FORWARD: A Model for FOrecasting Reservoir WAteR Dynamics Using Spatial Bayesian Network (SpaBN)

Monidipa Das, *Student Member, IEEE*, Soumya K. Ghosh, *Member, IEEE*, Pramesh Gupta, V. M. Chowdary, Ravoori Nagaraja, and V. K. Dadhwal

Abstract—Natural systems, like the hydrological, climatological, atmospheric, or any other environmental processes, are extremely complex as well as dynamic in nature. It is therefore difficult to forecast, analyze, and quantify these processes by using simple empirical equations. Modeling and forecasting of reservoir water dynamics are not exceptions in this respect, as these involve various challenges due to the effect of meteorological factors, natural processes of stream flow, climatic change, and so on. The intent of our present work is to propose a novel forecasting model, FORWARD, that handles some of these issues in complex reservoir dynamics. FORWARD is based on a *variant of spatial Bayesian network (SpaBN)*, having inherent capability of modeling impact of spatial variability of meteorological factors over the river catchment. The forecasting efficiency of FORWARD has been compared with four other linear and non-linear techniques based on six different statistical performance measures. The experimental results show the superiority of FORWARD over the other techniques. Though FORWARD has been demonstrated with respect to a case study on forecasting reservoir live capacity, the model possesses a generic structure that can also be applied in other domains by introducing minimal augmentation.

Index Terms—Spatial Bayesian network, spatial importance, spatio-temporal system, probabilistic reasoning, forecasting

1 Introduction

FORECASTING any natural event or process is an extremely complex and challenging task involving various significant issues related to its spatio-temporal dependency, non-linearity, uncertainty and inherent chaos [1]. The reservoir water dynamics is not an exception in this regard. Reservoirs are effective water-storage areas which are mostly made by constructing dams across rivers or at the outlet of a natural lake (refer Fig. 1). Accurate prediction of reservoir water level, live capacity etc. are important for handling various water management issues including water supply for hydroelectric energy production and irrigation, flow control during floods and droughts, and so on. However, the reservoir water dynamics is not merely a simple periodic event that shows the highest degree of change during summer and rainy season, and the least change during winter. In fact, it is a result of complex interplay among various water balance components, like the flow of incoming or outgoing rivers and streams, seepage of water from or into the groundwater-bed etc. [2], [3]. Above all, the meteorological factors, including precipitation over the catchment area [4], evaporation from the water-surface [2], [5], wind velocity, humidity, and temperature in the adjacent lower atmosphere play significant roles in determining the live capacity of a reservoir. It is therefore crucial to properly understand the spatial variability of various meteorological factors in reservoir-catchment area and their hydrological influence on reservoir water dynamics.

Throughout the last few decades various Artificial Intelligence (AI) based techniques, such as artificial neural networks (ANN) [6], Genetic Algorithms (GA), Gene Expression Programming (GEP) [7], fuzzy theory etc., have increasingly been applied to tackle many of these issues related to water resource systems (e.g., reservoir) [8]. Majority of these works are based on ANN [9], [10] or a combination of ANN with other intelligent methods [11], [12]. For example, in order to model the impact of meteorological variables on reservoir dynamics, Ondimu and Murase [13] applied ANN considering a feature set comprising of rainfall, evaporation rate, river-discharges, and the water levels. Coulibaly [9] has employed reservoir computing technique, based on a special kind of recurrent neural network (termed as echo state network or ESN), for forecasting water level in lake. Adaptive network-based fuzzy inference system (ANFIS), proposed by Chang and Chang [11], has been able to show highly accurate and reliable prediction for a very short duration. Bazartseren et al. [12] have proposed a system, based on ANN and neuro-fuzzy technique, which has

- *M. Das, S.K. Ghosh, and P. Gupta are with the Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur 721302, India. E-mail: monidipadas@hotmail.com, skg@iitkgp.ac.in, pramesh.gupta94@gmail.com.*
- *V.M. Chowdary is with the National Remote Sensing Centre, Indian Space Research Organization, Kolkata 700156, India. E-mail: chowdary_isro@yahoo.com.*
- *R. Nagaraja and V.K. Dadhwal are with the National Remote Sensing Centre, Indian Space Research Organization, Hyderabad 500 037, India. E-mail: nagaraja_r@nrsc.gov.in, dadhwalvk@hotmail.com.*

Manuscript received 21 Mar. 2016; revised 21 Nov. 2016; accepted 15 Dec. 2016. Date of publication 4 Jan. 2017; date of current version 3 Mar. 2017. Recommended for acceptance by C. Ordonez.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2016.2647240

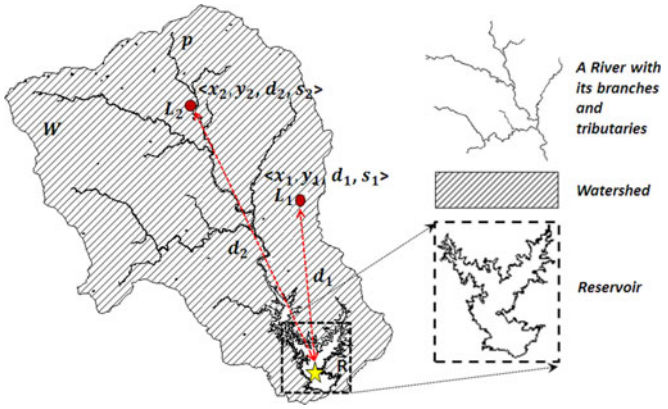


Fig. 1. A typical watershed (W), and a reservoir (R) on the river *p*.

proved to be considerably effective in short term water level forecast. In a comparative study over fifteen ephemeral catchments, Daliakopoulos and Tsanis [14] have found that ANN can be superior than the conventional conceptual models for modeling the complex hydrological processes.

In all these cases, though ANN has proved itself to be a useful technique to determine more or less exact pattern between the input and output variables, its effectiveness is highly influenced by proper understanding of the inter-variable dependency and the extent of knowledge regarding functionality of neural network [15]. So, according to Maier et al. [16], there is a need for incorporating robustness in those ANN-based approaches. On the other hand, our knowledge regarding the natural hydrological systems is limited over space as well as time. Therefore, it is also necessary to simulate the complex interactions between reservoir water and other influencing factors using some physical models, so that these can further aid in quantitative analysis and prediction process.

Our present work aims at proposing a computational intelligence framework (FORWARD), based on the spatial extension of standard/classical Bayesian network. It mainly concentrates on modeling the spatial influence of meteorological variables in the catchment area and the sensitivity of water dynamics in the respective reservoir. The dynamic behaviors of the hydrological processes in a reservoir depend not only on the meteorological features at the reservoir location, but also on those features associated with the locations over whole catchment (or watershed region). Moreover, based on the diverse topographical characteristics of the locations, these features show significant *spatial variation* and thereby affect the reservoir dynamics in different ways. Thus, the two main objectives in this work are:

- Modeling the influence of spatial variability of various meteorological variables on the hydrological process in a reservoir.
- Spatio-temporal prediction of reservoir dynamics by utilizing the knowledge of spatial variability obtained from the developed model.

1.1 Problem Statement, Challenges, and Contributions

Given a set of x number of meteorological variables $M = \{v_1, v_2, \ldots, v_x\}$ $(x > 1)$, influencing the natural hydrological process in a reservoir at location $L(lat, long)$. Also given the corresponding historical data series $H(l_i) = \{\{v_1^1, v_1^2, \ldots, v_1^t\}, \{v_2^1, v_2^2, \ldots, v_2^t\}, \ldots, \{v_x^1, v_x^2, \ldots, v_x^t\}\}$ for a set of k locations $\{l_1, l_2, \ldots, l_k\}$ in the reservoir catchment area, where t is the total number of temporal observations. The problem is to forecast the reservoir water dynamics for a given time instant with consideration to the impact of the spatial variability of the variables in M.

Hydrological processes in a reservoir depend not only on meteorological variables (e.g., temperature, rainfall) at the reservoir location, but also on other factors (e.g., soil, slope, and land use/cover distribution in the entire watershed), which influence these variables. One such example is depicted in Fig. 1. In the figure, W is the watershed/catchment area of a river, say p, and R is the location of the reservoir on p. L_1, and L_2 are two sample locations in the watershed. Each of them has been expressed in terms of quadruple $< x_i, y_i, d_i, s_i >$, where, (x_i, y_i) is the geo-spatial coordinate of the location L_i, d_i is the spatial distance (SD) of L_i from the reservoir R, and s_i is the soil category in the location. It has been assumed that the topographical slope is same for L_1 and L_2, and the locations are characterized by open forest associated with sandy soil texture, and built-up associated with clay soil, respectively. Therefore, for similar rainfall conditions at both locations, L_2 is likely to produce more surface runoff because of favorable runoff characteristics and will influence the reservoir inflows and reservoir capacity. Thus, the spatial variation of geographic features and meteorological factors has significant effect on hydrological processes in the watershed as well as on the reservoir water dynamics. Although the runoff generated from L_2 is higher than L_1, due to its proximity to reservoir location, runoff generated from the location L_1 will reach the reservoir earlier than location L_2. Thus, not only land use/cover, soil, and slope influence reservoir water dynamics, but also distances of the contributing areas play important role.

In this study, in order to account for spatial variability in the watershed, NRCS-curve number (CN) has been assigned to different soil and land use/land cover associations. The NRCS-CN is a dimensionless runoff index, which is a function of hydrologic soil group (HSG), land use, land treatment, hydrologic conditions, and antecedent moisture conditions (AMC). AMC is an indicator of watershed wetness and availability of soil storage prior to a storm. In general, three levels of AMC [17], namely AMC-I, AMC-II and AMC-III indicate dry, normal and wet conditions in the watershed, respectively. In the present work, CNs for AMC-II condition have been considered. Although surface runoff is slope dependent, most of the hydrological studies that involved NRCS-CN based approaches often ignored slope. In this study, NRCS-CN has been modified using Fuzzy Inference System (FIS), taking topographic slope into account. Further, the assumption of a uniform rainfall over the entire watershed may result in over/under estimation of reservoir inflows at gauge stations. Hence, the study area has been categorized into different clusters based on the meteorological conditions, where each cluster indicates single representative time series meteorological factor. In order to deal with the response times of contributing areas, the normalized inverse spatial distances have been computed for different clusters.

A novel framework, termed as FORWARD, has been proposed to accommodate the above-discussed tactics for modeling and forecasting the reservoir dynamics. The

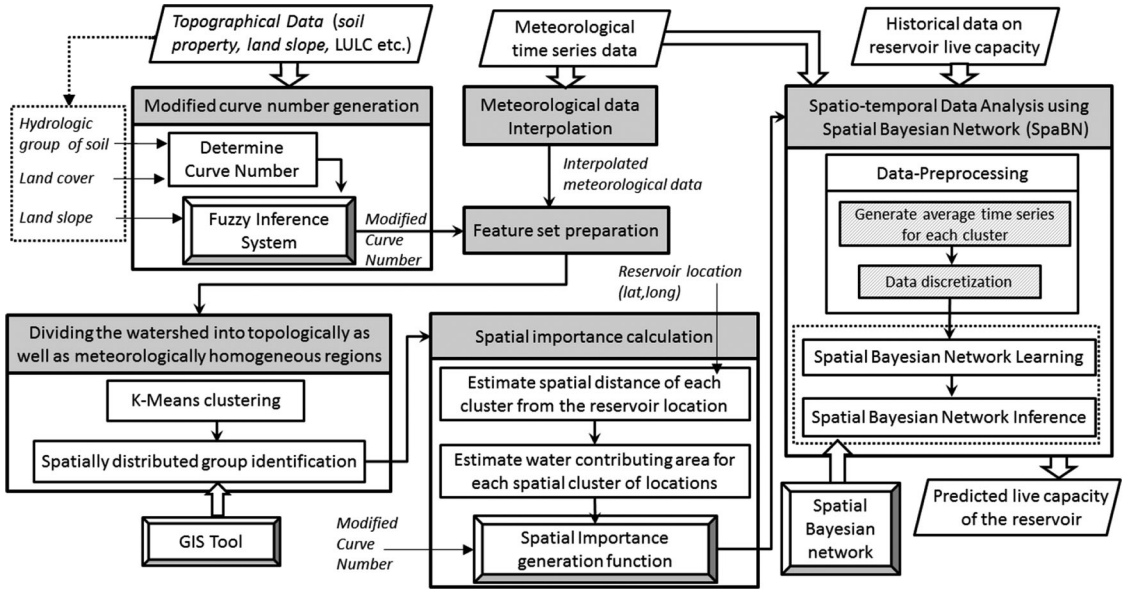


Fig. 2. Workflow for the proposed FORWARD model.

FORWARD is based on a *spatial Bayesian network* based analysis technique and is inherently capable of modeling *spatial impact* of such topographical and meteorological factors on hydrological processes in reservoir, assuming no human intervention in the present context. FORWARD has been evaluated with a case study on forecasting the daily live capacity of Mayurakshi reservoir in Jharkhand, India (refer Fig. 5) for the period of 1998-2001. Comparison with the results of classical/ standard Bayesian network, ANN and other techniques has established FORWARD as the most effective one in this regard.

The major contributions in this work can be stated as follows:

- Incorporating *spatial information* with the standard/ classical Bayesian network to *model the spatial variability* of the factors influencing prediction variable.
- Developing a *spatial Bayesian network* (SpaBN) based novel approach (FORWARD) for *modeling* and *forecasting* the dynamics in water resource system (e.g., reservoir).
- Applying *fuzzy inference system* based technique to generate the *slope-adjusted modified curve number* as a new runoff prediction parameter.
- Evaluating the *proposed FORWARD model* in comparison with the existing time series forecasting techniques including the traditional statistical models (ARIMA), standard BN, and ANN (feed forward back propagation network).

The rest of the paper is organized as follows. The proposed FORWARD model has been illustrated in Section 2 with a central concentration to our newly defined spatial Bayesian network, followed by analysis of the same in Section 3. A brief description of the case study has been provided in Section 4 along with the details of study area, data sets, and experimental setup. The results of experimentation on predicting reservoir live capacity using FORWARD framework has been thoroughly discussed in Section 5. Finally, we conclude in Section 6.

2 FORWARD: Proposed Model for Forecasting Reservoir Water Dynamics Using Spatial Bayesian Network

This section presents the theoretical description of proposed FORWARD model along with the details of our newly defined spatial Bayesian network analysis technique.

The workflow in FORWARD has been depicted in Fig. 2. From the figure it can be noted that the overall approach takes various topographical, meteorological and historical reservoir data as input, and as an output it produces the predicted value of daily reservoir data for a given prediction year. The entire model is composed of six main steps: i) *Interpolation of meteorological data*, ii) *Modified curve number generation*, iii) *Feature set preparation*, iv) *Dividing the watershed into homogeneous regions*, v) *Spatial importance calculation*, and vi) *Spatio-temporal Data Analysis using Spatial Bayesian Network*. Each of these steps has been thoroughly discussed in the following sections.

2.1 Interpolation of Meteorological Data

The objective of this step is to compensate the scarcity of high resolution meteorological data over the associated river watershed/catchment area. For this purpose, the whole watershed is first placed on a higher resolution grid and then for each of the intersecting grid locations, say l, the meteorological data is interpolated using the well-established interpolation techniques, e.g., IDW, Kriging [18] etc. One may also use Thiessen polygon [19] based raingauge network, for generating average daily precipitation time series for all the locations over the watershed. In the present study, the IDW technique has been used for interpolation purpose.

2.2 Modified Curve Number Generation

Curve number [20], also called runoff curve number, is an empirical measure of run-off prediction from the rainfall excess. It can be used as a good indicator of *land cover characteristics* and *hydrologic soil group*, i.e., two of the major topographical factors influencing the hydrological process in a watershed. However, there are several other factors (e.g.,

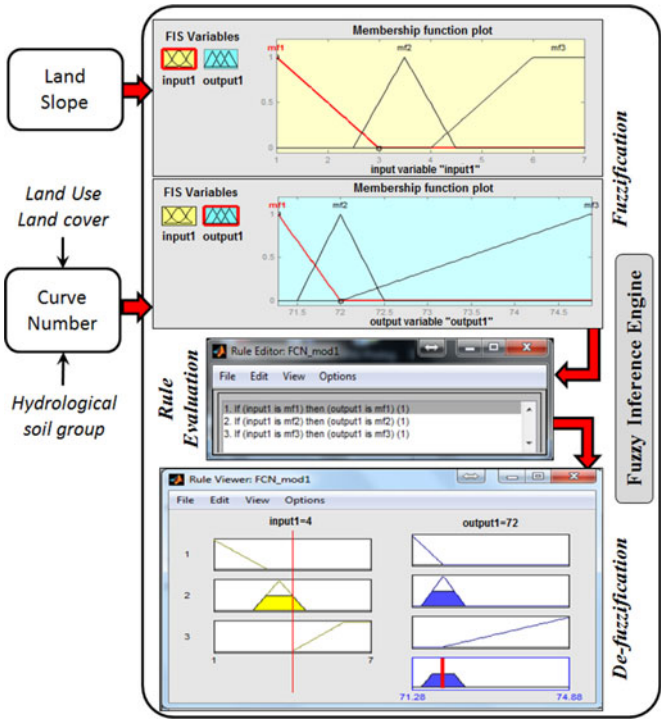


Fig. 3. Modified curve number (MCN) generation using fuzzy inference system.

land slope), which can significantly affect the run-off potentiality of a given region with known land cover category and hydrologic soil group. Therefore, to incorporate these information as well, the FORWARD model generates a variant of curve number, named as *modified curve number* (MCN), in its intermediate step. The internal process of modified curve number generation has been pictorially represented in Fig. 3. The process is based on a fuzzy inference system, which takes as input the fuzzy membership functions for the *land slope* and *curve number* respectively. Then, based on the rules defined in the rule base, it generates the modified curve number for a given slope value.

2.3 Feature Set Preparation

This step of FORWARD model generates the feature set for characterizing each of the grid locations l_i in the watershed region. The feature set is comprised of the mean monthly values for meteorological factors, e.g., temperature, rainfall, and the modified curve number, associated with the location. That is, for any location l, the feature set can be represented as follows:

$$< mf_1^1, mf_2^1, \ldots, mf_{12}^1, \ldots, mf_1^x, mf_2^x, \ldots, mf_{12}^x, MCN >,$$

where, mf_j^i is the mean value of the ith meteorological factor for the jth month, and MCN is the *modified curve number* value corresponding to the location l.

2.4 Dividing the Watershed into Homogeneous Regions

In this step, all the grid locations in the watershed are clustered into K number of groups $\{C_1, C_2, C_3, \ldots, C_K\}$, based on the feature set as prepared previously. Each cluster of locations, thus generated, becomes homogeneous with respect to both topographical and meteorological property. The step starts with a classical clustering technique (e.g., K-means), which basically generates non-spatial clusters, i.e., same cluster can be distributed over the space. In order to identify spatially distributed groups, the clustering is followed by spatial group identification process, performed using GIS software such as ArcGIS. After this processing, each newly identified group becomes a continuous region. The main objective here is to simplify the spatial analyses in following steps by grouping the similar category locations into single cluster, so that each of the group/cluster can be treated as a unique representative of a number of locations.

2.5 Spatial Importance Calculation

Once the spatial clusters are identified, each of the spatially distributed cluster C_i is assigned appropriate weight (SW_i) based on its significance in determining the hydrological processes in the reservoir. This weight is also termed as *spatial importance*. Three main parameters have been considered for this purpose:

- *Modified Curve Number:* As discussed earlier, MCN takes into account the land cover characteristics, land slope, and hydrological property of soil. The higher is the run-off potentiality, i.e., the MCN value of the locations in the cluster, the more is its contribution to the water dynamics in the reservoir.
- *Spatial Distance:* Influence of a cluster on the reservoir dynamics is inversely proportional to the spatial distance of the cluster from the reservoir location.
- *Water Contributing Area (WCA):* It takes into account the portions of the associated river or its tributary in the cluster. If the water-spread area is high in a particular cluster, then its effect will be higher during heavy rainfall conditions.

Since the range of values of these above-mentioned parameter may vary significantly, each of them are normalized before estimating the spatial importance of a cluster. Thus, the spatial weight (or, spatial importance) for a cluster C_i is determined as follows:

$$SW_i = \frac{NMCN_i + NISD_i + NWCA_i}{\sum_{j=1}^{K}(NMCN_j + NISD_j + NWCA_j)}, \quad (1)$$

where, $NMCN_i$ (normalized value of modified curve number for C_i) $= \frac{MCN_i}{\sum_{j=1}^{K} MCN_j}$, $NISD_i$ (normalized inverse spatial distance of C_i from the reservoir) $= \frac{1/SD_i}{\sum_{j=1}^{K} 1/SD_j}$, $NWCA_i$ (normalized value of water contributing area in C_i) $= \frac{WCA_i}{\sum_{j=1}^{K} WCA_j}$, and K is the total number of generated clusters.

The *spatial weights*, thus calculated, help in modeling the *spatial impact* of each cluster on the reservoir dynamics.

2.6 Spatio-Temporal Data Analysis Using Spatial Bayesian Network

This is the final and the most significant step in the proposed FORWARD model. The objective here is to predict the daily hydrological process in a typical reservoir (for a given year in future), with consideration to the spatial impact of each cluster over the corresponding watershed/catchment area. The whole analysis is based on a novel variation of *spatial Bayesian network*, which is *one of our major*

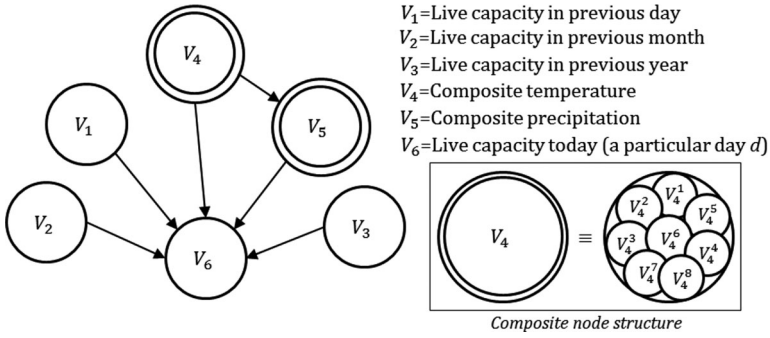


Fig. 4. Example DAG for spatial Bayesian network: SpaBN.

contributions in this work. The overall process of data analysis and forecast using SpaBN is illustrated in the subsequent parts of this section.

2.6.1 Data Pre-Processing

Before applying the spatio-temporal analysis using spatial Bayesian network, we perform pre-processing on the available data in two steps. The pre-processing is carried out to generate average time series per each cluster, and to discretize the data for making it fit for discrete Bayesian analysis.

1) *Generating average time series per each cluster:* In this step, the meteorological time series data corresponding to each location under a particular cluster is averaged out to produce a single representative time series for that cluster, and this is done considering each meteorological factor separately. Since all the locations under a same cluster are both topographically and meteorologically similar, this avoids any kind of inconsistency generated due to averaging.
2) *Data discretization*: This step discretizes the meteorological as well as the historical reservoir data (e.g., live capacity, water level) so as to make these suitable for spatial Bayesian network training and inference process in the following steps. The discretization is performed based on the maximum and minimum values observed in the training data and finally dividing the whole data range into suitable number of bins/intervals. The size of each interval is determined in the following manner: If, for any variable v_i, the maximum observed value is $max(v_i)$ and the minimum observed value is $min(v_i)$, then the interval size becomes

$$intervalSize(v_i) = \frac{[max(v_i) - min(v_i) + 1]}{R},$$

where, R is the total number of intervals or discretized range of v_i. The value of R may be pre-defined intuitively, or can be determined empirically so that it leads to optimum result with respect to prediction accuracy as well as execution time.

2.6.2 Spatial Bayesian Network Based Analysis

Spatial Bayesian network is a variant of classical/standard Bayesian network which has an intrinsic property of capturing spatial influence over the variables in the network. Unlike the standard Bayesian network, SpaBN contains *composite nodes* along with the standard/classical nodes in the directed acyclic graph (DAG). One such example network structure (or DAG) is shown in Fig. 4. Here, the nodes, denoted by double lined circles, are the composite nodes. A composite node is a composition of a number of standard/classical nodes associated with same but spatially distributed variable. For example, composite node V_4, as depicted in Fig. 4, consists of eight standard/classical nodes $V_4^1, V_4^2, V_4^3, \ldots V_4^8$, where V_4^i indicates the variable V_4 at the ith spatial location (or region). The purpose of introducing composite node is to reduce the learning time and space complexity in spatial Bayesian network. If instead of each single composite node, the constituting nodes were being used as standard/classical node, then it would introduce one or more edges for each such node leading to exponentially very high time and space requirement [21], [22]. Replacing all these nodes with a single composite node drastically reduces the algorithmic complexity in SpaBN. Moreover, being a variation of the classical Bayesian network, SpaBN retains the inherent property of uncertainty management [23], [24]. The learning and inference process for spatial Bayesian network is explained below.

2.7 Spatial Bayesian Network Learning

Consider a similar directed acyclic graph $G(V_s, V_c, E)$, as shown in the Fig. 4, where $V_s = \{V_1, V_2, V_3, V_6\}$ is the set of standard/classical nodes; $V_c = \{V_4, V_5\}$ is the set of composite nodes; and E is the set of edges $\{V_1 \rightarrow V_6, V_2 \rightarrow V_6, V_3 \rightarrow V_6, V_4 \rightarrow V_6, V_5 \rightarrow V_6, V_4 \rightarrow V_5\}$. An edge from V_i to V_j can be interpreted as V_i influences V_j. Let us also assume that the variable associated with the composite nodes are spatially distributed over K (=8 in this case) number of regions (each cluster, as generated previously, can represent a region).

Now, as per the principle of SpaBN, the marginal probabilities of the composite nodes $\in V_c$ are calculated as follows:

$$P(V_4) = \gamma \cdot \left[\sum_{i=1}^{K} P(V_4^i) \cdot SW_i\right] \tag{2}$$

$$P(V_5) = \gamma \cdot \left[\sum_{i=1}^{K} P(V_5^i) \cdot SW_i\right], \tag{3}$$

where, γ is the normalizing constant, $P(V_4^i)$ is the marginal probability of singular component V_4^i in V_4, $P(V_5^i)$ is the marginal probability of singular component V_5^i in V_5, and SW_i is the spatial weight/importance of the ith region.

Similarly, the conditional probabilities, involving composite nodes $\in V_c$, can be calculated as follows:

$$P(V_5|V_4) = \gamma \cdot \left[\sum_{i=1}^{K} \frac{n(V_4^i, V_5^i)}{n(V_4^i)} \cdot SW_i\right] \quad (4)$$

$$P(V_6|V_1, V_2, V_3, V_4, V_5) = \gamma \cdot \left[\sum_{i=1}^{K} \frac{n(V_1, V_2, V_3, V_4^i, V_5^i, V_6)}{n(V_1, V_2, V_3, V_4^i, V_5^i)} \cdot SW_i\right], \quad (5)$$

where, $n(< \cdot >)$ denotes the total count of observation for the combination $< \cdot >$.

Now, in order to capture the large scale temporal change in variable characteristics, the learning is performed for each instant (say a year) T_i in the time scale, and the final probability (whether marginal or conditional) is generated by honoring the temporal auto-correlation property [25], so that the final probability (marginal or conditional) P_f becomes

$$P_f = \sum_{i=1}^{t} \left(P_{T_i} \times \frac{1/d_i}{\sum_{j=1}^{t} 1/d_j}\right), \quad (6)$$

where, d_i is the temporal distance of the time instant T_i from the prediction time T_p, and t is the total number of time instants considered for training.

2.8 Spatial Bayesian Network Inference

Now, in order to explain the inference generation principle in SpaBN, let's consider a case, where the observed/ evidence variables are: $V_1, V_2, V_3, V_4^1, V_4^2, \ldots, V_4^K, V_5^1, V_5^2, \ldots, V_5^K$, from which the value of V_6 is to be inferred.

Then, as per the principle in FORWARD,

$$\text{Inferred value of } V_6 = \sum_{i=1}^{K} P(V_6|V_1, V_2, V_3, V_4^i, V_5^i) \cdot SW_i, \quad (7)$$

where the value for $P(V_6|V_1, V_2, V_3, V_4^i, V_5^i)$ can be determined from the conditional probability table for the variable V_6.

Among these inferred values of V_6, the predicted value becomes the one corresponding to the maximum probability. Since, in the data pre-processing step, the variable values are discretized into certain number of ranges, the predicted value of V_6 is also obtained in the form of a range. In order to obtain a single continuous value, the mean of the range is assigned to the predicted variable.

3 Algorithmic Complexity Analysis for Proposed Spatial Bayesian Learning

In this section, we analyze the computational cost of learning the proposed spatial Bayesian network in FORWARD model. The computational cost has been measured with respect to parameter learning in terms of both time and space.

Let $G(V_s, V_c, E)$ be a directed acyclic graph of spatial Bayesian network containing n number of *nodes*, where V_c is the set of *composite nodes* and V_s is the set of *standard/classical nodes*. Likewise, assume that the variables corresponding to the composite nodes are distributed over K number of *spatial regions*. Moreover, let the *maximum number of parents* of any node in G is P and the *domain size* of any variable is D.

3.1 Time Complexity for SpaBN Parameter Learning

As per the network learning equations for SpaBN (refer Section 2), total number of iterations required for learning/ updating a composite node, having i number of parents, is $(D-1).D^i.K$, and that for learning/updating a standard/ classical node, having i number of parents, is $(D-1).D^i$.

Now, if we consider that the number of composite node having i $(0 \leq i \leq P)$ number of parents is nc_i, and the number of standard/classical node having i $(0 \leq i \leq P)$ number of parents is ns_i, then computational cost for learning parameters of all the composite nodes is

$$TC_{composite}(G) = \sum_{i=0}^{P} (D-1).nc_i.D^i.K. \quad (8)$$

Similarly, computational cost for learning parameters of all the standard/classical nodes is

$$TC_{standard}(G) = \sum_{i=0}^{P} (D-1).ns_i.D^i, \quad (9)$$

where, $\sum_{i=0}^{P} nc_i = |V_c|, \sum_{i=0}^{P} ns_i = |V_s|$, and $\sum_{i=0}^{P} (nc_i + ns_i) = n$

Therefore, the overall time complexity for learning parameters in SpaBN is

$$TC(G) = \sum_{i=0}^{P} (D-1).nc_i.D^i.K + \sum_{i=0}^{P} (D-1).ns_i.D^i \quad (10)$$

$$\begin{aligned}
&\leq \sum_{i=0}^{P} K.(D-1).D^i.(nc_i + ns_i) \\
&= K.(D-1).\big[D^P.(nc_P + ns_P) + D^{P-1}.(nc_{P-1} + ns_{P-1}) \\
&\qquad + \cdots + D^0.(nc_0 + ns_0)\big] \\
&\leq K.(D-1).\big[D^P.n + D^{P-1}.(nc_{P-1} + ns_{P-1}) + \cdots + \\
&\qquad D^0.(nc_0 + ns_0)\big] \\
&\qquad [\because (nc_P + ns_P) \leq n \text{ Always}] \\
&= O(n.K.D^{P+1}).
\end{aligned} \quad (11)$$

Now, since the number of region K tends to be limited within a certain small range, it can be treated as a constant, and therefore the time complexity of SpaBN becomes $TC(G) = O(n.D^{P+1})$, which is similar to that of standard/ classical Bayesian network containing no spatially distributed variable.

3.2 Space Complexity for SpaBN Parameter Learning

For any Bayesian network, the minimum amount of space requirement for any node x= $(|D(x)| - 1).\prod_i |D(Pa_i)|$, where $|D(Pa_i)|$ denotes the domain size of the ith parent Pa_i of x, and $|D(x)|$ is the domain size for the variable x. Therefore, in SpaBN, space required for learning/updating a composite node, having i number of parents, becomes $(D-1).D^i.K$, and that for learning/updating a standard/classical node, having i number of parents, becomes $(D-1).D^i$.

Now, if we consider that the number of composite nodes having i $(0 \leq i \leq P)$ number of parents is nc_i, and

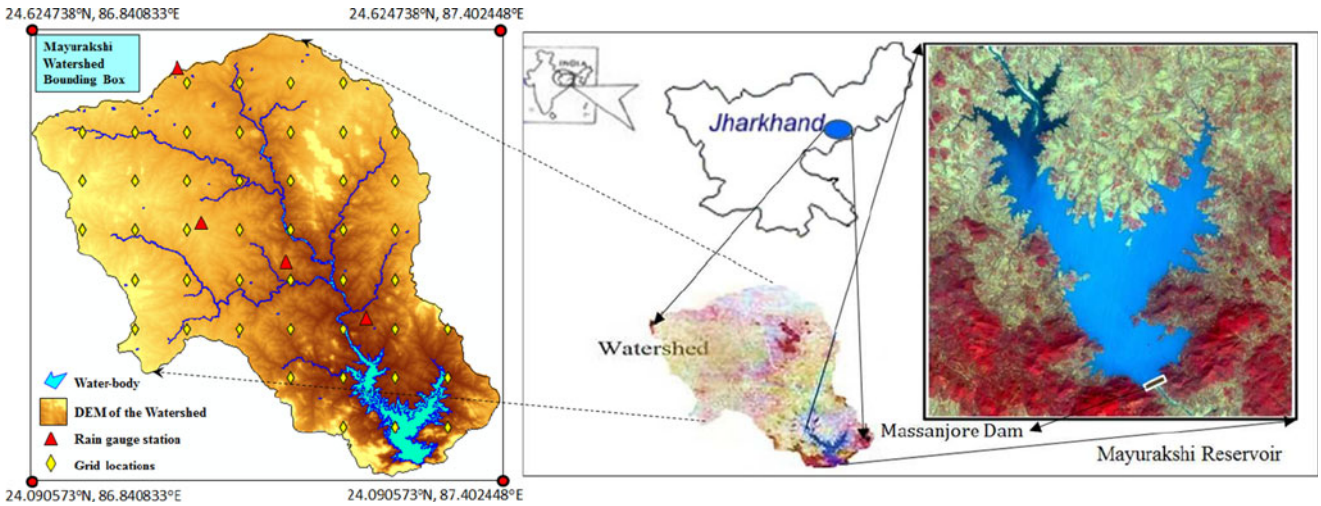


Fig. 5. Study area: Mayurakshi river watershed, Jharkhand, India.

the number of standard/classical node having i $(0 \leq i \leq P)$ number of parents is ns_i, then space requirement for learning parameters of all the composite nodes becomes

$$SC_{composite}(G) = \sum_{i=0}^{P} (D-1).nc_i.D^i.K. \tag{12}$$

Similarly, space requirement for learning parameters of all the standard/classical nodes becomes

$$SC_{standard}(G) = \sum_{i=0}^{P} (D-1).ns_i.D^i, \tag{13}$$

where, $\sum_{i=0}^{P} nc_i = |V_c|$, $\sum_{i=0}^{P} ns_i = |V_s|$, and $\sum_{i=0}^{P} (nc_i + ns_i) = n$

Therefore, the overall space complexity for learning parameters in SpaBN becomes

$$\begin{aligned} SC(G) &= \sum_{i=0}^{P} (D-1).nc_i.D^i.K + \sum_{i=0}^{P} (D-1).ns_i.D^i \qquad (14) \\ &\leq \sum_{i=0}^{P} K.(D-1).D^i.(nc_i + ns_i) \\ &\leq K.(D-1).\big[D^P.n + D^{P-1}.(nc_{P-1} + ns_{P-1}) + \cdots + \\ &\qquad D^0.(nc_0 + ns_0)\big] \\ &\qquad [\because (nc_P + ns_P) \leq n \text{ Always}] \\ &= O(n.K.D^{P+1}). \qquad (15) \end{aligned}$$

Now, since the number of region K tends to be limited within a certain small range, it can be treated as a constant, and therefore the space complexity of SpaBN becomes $SC(G) = O(n.D^{P+1})$, which is similar to that of standard/ classical Bayesian network containing no spatially distributed variable.

Therefore, in spite of being spatial extension of standard/ classical Bayesian network, the proposed SpaBN does not show any degradation with respect to computational complexity.

4 Case Study

In the present work, we have proposed a *spatial Bayesian network* based *forecast model* (FORWARD), that incorporates the spatial impact of meteorological factors while forecasting the water dynamics in a reservoir. In order to validate the performance of the proposed FORWARD model, we have considered a case study on forecasting the *daily live capacity* of a reservoir. The details of study area, data sets, and experimental setup have been discussed in the subsequent part of this section.

4.1 Data Sets and Study Area

In this work, the watershed [26] of river *Mayurakshi* in Jharkhand, India and its associated reservoir (*Mayurakshi reservoir*), has been considered as the area of case study. Geographically, the reservoir is located at $24°6.6'N$ latitude and $87°18.9'E$ longitude (refer Fig. 5) and the entire watershed is composed of nearly 1,866 sq. km area (Bottom-Left: $[24.09°N, 86.84°E]$, Top-Right: $[24.62°N, 87.40°E]$). The region belongs to the tropical climate zone and shows three well defined seasons, namely (i) *summer season* from March to June, (ii) *rainy season* from July to October, and (iii) *winter season* from November to February.

For experimental purpose, the whole watershed region has been divided based on 10×10 grid with each cell comprising approximately 33 sq. km area. The detailed specifications of the data collected from the study region are described below:

- *Daily rainfall data*: The daily data of rainfall for each of these gridded locations in the watershed has been interpolated for a span of 10 years (1st January, 1991 to 31st December 2000) from the daily data of four rain gauge stations (situated in Dumka ($24.28°N$, $87.24°E$), Jama ($24.35°N$, $87.15°E$), Jharmundi ($24.40°N$, $87.05°E$), and Sariyahat ($24.58°N$, $87.01°E$) and also using $0.5° \times 0.5°$ gridded rainfall data from *Indian Meteorological Department* (IMD).
- *Daily temperature data*: The daily data of temperature for each of the gridded locations has been interpolated from IMD high resolution $1° \times 1°$ gridded temperature data.

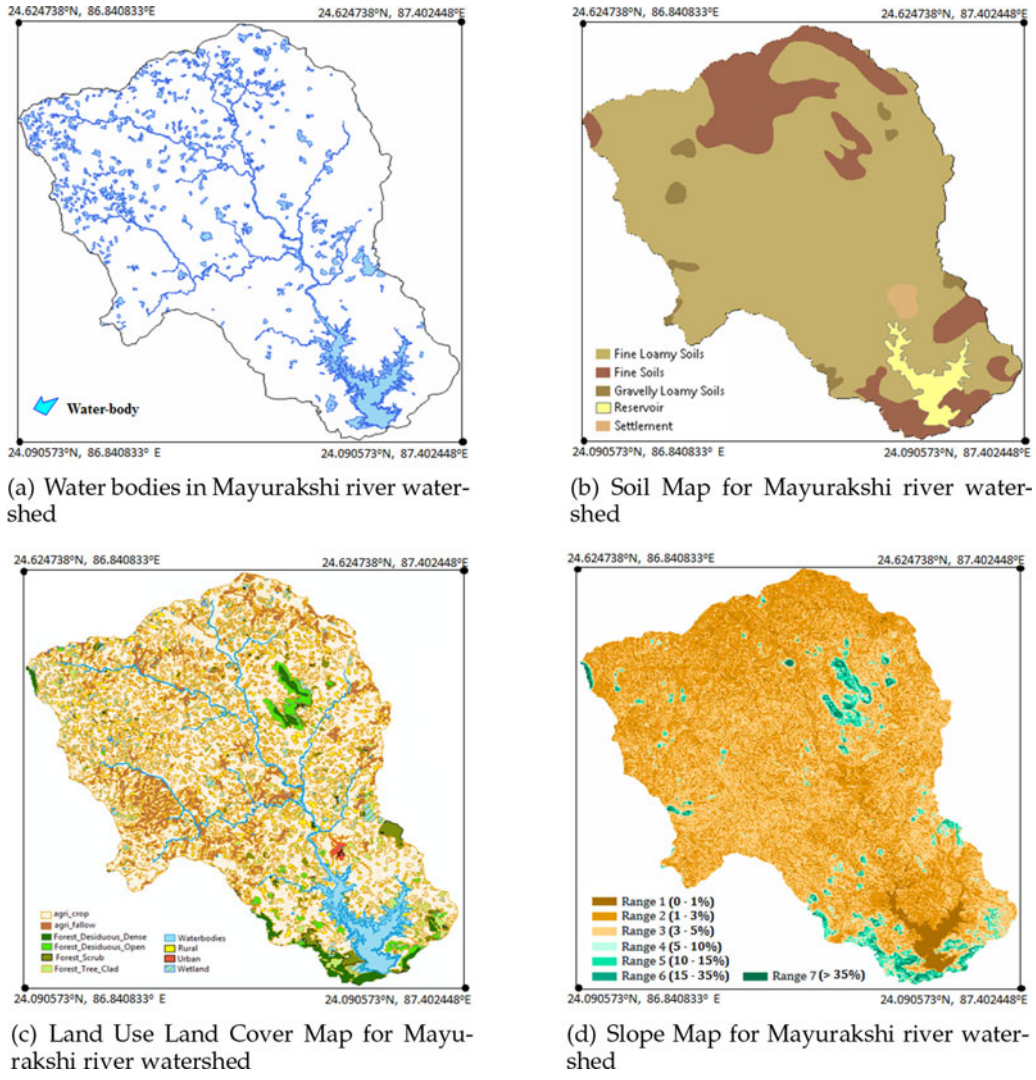


Fig. 6. Topographical features in Mayurakshi river watershed, Jharkhand, India.

- *Daily live capacity data of the reservoir*: The live capacity data of the reservoir has been collected for the same time span (1st January, 1991 to 31st December 2000) from the office of *Irrigation and Waterways Dept. Govt. of West Bengal*, Kolkata, India.
- *Topographical data*: The topographical data, namely the *soil map* (Fig. 6b), *land slope map* (Fig. 6d), and spatial distribution of *land use land cover (LULC)* (Fig. 6c) for the entire watershed region, have been collected from the*National Bureau of Soil Survey and Land Use Planning, Govt of India*, and the *Bhuvan portal* [27], respectively. From these maps, it is found that our study watershed is rich in diverse categories of LULC and soil, where almost 66 percent of the total area contains agricultural crop land, and about 74 percent of the region is covered with fine loamy soil.

4.2 Experimental Setup

The proposed model, FORWARD, has been implemented using *R-tool* version 3.2.2 (64 bit) [28] and *MATLAB* 7.12.0 (R2011a) [29] in Windows 7 (64 bit Operating System, 3.10 GHz CPU, 4.00 GB RAM). R-tool has been used for data pre-processing, interpolation, feature set preparation, cluster analysis, and spatial importance calculation purpose. MATLAB has been utilized to implement the spatial Bayesian network-based spatio-temporal analysis part of FORWARD model. Moreover, the ArcGIS tool [30] has been used for generating spatially distributed clusters/ regions. The meteorological data over rainfall has been interpolated for 45 locations (in the Mayurakshi watershed), placed over a 10×10 grid within the watershed bounding box.

The proposed approach has been evaluated in comparison with other popular linear as well as nonlinear time series forecasting approaches like exponential model (Holt-winters approach), automated ARIMA, standard BN (SBN), and ANN. MATLAB (NNTool) has been utilized to perform time series forecast of live capacity using artificial neural network (ANN), and implementing the standard BN technique. As the model of ANN, we have considered the *feed-forward neural network*, trained with the *Levenberg-Marquardt algorithm* [31], and as the Bayesian network inference technique, we have used the *exact method* [32]. On the other side, the R-tool has been used for forecasting water level using *Holt-Winters method* [33], and *Automated ARIMA* [34].

In the present work, the simulation has been carried out for three conditions, representing (i) *normal rainfall year* (1998), (ii) *high rainfall years* (1999, 2000), and (iii) *low rainfall year* (2001). Same combinations of input data have been

TABLE 1
Combination of *Training Years* and *Prediction Year*

Training years:	1991-1997	1991-1998	1991-1999	1991-2000
Prediction year:	1998	1999	2000	2001

used for the proposed approach and all the other methods for carrying out the comparative study. The various combination of *training years* and the corresponding *prediction year* is presented in Table 1. However, our proposed model FORWARD is flexible enough to adjust with any other combinations of *training* and *testing year* as well.

5 Results and Discussion

In this section, we have discussed the details of various outcomes of our experimentation on forecasting *daily live capacity* of the *Mayurakshi reservoir*. The overall results of forecasting are found to be promising.

5.1 Modified Curve Number Generation and Feature Set Preparation

In order to generate modified curve number, the slope data in the study region has been fuzzified into three linguistic variables: *Low*, *Medium*, and *High*. Besides, each time a curve number is modified, it is fuzzified into three different linguistic variables in similar manner. The process of MCN generation for the CN value 72 is shown in Fig. 3. Moreover, the generated MCNs associated with all the 45 grid locations are presented in the Table 2.

Once the modified curve numbers have been generated, the feature set is prepared based on the mean temperature and rainfall data for each month, and the modified curve number value. A typical form of feature set for any location l in the study watershed region is given below

$$F_l = \{mt_1, mt_2, \ldots, mt_{12}, mp_1, mp_2, \ldots, mp_{12}, mcn\},$$

where, mt_i and mp_i are the mean temperature and mean rainfall for the ith month respectively, and mcn is the modified curve number corresponding to the location l.

5.2 Cluster Generation and Spatial Importance Calculation

K-means clustering technique has been adopted to segregate the study watershed area (45 locations) into six groups based on the feature set. The results thus produced are non-spatial clusters, which are further processed in ArcGIS to identify *eight* spatially distributed clusters as described in Table 3 and depicted in Fig. 7. Once the spatially distributed clusters are identified, we determine the average value of modified curve number, average water contributing area, and mean spatial distance for each of the clusters (Table 3). Then, the spatial weight/ importance for each cluster is determined on the basis of average MCN, WCA, and SD values. The percentage (%) spatial importance value, thus obtained for each cluster, is graphically shown in Fig. 7.

5.3 Forecasting Reservoir Live Capacity Using FORWARD Approach

In the proposed FORWARD model, the spatio-temporal analysis has been performed based on the spatial Bayesian network. The structure of the network used for this purpose is shown in Fig. 4. Before performing the parameter learning, all the historical meteorological data (on temperature, and rainfall), and the reservoir live capacity data has been discretized into *eight* and *nine* classes, respectively. Once the learning has been performed, the inference is generated for live capacity for each day in the prediction year. The mean of the inferred live capacity range, producing the highest probability value, is considered as the predicted value in each case. The results of forecasting with respect to various performance measures have been discussed next, in comparison with various linear and non-linear approaches.

5.3.1 Model Evaluation Criteria

In order to evaluate the effectiveness of the proposed FORWARD approach, six statistical indicators (goodness-of-fit criteria), namely *Normalized Root Mean Square Deviation (NRMSD)*, *Nash-Sutcliffe efficiency (NSE)* [35], [36], *Mean percent deviation (D_v)* [37], *Percent standard error of prediction (SEP)* [36], *Coefficient of determination or R-squared (R^2)* [37], and *Pearson's correlation coefficient (CC)*, have been used in the present study. The respective formulations for each of these measures are as follows:

$$NRMSD = \frac{1}{(O_{max} - O_{min})}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(V_{o_i} - V_{p_i})^2} \tag{16}$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(V_{o_i} - V_{p_i})^2}{\sum_{i=1}^{N}(V_{o_i} - \overline{V_o})^2} \tag{17}$$

$$D_v = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{(V_{p_i} - V_{o_i})}{V_{o_i}}\right) \times 100 \tag{18}$$

$$SEP = \left(\frac{1}{\overline{V_o}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(V_{o_i} - V_{p_i})^2}\right) \times 100 \tag{19}$$

$$R^2 = \frac{\left[\sum_{i=1}^{N}(V_{o_i} - \overline{V_o})(V_{p_i} - \overline{V_p})\right]^2}{\sum_{i=1}^{N}(V_{o_i} - \overline{V_o})^2 \cdot \sum_{i=1}^{N}(V_{p_i} - \overline{V_p})^2} \tag{20}$$

$$CC = \frac{\sum_{i=1}^{N}(V_{o_i} - \overline{V_o})(V_{p_i} - \overline{V_p})}{\sqrt{\sum_{i=1}^{N}(V_{o_i} - \overline{V_o})^2 \sum_{i=1}^{N}(V_{p_i} - \overline{V_p})^2}}, \tag{21}$$

where, O_{max} is the maximum observed value of reservoir live capacity, O_{min} is the minimum observed value of reservoir live capacity, V_{o_i} is the observed value of reservoir live capacity on the ith day, V_{p_i} is the predicted value of reservoir live capacity for the ith day, $\overline{V_o}$ is the mean of observed values of reservoir live capacity, $\overline{V_p}$ is the mean of predicted values of reservoir live capacity, and N is the total number of prediction day considered.

The best-fit between observed and predicted live capacity under ideal conditions yields NRMSD = 0, NSE = 1, D_v = 0, SEP = 0, R^2 = 1, and CC = 1.

5.4 Performance Evaluation

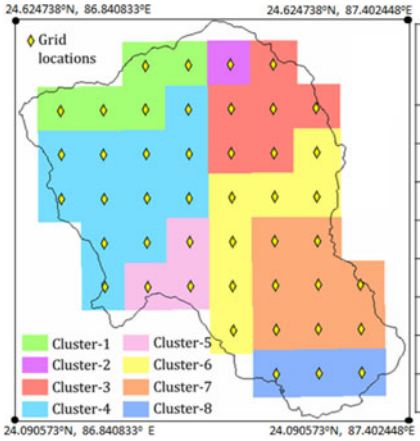
The results for the prediction years 1998, 1999, 2000, and 2001 have been tabulated in comparison with various other approaches (refer Tables 4, 5, 6, and 7). By analyzing the different outcomes, as shown in the tables and in Fig. 8, the

TABLE 2
Modified Curve Number (MCN) Associated with the Grid Locations in the Study Area

Grid Point ID	Lat	Long	Soil Runoff Potential	Land-Cover	Slope Range/ Categories (%)	Modified Curve Number
1	24.57	87.03	High	Agricultural Land-Crop	3	78.00
2	24.57	87.09	Moderate	Agricultural Land-Crop	3	72.00
3	24.57	87.15	Moderate	Wasteland	3	80.00
4	24.57	87.22	High	Wasteland	2	84.47
5	24.51	86.90	Moderate	Wasteland	3	80.00
6	24.51	86.97	Moderate	Wasteland	3	80.00
7	24.51	87.03	High	Rural	1	87.40
8	24.51	87.09	Moderate	Agricultural Land-Crop	2	71.55
9	24.51	87.15	Moderate	Wasteland	2	79.50
10	24.51	87.22	Moderate	Wasteland	2	79.50
11	24.51	87.28	Moderate	Agricultural Land-Crop	3	72.00
12	24.45	86.90	Moderate	Agricultural Land-Crop	2	71.55
13	24.45	86.97	Moderate	Agricultural Land-Crop	2	71.55
14	24.45	87.03	Moderate	Agricultural Land-Crop	2	71.55
15	24.45	87.09	Moderate	Tree Clad Area	2	39.75
16	24.45	87.15	Moderate	Agricultural Land-Crop	2	71.55
17	24.45	87.22	Moderate	Rural	2	82.48
18	24.45	87.28	Moderate	Agricultural Land-Crop	2	71.55
19	24.39	86.90	Moderate	Agricultural Land-Crop	3	72.00
20	24.39	86.97	Moderate	Agricultural Land-Crop	2	71.55
21	24.39	87.03	Moderate	Agricultural Land-Crop	4	72.00
22	24.39	87.09	Moderate	Agricultural Land-Crop	3	72.00
23	24.39	87.15	Moderate	Agricultural Land-Crop	2	71.55
24	24.39	87.22	Moderate	Agricultural Land-Crop	3	72.00
25	24.39	87.28	Moderate	Agricultural Land-Fallow	3	79.00
26	24.33	86.97	Moderate	Agricultural Land-Fallow	6	81.12
27	24.33	87.03	Moderate	Agricultural Land-Crop	3	72.00
28	24.33	87.09	Moderate	Agricultural Land-Crop	3	72.00
29	24.33	87.15	Moderate	Agricultural Land-Fallow	2	78.51
30	24.33	87.22	Moderate	Agricultural Land-Crop	1	71.51
31	24.33	87.28	Moderate	Wasteland	2	79.50
32	24.27	86.97	High	Agricultural Land-Crop	3	72.00
33	24.27	87.03	Moderate	Agricultural Land-Fallow	3	79.00
34	24.27	87.09	Moderate	Agricultural Land-Crop	2	71.55
35	24.27	87.15	Moderate	Rural	2	82.48
36	24.27	87.22	Moderate	Agricultural Land-Crop	2	71.55
37	24.27	87.28	Moderate	Agricultural Land-Crop	3	72.00
38	24.27	87.34	High	Rural	2	87.45
39	24.21	87.15	Moderate	Agricultural Land-Crop	3	72.00
40	24.21	87.22	Moderate	Agricultural Land-Crop	2	71.55
41	24.21	87.28	Moderate	Agricultural Land-Crop	3	72.00
42	24.21	87.34	Moderate	Agricultural Land-Crop	2	71.55
43	24.15	87.22	High	Forest-Scrub	7	65.72
44	24.15	87.28	High	Waterbodies	2	94.41
45	24.15	87.34	High	Agricultural Land-Crop	2	77.51

TABLE 3
Details of the Clusters Generated Using Modified Curve Number (MCN)

Spatial Cluster ID	Clustered Grid Locations (longitude in $^\circ E$, latitude in $^\circ N$)	Average Modified Curve Number	Water Containing Area (sq. km)	Average Spatial Distance (km)
1	(87.03,24.57), (87.09 ,24.57), (86.90,24.51), (86.96, 24.51), (87.03 , 24.51)	79.480	21.237	52.039
2	(87.15, 24.57)	80.000	05.334	47.718
3	(87.21,24.57), (87.15,24.51), (87.21,24.51), (87.28,24.51), (87.15,24.45), (87.21,24.45)	78.250	07.297	39.239
4	(87.09,24.51), (86.90,24.45), (86.96,24.45), (87.03,24.45), (87.09,24.45), (86.90,24.39), (86.96,24.39), (87.03,24.39), (87.09,24.39), (86.96,24.39), (87.03,24.33), (86.96,24.27)	69.885	41.342	41.536
5	(87.09,24.33), (87.03,24.27), (87.09,24.27)	74.183	04.299	27.735
6	(87.28,24.45), (87.15,24.39), (87.21,24.39), (87.28,24.39), (87.15,24.33), (87.15,24.27), (87.15,24.21)	75.299	12.922	24.761
7	(87.21,24.33), (87.28,24.33), (87.21,24.27), (87.28,24.27), (87.34,24.27), (87.21,24.21), (87.28,24.21), (87.34,24.21)	74.639	44.438	13.547
8	(87.21, 24.15), (87.28, 24.15), (87.34, 24.15)	79.212	39.573	04.576



Cluster ID	Spatial importance/ weights (%)
Cluster-1	09.77
Cluster-2	06.94
Cluster-3	07.45
Cluster-4	13.39
Cluster-5	07.50
Cluster-6	09.47
Cluster-7	17.80
Cluster-8	27.70

Fig. 7. Spatial importance/ weights (in percentage) for different clusters.

following inferences can be drawn about the proposed FORWARD model:

i) From the Tables 4, 5, 6, and 7, it is evident that the proposed SpaBN-based approach (FORWARD) has resulted in the highest value of NSE compared to the standard BN, statistical ARIMA and ANN models. Moreover, the value of NSE in almost all the cases is $\sim$ 1, indicating a highly accurate forecast made by FORWARD. On the other side, the NSE values corresponding to other prediction models, including standard BN, highly deviate from 1. *This proves the pre-eminence of FORWARD over the other prediction models.*

ii) The lower values of NRMSD (0.07-0.15), computed for all the prediction years, indicate the efficacy of proposed approach compared to the other techniques (Tables 4, 5, 6, and 7). This also *ensures that the incorporation of spatial information has improved the accuracy for FORWARD.*

iii) From the D_v and SEP values in Tables 4, 5, 6, and 7, it can also be inferred that the proposed approach is on average more than 55 percent better than the statistical forecasting models (ARIMA models), and almost 25 percent better than the artificial neural network (ANN)-based prediction technique. Moreover, *the performance of the SpaBN-based approach has improved about 13 percent with respect to the standard Bayesian network handling no spatial information.*

iv) In order to get the view of fitness of the forecasting methods, the R^2 and CC values have also been estimated as displayed in the Tables 4, 5, 6, and 7. R^2 and CC value (magnitude) ranges from 0 to 1, and the higher the value of R^2 and CC, the better the model fits for prediction. From the tables, it may be observed that *in most of the cases of prediction, the proposed FORWARD approach provides a high R-squared value $\sim$1*, whereas the R-squared value for ANN-based model is $\sim$ 0.3, and that for the standard BN and ARIMA models are $\sim$ 0.6 and 0.0 respec-

TABLE 4
Prediction Year 1998: Comparative Study of Proposed Approach (FORWARD) with Existing Prediction Techniques

Prediction Techniques	Prediction Year 1998					
	NRMSD	NSE	D_v	SEP	R^2	CC
Exponential Model [Holt-Winters Approach]	0.315	00.000	004.17	19.82	0.000	0.000
Automated ARIMA	0.518	−1.703	−22.79	32.58	0.000	0.000
ANN (feed-forward back propagation)	0.315	−0.003	002.61	19.85	0.575	0.758
Standard BN (SBN)	0.192	00.630	004.85	12.06	0.663	0.814
Proposed Approach **(FORWARD)**	**0.157**	**00.751**	**005.39**	**09.89**	**0.805**	**0.897**

TABLE 5
Prediction Year 1999: Comparative Study of Proposed Approach (FORWARD) with Existing Prediction Techniques

Prediction Techniques	Prediction Year 1999					
	NRMSD	NSE	D_v	SEP	R^2	CC
Exponential Model [Holt-Winters Approach]	0.475	−1.036	059.20	46.21	0.000	0.000
Automated ARIMA	0.474	−1.029	059.08	46.14	0.061	0.247
ANN (feed-forward back propagation)	0.462	−0.926	−16.95	44.94	0.286	0.534
Standard BN (SBN)	0.203	00.626	016.65	19.80	0.679	0.824
Proposed Approach **(FORWARD)**	**0.099**	**00.910**	**004.96**	**09.71**	**0.940**	**0.969**

TABLE 6
Prediction Year 2000: Comparative Study of Proposed Approach (FORWARD) with Existing Prediction Techniques

Prediction Techniques	Prediction Year 2000					
	NRMSD	NSE	D_v	SEP	R^2	CC
Exponential Model [Holt-Winters Approach]	0.550	−2.705	168.55	101.42	0.000	0.000
Automated ARIMA	0.550	−2.706	168.56	101.42	0.166	0.409
ANN (feed-forward back propagation)	0.419	−1.152	052.07	077.29	0.073	0.270
Standard BN (SBN)	0.168	00.654	012.12	031.00	0.719	0.848
Proposed Approach **(FORWARD)**	**0.068**	**00.942**	**004.09**	**012.65**	**0.973**	**0.986**

TABLE 7
Prediction Year 2001: Comparative Study of Proposed Approach (FORWARD) with Existing Prediction Techniques

Prediction Techniques	Prediction Year 2001					
	NRMSD	NSE	D_v	SEP	R^2	CC
Exponential Model [Holt-Winters Approach]	0.266	0.458	45.92	53.07	0.000	0.000
Automated ARIMA	0.266	0.459	45.66	53.05	0.001	0.032
ANN (feed-forward back propagation)	0.256	0.500	12.49	50.98	0.178	0.422
Standard BN (SBN)	0.254	0.506	34.40	50.68	0.357	0.598
Proposed Approach **(FORWARD)**	**0.097**	**0.928**	**08.98**	**19.39**	**0.937**	**0.968**

tively. Similar observations can be found for the *CC* value as well.

v) Time series of the observed daily reservoir live capacities and the model forecasts for the five principal prediction models for the validation period 1998-2001 are shown in the Fig. 8. From the figure, it is clear that *the outcome from the proposed SpaBN-based forecast model, FORWARD, is matching well with the actual value of live capacity in all the prediction years (1998-2001), thus indicating better model efficiency*.

Though in few cases the final prediction value is considerably an over/under-estimation, it is evident from the figure that, whenever there is

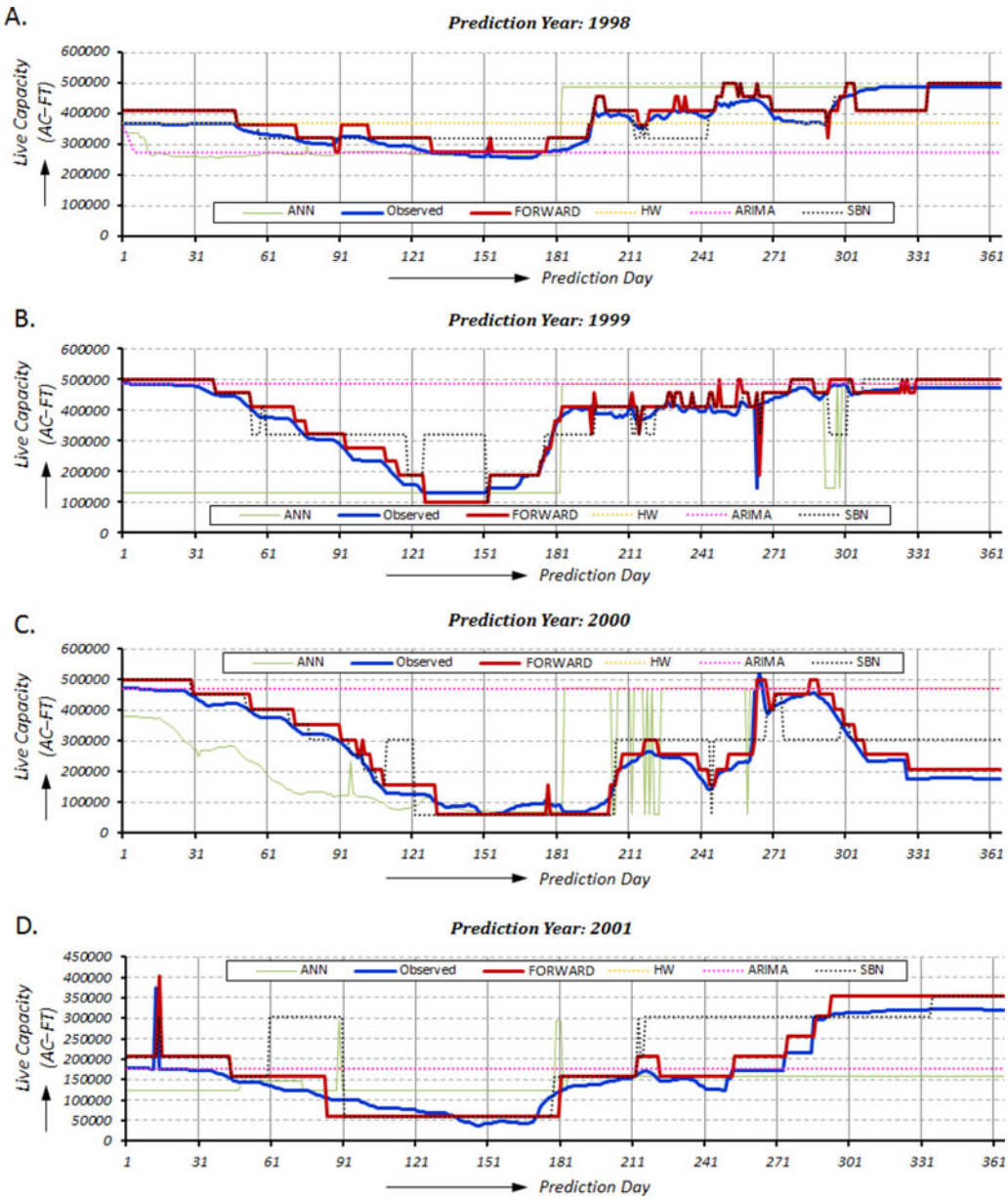


Fig. 8. Graphical plot for predicted and observed series: A. Year 1998, B. Year 1999, C. Year 2000, D. Year 2001.

over/under-estimation produced by standard BN, the proposed SpaBN has a notable tendency to improve it by making it as near to the observed value as possible. Therefore, the proposed approach is found to be more versatile than the standard BN. Moreover, the problem of over/under-estimation can be reduced to a greater extent by discretizing the variable ranges into smaller sub-ranges. Inclusion of additional predictors like evapotranspiration (ET) and evaporation from the watershed and reservoir water surface may further improve the proposed methodology.

6 Conclusion

A proper assessment of reservoir water dynamics is of utmost importance, since it has a significant impact on the industrial, agricultural, and socio-economic development of any region. Various meteorological variables, e.g., rainfall, temperature etc., are the key factors which influence the natural hydrological processes in reservoir, and for better understanding of the same, a proper modeling is necessary. The present work proposes a spatial Bayesian network based approach (FORWARD), considering various meteorological and spatial parameters, for forecasting the reservoir dynamics on a daily basis. FORWARD can intrinsically model the impact of spatial variability of various influencing factors over the associated river-catchment area. The performance of FORWARD has been evaluated over a case study on forecasting daily live capacity of *Mayurakshi reservoir* for a duration of four years (1998-2001). Six popular statistical parameters, namely *Normalized Root Mean Square Deviation (NRMSD)*, *Nash-Sutcliffe efficiency (NSE)*, *Mean percent deviation* (D_v), *Percent standard error of prediction (SEP)*, *Coefficient of determination* (R^2), and *Correlation coefficient (CC)*, have been used as the measures of goodness-of-fit criteria. The overall estimates (*NRMSD*: 0.1 ± 0.05, *NSE*: 0.88 ± 0.13, D_v: 5.8 ± 3.1, *SEP*:12.91 ± 0.48, R^2: 0.91 ± 0.1, *CC*: 0.95 ± 0.05) have proved the efficacy as well as preeminence of FORWARD in comparison with several other existing methods.

Though in this work, FORWARD has been illustrated with respect to reservoir live capacity prediction, the generic structure of this model can easily be extended to various domains by incorporating appropriate domain knowledge. Moreover, the SpaBN can be treated as a *generic machine learning technique*, which can be used for long-range dependency analysis by modeling the spatial influence of variables from neighboring locations in a large spatial region. It can be applied as a *space-time model* in a wide range of applications, including meteorological prediction, surveillance of epidemics [38], traffic flow modeling [39], [40], and so on.

References

[1] M. Das and S. K. Ghosh, "Short-term prediction of land surface temperature using multifractal detrended fluctuation analysis," in *Proc. Annu. IEEE India Conf.*, 2014, pp. 1–6.

[2] B. Bates, et al., *Climate Change and Water*. Budapest, Hungary: Intergovernmental Panel Climate Change, 2008.

[3] S. Piao, et al., "The impacts of climate change on water resources and agriculture in China," *Nature*, vol. 467, no. 7311, pp. 43–51, 2010.

[4] D. Legesse, C. Vallet-Coulomb, and F. Gasse, "Hydrological response of a catchment to climate and land use changes in Tropical Africa: Case study South Central Ethiopia," *J. Hydrology*, vol. 275, no. 1, pp. 67–85, 2003.

[5] N. S. Christensen, A. W. Wood, N. Voisin, D. P. Lettenmaier, and R. N. Palmer, "The effects of climate change on the hydrology and water resources of the Colorado River basin," *Climatic Change*, vol. 62, no. 1–3, pp. 337–363, 2004.

[6] P. Coulibaly, F. Anctil, and B. Bobee, "Multivariate reservoir inflow forecasting using temporal neural networks," *J. Hydrologic Eng.*, vol. 6, no. 5, pp. 367–376, 2001.

[7] P. Chaves and T. Kojiri, "Deriving reservoir operational strategies considering water quantity and quality objectives by stochastic fuzzy neural networks," *Advances Water Resources*, vol. 30, no. 5, pp. 1329–1341, 2007.

[8] M. Das, S. K. Ghosh, V. Chowdary, A. Saikrishnaveni, and R. Sharma, "A probabilistic nonlinear model for forecasting daily water level in reservoir," *Water Resources Manage.*, vol. 30, no. 9, pp. 3107–3122, 2016.

[9] P. Coulibaly, "Reservoir computing approach to Great Lakes water level forecasting," *J. Hydrology*, vol. 381, no. 1, pp. 76–88, 2010.

[10] O. Kisi, J. Shiri, and B. Nikoofar, "Forecasting daily lake levels using artificial intelligence approaches," *Comput. Geosciences*, vol. 41, pp. 169–180, 2012.

[11] F.-J. Chang and Y.-T. Chang, "Adaptive neuro-fuzzy inference system for prediction of water level in reservoir," *Advances Water Resources*, vol. 29, no. 1, pp. 1–10, 2006.

[12] B. Bazartseren, G. Hildebrandt, and K.-P. Holz, "Short-term water level prediction using neural networks and neuro-fuzzy approach," *Neurocomputing*, vol. 55, no. 3, pp. 439–450, 2003.

[13] S. Ondimu and H. Murase, "Reservoir level forecasting using neural networks: Lake Naivasha," *Biosystems Eng.*, vol. 96, no. 1, pp. 135–138, 2007.

[14] I. N. Daliakopoulos and I. K. Tsanis, "Comparison of an artificial neural network and a conceptual rainfall–runoff model in the simulation of ephemeral streamflow," *Hydrological Sci. J.*, vol. 61, pp. 2763–2774, 2016.

[15] M. Mustafa, M. Isa, and R. Rezaur, "Artificial neural networks modeling in water resources engineering: Infrastructure and applications," *World Academy Sci. Eng. Technol.*, vol. 62, pp. 341–349, 2012.

[16] H. R. Maier, A. Jain, G. C. Dandy, and K. P. Sudheer, "Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions," *Environ. Model. Softw.*, vol. 25, no. 8, pp. 891–909, 2010.

[17] M. Huang, J. Gallichand, Z. Wang, and M. Goulet, "A modification to the soil conservation service curve number method for steep slopes in the Loess Plateau of China," *Hydrological Processes*, vol. 20, no. 3, pp. 579–589, 2006.

[18] R. R. Vatsavai, A. Ganguly, V. Chandola, A. Stefanidis, S. Klasky, and S. Shekhar, "Spatiotemporal data mining in the era of big spatial data: Algorithms and applications," in *Proc. 1st ACM SIGSPATIAL Int. Workshop Analytics Big Geospatial Data*, 2012, pp. 1–10.

[19] J. Bothwell and M. Yuan, "A kinematics-based GIS methodology to represent and analyze spatiotemporal patterns of precipitation change in IPCC A2 scenario," in *Proc. 19th ACM SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2011, pp. 152–161.

[20] S. USDA, *National Engineering Handbook, Section 4: Hydrology*. Washington, DC, USA: Univ. Minnesota, 1972.

[21] L. Jiang, H. Zhang, and Z. Cai, "A novel bayes model: Hidden naive bayes," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 10, pp. 1361–1371, Oct. 2009.

[22] A. Fernández, M. Morales, C. Rodríguez, and A. Salmerón, "A system for relevance analysis of performance indicators in higher education using Bayesian networks," *Knowl. Inf. Syst.*, vol. 27, no. 3, pp. 327–344, 2011.

[23] Y. Li, J. Chen, and L. Feng, "Dealing with uncertainty: A survey of theories and practices," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 11, pp. 2463–2482, Nov. 2013.

[24] T.-L. Wong and W. Lam, "Learning to adapt Web information extraction knowledge and discovering new attributes via a Bayesian approach," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 4, pp. 523–536, Apr. 2010.

[25] M. Das and S. K. Ghosh, "A probabilistic approach for weather forecast using spatio-temporal inter-relationships among climate variables," in *Proc. 9th IEEE Int. Conf. Ind. Inf. Syst.*, 2014, pp. 1–6.

[26] V. Chowdary, D. Ramakrishnan, Y. Srivastava, V. Chandran, and A. Jeyaram, "Integrated water resource development plan for sustainable management of Mayurakshi Watershed, India using remote sensing and GIS," *Water Resources Manage.*, vol. 23, no. 8, pp. 1581–1602, 2009.
[27] Bhuvan, "Indian Geo-Platform of ISRO," 2015. [Online]. Available: http://bhuvan.nrsc.gov.in/bhuvan_links.php#, Accessed on: Nov. 8, 2015.
[28] R, "R-3.2.2 for windows (32/64 bit)," 2015. [Online]. Available: https://cran.r-project.org/bin/windows/base/old/3.2.2/ Accessed on: Aug. 27, 2015.
[29] MATLAB, "Mathworks," 2014. [Online]. Available: http://in.mathworks.com/products/matlab/?requestedDomain=www.mathworks.com, Accessed on: Oct. 19, 2014.
[30] ESRI, "Arcgis for desktop," 2015. [Online]. Available: http://www.esri.com/software/arcgis/arcgis-for-desktop, Accessed on: Aug. 11, 2015.
[31] I. N. Daliakopoulos, P. Coulibaly, and I. K. Tsanis, "Groundwater level forecasting using artificial neural networks," *J. Hydrology*, vol. 309, no. 1, pp. 229–240, 2005.
[32] S. J. Russell, P. Norvig, J. F. Canny, J. M. Malik, and D. D. Edwards, *Artificial Intelligence: A Modern Approach*, vol. 2. Upper Saddle River, NJ, USA: Prentice Hall, 2003.
[33] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *Int. J. Forecasting*, vol. 20, no. 1, pp. 5–10, 2004.
[34] A. J. Conejo, M. A. Plazas, R. Espinola, and A. B. Molina, "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 1035–1042, May 2005.
[35] J. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models part I—a discussion of principles," *J. Hydrology*, vol. 10, no. 3, pp. 282–290, 1970.
[36] I. Pulido-Calvo and J. C. Gutiérrez-Estrada, "Improved irrigation water demand forecasting using a soft-computing hybrid model," *Biosystems Eng.*, vol. 102, no. 2, pp. 202–218, 2009.
[37] S. Mohanty, M. K. Jha, A. Kumar, and D. Panda, "Comparative evaluation of numerical model and artificial neural network for simulating groundwater flow in Kathajodi–Surua Inter-Basin of Odisha, India," *J. Hydrology*, vol. 495, pp. 38–51, 2013.
[38] Y. Matsubara, Y. Sakurai, W. G. van Panhuis, and C. Faloutsos, "FUNNEL: Automatic mining of spatially coevolving epidemics," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 105–114.
[39] W. Shen, et al., "Traffic velocity prediction using GPS data: IEEE ICDM contest task 3 report," in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2010, pp. 1369–1371.
[40] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing, "Discovering spatio-temporal causal interactions in traffic data streams," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1010–1018.

Monidipa Das received the ME degree in computer science and engineering from the Indian Institute of Engineering Science and Technology (IIEST), Shibpur, India, in 2013. She is currently working toward the PhD degree in the Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Kharagpur, India. Her research interests include spatial and spatio-temporal data mining, soft computing, and machine learning. She is a student member of the IEEE.

Soumya K. Ghosh is a professor in the Department of Computer Science and Engineering, IIT Kharagpur. Before joining IIT Kharagpur, he worked for the Indian Space Research Organization in the area of Satellite Remote Sensing and GIS. His research interests include spatial informatics and spatial web services. He has more than 200 research papers in reputed journals and conference proceedings. He is a member of the IEEE.

Pramesh Gupta received the BTech degree in computer science and engineering from the Indian Institute of Technology Kharagpur, India, in 2016. He has previously worked at Directi, Bengaluru, as a software engineering intern and IIIT Hyderabad as a research intern. His current research is primarily focused on spatio-temporal analysis of geographic information system (GIS) data.

V. M. Chowdary is currently a scientist and head (applications) with the Regional Remote Sensing Centre-East, Kolkata, National Remote Sensing Centre (NRSC), ISRO, India. His areas of interest include: application of geospatial technologies, multi-criteria analysis and soft computing tools for agricultural water management, integrated watershed management, hydrological modeling, and land use/cover changes. He has published widely in peer reviewed journals. He is the recipient of the Eminent scientist award from NESA and Team excellence award from ISRO.

Ravoori Nagaraja was a former chief general manager with Regional Centres, NRSC, ISRO, India. He was responsible for executing national projects like National Wasteland Mission, National Land use Mission, Large Scale Digital Database, and Web based Information System. He was the project director of Space based Information Support for Decentralized Planning (SIS-DP) for generation of High Resolution Ortho-Image base and large scale natural resources database of the country. He received several prestigious awards, namely Indian Geographical Society (IGS) (1997), Indian National Remote Sensing Award (1999), ISRO Team Excellence Award (2009), Astronautical Society of India (ASI) Team Achievement Award (2009), National Geomatic Award (2012), and ISRO Team Leader Award (2015).

V. K. Dadhwal is currently director in the Indian Institute of Space Science and Technology, India, and former director, NRSC (ISRO), India. He has made significant scientific contributions in diverse applications of EO data including agriculture, biogeochemical cycle of carbon, climatology, forestry, geo-hazards, geosciences, hydrology, land surface processes, land use and land transformation, meteorology, and oceanography. He has coauthored more than 250 peer-reviewed journal articles and is the editor of the *Journal of the Indian Society of Remote Sensing*. He is a fellow of International Academy of Astronautics, National Academy of Agricultural Sciences, and ISRS. He received the Bhaskara Award (2013).