Hydrol. Earth Syst. Sci., 13, 1607–1618, 2009
www.hydrol-earth-syst-sci.net/13/1607/2009/

Hydrology and Earth System Sciences

River flow forecasting with artificial neural networks using satellite observed precipitation pre-processed with flow length and travel time information: case study of the Ganges river basin

M. K. Akhtar[2], **G. A. Corzo**[1], **S. J. van Andel**[1], **and A. Jonoski**[1]

[1]UNESCO-IHE Institute for Water Education, Dept. of Hydroinformatics and Knowledge management, P.O. Box 3015, 2601 Delft, The Netherlands
[2]University of western Ontario, Dept. of Civil and Environmental Engineering, Spencer Engineering Building, London, Ontario, N6A 5B9, Canada

Received: 8 April 2009 – Published in Hydrol. Earth Syst. Sci. Discuss.: 24 April 2009
Revised: 13 August 2009 – Accepted: 18 August 2009 – Published: 10 September 2009

Abstract. This paper explores the use of flow length and travel time as a pre-processing step for incorporating spatial precipitation information into Artificial Neural Network (ANN) models used for river flow forecasting. Spatially distributed precipitation is commonly required when modelling large basins, and it is usually incorporated in distributed physically-based hydrological modelling approaches. However, these modelling approaches are recognised to be quite complex and expensive, especially due to the data collection of multiple inputs and parameters, which vary in space and time. On the other hand, ANN models for flow forecasting are frequently developed only with precipitation and discharge as inputs, usually without taking into consideration the spatial variability of precipitation. Full inclusion of spatially distributed inputs into ANN models still leads to a complex computational process that may not give acceptable results. Therefore, here we present an analysis of the flow length and travel time as a basis for pre-processing remotely sensed (satellite) rainfall data. This pre-processed rainfall is used together with local stream flow measurements of previous days as input to ANN models. The case study for this modelling approach is the Ganges river basin. A comparative analysis of multiple ANN models with different hydrological pre-processing is presented. The ANN showed its ability to forecast discharges 3-days ahead with an acceptable accuracy. Within this forecast horizon, the influence of the pre-processed rainfall is marginal, because of dominant influence of strongly auto-correlated discharge inputs. For forecast horizons of 7 to 10 days, the influence of the pre-processed rainfall is noticeable, although the overall model performance deteriorates. The incorporation of remote sensing data of spatially distributed precipitation information as pre-processing step showed to be a promising alternative for the setting-up of ANN models for river flow forecasting.

Correspondence to: G. A. Corzo
(corzogac@yahoo.es)

1 Introduction

Many of the activities associated with the planning and operation of the components of a water system require forecasts of future events. There is a need for both short-term and long-term forecasts of stream flow, in order to optimise the water resources system. Moreover, operational river management strongly depends on accurate and reliable flow forecasts. Such forecasting of river flow provides warnings of approaching floods and assists in regulating reservoir outflow during low river flows for water resources management.

Next to the widely applied distributed (semi) physically-based based hydrological models, data driven techniques are increasingly being applied for flow forecasting. In particular, flow forecasting with artificial neural network (ANN) models has been accepted as a good alternative to forecasting with hydrological and hydrodynamic models (ASCE, 2000a,b). ANN models extract the relationship between the inputs and outputs of a process, without the physics being explicitly provided. These models need only a limited number of input variables, such as discharge and rainfall, while, distributed (semi) physically-based based models need a large number of additional parameters to be provided, such as flow resistance, cross-sections, groundwater flow characteristics, etc. these parameters are difficult to measure or to estimate, mainly because of strong spatial and temporal variability. In addition to this, ANN models are computationally fast and reliable,

which makes them very suitable for real-time applications, such as flood forecasting and early warning. Their disadvantages are related to the interpretation of the ANN structure ("black box"), and on their extrapolation capacity (Minns and Hall, 1996). It is important to highlight that the ANN solution is obtained through an optimisation process validated through trials and errors (ASCE, 2000a,b; Brath et al., 2002; Brath and Rosso, 1993). Recently, researchers have been exploring the use of different pre-processing approaches for inclusion of additional hydrological knowledge as input to ANN models to improve the hydrological representation and generalisation (Corzo and Solomatine, 2007a,b; Corzo et al., 2009).

The ANN models can be setup with limited number of input variables, but comprehensive number of records is needed. This is required, because data-driven methods have limited capability to provide accurate forecasts of events that are outside the range of the data set. On the other hand, when excessive numbers of variables are used as input the most correlated variables dominate the model and therefore it is not possible to use all the physical knowledge or measurements available. This is normally solved by pre-processing techniques aimed at reduction of the input space by selecting the most sensitive variables (Bowden et al., 2005a,b). A problem in the implementation of a big river basin is the high number of variables that an ANN model should manage, and therefore most of the studies found seem to deal only with small river basins. However, the recent work of Lin et al. (2006) shows the potential of ANN models when applied to large scale hydrological prediction.

The use of distributed rainfall input in ANN models is not new, as demonstrated by several examples from literature. Campolo et al. (2003) used distributed rainfall measured at several rain gauges; whereas Dawson et al. (2006) used a set of peripheral catchment weather station records. The modelling approach presented in this paper follows the principle of exploring different ways of using and adapting spatial precipitation in order to analyze the ANN model results in forecasting flows.

The pre-processing applied includes different methods of spatial and time integration of the rainfall data, on the basis of flow path and travel time information. This analysis is done for flood forecasting in the Ganges river basin.

2 Artificial Neural Networks (ANNs)

This study is based on the application of ANN multi-layer perceptron (MLP) networks trained with gradient based methods. The basic structure of the ANN model used can be seen in Fig. 1, where "neurons" represent linear or non-linear combinations of the input and weights. The mapping of input output requires to find the right weights in the neurons structure. The optimization of the weights is done by minimizing the mean square error of the difference between

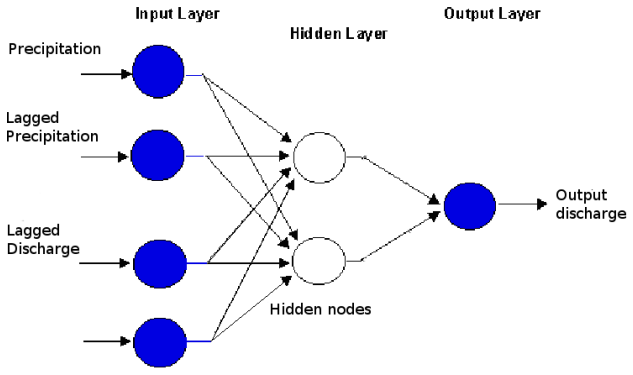

Fig. 1. Multi-layer perceptron (source: Abrahart and See, 2000).

the results of the ANN and the observed information. A detailed description of ANN modelling can be found on the publication made by the ASCE in the year 2000.

The determination of the weights in the ANN models ("training" phase), is done by minimising the mean square error between the measured discharge and the forecasted by the ANN model. In this study the Levenberg-Marquardt (LM) algorithm is used (Levenberg, 1944). This algorithm is an iterative technique that locates the minimum of a multivariate function that is expressed as the sum of squares of non-linear real-valued functions. The LM algorithm is a blend of gradient decent and Gauss-Newton iteration.

The data sets were divided in training (321 data sets) and a validation set (118 data sets). These training and validation data sets had small variations in the different experiments according to the lags of the variables. In addition to the previously observed discharge, the spatially distributed precipitation input was formulated based on the travel time and flow length information, as described in the following section. Although there are many variables that seem to be part of the physics, a selection following the ideas of Bowden et al. (2005a) was applied.

3 Travel time and flow length information

The input data to the set of ANN models of the Ganges river, explored here, consists of discharge and precipitation. The precipitation data cannot be applied directly, because the large spatial extent of the Ganges basin introduces great time lags between the rainfall occurrence and the moment it contributes to the river flow close to the border between India and Bangladesh, which is the target location for flow forecasting in this work. Pre-processing experiments to introduce these time lags are explained in the following section. This pre-processing is based on GIS analysis to estimate flow length and travel time.

The flow length describes the distance from any point in the river basin to the basin outlet. Such distance is measured along the flow paths determined from the topography.

In GIS, the flow length of an arbitrary pixel is determined by summing the incremental distances from centre-to-centre of pixel along the flow path from the selected pixel to the outlet pixel. The concept of flow length is an important issue to hydrologists. When it rains, a drop of water landing somewhere in the basin must first travel some distance before reaching the outlet. Assuming constant flow velocities the pixel with the greatest flow length to the outlet represents the hydrologically most remote pixel. So, the time of concentration can be obtained through flow length divided by the flow velocity. Therefore the time of concentration indicates how much time is required for the entire basin to contribute to surface flow at the outlet, after a certain amount of rainfall. In watershed hydrology, there are various formulations (Izzard formula, Kerby formula, Kirpich formula, Bransby Williams equation, National Resources Conservation Service, Kinematic wave formula and etc.) to calculate time of concentration based on the nature of flow as well as availability of information and scope of work (Wanielista, 1996).

The general assumption of calculating the travel time is that a uniform velocity sustains throughout the basin, which can be interpreted as the Instantaneous Unit Hydrograph (IUH) function. IUH is defined as the flow response that would be observed at the basin outlet if a unit pulse of water were instantaneously placed uniformly over the entire river basin at a given instant. With the travel lengths known and a single uniform velocity of flow observed throughout the watershed, the travel time (t_{ti}) to the outlet for any randomly chosen pixel, i, would be given by:

$$t_{ti} = d_i / v \tag{1}$$

Where, d_i is the distance from ith pixel to the watershed outlet and v is the uniform flow velocity. This concept is exploited by using the Digital Elevation Models (DEMs) to discern the flow organisation of the watershed and its unique hydrologic signal (IUH) which is dependent on the watershed size, shape, and connectivity.

The mentioned equation (Eq. 1) neglects the velocity difference between overland flow and river flow. An improvement to this approach is expected to come from introduction of different velocities for the overland and the river. This velocity differences can be conceptually understood as coming from differences in the Mannings roughness (n) of the flow-surface encountered on the watershed versus the river channels. Velocity differences could be easily such that the river flow velocities are 10 to 100 times larger then the overland flow velocities (Moglen and Maidment, 2005). With this approach, a modified travel time can be applied:

$$t_{ti} = \frac{d_{H,i}}{v_H} + \frac{d_{C,i}}{v_C} \tag{2}$$

where, $d_{H,i}$ is the flow distance for pixel i, along the basin, $d_{C,i}$ is the flow distance for pixel i, along the river channel, v_H is the overland velocity along the watershed, v_C is the velocity along the river channel. If we consider the river basin as a system and its objective it is to drain water as quickly as possible, then the presence of channel pixels with high travel velocities represents efficiency within the system, as indicated by the small travel times associated with these pixels.

An important consideration is the expected reduction in the ANN processes to be represented due to the preprocessing transformation of the input precipitation.

4 Case Study: flood forecasting in Bangladesh

4.1 Study area and problem description

Bangladesh is a low-lying country located at the confluence of three major rivers: the Ganges, the Brahmaputra and the Meghna (Fig. 2). About 92% of the catchment area of these rivers is located outside the country (Jakobsen and Bhuiyan, 2005) and 80% of the annual rainfall occurs in the monsoon season from June to September (Mirza, 2002). Thus huge cross-border monsoon flows, in addition to discharges from local rainfall are drained through Bangladesh into the Bay of Bengal. In many occasions, the volume of generated runoff exceeds the capacity of the rivers, causing serious flooding in Bangladesh.

The river Ganges originates as the Bhagirathi from the Gangotri Glacier in the Uttaranchal Himalaya and joins the Alaknanda near Deoprayang to form the Ganga. From there, the Ganges flows across the large plains of north India and empties in to the bay of Bengal after dividing up into many distributaries (Fig. 2). The source of the Ganges is at an elevation of 7010 m. The river has a bed slope of about 1:10 000 in the stretch between the Allahabad and the Banaras. From the Banaras onward to the Calcutta the bed slope changes from 1:12 000 to about 1:20 000. Along its great length, the Ganges passes through Bangladesh, which is almost everywhere flat. The total length of the river is about 2507 km. The Ganges has nine main sub-basins (Chambal, Betwa, Yamuna, Ramganga, Sone, Karnali, Gandak, Bagmati and Kosi). The total basin area is 907×10^3 km^2. The first five tributaries originate in India and the last four tributaries join the Ganges from Nepal. Among these nine tributaries, the Yamuna is the most important one, which drains about one–third of the entire Ganges basin. The Kosi and Karnali drain about 12% and 9% of the basin, respectively (Mirza, 1997). In 1987, 1988, 1998 and 2004, several serious floods occurred in Bangladesh, which are good examples of the need for a forecasting and warning system as an essential tool to reduce flood damage.

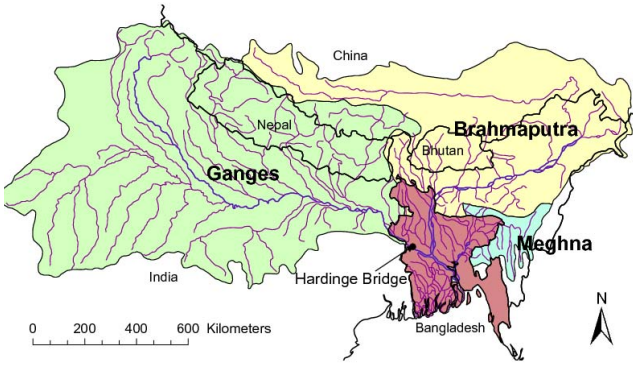

Fig. 2. Ganges-Brahmaputra-Meghna Basin.

In Bangladesh, there are about 52 forecasting stations where 24, 48, and 72-h forecasts are made every day (FFWC, 2007). The lead-time of the model for the Northern part of the country is shorter. Improved model performance, in principle, can be achieved through regional cooperation among the countries that share the river basins in question, particularly for exchanging flood information and data sharing. The actual model in Bangladesh consists of three modules: (a) a rainfall-runoff modules (NAM), (b) a one-dimensional finite difference hydrodynamic model (HD) based on St. Venant equations, (c) an updating module. The updating module analyses measured and simulated water levels and discharges up to the time of forecasting in order to eliminate amplitude and phase errors which could influence the forecast results (Chowdhury, 2000). However the developed model can only forecast water level 3-day ahead for the Southern part of the country.

Due to regional limitation on the availability of data the current physical based modelling of the region is more complex. The regional initiative of the World meteorological Organisation (WMO) and International Centre for Integrated Mountain Development (ICIMOD) is establishing a regional data exchange in the Hindukush Himalayan counties (Bangladesh, India, Bhutan, Nepal, and Pakistan). However, until now there is no significant improvement in data shearing, which hampers the expansion of the model boundary further upstream of the Brahmaputra and the Ganges.

Therefore, the major contribution of this case study is to explore the possibility for provision of accurate flood forecasts for the river Ganges, close to its entry point from India into Bangladesh. If ANN models can provide sufficiently accurate forecasts several days ahead at this location, the lead-time for flood forecasting and warning within Bangladesh can be extended, and the subsequent flood emergency measures can be better planned and executed. The setup of the ANN models is done by making use of freely available remotely sensed (satellite) rainfall data and water level measurement records.

4.2 Tropical Rainfall Measurement Mission (TRMM)

Inputs to the ANN model are of critical importance. Satellite derived rainfall data of Tropical Rainfall Measurement Mission (TRMM, NASDA, 2001) is providing 3 hourly rainfall, which is very promising. The reliability of the remotely sensed data is always facing challenges, but it is found from various validation projects that precipitation radar of TRMM is producing error within acceptable range. However the accuracy of such data, when compared with rainfall observed from ground stations varies from place to place and it has already been tested over Bangladesh. The study proves that the correlation coefficient (R) is more than 0.773, which is sensible to certain extent (Akhtar, 2006).

4.3 Data preparation

In this study, ANN is used to predict the river flow at Hardinge Bridge (close to the entry point of the Ganges into Bangladesh, see Fig. 2), utilising (i) the calculated discharge from water level gauge which is located at the same location (with known rating curve for conversion of water levels into discharges) and (ii) satellite based rainfall for the entire catchment.

The documentation of satellite-derived rainfall is provided in Huffman et al. (2007) (ftp://meso-a.gsfc.nasa.gov/pub/trmmdocs/) and data from the Tropical Rainfall Measurement Mission are available in a regular 0.25 degree lat/long grid (TRMM V6.3B42). The extracted data has resolution of $0.25° \times 0.25°$ (approximately: Lat. 27.7 km, Lon. 25.2 km), with a temporal resolution of three hour, which is accumulated to daily data for the period from 2001 to 2005. A data processing tool has been developed to generate time series for each pixel or areal average rainfall depending on the users requirement.

A DEM (digital elevation model) has been extracted from the website of Shuttle Radar Topography Mission (SRTM, http://www2.jpl.nasa.gov/srtm/) with a grid size of 1.0 km. The extracted DEM is then smoothened with ArcGIS and then modified by FAP-19 (Flood Action Plan) information. Sinks (and peaks) are isolated grids with missing or abnormal values that often occur due to the resolution of the data or rounding of elevation to the nearest integer value. Sinks can generate discontinuity on the process of drainage network derivation. The DEM has also been adjusted with known river lines and catchment boundaries.

Five years (from 2001 to 2005) of daily river flow data was generated using rating equations based on the continuous observed water level and validation discharge measurements of Bangladesh Water Development Board (BWDB).

Both discharge and rainfall data from 2001 to 2003 were selected for training, the data from 2004 to 2005 were selected for verification. Only high flow data were used to reduce the influence of low flow condition as the study aims

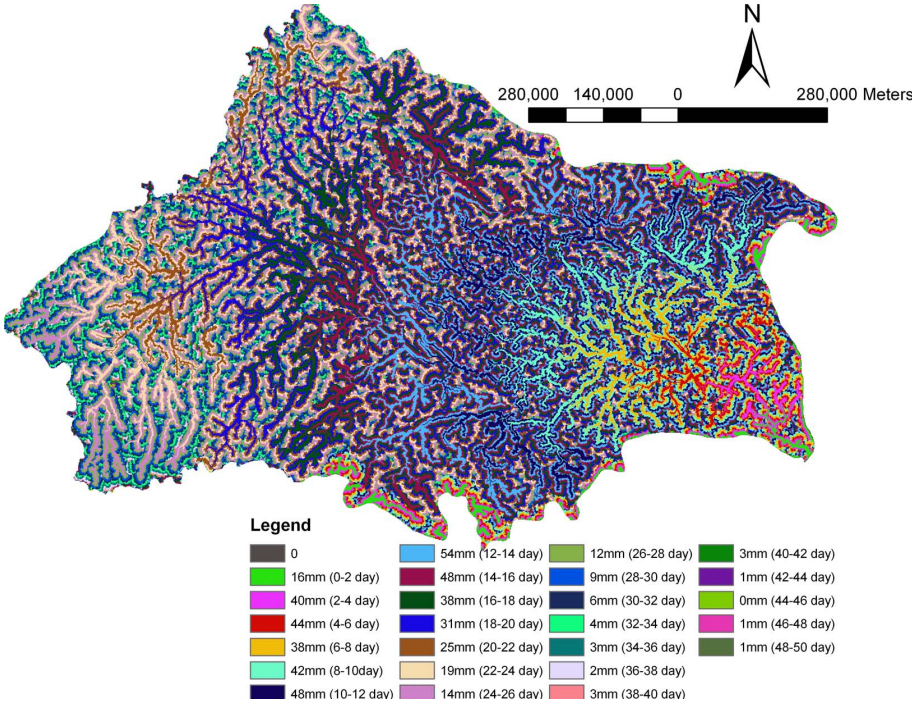

Fig. 3. Example of precipitation cluster-area analysis for 9th of July 2003.

to develop flood forecasting model. It is important to highlight that low flow are less dominated by the precipitation acting on previosu days and instead determined by the groundwater dynamics in the region. Therefore, the selection of data is reduced. The number of samples used for training to 321 and 118 for verification. Statistical comparisons of the mean, standard deviation and probability distribution of the data sets has been performed. The results showed a good agreement between the training and verification data sets.

In order to take into account the spatiotemporal distribution of rainfall as an input to the ANN model, a GIS-based analysis of the satellite rainfall data has been carried out which resulted in areal clusters of rainfall data time-series, each with different lag time. The areal clusters have been defined according to their calculated travel time to the outlet of the catchment. In conventional approaches to such clustering usually one average velocity is assumed for both overland and channel flow. A new method has been applied here that assumed different velocities for these two flow components (Sect. 3, Eq. 2). The velocity along the river channel (v_C) is assumed 1 m/s and the overland velocity along the watershed is assumed 40 times smaller (v_H=1/40 m/s), which is within the range indicated by Moglen and Maidment (2005), (10–100 times smaller).

Comparison of the areal clusters between the conventional and the new method that has been applied here is presented in Fig. 4a and b. The spatial distribution of equal travel time areas is obviously coinciding with the drainage pattern by using the new method.

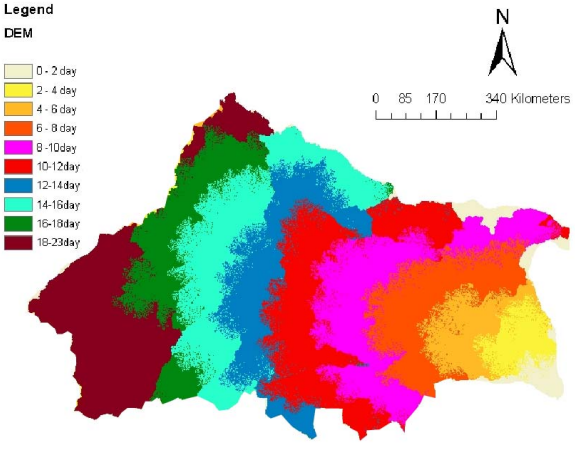

Fig. 4a. Catchment delineation by travel time: velocity in channel and flood plain.

In the process of building the best ANN model, a 10-fold cross validation is performed. Figure 5 shows the training and validation data sets, together with one of the 10 fold cross validation data sampled. It can be seen that the maximum and minimum values are in the same bounds. Although the distribution seems shifted, its shape shows good agreement with the training data set.

Subsequently to this analysis, a composite of the rainfall time series was also generated by adding all the individual areal rainfall time series, keeping in mind their lag time

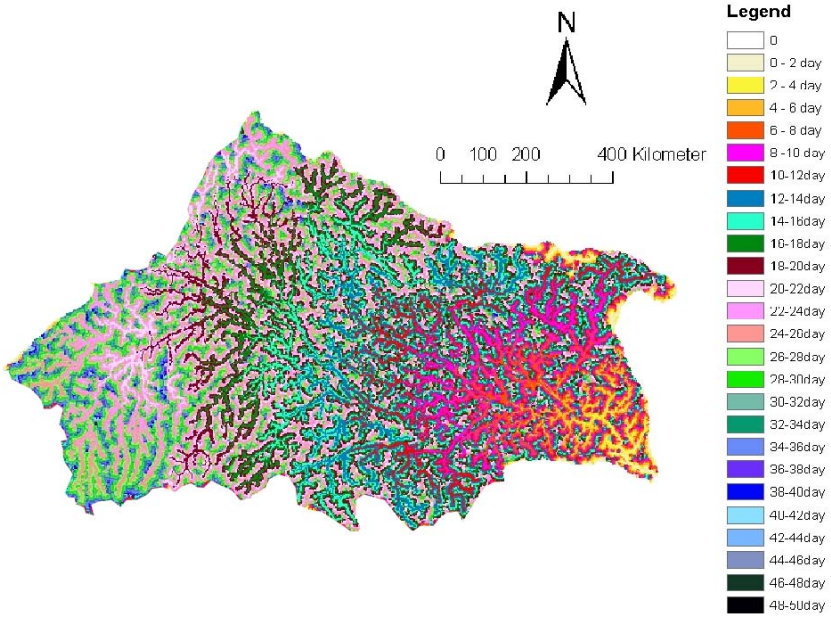

Fig. 4b. Catchment delineation by travel time: conventional average velocity method.

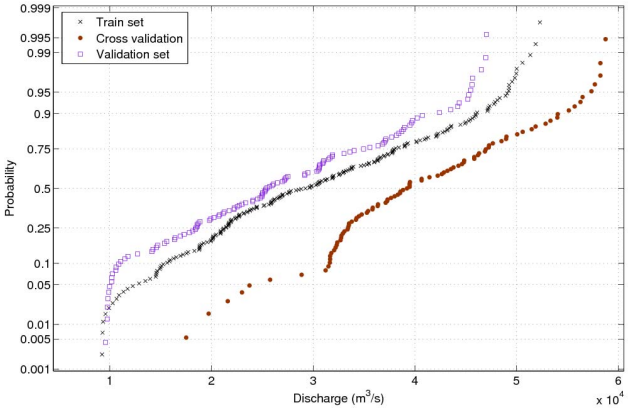

Fig. 5. Probability distribution of training, cross validation and validation data sets.

(following the calculated travel times). It has been found from correlation analysis that composite rainfall derived from the new two-velocity travel time approach, demonstrates better correlation with the outlet discharge compared to the conventional one.

Furthermore, in order to improve the performance of the ANN model the usual practice is to consider previous discharges, as they contain more information than rainfall for larger basins. Further correlation analysis has been carried out to select the number of previous discharges, which showed that 1 day previous discharge should be included in the input as it contains 99% information of the present discharge.

4.4 Modelling

4.4.1 ANN Setup

A number of scripts have been prepared to pre-process and analyse the models using the ANN toolbox of Matlab. The optimal structure of the model was analysed by testing the training data set with different hidden nodes, ranging from 1 to 10. Various combinations of input data are tested to compare and evaluate the sensitivity of the ANN. Fifteen different options were tested in an extensive analysis carried out by Akhtar (2006). Most of the result that excluded precipitation had the disadvantage of performance reduction on the high flow situations. The results found are in accordance with the studies done by Toth (2008); Elshorbagy et al. (2009). In this paper the options with most important results are discussed as follows:

[A] Only Discharge is used as input data (Only Q). Two discharge time series, one of the present day and one of the day before, are used as input data for the ANN.

[B] Two discharges and 25 average areal rainfall with lagged time as input data (RF+Q). The area-average rainfall time series are used with their respective lag time along with two discharge (present and 1 day before) time series.

[C] Two discharges, and lagged sum of the rainfall as input data (TRF+Q). All the 25 lagged area-average rainfall time series are added to form one, composite rainfall time series. Again two discharge (present and 1 day before) time series are used. The idea behind this reduction of the number of rainfall time series is to test if ANN may be able to perform better with fewer input data series.

Table 1. Root mean square error of the verification results for different options.

Option	Description	1 day	2 day	3 day	4 day	5 day
A	Only Q	952	2108	3281	4432	5375
B	RF+Q	1117	2578	3730	6211	6244
C	TRF+Q	968	2223	3250	4322	5290
D	TRF+Q+Act.ET	968	2176	3390	4543	5427

[D] Two discharges, actual evapotranspiration and lagged sum of rainfall as input data (TRF+Q+Act.Evp). Actual evapotranspiration is used in this option as an input along with the other inputs of option [C], to take into account the evaporation losses. The actual evapotranspiration data were generated by a quasi-physically based model built with the Soil and Water Assessment Tool (SWAT).

4.4.2 Performance analysis

Training and verification has been performed for all different ANN model setups (option A to D). To measure the performance of the models, four criteria are selected, which are Root Mean Square Error (RMSE), Normalised Root Mean Square Error (NRMSE), Mean Average Error (MAE) and Correlation coefficient (CoE, Nash and Sutcliffe, 1970). Aditionally the PERS index, which is a more conventional measure for time series is included (Eq. 6). Their values are supplied in the following tables (Tables 1 to 6). The errors calculated numerically are supplemented by visual inspection of the hydrographs (Fig. 6) based on verification set. Root mean square error is calculated as:

$$\mathrm{RMSE} = \sqrt{\frac{\mathrm{SSE}}{n}} \tag{3}$$

$$\mathrm{SSE} = \sum_{t=1}^{n} \left(Q_{\mathrm{est,t}} - Q_{\mathrm{obs,t}}\right)^2 \tag{4}$$

where Q_{obs} and Q_{est} are the values of the observed and estimated discharge, respectively. The total number of samples is represented by n and the SSE is the abbreviation for the sum of square errors. Equation (3) is used to answer what is the average magnitude of the forecast errors.

Sometimes it is important to compare two time series using a reference of statistical properties of measurements. Therefore, here we use root relative squared error (Witten and Frank, 2000), which compares the root square of the mean of squared errors with the standard deviation of measurement. This means that we can see if the average errors are outside of the standard deviation of measurements. This measure is sometimes expressed as percentage, so a value of

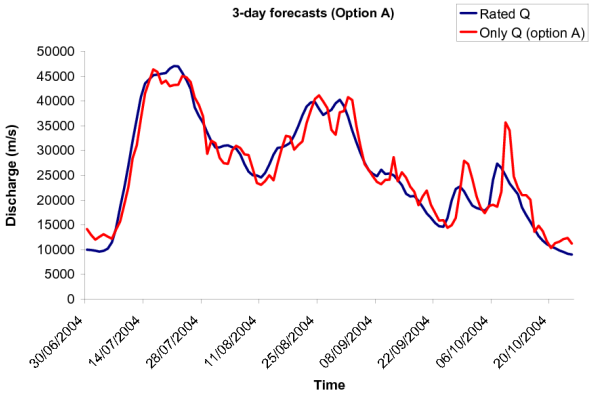

Fig. 6a. Option A, only discharge Q.

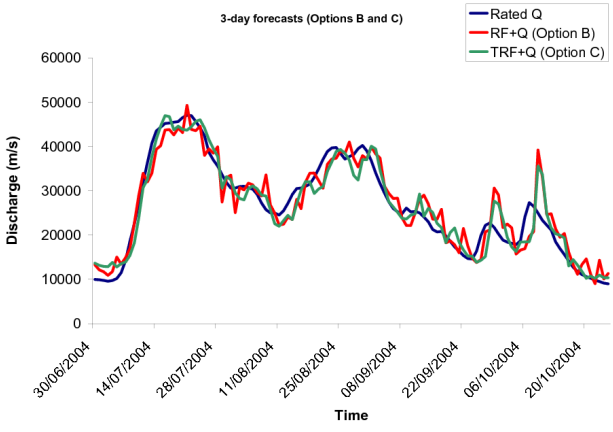

Fig. 6b. Option B (distributed rainfall and discharge (RF+Q)) and C (composite rainfall and discharge (TRF+Q)).

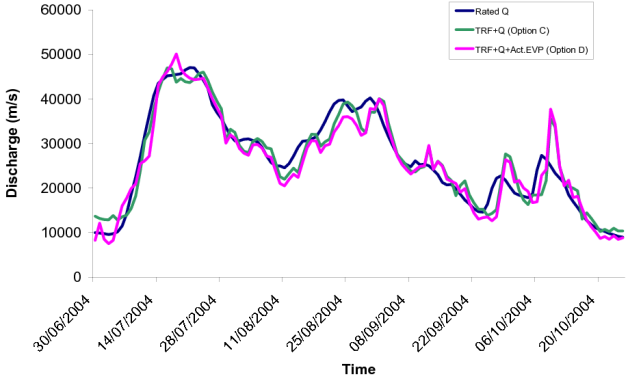

Fig. 6c. Option C (composite rainfall and discharge (TRF+Q)) and D (discharge and actual evapotranspiration, (TRF+Q+Act. Evp)).

Table 2. Normalised root mean square error of the verification results for different options.

Option	Description	1 day	2 day	3 day	4 day	5 day
A	Only Q	8.937	19.76	30.7	41.4	50.1
B	RF+Q	10.49	24.17	34.91	58.01	58.2
C	TRF+Q	9.087	20.84	30.41	40.38	49.31
D	TRF+Q+Act.ET	9.09	20.4	31.73	42.43	50.59

Table 3. Mean abasolute error of the verification results for different options.

Option	Description	1 day	2 day	3 day	4 day	5 day
	Only Q	681.4	1602	2592	3398	4288
B	RF+Q	847.7	1965	2948	4939	4891
C	TRF+Q	688.9	1676	2535	3421	4292
D	TRF+Q+Act.ET	687.9	1641	2564	3507	4443

100% means that the RMSE is in the bound of the standard deviation. If the errors are much higher than these bound values the root relative squared error will be above 100%. In this sense the root relative error is a Normalized Root Mean Square, term used in this thesis (NRMSE, Eq. 5).

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{\mathrm{SSE}}{n}}}{\sigma_{\mathrm{obs}}} \tag{5}$$

Where the value of σ is the standard deviation of the measured or observed discharges.

The persistence index (PERS) focuses on the relationship of the model performance and the performance of the naïve ("no-change") model which assumes that the forecast at each time step is equal to the current value (Kitanidis and Bras, 1980):

$$\mathrm{PERS} = 1 - \frac{\mathrm{SSE}}{\mathrm{SSE}_n} \tag{6}$$

$$\mathrm{SSE}_n = \sum_{t=1}^{n} \left(Q_{\mathrm{obs,t+L}} - Q_{\mathrm{obs,t}}\right)^2 \tag{7}$$

SSE_n is a scaling factor based on the performance of the naïve model; $Q_{\mathrm{est,t}}$ is the DDM forecast or a process-based model simulation of the next time step, $Q_{\mathrm{obs,t}}$ is the observed discharge at time t where t=1, 2, . . . , n; L is the lead time (L=1 for one day ahead forecast); and n is the number of steps for which the model error is to be calculated.

PERS is a unit that is relative to the naïve model. It can range between 1 and minus infinite (i.e. it is degrading the provided information), values above 0 indicate that the considered model is better than the naïve model (where the

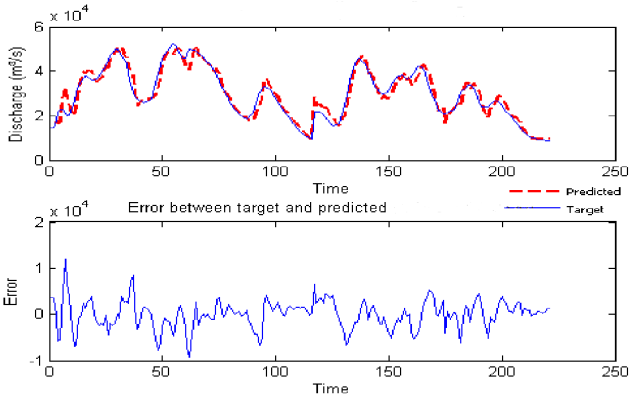

Fig. 7a. Comparison between target and predicted values together with errors (NRMSE) for Option C in training stage.

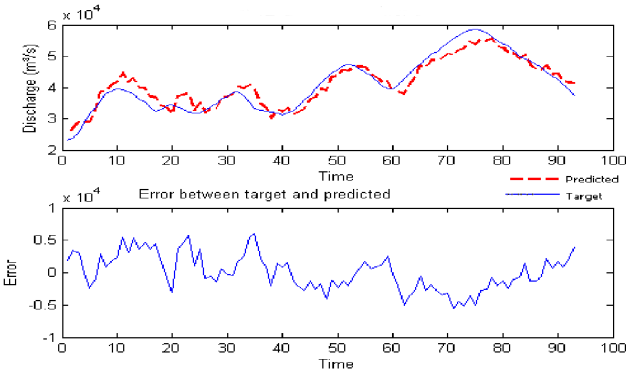

Fig. 7b. Comparison between target and predicted values together with errors (NRMSE) for Option C in cross-validation stage.

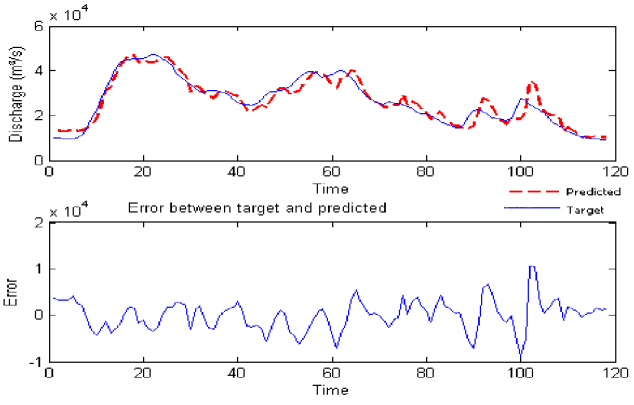

Fig. 7c. Comparison between target and predicted values together with errors (NRMSE) for Option C in verification stage.

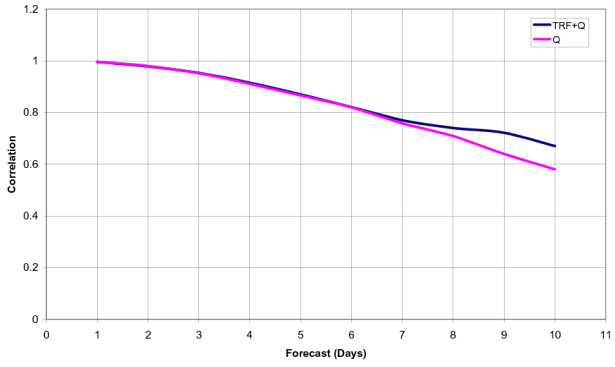

Fig. 8. Comparison of correlation coefficients for Options A (only Q) and Option C (TRF+Q) for extended forecast horizon up to 10 days.

Table 4. Correlation coefficient of the verification results for different options.

Option	Description	1 day	2 day	3 day	4 day	5 day
A	Only Q	0.996	0.98	0.952	0.911	0.866
B	RF+Q	0.995	0.971	0.938	0.821	0.817
C	TRF+Q	0.996	0.978	0.953	0.915	0.87
D	TRF+Q+Act.ET	0.996	0.979	0.948	0.906	0.863

Table 5. Mean of the forecast predictions.

Option	Description	1 day	2 days	3 days	4 days	5 days
A	Only Q	2.6544	2.6635	2.6508	2.6495	2.654
B	RF+Q	2.6536	2.6598	2.6941	2.723	2.6018
C	TRF+Q	2.6521	2.6474	2.6522	2.6517	2.6512
D	TRF+Q+Act. ET	2.652	2.6523	2.5773	2.6202	2.6715

closer to 1 the better), and negative values show less performance than the naïve model. Lauzon et al. (2006) suggest using PERS in cases when the discharge forecast is made on the basis of previous values.

For each option (A–D), five different simulations are performed to check the performance of the model for a 1-day, 2-day, 3-day, 4-day and 5-day forecast horizon. The results indicate that the model performance is showing decreasing trend with the increase of lead-time. However, from the simulated hydrographs and performance tables, it is established that forecasting performance is acceptable up to 3-days. Beyond this period, results are getting worse. To keep the discussion within limits and also to review the results meticulously, 3-day forecast (verification) results are discussed here. The results of the 3-day forecasts are shown in Fig. 6.

Option [A], with only Q time series as input, does not show large differences from the options where rainfall time series are included (Fig. 6a). Option [C] exhibits some improvement in the forecasting performances compared to Option [B], which can also be seen from the performance criteria tables (Tables 1 to 6). This indicates that large number of input parameters is not suitable to develop an ANN for a basin area like the Ganges. After inclusion of the actual evapotranspiration as a separate time series in Option [D], Fig. 6c and the performance analysis criteria (Table 1 to 6) indicate deterioration of model result. This is most likely due to very high uncertainty of the SWAT model results which were used for generating the time series of actual evapotranspiration. Assessment of the values of the performance analysis tables indicates that option [C] is most suitable for flood forecasting of Ganges basin with a 3-day forecast horizon.

Figure 7 has been introduced to visualise the performance of the model (option C) by comparison plots of training, cross-validation and verification as well as errors analysis in terms of NRMSE. It shows that there is an excellent agreement between the observed and simulated data for the training phase but the performance deteriorates in the cross-validation stage. However, model performance in the verification stage is satisfactory, as several peaks show good resemblance with observation.

Moreover, accurate timing is also important and is a critical factor in operational management and decision-making activities related to high magnitude flood events. Timing errors (phase lag) of the model results have, however, been identified from all the options. This is a common problem in neural network rainfall-runoff models and causes are still under investigation by neuro-hydrologists. One approach to this problem (as suggested by Abrahart et al., 2007) is to use a time-error correction procedure as an integrated part of the neural network optimisation process. However, at the time of this writing the full description of this procedure was not available for testing.

Note that Fig. 7a is not completely continuous in time and in sample 116, 28 October 2001, there is a gap of low flows, so sample 117 corresponds to 9 July 2002. Figures 7b and c are continous in time for all the time series plot. Their time frame is between 5 July 2003 till 5 October 2003 and 1 July 2004 till 26 October 2004, respectively.

4.4.3 Expanded forecast horizons

The similar performance of option [A] to options [B,C,D] confirms that for short- to medium-term forecasts and for large rivers, ANN can provide good forecasts based only on current and previous discharge measurements. For long-term forecasts it is expected that this predictability on the basis of real-time discharge measurements decreases. In that case rainfall information and rainfall-runoff modelling would become more important. To investigate whether this applies to the Ganges case study, forecasting horizon is increased from 5 days to 10 days for option [A] (only Q), and option [C] (TRF + Q). The correlation performance of the two ANNs is presented in Fig. 8. It shows that for forecast horizons from 7–10 days the inclusion of the rainfall as an input to

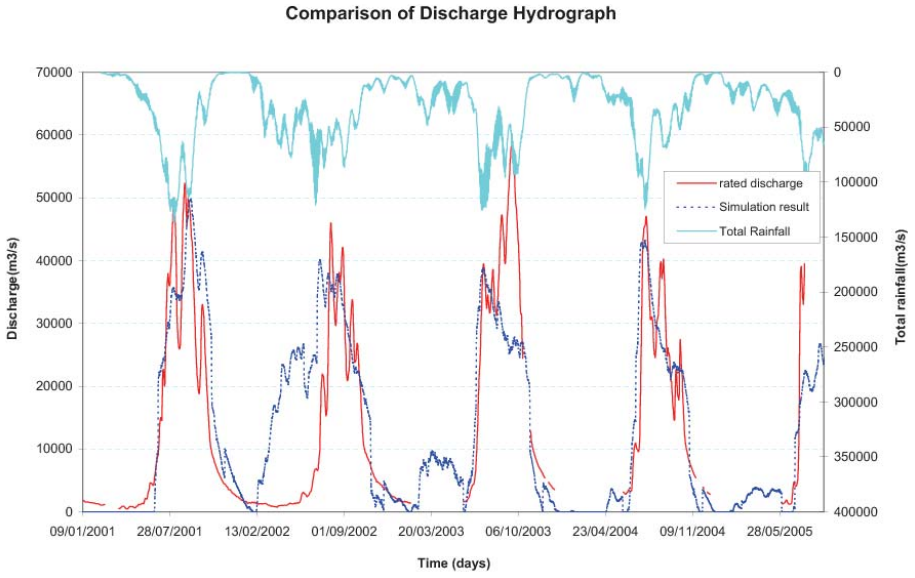

Fig. 9. Hydrograph comparison between SWAT model simulation results and the rated discharge.

Table 6. PERS index of the verification results for different options.

Option	Description	1 day	2 days	3 days	4 days	5 days
A	Only Q	0.647	−0.7107	−3.3473	−7.4138	−11.5369
B	RF+Q	0.5783	−1.2459	−3.6998	−2.0217	−12.1628
C	TRF+Q	0.6836	−0.6696	−2.5676	−5.3076	−8.4465
D	TRF+Q+Act. ET	0.6832	−0.6	−3.23	−5.615	−8.9429

the ANN (option [C]), improves the forecasting performance. Note that with the increase of forecasting horizon, the performance of the model is getting worse and the performance of the forecast beyond three days is not acceptable, whatever improvement can be seen by including the rainfall. However this exercise proves that composite rainfall (which is satellite driven) along with previous discharge can help to build better ANN, for longer forecasting lead times.

Figure 9 shows the rated and simulated discharge together with total rainfall over the catchment. This simulated discharge is the best performed simulation output of our SWAT model setup. The results presented in this paper were compared to the Soil Water Assessment Tool (SWAT) model of the basin. Although a number of complex optimization algorithms were tested, the SWAT model results did not achieve the performance of the ANN model results presented here (van Griensven et al., 2007). The errors for the SWAT model results were a RMSE of 11600, a MAE of 8930 and a Correlation coefficient of 0.359. This is most likely due to the large spatial extension of the basin and lack of information on catchment parameters data, as required for the SWAT model.

5 Conclusions

An ANN flow forecasting model that makes use of spatial precipitation obtained from pre-processing based on hydrological concepts of travel time and flow length has been developed for the Ganges river basin. This was done by combining ground station flow measurements with satellite derived rainfall and DEMs, and hydrological GIS analyses. A new method for estimation of travel time has also been applied and tested with artificial neural networks.

From the analysis of various options for input data, it was revealed that the forecasted discharge is highly influenced by the previous discharge input data, because of their strong correlation. This was expected, particularly because of the exceptionally large spatial scale of the river. For this reason, different combinations of rainfall input did not influence the model much for the short forecast horizons, For forecast horizons of 7 to 10 days inclusion of rainfall information in addition to discharge data, improves the ANN model performance.

Accurate timing is a critical factor in operational management and decision-making activities related to high magnitude flood events. Timing errors (phase lag), however, have been identified, which is a common problem in ANN rainfall-runoff models. Inclusion of a time-error correction procedure as an integral part of the ANN optimisation process can improve the model performance.

The finally selected ANN model shows some disagreement with the observed values, especially during the peak discharge. The causes may be hidden within the unknown processes of the catchment, unverified rainfall data and rated discharge, etc. For further improvement of the model, it is essential to investigate the target value (rated discharge),

where discharge is generated from a conventional rating curve.

The method that has been used to calculate travel-time and delineate the clustered areas, from which water can reach the outlet of the basin within a certain range of time, can be further improved. In this study only two different velocities were assumed, one for channels and the other for land surface flow, but in reality the velocity is not the same in all rivers, even velocity differs from reach to reach of a river. Surface-runoff velocity is also considered constant for all over the basin, irrespective of land use and land slope, which is contrary of the physical conditions. More detailed velocity estimates by considering the land use characteristics can help to improve the model performance.

From the overall analysis it is found that one-day previous discharge along with composite rainfall (derived from GIS-based travel time calculation) gives the best result compared to other options. This shows that remote sensing techniques and data driven modelling can be combined successfully to prepare a spatially distributed ANN for flow forecasting of large-scale river basins like the Ganges.

The study presented is a particular case and the findings here can be tested and extended in further research. This methodology will be explored and benchmarked together with other methodologies reported in literature, specially on smaller catchments in order to see the effect of distributed rainfall input more clearly.

Acknowledgements. This study was possible due to the financial support provided by the Delft Cluster project in the Netherlands. Many thanks to Editor N. Verhoest for his careful editing which greatly improves the manuscript.

Edited by: N. Verhoest

References

Abrahart, R. J. and See, L.: Comparing neural network and autoregressive moving average techniques for the provision of continous river flow forecasts in two contrasting catchments, Hydrol. Process., 14, 2157–2172, 2000.

Abrahart, R. J., Heppenstall, A. J., and See, L. M.: Timing error correction procedure applied to neural network rainfall-runoff modelling, Hydrolog. Sci. J., 52, 414–431, 2007.

Akhtar, M. K.: Flood Forecasting for Bangladesh with satellite Data, Msc Thesis, UNESCO-IHE, Delft, the Netherlands, 134 pp., 2006.

ASCE: Task Committee on Application of Artificial Neural Networks in Hydrology, Artificial Neural Networks in Hydrology, II:Hydrologic Application, J. Hydrol. Eng., 5, 124–136, 2000a.

ASCE: Task Committee on Application of Artificial Neural Networks in Hydrology, Artificial Neural Networks in Hydrology. I: Preliminary Concepts, J. Hydrol. Eng., 5, 115–123, 2000b.

Bowden, G. J., Dandy, G. C., and Maier, H. R.: Input determination for neural network models in water resources applications. Part 1-background and methodology, J. Hydrol., 301, 75–92, 2005a.

Bowden, G. J., Dandy, G. C., and Maier, H. R.: Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river, J. Hydrol., 301, 93–107, 2005b.

Brath, A. and Rosso, R.: Adaptive calibration of a conceptual model for flash flood forecasting, Water Resour. Res., 29, 2561–2572, 1993.

Brath, A., Montanari, A., and Toth, E.: Neural networks and non-parametric methods for improving real-time flood forecasting through conceptual hydrological models, Hydrol. Earth Syst. Sci., 6, 627–639, 2002, http://www.hydrol-earth-syst-sci.net/6/627/2002/.

Campolo, M., Soldati, A., and Andreussi, P.: Artificial neural network approach to flood forecasting in the River Arno/Une approche à base de réseau de neurones artificiels pour la prévision des crues du fleuve Arno, Hydrolog. Sci. J., 48, 381–398, 2003.

Chowdhury, M.: An assessment of flood forecasting in Bangladesh: the experience of the 1998 flood, Nat. Hazards, 22, 139–163, 2000.

Corzo, G. and Solomatine, D.: Knowledge-based modularization and global optimization of artificial neural network models in hydrological forecasting, Neural Networks, 20, 528–536, 2007a.

Corzo, G., Solomatine, D., Hidayat, de Wit, M., Werner, M., Uhlenbrook, S., and Price, R.: Combining semi-distributed process-based and data-driven models in flow simulation: a case study of the Meuse river basin, Hydrol. Earth Syst. Sci. Discuss., 6, 729–766, 2009, http://www.hydrol-earth-syst-sci-discuss.net/6/729/2009/.

Corzo, G. A. and Solomatine, D. P.: Baseflow separation techniques for modular artificial neural networks modelling in flow forecasting, Hydrolog. Sci. J., 52, 491–507, 2007b.

Dawson, C., See, L., Abrahart, R., and Heppenstall, A.: Symbiotic adaptive neuro-evolution applied to rainfall–runoff modelling in northern England, Neural Networks, 19, 236–247, 2006.

Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D.: Experimental investigation of the predictive capabilities of soft computing techniques in hydrology., Technical Rep., 49 pp., 2009.

FFWC: Flood Forecasting and Warning Centre; Annual Flood Report 2007, Technical Rep., Dhaka, Bangladesh, 4, 86 pp., 2007.

Huffman, G., Adler, R., Curtis, S., Bolvin, D., and Nelkin, E.: Global rainfall analyses at monthly and 3-hr time scales, Measuring Precipitation from Space: EURAINSAT and the Future, Springer, Dordrecht, The Netherlands, 28, 291–305, 2007.

Jakobsen, F., Hoque, A. K. M. Z. , Paudyal, G. N., and Bhuiyan, S.: Evaluation of the Short-Term Processes Forcing the Monsoon River Floods in Bangladesh, Hydrolog. Sci. J., 30, 389–399, 2005.

Kitanidis, P. K. and Bras, R. L.: Real-Time Forecasting With a Conceptual Hydrologic Model: Analysis of Uncertainty, Water Resour. Res., 16, 1025–1033, 1980.

Lauzon, N., Anctil, F., and Baxter, C. W.: Clustering of heterogeneous precipitation fields for the assessment and possible improvement of lumped neural network models for streamflow forecasts, Hydrol. Earth Syst. Sci., 10, 485–494, 2006, http://www.hydrol-earth-syst-sci.net/10/485/2006/.

Levenberg, K.: A method for the solution of certain problems in least squares, Quart. Appl. Math, 2, 164–168, 1944.

Lin, J., Cheng, C., and Chau, K.: Using support vector machines for

long-term discharge prediction, Hydrolog. Sci. J., 51, 599–612, 2006.

Minns, A. W. and Hall, M.: Artificial Neural Networks as rainfall-runoff models, Hydrolog. Sci. J., 41, 399–417, 1996.

Mirza, M.: The runoff sensitivity of the Ganges river basin to climate change and its implications, J. Environ. Hydrol., 5, 1–13, 1997.

Mirza, M.: Global warming and changes in the probability of occurrence of floods in Bangladesh and implications, Global Environmental Change, 12, 127–138, 2002.

Moglen, G. E. and Maidment, D. R.: Digital Elevation Model Analysis and Geographic Information Systems, Encyclopedia of Hydrological Sciences, Part 2., Hydroinformatics, Vol. 1, John Wiley & Sons, Ltd., Chichester, England, UK, 239–255, 2005.

NASDA: TRMM data users Handbook, Technical Rep., 226 pp., 2001.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models Part 1- A Discussion Principles, J. Hydrol., 10, 282–290, 1970.

Toth, E.: Data-Driven Streamflow Simulation: The Influence of Exogenous Variables and Temporal Resolution, in: Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications, edited by: Abrahart, R. J., Linda, M. S., and Dimitri, P., Solomatine, Berlin Heidelberg, Germany, 2008.

van Griensven, A., Akhtar, M. K., A., Haguma, D., Sintayehu, R., Schuol, J., Abbaspour, K., van Andel, S., and Price, R.: CATCHMENT Modeling using Internet-Based Global Data, 4th SWAT conference UNESCO-IHE Delft, The Netherlands, 2007.

Wanielista, M. P., Kersten, E., and Robert, R.: Hydrology: Water Quantity and Quality Control, John Wiley and Sons, Ltd., New York, USA, 567 pp., 1996.

Witten, I. H. and Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 525 pp., 2000.