METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **22**: 610–616 (2015)
Published online 27 March 2015 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/met.1495

Using the mutual information technique to select explanatory variables in artificial neural networks for rainfall forecasting

Mukand S. Babel,* Girish B. Badgujar and Victor R. Shinde

Water Engineering and Management, Asian Institute of Technology, Klong Luang, Pathum Thani, Thailand

ABSTRACT: The artificial neural network (ANN), a data-driven approach, is a powerful tool for forecasting rainfall. However, selecting the appropriate explanatory variables in order to develop ANN models for this purpose is a major challenge. Recent studies in various fields have highlighted the usefulness of the mutual information (MI) technique in identifying explanatory variables for application in non-linear problems, which, however, has largely been unexplored in forecasting rainfall. The present study was carried out to fill this knowledge gap. Three ANN models were developed, with different explanatory variables, to forecast the rainfall in Mumbai, India. Model A used temporal data of past rainfall events, Model B used selected meteorological data apart from rainfall and Model C used those variables identified by the MI technique. When the results of Model C were compared with those of Models A and B, a reduction of 5.79 and 4.11% in normalized mean square error, respectively, 16.66 and 12.90% improvement in efficiency index, respectively, and 3.22 and 4.24% reduction in the root mean square error, respectively, were observed. Thus, this study highlights the superiority of the MI technique in selecting explanatory variables for ANN modelling, not only because of the enhanced performance of the model with respect to various indicators but also because this performance has been achieved with a simple ANN architecture.

KEY WORDS artificial neural networks; explanatory variables; meteorological variables; modelling; mutual information; rainfall forecasting

Received 27 September 2013; Revised 26 November 2014; Accepted 3 December 2014

1. Introduction

Rainfall forecasting is an integral component in the development of all water-related disaster response mechanisms, and more so for floods and droughts. While droughts are a creeping disaster insofar that their effects are felt over a long period, floods are usually more rapid in nature and can cause considerable destruction, especially in urban areas where the ground surface is largely impervious. In the decade prior to 2011, flooding was the most common type of disaster globally, responsible for almost half of all the victims of natural disasters, and for economic losses of nearly US $185 billion (EM-DAT, 2011). Apart from the economic losses, both direct and indirect, floods also hamper daily activities because of their tendency to disrupt traffic and transportation systems. Because the population density in urban areas is typically quite high, even a small flood can cause significant damage. In order to mitigate/prevent flood hazards, it is important to have an appropriate early flood warning system. Such a warning system typically has three components: forecasting, transforming the forecast into a warning and transmitting the warning to local decision makers, and converting the warning into remedial action (United Nations Inter-Agency Secretariat of the International Strategy for Disaster Reduction (UN/ISDR), 2004).

Forecasting rainfall is one of the most difficult, while at the same time integral, processes of a flood warning system (French *et al.*, 1992; Hung *et al.*, 2009). The difficulty arises because a rainfall event depends on a large number of variables, including pressure, temperature, wind speed, wind direction and relative humidity. Rain is the outcome of different types of physical interactions among these variables. Approaches to forecasting rainfall have evolved over time and can generally be classified into two groups: ‘physical modelling’ wherein the rainfall process is studied in order to model the underlying physical laws and ‘systems theoretical modelling’ that attempts to recognize the rainfall patterns based on various features of the system (Luk *et al.*, 2000). While the former is considered cumbersome because of the vast range of data and the sophisticated mathematical tools that are required for calibration and computation (Hong, 2008), the latter is increasingly becoming popular. One of the most widely used tools for pattern recognition is the artificial neural network (ANN), and in recent years, various studies have proved the strength and suitability of the ANN in forecasting rainfall (for example Nasseri *et al.*, 2008; Hung *et al.,* 2009; Srivastava *et al.*, 2010; Wu *et al.*, 2010; Wu and Chau, 2013).

The ANN is capable of recognizing a relationship from a given pattern, and because of this property, it is of immense use in non-linear modelling, pattern recognition and classification problems. The ASCE (2000a) has published a comprehensive review of various ANN applications in hydrology, and the study concluded that the ANN has had a significant impact in solving a variety of real-time problems. The application of ANN in various fields of hydrology such as rainfall-runoff modelling (e.g. Adamowski *et al.*, 2013), water quality modelling (e.g. Jiang *et al.*, 2013), groundwater studies (e.g. Mohanty *et al.*, 2013), sediment yield estimation (e.g. Mount and Abrahart, 2011), reservoir operations (e.g. Sattari *et al.*, 2012) and water demand forecasting (e.g. Babel and Shinde, 2011) has been on the rise in recent years. In one of the earliest studies on rainfall forecasting, French *et al.* (1992) developed an ANN model to forecast the

* Correspondence: M. S. Babel, Water Engineering and Management, Asian Institute of Technology, P. O. Box No. 4, Klong Luang, Pathum Thani 12120, Thailand. E-mail: msbabel@ait.asia

rainfall intensity field with a lead time of 1 h. The intensity field at the current time step was used as an explanatory variable to forecast the intensity field at the next time step. This study showed that although the ANN models performed slightly better than the persistence forecasting models for the training data set, the performance with the testing data set was not satisfactory. Despite this, other researchers continued to use the ANNs, with different explanatory variables, to forecast complex rainfall processes. Navone and Ceccatto (1994) developed an hybrid ANN model to predict the summer monsoon rainfall in India using the relevant parameters corresponding to (1) the life cycle of the Southern Oscillation and (2) the seasonal transition of mid-tropospheric circulation over India. The model provided 40% more accurate results than the best linear statistical method, using the same data. Further, this hybrid ANN model also outperformed a more complex statistical model that used a larger number of predictors. Moving away from the conventional norm, Luk *et al.* (2000) investigated the effect of spatial and temporal explanatory variables on ANN models that could forecast rainfall with a 15 min lead time. Interestingly, the results of this study showed that there only exists a certain optimal limit of temporal and spatial information, which is useful in ANN modelling, and any additional input beyond this limit only adds to noise in the network. This assessment was validated by Lin and Chen (2005), who developed an ANN model for short-term typhoon forecasting; they used typhoon characteristics with lags of 1, 2 and 3 h as input for the model and observed that the model with the 2 h lag performed best. This study reiterates the notion that the ability of ANN models to generalize is hampered if the temporal lag considered for the modelling is too long.

Broadly speaking, as suggested by Bowden *et al.* (2005), the inclusion of a larger number of explanatory variables in the ANN models is disadvantageous because:

- the requirement of computational memory and computational complexity increases;
- learning becomes more difficult with irrelevant explanatory variables;
- irrelevant explanatory variables may result in poor accuracy and mis-convergence;
- understanding complex models is more difficult than understanding simple models, especially when both offer comparable results.

In light of the aforementioned points, the selection of the most pertinent explanatory variables for the model development is crucial. Unfortunately, there is no set methodology for selecting the appropriate explanatory variables for the ANN models. Bowden *et al.* (2005), after an extensive literature review, reported that the methods for selecting the explanatory variables for the ANN models in water resources applications can be broadly classified into five groups:(1) methods that rely on the use of prior knowledge of the system being modelled, (2) methods based on linear cross-correlation, (3) methods that use a heuristic approach, (4) methods that extract knowledge contained within trained ANNs and (5) methods that use various combinations of the previous four approaches. In more recent studies (related to rainfall forecasting), techniques such as principal component analysis (Shukla *et al.*, 2011), fuzzy ranking algorithms (Srivastava *et al.*, 2010) and linear correlation matrix (Wu *et al.*, 2010) have been used to identify the relevant explanatory variables.

The present study seeks to explore the use of the mutual information (MI) technique for selecting the appropriate number and type of explanatory variables in order to forecast the rainfall. Literature on the use of this technique in hydrological and environmental studies is very limited: for instance, Maier *et al.* (2006) used the MI technique for clustering ecological data to assess the health of Australian rivers and streams; Dhamge *et al.* (2012) used this technique to develop ANN models for predicting the runoff in a catchment in India. However, the MI technique has been widely used for feature selection in a broad range of other studies such as cartography (Kerroum *et al.*, 2011), tomographic colonography (Ong and Seghouane, 2011), electrical systems (Devaraj and Roselyn, 2011) and spectrophotometry (Rossi *et al.*, 2006), among others. All these studies have reported the success of the MI technique in identifying the explanatory variables for non-linear problems. The main objectives of the current study, hence, were to (1) evaluate the application of the MI technique in identifying those explanatory variables needed to develop an ANN model for forecasting the rainfall and (2) compare the results of this model with other ANN models developed with traditional explanatory variables as reported in the literature. There have been no previous studies conducted on using the MI technique for explanatory variable selection for forecasting rainfall.

For this study, three ANN models were developed with different sets of explanatory variables to forecast the rainfall at the Santa Cruz weather station in Mumbai, India. The variables for the first model were in the form of temporal data of past rainfall events. Selected meteorological variables were used as explanatory variables for the second model, while the variables for the third model were identified using the MI technique.

2. Mutual information (MI)

MI is defined as a measure that quantifies the stochastic dependency between two random variables without making any assumptions (e.g., linearity) about the nature of their relation (Steuer *et al.*, 2002). In other words, MI evaluates the dependencies between random variables.

A detailed description of the MI technique has been reported by Rossi *et al.* (2006), parts of which are presented here to help elaborate the science behind this technique. Consider a system X with M_X possible states: that is, a measurement performed on X will yield one of the possible values $x1$, $x2 \dots$, xMx, each with a corresponding probability $p(x_i)$. The average amount of information gained from a measurement that defines one particular value x_i is given by the entropy $H(X)$ of the system. Hence, entropy is a measure of the uncertainty of any random variable, required on an average to describe the random variable (Cover and Thomas, 1991):

$$H(X) = -\sum_{i=1}^{M_x} p\left(x_i\right) \log p\left(x_i\right) \tag{1}$$

The joint entropy $H(X, Y)$ of two discrete systems, X and Y, is defined analogously as:

$$H(X, Y) = -\sum_{i=1}^{M_X} \sum_{J=1}^{M_Y} p\left(x_i, y_j\right) \log p\left(x_i, y_j\right) \tag{2}$$

Here, $p(x_i,y_j)$ denotes the joint probability that X is in state x_i and Y is in state y_j. The number of possible states, M_X and M_Y, may be different.

Mutual information $I(X, Y)$ between systems X and Y is defined as:

$$I(X, Y) = H(X) + H(Y) - H(X, Y), \geq 0 \tag{3}$$

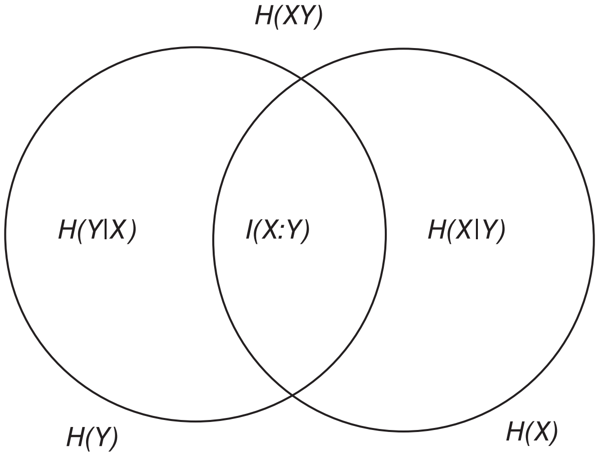

Figure 1. Pictorial representation of mutual information.

After combining Equations (1)–(3)

$$I(X,Y)=\sum_{i=1}^{M_X}\sum_{j=1}^{M_Y}p\left(x_i,y_j\right)\log\left[\frac{p\left(x_i,y_j\right)}{p\left(x_i\right)p\left(y_j\right)}\right] \quad (4)$$

Figure 1 provides a pictorial representation of the MI technique. It is apparent that the MI technique measures the information about Y shared with X (common information). In Equation (4), three probability distributions are used to describe the relationship between an input variable X and an output variable Y; both $p(x_i)$ and $p(y_i)$ are univariate probability distributions developed for each value separately, and $p(x_i,y_i)$ is the joint probability distribution when both values are considered simultaneously. An accurate determination of these probability distributions is very important for the precise computation of the MI. In practical setups, the underlying distribution p of the variables is unknown. Therefore, entropy H cannot be computed directly; rather, it requires estimation. In this study, entropy was estimated using an empirical approach. For further details of this computation, and for a more comprehensive understanding of the MI technique, readers are referred to Cover and Thomas (1991).

3. Study area and data collection

Located on the west coast of India, Mumbai is the capital of the nation's second most populous state, Maharashtra. It is also one of the most populous cities in the world, with a population of ~12.94 million (Government of Maharashtra, 2013). Mumbai is also the commercial and entertainment centre of India and makes a large contribution to India's economy. It has a typical monsoon climate and experiences hot, rainy and cold weather seasons. The city of Mumbai has two weather stations that are located in Santa Cruz and Colaba areas. The trend of average monthly rainfall, as recorded at the Santa Cruz weather station, is shown in Figure 2. The average annual rainfall is 2146.6 mm; however, virtually, all the rainfall occurs in the months between June and September, which renders the city susceptible to flooding during these months. In addition, unprecedented changes in rainfall patterns, rapid urbanization and inadequate city management and planning have further increased the city's vulnerability to floods. Each year, the southwest monsoon brings high intensity precipitation over Mumbai, and this coupled with a poor drainage network leads to frequent flooding in various areas of the city. For instance, on 26 July 2005, unusually heavy rains battered suburban Mumbai, causing one of the worst floods (940 mm in 24 h) in the history of the state (Government of Maharashtra, 2005). Such instances prove that there is a dire need for an early flood warning system for the city, and such a system requires reliable forecasts of rainfall.

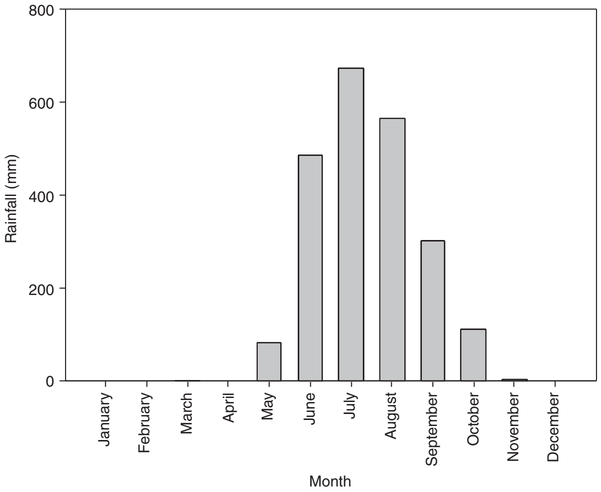

Figure 2. Average monthly rainfall at Santa Cruz weather station, Mumbai.

Although the ANN is a data-driven approach, variables that have some theoretical relation to the desired output are more useful in the model development than any random input. As discussed previously, earlier rainfall forecasting studies that used the ANN used a variety of explanatory variables as the model input. Input variables included historical rainfall data, cloud images and meteorological variables such as humidity and wind speed. Accordingly, for the present study, five explanatory variables were considered: rainfall (R); atmospheric pressure (P); dry bulb temperature (T); relative humidity (RH) and wind speed (W). Data were collected from both the Santa Cruz ($_{\mathrm{SC}}$) and the Colaba ($_{\mathrm{CL}}$) weather stations (operated by the Indian Meteorological Department) located in Mumbai's suburban regions, as shown in Figure 3. For the ease of understanding, from here onwards, these stations will be subscripted when a particular meteorological variable at any station is being described. For example, rainfall at the Santa Cruz and Colaba stations will be represented as R_{SC} and R_{CL}, respectively. The duration of data collection was from 1998 to 2006, but only selected data (described in the next section of the paper) were used in the model development, so as to reduce the computation time.

4. Model development

4.1. Selection of explanatory variables

Three data sets were prepared to develop three different ANN models: Models A, B and C. The explanatory variables used for each of the three models, and the rationale for selecting these variables, are listed below:

- Explanatory variables for Model A: Model A was developed to check the memory characteristics of the rainfall time series recorded at the Santa Cruz weather station. This is in line with the study by Luk *et al.* (2000), which investigated the effect

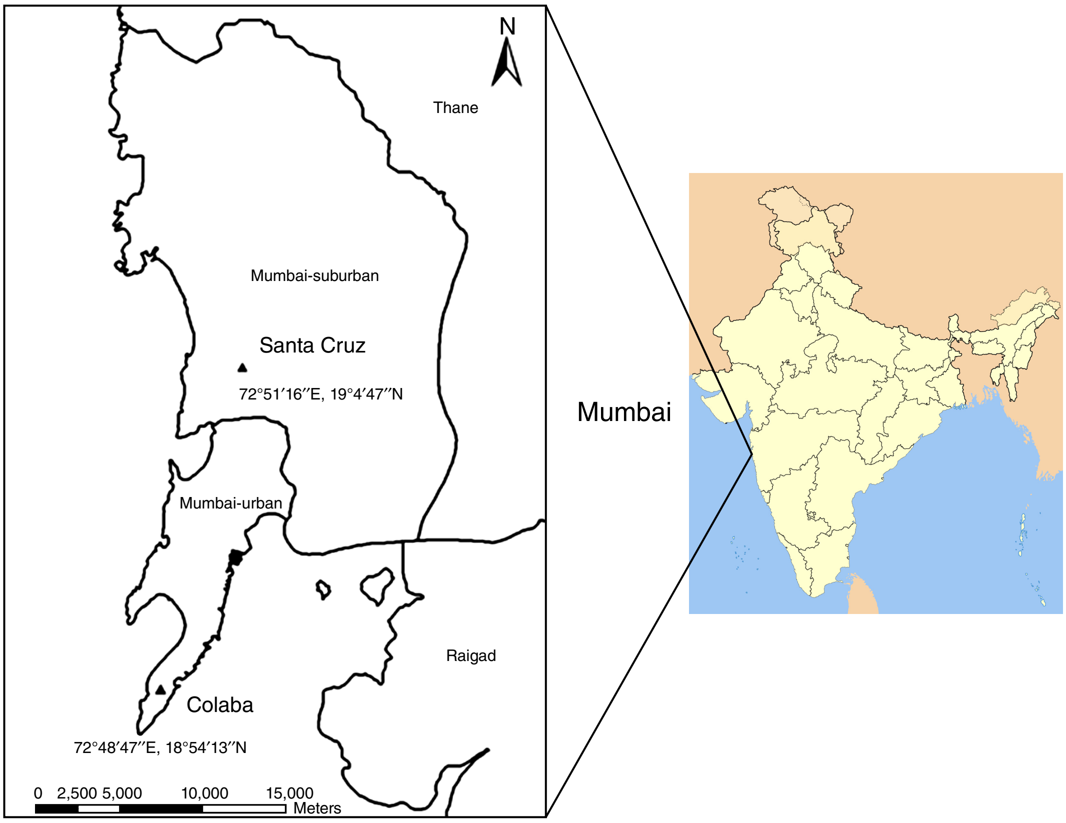

Figure 3. Location of weather stations in Mumbai.

of temporal inputs and concluded that there exists an optimum limit of temporal variables that is useful in modelling. Rainfall depth for the current time step (t) and five previous time steps (t-1, t-2, t-3, t-4 and t-5) were taken as input for the model to forecast the rainfall recorded at the next time step ($t+1$; 1 h lead time).

- Explanatory variables for Model B: In line with Hung *et al.* (2009), who used meteorological variables (temperature, relative humidity, wind speed, atmospheric pressure) and rainfall data to forecast rainfall in Bangkok, Model B was developed using rainfall and other meteorological variables (T, P, RH and W) as explanatory variables. These variables were observed at the Santa Cruz weather station. The purpose of selecting these variables was to determine the influence of the prevailing weather conditions on forecasting rainfall. The current time step (t) was used for all the explanatory variables to forecast the rainfall for the next time step ($t+1$).
- Explanatory variables for Model C: Model C was developed with those variables that were identified using the MI technique, based on three key hypotheses:

(i) rainfall at the next time step is influenced by rainfall at the current and previous time steps recorded at the station/location under consideration;
(ii) rainfall at the next time step is also influenced by weather conditions at the current and previous time steps observed at the station/location under consideration;
(iii) rainfall at the next time step is also influenced by rainfall as well as weather conditions at the current and previous time steps recorded at the surrounding locations.

In order to incorporate these conditions in Model C, all the five variables (R, P, T, W and RH) from both Santa Cruz and Colaba weather stations, at the current and five previous time steps (5 variables × 2 stations × 6 time steps = 60 variables), were considered as the potential explanatory variables. The MI technique was used to select the more pertinent variables from this set of 60 potential variables. The MI between the potential explanatory variables and the model output (rainfall at the next time step, $t+1$) was determined using the equations previously presented in Section 2. After calculating the MI score, the minimum redundancy and maximum relevance (MRMR) algorithm, as described by Parviz *et al.* (2008), was used to identify the six most influential explanatory variables. Only six variables were chosen so as to have a simple ANN structure, and so that comparisons could be made with Models A and B, which have six and five explanatory variables, respectively. The MRMR is a scheme in variable selection that helps to select those variables that have the strongest correlation with the classification variable.

Table 1 presents the six explanatory variables that have the highest MI with the output $R_{SC}(t+1)$, and the corresponding MRMR scores. It can be seen that the variable $T\mathrm{sc}(t)$ has the highest MI (0.215) with the output, and a high MRMR score (0.0858), as well. Similarly, based on the magnitudes of the MI and the MRMR scores, variables $R_{CL}(t)$, $R_{SC}(t)$, $R_{CL}(t\text{-}4)$, $W_{CL}(t\text{-}3)$ and $R_{SC}(t\text{-}1)$ were also made a part of the input data set used to develop Model C. It is interesting to note that the selected variables are a good mix: they are from both stations and with different lags. Thereby, they facilitate the testing of the hypothesis formulated for this category of models.

A majority of the studies related to rainfall forecasting using the ANN have used an event-based forecasting approach (Luk *et al.*, 2000; Lin and Chen, 2005; Nasseri *et al.*, 2008). In such studies, rainfall events exceeding only a certain duration and

Table 1. Explanatory variables for Model C based on MI and MRMR scores.

Model output	Explanatory variable	MI	MRMR score
$R_{SC}(t+1)$	$T_{sc}(t)$	0.215	–
	$R_{CL}(t)$	0.110	0.086
	$R_{SC}(t)$	0.159	0.035
	$R_{CL}(t\text{-}4)$	0.066	0.034
	$W_{CL}(t\text{-}3)$	0.055	0.026
	$R_{SC}(t\text{-}1)$	0.105	0.017

CL, Colaba weather station; MI, mutual information; MRMR, minimum redundancy and maximum relevance; *R*, rainfall; SC, Santa Cruz weather station; *T*, dry bulb temperature; *W*, wind speed; *t*, time step.

Table 2. Explanatory and desired output variables for Models A, B and C.

Model	Explanatory variables	Output variables
A	$R_{SC}(t)$, $R_{SC}(t\text{-}1)$, $R_{SC}(t\text{-}2)$, $R_{SC}(t\text{-}3)$, $R_{SC}(t\text{-}4)$, $R_{SC}(t\text{-}5)$	$R_{SC}(t+1)$
B	$R_{SC}(t)$, $T_{SC}(t)$, $P_{SC}(t)$, $\mathrm{W}_{SC}(t)$, $RH_{SC}(t)$	$R_{SC}(t+1)$
C	$T_{SC}(t)$, R_{SC} (t), $R_{SC}(t\text{-}1)$, $R_{CL}(t)$, $W_{CL}(t)$, $R_{CL}(t\text{-}4)$	$R_{SC}(t+1)$

CL, Colaba weather station; *P*, pressure; *R*, rainfall; *RH*, relative humidity; SC, Santa Cruz weather station; *T*, dry bulb temperature; *t*, time step; *W*, wind speed.

intensity were selected for modelling. In most cases, non-rainy days were not considered for the development of the model. However, in a recent study, Hung *et al.* (2009) used continuous time series input data, consisting of dry and wet periods, to forecast rainfall in Bangkok, and found that the ANN model was able to learn from continuous data. Because this approach is useful in real-time forecasting, where both dry and wet periods are likely to occur, continuous time series data were used to develop the ANN models in the present study as well. Given that the study area (Mumbai) receives most of its rainfall between June and September, and that forecasting is more crucial for this period, continuous data from June to September were used. The explanatory variables used for Models A, B and C are presented in Table 2.

4.2. ANN models

A comprehensive description of the ANN, its structures and terminologies used in the model development can be found in Appendix S1 of the Supporting information (adapted from ASCE, 2000b). The ANN is an information processing system designed to mimic certain aspects of the human brain, i.e. learning to recognize patterns and trends. It typically consists of a number of Processing Elements (PE), also called neurons, which are arranged in three types of layers: an input layer, one or more hidden layers and an output layer. The structure of the ANN is called the neural network architecture, which depends upon a number of factors as described in the Appendix S1. Among the most common of these are the multilayered perceptron (MLP) and generalized feed-forward (GFF) networks, which have been used in this study. Furthermore, two non-linear transfer (activation) functions, the hyperbolic tangent and the sigmoid, were used to convert input to output mathematically. The hyperbolic tangent is a modified form of the sigmoid function and normalizes the data between the range of −1 and +1, while the sigmoid function scales the output data between 0 and +1. The NeuroSolutions (2008) software (developed by NeuroDimension Inc., Gainesville, FL) was used to develop the ANN models in this study.

The normalized mean square error (NMSE), the Nash–Sutcliffe Model Efficiency Co-efficient or the efficiency index (EI), and the root mean square error (RMSE) were used to evaluate the model performances. The NMSE is an estimator of the overall deviations between predicted (*P*) and measured (*M*) values, as described in Equation (5):

$$\mathrm{NMSE} = \frac{1}{n}\sum_{i}\frac{\left(P_i - M_i\right)^2}{\overline{PM}} \tag{5}$$

The normalization by the product $\overline{PM}$ (averages of *P* and *M*, respectively) in Equation (5) ensures that the NMSE is not biased towards models that over- or under-predict. A value of NMSE equal to zero indicates the most perfect fit between the modelled and observed data, while NMSE equal to infinity indicates the poorest fit. The EI and the RMSE are commonly used indices and are often used in various kinds of studies.

5. Results and discussion

Three ANN models, A, B and C, were developed using the variables presented in Table 2, and by following the procedure described earlier. Details about the models (network, transfer function, number of PEs in each layer) are presented in the first four columns of Table 3, while the performance evaluation indicators (NMSE, EI and RMSE) for the models are presented in the last three columns.

The best-fit architecture for Model A used the MLP network with the sigmoid transfer function. It has 18 hidden PEs, arranged in two hidden layers. The NMSE, the EI and the RMSE for this model are 0.69, 30% and 2.79 mm, respectively. This model was developed using only temporal lags of rainfall at the Santa Cruz weather station. The best-fit architecture of Model B, developed using the meteorological data of the Santa Cruz weather station, used the GFF network and the hyperbolic tangent transfer function. The forecasting results improved with the inclusion of the meteorological variables, the NMSE reduced from 0.69 to 0.68, and there was a slight improvement in the EI from 30% to 31%. However, the RMSE of Model B increased by 0.03 mm when compared to that of Model A. The best-fit architecture of Model C, developed using the MI-based explanatory variables, used the GFF network and the hyperbolic tangent function with 19 hidden PEs arranged in two layers. Compared to Models A and B, this model had the lowest NMSE (0.65) and RMSE (2.70 mm), while at the same time, it exhibited the highest EI (35%). Overall, there is a 5.79 and 4.41% reduction in error in NMSE when the results of Model C are compared with those of Models A and B, respectively. Similarly there is a 16.66 and 12.90% improvement in the EI, respectively, and a 3.22 and 4.24% reduction in RMSE, respectively. Figure 4(a) and (b) shows the observed and forecasted trend of rainfall of Model C, over a 24 h period, for both the training and testing data sets.

The results of the modelling suggest that the rainfall time series for Mumbai has short-term memory characteristics, which means that rainfall at $t+1$ has very little or no significant relation with the rainfall recorded at *t*, *t*-1, *t*-2, *t*-3, *t*-4 and *t*-5. The inclusion of five lags of rainfall may have been the source of unnecessary noise, which resulted in relatively poorer forecasts, as indicated by the performance of Model A. These findings are

Table 3. Performance Indicator of Models A, B and C for 1 h lead time forecasts.

Model	Network	Transfer function	Structure	NMSE	EI (%)	RMSE (mm)
A	MLP	Sigmoid	6-12-6-1	0.69	30	2.79
B	GFF	Hyperbolic tangent	5-15-10-1	0.68	31	2.82
C	GFF	Hyperbolic tangent	6-14-5-1	0.65	35	2.70

EI, efficiency index; GFF: generalized feed-forward network; MLP, multilayered perceptron; NMSE, normalized mean square error; RMSE, root mean square index.

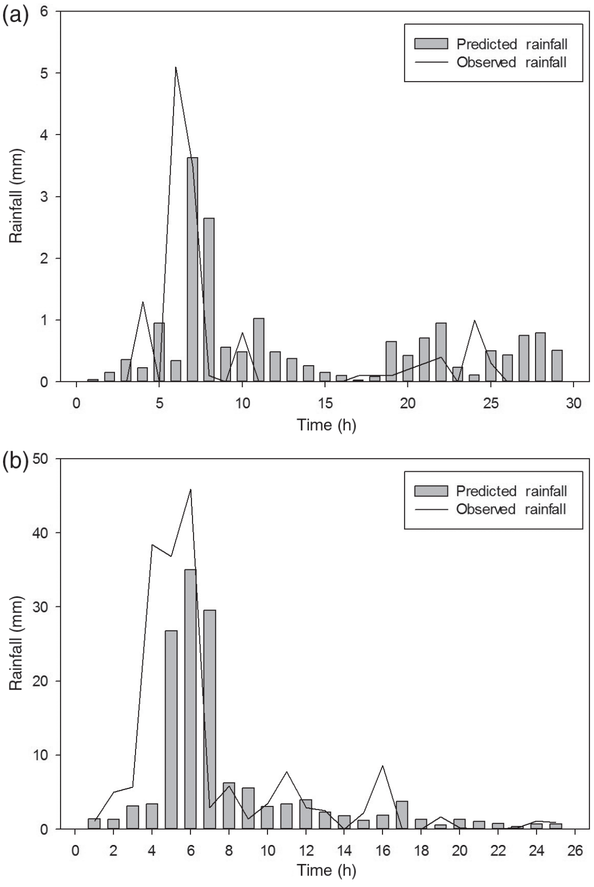

Figure 4. (a,b) Observed *versus* modelled rainfall with Model C during training and testing activities, respectively.

in good agreement with those of Luk *et al.* (2000), who found that increasing the number of lags of historic data reduces the forecasting accuracy.

As was expected, the performance improved when meteorological variables, *T*, *P*, *RH* and *W*, were used (in Model B). This finding is in agreement with studies reported by Lin and Chen (2005) and Hung *et al.* (2009). While the former study found that typhoon rainfall forecasting results improve when other typhoon characteristics are also used as input for the ANN model, the latter study improved the forecasting accuracies by using meteorological variables along with rainfall time series as explanatory data.

Among the three models, Model C, which was developed using the explanatory variables selected by the MI technique, provided the best results. This indicates the superiority of this technique in ANN modelling to forecast rainfall. The greatest advantage of this technique is that a wider range of potential explanatory variables can be considered for selection, a range that will ultimately be downsized to a smaller number for inclusion in the model development. Hence, there is more versatility in selection, without increasing the complexity of the model itself. Furthermore, as can be gleaned from Figure 4(a) and (b), not only is the model able to replicate the trend of observed rainfall well, it can also forecast both ‘rain’ and ‘no rain’ events. It is also important to note that compared to Models A and B, better results were achieved with Model C, which uses a simpler ANN structure. For example, there are fewer PEs in the hidden layers of Model C than in those of Model B. Also, although Model C has the best performance, it does not necessarily use more explanatory variables (the number of explanatory variables used in each model is almost similar: 6, 5 and 6 in A, B and C, respectively). Hence, it can be averred that the explanatory variables selected by the MI technique form a more efficient model. A more efficient model ensures reduced requirement of computational efforts in terms of lesser time, and it also requires fewer computational system hardware.

6. Conclusions

This study was carried out to examine the effectiveness of the mutual information (MI) technique to select the explanatory variables for artificial neural network (ANN) modelling in rainfall forecasting. The results of the ANN model, which used the explanatory variables selected by the MI technique (Model C), were compared with the results of the models that used historical lags of rainfall data (Model A) and selected meteorological data (Model B) as explanatory variables. A reduction of 5.79 and 4.34% in the normalized mean square error was observed when the results of Model C were compared with those of Models A and B, respectively. Similarly, there was a 16.66 and 13.33% improvement in the efficiency index, respectively, and a 3.22 and 4.25% reduction in the root mean square error, respectively, when Model C was compared to Models A and B. This study highlights the superiority of the MI technique in selecting the explanatory variables for ANN modelling, not just because of better performance with respect to various indicators but also because this performance is achieved with a simple ANN architecture. The MI measures the general dependence of random variables without making any assumptions about their relationships, making this technique an effective tool for selecting explanatory variables that can be used for forecasting the complex and non-linear processes of rainfall with better accuracy, as compared with other variables selection techniques.

Supporting information

The following material is available as part of the online article:

Appendix S1. Artificial Neural Networks and ANN model development.

References

Adamowski J, Chan HF, Prasher SO, Sharda VN. 2013. Comparison of multivariate adaptive regression splines with coupled wavelet transform artificial neural networks for runoff forecasting in Himalayan micro-watersheds with limited data. *J. Hydroinform.* **14**(3): 731–744.

ASCE [Task committee on application of artificial neural networks in hydrology]. 2000a. Artificial neural networks in hydrology, II: hydrologic application. *J. Hydrol. Eng.* **5**(2): 124–137.

ASCE [Task committee on application of artificial neural networks in hydrology]. 2000b. Artificial neural networks in hydrology, I: Preliminary Concepts. *J. Hydrol. Eng.* **5**(2): 115–123.

Babel MS, Shinde VR. 2011. Identifying prominent explanatory variables for water demand prediction using artificial neural networks: a case study of Bangkok. *Water Resour. Manage.* **25**: 1653–1676.

Bowden GJ, Dandy GC, Maier HR. 2005. Input determination for neural network models in water resources applications. Part 1 – Background and methodology. *J. Hydrol.* **301**: 75–92.

Cover T, Thomas J. 1991. *Elements of Information Theory*. John Wiley and Sons: New York, NY.

Devaraj D, Roselyn JP. 2011. On-line voltage stability assessment using radial basis function network model with reduced input features. *Int. J. Elect. Power* **33**(9): 1550–1555.

Dhamge NR, Atmapoojya SL, Kadu MS. 2012. Genetic algorithm driven ANN model for runoff estimation. *Procedia Technol.* **6**: 501–508.

EM-DAT. 2011. Disaster Profiles, The OFDA/CRED International Disaster Database. http://www.emdat.be/database (accessed 20 September 2011).

French MN, Krajewski WF, Cuykendall RR. 1992. Rainfall forecasting in space and time using a neural network. *J. Hydrol.* **137**: 1–31.

Government of Maharashtra. 2005. Maharashtra floods 2005, Relief & rehabilitation. http://mdmu.maharashtra.gov.in/pdf/Flood/statusreport.pdf (accessed 16 November 2012).

Government of Maharashtra. 2013. Economic survey of Maharashtra 2012–13. Directorate of economics & statistics, planning department, Government of Maharashtra, Mumbai, mahades.maharashtra.gov.in/files/publication/esm_2012–13_eng.pdf (accessed 12 November 2013).

Hong W-C. 2008. Rainfall forecasting by technological machine learning models. *Appl. Math. Comput.* **200**(1): 41–57.

Hung NQ, Babel MS, Weesakul S, Tripathi NK. 2009. An Artificial neural network model for rainfall forecasting in Bangkok, Thailand. *Hydrol. Earth Syst. Sci.* **13**: 1413–1425.

Jiang Y, Nan Z, Yang S. 2013. Risk assessment of water quality using Monte Carlo simulation and artificial neural network method. *J. Environ. Manage.* **122**: 130–136.

Kerroum MA, Hammouch H, Aboutajdine D. 2011. Textural feature selection by joint mutual information based on Gaussian mixture model for multispectral image classification. *Pattern Recogn. Lett.* **31**(10): 1168–1174.

Lin GF, Chen LH. 2005. Application of an artificial neural network to typhoon rainfall forecasting. *Hydrol. Process.* **19**: 1825–1837.

Luk KC, Ball JE, Sharma A. 2000. A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting. *J. Hydrol.* **227**: 56–65.

Maier HR, Zecchin AC, Radbone L, Goonan P. 2006. Optimising the mutual information of ecological data clusters using evolutionary algorithms. *Math. Comput. Model.* **44**(5–6): 439–450.

Mohanty S, Jha MK, Kumar A, Panda DK. 2013. Comparative evaluation of numerical model and artificial neural network for simulating groundwater flow in Kathajodi–Surua Inter-basin of Odisha, India. *J. Hydrol.* **495**: 38–51.

Mount NJ, Abrahart RJ. 2011. Load or concentration, logged or unlogged? Addressing ten years of uncertainty in neural network suspended sediment prediction. *Hydrol. Process.* **25**(20): 3144–3157.

Nasseri M, Asghari K, Abedini MJ. 2008. Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network. *Expert Syst. Appl.* **35**: 1415–1421.

Navone HD, Ceccatto HA. 1994. Predicting Indian monsoon rainfall: A neural network approach. *Clim. Dyn.* **10**: 305–312.

NeuroSolutions. 2008. NeuroSolutions user manual. NeuroDimension Inc.: Gainesville, FL.

Ong JL, Seghouane A-K. 2011. Feature selection using mutual information in CT colonography. *Pattern Recogn. Lett.* **32**(2): 337–341.

Parviz RK, Mozayani N, Jahed Motlagh MR. 2008. Mutual information based explanatory variable selection algorithm and wavelet neural network for time series prediction. In *proceedings of ICANN 2008, 18th International Conference on Artificial Neural Networks*, Part 1; 798–807.

Rossi F, Lendasse A, Francois D, Wertz V, Verleysen M. 2006. Mutual Information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometr. Intell. Lab.* **80**: 215–226.

Sattari MT, Yurekli K, Pal M. 2012. Performance evaluation of artificial neural network approaches in forecasting reservoir inflow. *Appl. Math. Model.* **36**(6): 2649–2657.

Shukla RP, Tripathi KC, Pandey AC, Das IML. 2011. Prediction of Indian summer monsoon rainfall using Niño indices: a neural network approach. *Atmos. Res.* **102**(1–2): 99–109.

Srivastava G, Panda SN, Mondal P, Liu J. 2010. Forecasting of rainfall using ocean-atmospheric indices with a fuzzy neural technique. *J. Hydrol.* **395**(3–4): 190–198.

Steuer R, Kurths J, Daub C, Weise J, Selbig J. 2002. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18**(2): S231–S240.

United Nations Inter-Agency Secretariat of the International Strategy for Disaster Reduction. 2004. *Living with Risk: A Global Review of Disaster Reduction Initiatives,* United Nations Publications: Geneva, Switzerland; 359 pp.

Wu CL, Chau KW. 2013. Prediction of rainfall time series using modular soft computing methods. *Eng. Appl. Artif. Intell.* **26**(3): 997–1007.

Wu CL, Chau KW, Fan C. 2010. Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. *J. Hydrol.* **389**(1–2): 146–167.