural network approach to flood forecasting in the river Arno. Hydrological Sciences Journal 48 (3), 381–398.

Central Water Commission., 1989. Manual on Flood Forecasting. River Management Wing, New Delhi.

Coulibaly, P., Anctil, F., Bobee, B., 2001. Multivariate reservoir inflow forecasting using temporal neural network. Journal of Hydrological Engineering, ASCE 6 (5), 367–376.

Cover, K.A., Unny, T.E., 1986. Application of computer intensive statistics to parameter uncertainty in streamflow synthesis. Water Resources Bulletin 22 (3), 495–507.

Dawson, C.W., Wilby, R.L., 2001. Hydrological modelling using artificial neural networks. Progress in Physical Geography 25, 80–108.

Dybowski, R., Roberts, S.J., 2000. Confidence and prediction intervals for feed forward neural networks. In: Dybowski, R., Gant, V. (Eds.), Clinical Applications of Artificial Neural Networks. Cambridge University Press.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7, 1–26.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, London, UK.

Govindaraju, R.S., Rao, A.R., 2000. Artificial Neural Networks in Hydrology. Kluwer, The Netherlands.

Hagan, M.T., Menhaj, M.B., 1994. Training feed forward techniques with the Marquardt algorithm. IEEE Transactions on Neural Networks 5 (6), 989–993.

Han, D., Kwong, T., Li, S., 2007. Uncertainties in real-time flood forecasting with neural networks. Hydrological Processes 21 (2), 223–228.

Haykin, S., 1998. Neural Networks: A Comprehensive Foundation. Prentice-Hall, New Jersey.

Heskes, T., 1997. Practical confidence and prediction intervals. Advances in Neural Information Processing Systems 9, 466–472.

Hsu, K., Gupta, H.V., Sorooshian, S., 1995. Artificial neural network modeling of the rainfall–runoff process. Water Resources Research 31 (10), 2517–2530.

Jain, A., Ormsbee, L.E., 2002. Evaluation of short-term water demand forecast modeling techniques: conventional methods versus AI. Journal of American Water Works Association 94 (7), 64–72.

Jain, A., Srinivasulu, S., 2004. Development of effective and efficient rainfall–runoff models using integration of deterministic, real-coded genetic algorithms and artificial neural network techniques. Water Resources Research 40, W04302.

Jain, A., Varshney, A.K., Joshi, U.C., 2001. Short-term water demand forecast modeling at IIT Kanpur using artificial neural networks. Water Resources Management 15 (5), 299–321.

Jeong, D., Kim, YO., 2005. Rainfall–runoff models using artificial neural networks for ensemble streamflow prediction. Hydrological Processes 19 (19), 3819–3835.

Jia, Y., Culver, T.B., 2006. Bootstrapped artificial neural networks for synthetic flow generation with a small data sample. Journal of Hydrology 331, 580–590.

Karunanithi, N., Grenney, W.J., Whitley, D., Bovee, K., 1994. Neural networks for river flow prediction. Journal of Computing in Civil Engineering 8 (2), 201–220.

Lall, U., Sharma, A., 1996. A nearest neighbor bootstrap for resampling hydrologic time series. Water Resources Research 32 (3), 679–693.

LeBaron, B., Weigend, A.S., 1998. A bootstrap evaluation of the effect of data splitting on financial time series. IEEE Transactions on Neural Networks 9, 213–220.

Legates, D.R., McCabe, G.J., 1999. Evaluating the use the goodness-of-fit measure in hydrologic and hydroclimatic model validation. Water Resources Research 35, 233–241.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling and Software 15, 101–124.

Muller, B., Reinhardt, J., 1991. Neural Networks – An Introduction. Springer-Verlag, Berlin.

Nash, J.E., Shutcliff, J.V., 1970. River flow forcasting through conceptual models: I. Journal of Hydrology 10, 282–290.

Nayak, P.C., Sudheer, K.P., Rangan, D.M., Ramasastri, K.S., 2005. Short-term flood forecasting with a neurofuzzy model. Water Resources Research 41, W04004.

Rumelhart, D.E., McClelland, J.L., 1986. Parallel Distributed Processing 1. MIT Press, Cambridge, MA.

Sharma, A., Tarboton, D.G., Lall, U., 1997. Streamflow simulation: a nonparametric approach. Water Resources Research 33 (3), 291–308.

Srivastav, R.K., Sudheer, K.P., Chaubey, I., 2007. A simplified approach to quantifying predictive and parametric uncertainty in artificial neural network hydrologic models. Water Resources Research 43, W10407.

Stuart, G., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. Neural Computations 4, 1–58.

Sudheer, K.P., Gosain, A.K., Ramasastri, K.S., 2002. A data-driven algorithm for constructing artificial neural network rainfall–runoff models. Hydrological Processes 16, 1325–1330.

Sudheer, K.P., Jain, A., 2004. Explaining the internal behaviour of artificial neural network river flow models. Hydrological Processes 18 (4), 833–844.

Tasker, G.D., Dunne, P., 1997. Bootstrap position analysis for forecasting low flow frequency. Journal of Water Resources Planning and Management 123 (6), 359–367.

Tibshirani, R., 1996. A comparison of some error estimates for neural network models. Neural Computation 8, 152–163.

Twomey, J.M., Smith, A.E., 1998. Bias and variance of validation methods for function approximation neural networks under conditions of sparse data. IEEE Transactions on Systems, Man, and Cybernetics C: Applications and Reviews. 28 (3), 417–430.

Woo, M.K., 1989. Confidence intervals of optimal risk based hydraulic design parameters. Canadian Water Resources Journal 14, 10–16.

Zio, E., 2006. A study of the bootstrap method for estimating the accuracy of artificial neural networks in predicting nuclear transient processes. IEEE Transactions on Nuclear Science 53 (3), 1460–1478.