Expert Systems With Applications 117 (2019) 211–227

Contents lists available at ScienceDirect

Expert Systems With Applications

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

FB-STEP: A fuzzy Bayesian network based data-driven framework for spatio-temporal prediction of climatological time series data

Monidipa Das, Soumya K. Ghosh*

Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, 721302, India

ARTICLE INFO

Article history:
Received 20 November 2017
Revised 15 August 2018
Accepted 31 August 2018
Available online 13 September 2018

Keywords:
Spatio-temporal analysis
Multivariate prediction
Computational intelligence
Fuzzy Bayesian network
Multifractal analysis
Climatic time series

ABSTRACT

With the recent development of computational intelligence (CI), data-driven models have gained growing interest to be applied in various scientific disciplines. This paper aims at proposing a hybrid CI-based data-driven framework as a complement for the physics-based models used in climatological prediction. The proposed framework, called *FB-STEP*, is based on a combination of *fuzzy Bayesian strategy* and *multifractal analysis technique*. The focus is to address *three* major research challenges in multivariate climatological prediction: (1) modeling complex spatio-temporal dependency among climatological variables, (2) dealing with non-linear, chaotic dynamics in climatic time series, and (3) reducing epistemic uncertainty in the data-driven prediction process. The present work not only explores Fuzzy-Bayesian modeling of spatio-temporal processes, but also presents an elegant approach of dealing with intrinsic chaos in time series, through a synergism between multifractal analysis and Bayesian inference mechanism. Similar concepts may also be successfully employed in developing expert or intelligent systems for wide range of applications, including reservoir-water dynamics modeling, flood monitoring, traffic flow modeling, chemical-mechanical process monitoring, and so on. Thus, the present research work carries a significant value not merely in the field of climate research, but also in the domains of AI and machine intelligence. The experimentation has been carried out to *spatio-temporally extrapolate* the climatic conditions of five different locations in *India*, with the help of historical data on *temperature, humidity, precipitation rate*, and *soil moisture*. A comparative study with popular linear and non-linear methods has validated the efficacy of the proposed data-driven approach for climatological prediction.

1. Introduction

Climatological prediction, defined as forecasting of environmental variables, like temperature, precipitation rate, wind speed, humidity etc. in a given geographic location, is challenging as well as important for adopting appropriate future mitigation and adaptation measures. In general, several climate models like, *Community Atmosphere Model (CAM), Community Climate System Model (CCSM), Hadley Centre Coupled Model (HadCM3)* etc. (Kirtman et al., 2012) are popularly used for monitoring and studying the climatological processes. However, these global/regional climate models are based on the *physics-based approaches* involving several differential equations on various physical processes, and suffer from two major limitations: (1) the models assume that all the physical systems are well understood, which may not be true in reality; (2) the models need to solve a number of differential equations, and thus, are computationally expensive (NIPCC, 2014). Therefore, in addition to these climate models, various *linear* and *non-linear data-driven approaches* have recently been emerged as the new paradigms, which mainly aim to extensively analyze the historical climate data for generating insights, and then utilize those in further studies. The *linear approaches* are mostly based on the auto regressive integrated moving average (ARIMA), whereas the *non-linear methods* are mainly based on artificial neural network (ANN), standard Bayesian network (BN), support vector machines (SVM) etc. computational intelligence (CI) techniques.

Now, the key challenges in climatological prediction with such *data-driven approaches* mainly arise due to the *inherent chaotic nature* of the climate data and the *complex, non-linear dynamics* of the climate system itself. Since the parameters underlying the *non-linear* and *deterministic* climate system dynamics are sometimes *unknown*, the system properties cannot be determined by proper analysis and simulations of the associated equations (Drignei, Forest, Nychka et al., 2008). So, given the historical data on climatic time series, there is a need to define a mechanism for capturing the rhythm in climate system dynamics to understand the clima-

* Corresponding author.
E-mail addresses: monidipadas@hotmail.com (M. Das), skg@cse.iitkgp.ac.in (S.K. Ghosh).

https://doi.org/10.1016/j.eswa.2018.08.057
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

tological processes in a better way. Moreover, the climatological data is a kind of *spatio-temporal data*, thus, unlike the classical data, these are *embedded in continuous space* and show *long range spatio-temporal dependency* with high autocorrelation (Faghmous & Kumar, 2014). These dependencies can be local in nature, involving spatial and temporal spans in a neighborhood, or there may be *long-range tele-connections* and *long memory time series effects*. All these dependencies in climate data cannot be effectively captured by conventional approaches which are often used to model local dependencies in various domains such as image, speech, video, and signal analysis etc. Therefore, another key challenge in data-driven climatological prediction is to define a model for *studying the complex spatio-temporal inter-relationships* among different climate variables to gain a better understanding of the climate system behavior. Because of this strongly non-linear, highly uncertain, and time-varying characteristics of the climate system, none of the linear data-driven methods (Box, Jenkins, & Reinsel, 2008; Chatfield, 2013; Holt, 2004; Riahy & Abedi, 2008) can be considered as a single superior model. These traditional linear statistical approaches are not only too simple to model complex climatological processes, but also suffer from backward looking problem, and therefore, often result in poor prediction performance by generating the same value as the output for the entire predicted time series. In order to overcome the shortcomings of linear models, a number of non-linear data-driven prediction models, especially based on Bayesian analysis and Artificial Neural Network or ANN, have been proposed in recent days. Although the existing ANN-based prediction models (Abhishek, Kumar, Ranjan, & Kumar, 2012; Nayak, Patheja, & Waoo, 2012; Nourani, Mogaddam, & Nadiri, 2008; Venkadesh, Hoogenboom, Potter, & McClendon, 2013) are fairly tractable and well-performing for time series prediction, these require large training time and also are not able to directly utilize spatial features for spatial/spatio-temporal dependency modeling for climatological data. Moreover, the ANN models are less explored to the uncertainty management issues and do not have any mechanism to explicitly handle the intrinsic chaos in climatological data. On the other side, though the Bayesian network based models (Aguilera, Fernández, Fernández, Rumí, & Salmerón, 2011; Cofıno, Cano, Sordo, & Gutierrez, 2002; Das & Ghosh, 2014a; Madadgar & Moradkhani, 2013; Nandar, 2009) are inherently capable of modeling uncertainty, these approaches suffer from exponential time and space requirement, and also lack natural-chaos handling property. Similar problem is also faced in the fuzzy-rule-based prediction system proposed by Awan and Awais (2011). Contrarily, though the time series prediction approach proposed by Das and Ghosh (2014b) attempts to model chaotic nature of the data, the approach is unable to handle spatial dependencies, and consequently, lacks spatial extrapolation capability.

The primary focus of the present paper is illustrated in Fig. 1. The objective is to exploit the innate potential of the computational intelligence (CI) techniques for developing an improved data-driven framework which attempts to address the three above-discussed challenges in climatological prediction, namely, (i) epistemic uncertainty; (ii) long range spatio-temporal dependency; and (iii) non-linear, chaotic nature of climate data. In our proposed framework (termed as FB-STEP), a new fuzzy Bayesian network based analysis mechanism has been introduced to address the first two issues. The mechanism also helps to reuse information, and assists in managing large dataset. Additionally, another module, performing multifractal analysis of the climatic time series data, has been incorporated to capture the intrinsic regularity which handles the third issue discussed above.

The novelty in this work lies in incorporating *spatial information* in *fuzzy Bayesian network*, and refining the network-inferred values of climatological variables by a data tuning process based on *multifractal analysis*. The empirical study in comparison with other data-driven methods proves the superiority of our proposed *FB-STEP* in climatological prediction. Incidentally, *FB-STEP* is applicable not only for climatological data, but also for other kinds of spatial time series, especially those having such inherent chaotic nature.

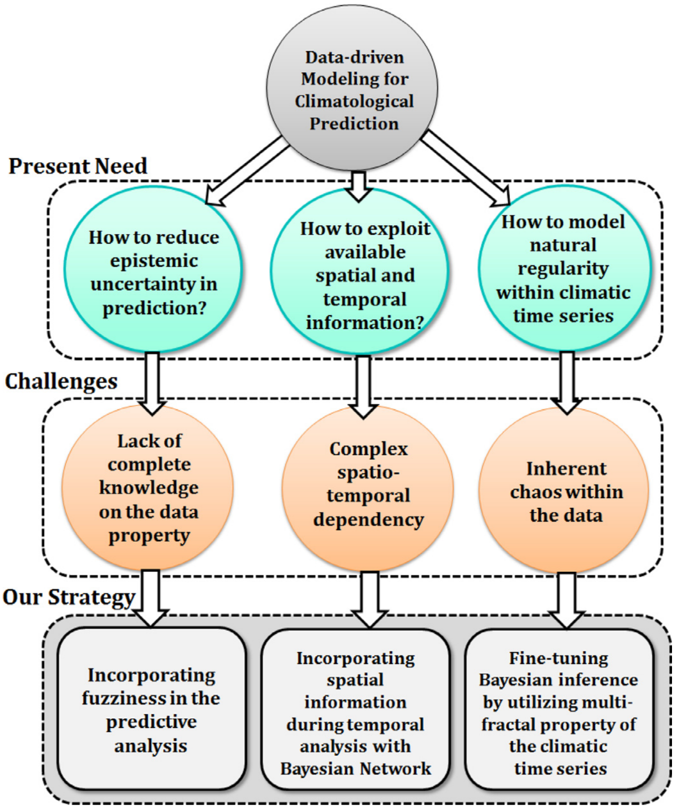

Data-driven Modeling for Climatological Prediction
Present Need
How to reduce epistemic uncertainty in prediction?
How to exploit available spatial and temporal information?
How to model natural regularity within climatic time series
Challenges
Lack of complete knowledge on the data property
Complex spatio-temporal dependency
Inherent chaos within the data
Our Strategy
Incorporating fuzziness in the predictive analysis
Incorporating spatial information during temporal analysis with Bayesian Network
Fine-tuning Bayesian inference by utilizing multi-fractal property of the climatic time series

Fig. 1. Conceptualization of the research problem addressed in the paper.

1.1. Problem statement and contributions

The broad objective of the present work is climatological time series prediction which can be formally stated as follows:

- Given, the historical *daily time series* dataset over n climate variables in $V=\{v_1, v_2, \cdots, v_n\}$, corresponding to a set of locations $Loc=\{loc_1, loc_2, \cdots\}$ for previous t years: $\{y_1, y_2, \cdots, y_t\}$. Also given, the spatial attribute information $SA=\left\{SA_1^{loc}, SA_2^{loc}, \cdots, SA_p^{loc}\right\}$ regarding each location $loc \in Loc$. The problem is to determine the daily climatic conditions of any location $x \in (Loc \cup Z)$ for future years $\left\{y_{(t+1)}, y_{(t+2)}, \cdots\right\}$, in terms of the state/values of the variables in V, when the spatial attributes of x is observed as $\left\{SA_1^x, SA_2^x, \cdots, SA_p^x\right\}$. Here, Z is a set of k new locations $\{z_1, z_2, \ldots, z_k\}$, such that $z_i \notin Loc$, for $i=1$ to k.

The problem, as stated above, is a kind of spatio-temporal extrapolation that needs to predict the future climatic condition of not only the set of training locations but for the other locations outside the training set as well. In this regard, the current work proposes a *multivariate, data-driven prediction framework (FB-STEP)* based on a *new fuzzy Bayesian approach* followed by *multifractal analysis*. The proposed fuzzy Bayesian approach extends our previous work (Das & Ghosh, 2014a), by including *spatial information* in the learning framework. Besides, it also overcomes the *cascading effect of prediction error* in the multifractal analysis based prediction approach proposed in our earlier work (Das & Ghosh, 2014b). Unlike the linear statistical (Box et al., 2008; Chatfield, 2013; Holt, 2004; Riahy & Abedi, 2008) and several other non-linear prediction models, the FB-STEP is highly suitable for applying in complex real-world scenarios. FB-STEP is generic and flexible enough to be applied not only on the climatological data, but also on the time series from various other domains of applications, with little modifications in its data pre-processing step. Further, in contrast to the existing Bayesian models (Aguilera et al., 2011; Das & Ghosh, 2014a; Madadgar & Moradkhani, 2014; Nandar, 2009), the proposed FB-STEP is able to capture and utilize the natural regularities within time series data, and thereby, provides notably better performance with low uncertainty in complex time series prediction. Moreover, since the ANN-based prediction models, proposed by Abhishek et al. (2012), Venkadesh et al. (2013) and Nayak et al. (2012), etc., neither have special treatment for the spatial features nor have the capability to handle intrinsic chaos in data, our proposed FB-STEP outperforms these models from both the perspectives of *accuracy* and *uncertainty* in prediction. Unlike the state-of-the-art space-time prediction model HBAR (Sahu & Bakar, 2012), FB-STEP also does not require special software for its exact implementation and realization purpose. Eventually, FB-STEP may become a more economical solution for real-world issues, such as prediction of weather condition using data collected from distributed weather-stations/sensor-network, assessment and monitoring of flood through prediction of water dynamics in natural reservoirs, and so on. The present work has been evaluated with respect to prediction of daily climatic conditions of *five* different locations in India during 2015–2016. The training has been performed with the historical datasets (Microsoft-Research, 2015) of *temperature, precipitation rate, humidity,* and *soil moisture*, corresponding to three locations (three cities in India), namely *Kolkata* (22.58°N, 88.36°E), *Raipur* (21.25°N, 81.63°E), and *Lucknow* (26.85°N, 80.91°E); whereas the prediction has been made for two more locations, namely *Baleshwar* (21.49°N, 86.93°E) and *Kharagpur* (22.33°N, 87.24°E), as well. The accuracy of the prediction demonstrates the efficacy of the proposed approach.

Thus, the major contributions in this work can be summarized as follows:

1. Proposing FB-STEP, a *data-driven* framework for *multivariate prediction* of climatological time series over *space as well as time*;
2. Introducing *fuzziness* in predictive analysis to *reduce* the *epistemic uncertainty* in prediction process;
3. Incorporating *spatial information* during *temporal analysis* with *fuzzy Bayesian network*, to model the *spatio-temporal interrelationships* among climate variables;
4. Modeling *intrinsic regularities* within climatic time series, with the incorporated mechanism based on *multifractal analysis*;
5. Verifying the effectiveness of the proposed framework using an empirical study on spatio-temporal prediction of *temperature, humidity, precipitation rate*, and *soil moisture*, for *five* different locations (*Kolkata, Raipur, Lucknow, Baleshwar,* and *Kharagpur*) in India.

The remainder of the paper is organized as follows: The proposed spatio-temporal prediction framework (*FB-STEP*) has been thoroughly discussed in Section 2. A detailed description of the experimentation with climatological data has been provided in Section 3. The section starts with the details of used datasets and study area, followed by an exhaustive analysis of the experimental results. Finally, the concluding remarks have been presented in Section 4.

2. FB-STEP: a fuzzy Bayesian network driven framework for spatio-temporal prediction

The overall framework for the proposed prediction approach along with the flow of entire process is shown in Fig. 2. As shown in the figure, the proposed prediction framework (FB-STEP) consists of three key modules corresponding to: (1) Capturing spatio-temporal inter-relationships among climate variables, (2) Measuring intrinsic regularities in each considered climatic time series, and (3) Incorporating the natural regularities in multivariate prediction. The details of each module are discussed in the following part of this section. The meanings of the various notations used throughout the paper are summarized in Table 1.

2.1. Module-1: capturing spatio-temporal inter-relationships

Be it observed or model-simulated, the climatological data has complex dependencies across space as well as time. One way to model these relationships/dependencies is to capture various features of these dependencies through statistical modeling, estimation, testing, and inference. In this respect, we have utilized *probabilistic analysis* with *fuzzy Bayesian network*. The fuzziness incorporated in the Bayesian network model also helps in reducing the epistemic uncertainty arising due to lack of knowledge over the typical properties of the data. The process of capturing spatio-temporal relationship consists of two major steps: (a) Data preprocessing, and (b) Relationship learning. It takes as input the historical data of past years, and a causal dependency graph of Bayesian network over considered variables. The output of this module is a trained Bayesian network along with the incorporated spatio-temporal relationships for the prediction year.

2.1.1. Data preprocessing

The historical data is processed to determine the interval size for different climatological variables for discretization purpose. The interval size is determined based on the maximum and minimum value observed in the training data of the variable. 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 size of the interval becomes:

$$I_s(v_i) = \frac{[max(v_i) - min(v_i) + 1]}{I} \quad (1)$$

where, I is the total number of discretized range value of v_i. The value of I may be predefined intuitively, or can be determined empirically so that it leads to optimum result with respect to prediction accuracy as well as execution time. In order to empirically determine the optimal number of discretized ranges for a particular variable, first, a threshold value of execution time is assumed. Then, the prediction accuracy (say in terms of root mean square error or RMSE) is studied with the increasing value of range count (I). The value of I, for which the error becomes minimum (and the execution time remains within the threshold), is considered to be the optimal range count.

The discretized ranges of values are then fuzzified to aid in accurate prediction and uncertainty management issue. In our proposed approach, the fuzzification has been done by assigning the membership values in intuitive manner. The procedure is dependent on the historical data and the respective domain knowledge from the experts. The step of fuzzification, used in our proposed methodology, deals with the uncertainty introduced through discretization of the climatological time series data. Whenever a time series data is discretized, problem arises with the crisp boundary values (Jun, Chung, Kim, & Kim, 2013) leading to introduction of some added impreciseness or uncertainty in the model. Following is an example, illustrating the same.

Suppose the variable temperature (T) in a particular region can take values between 15 °C and 40 °C. So, one may discretize the

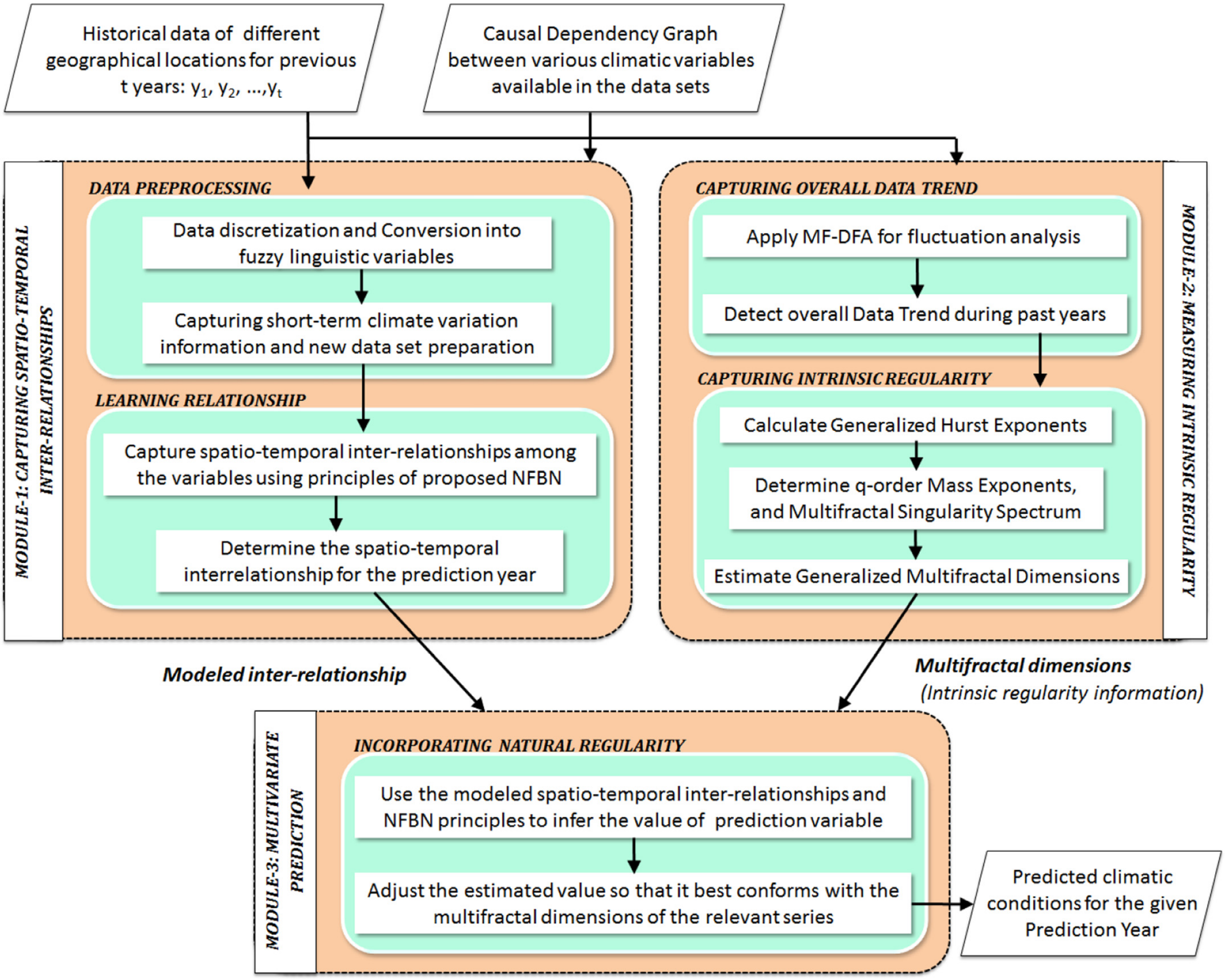

Historical data of different geographical locations for previous t years: y1, y2, …,yt
Causal Dependency Graph between various climatic variables available in the data sets
MODULE-1: CAPTURING SPATIO-TEMPORAL INTER-RELATIONSHIPS
DATA PREPROCESSING
Data discretization and Conversion into fuzzy linguistic variables
Capturing short-term climate variation information and new data set preparation
LEARNING RELATIONSHIP
Capture spatio-temporal inter-relationships among the variables using principles of proposed NFBN
Determine the spatio-temporal interrelationship for the prediction year
CAPTURING OVERALL DATA TREND
Apply MF-DFA for fluctuation analysis
Detect overall Data Trend during past years
CAPTURING INTRINSIC REGULARITY
Calculate Generalized Hurst Exponents
Determine q-order Mass Exponents, and Multifractal Singularity Spectrum
Estimate Generalized Multifractal Dimensions
MODULE-2: MEASURING INTRINSIC REGULARITY
Modeled inter-relationship
Multifractal dimensions
(Intrinsic regularity information)
MODULE-3: MULTIVARIATE PREDICTION
INCORPORATING NATURAL REGULARITY
Use the modeled spatio-temporal inter-relationships and NFBN principles to infer the value of prediction variable
Adjust the estimated value so that it best conforms with the multifractal dimensions of the relevant series
Predicted climatic conditions for the given Prediction Year

Fig. 2. Process flow of the proposed spatio-temporal prediction framework (FB-STEP).

values, say in five ranges, like that shown in Table 2. Once discretized, the problem may arise when we want to use these ranges to qualify the data. For example, say T_1 is low temperature, T_2 is moderately low temperature, and T_3 is average temperature. Then, this means that 24.999 °C temperature is moderately low but 25.001 °C is average temperature. That is, the boundary temperatures are treated strictly within one sub-range and not in others. In order to overcome this problem, each range is fuzzified in intuitive manner, based on the knowledge from the domain experts. The idea is to represent each discretized sub-range in terms of a trapezoidal fuzzy number. The mid-value of the sub-range is assigned a membership value of 1 and the other values (may be outside the range as well) are assigned suitable memberships ($\in [0, 1]$) depending on the characteristics of the associated variable, which can efficiently be suggested by the domain experts. Fig. 3 shows example membership functions for the variable "temperature". In the similar fashion, all the other variables are fuzzified.

Now, the prediction for a single day may not always require the dataset of whole year for training purpose, since the concerned variables may show short term (weekly, monthly, seasonal etc.) variation, for which the data of corresponding time duration is more suitable than the whole data to train with. Again, a training with previous year's data of only that particular day may not always be sufficient. Hence, there always remains a need of having an optimal training dataset. In order to achieve the same, this step utilizes the short term climatic variations within the historical time series data, and eventually helps in handling large training dataset through information reuse process.

Given a variable $v \in V$ and a training year y_i, in order to determine the short term variation within a period of d days, the corresponding daily time series ($series_v$) is first divided into L_d number of segments, each of size d, such that $L_d = \lfloor |series_v|/d \rfloor$. Here $|series_v|$ denotes the series length. Then, for each segment s, the series variance is measured as follows:

$$var(s, d) = \frac{1}{d}\sum_{j=1}^{d}\{series_v[(s-1)*d+j] - mean(s, d)\}^2 \qquad (2)$$

where, $mean(s, d) = \frac{1}{d}\sum_{j=1}^{d} series_v[(s-1)*d+j]$ is the series mean for the segment s. Therefore, for the entire series ($series_v$), the overall short-term variance within d days becomes:

$$shortVar(d) = \frac{1}{L_d}\sum_{s=1}^{L_d} var(s, d) \qquad (3)$$

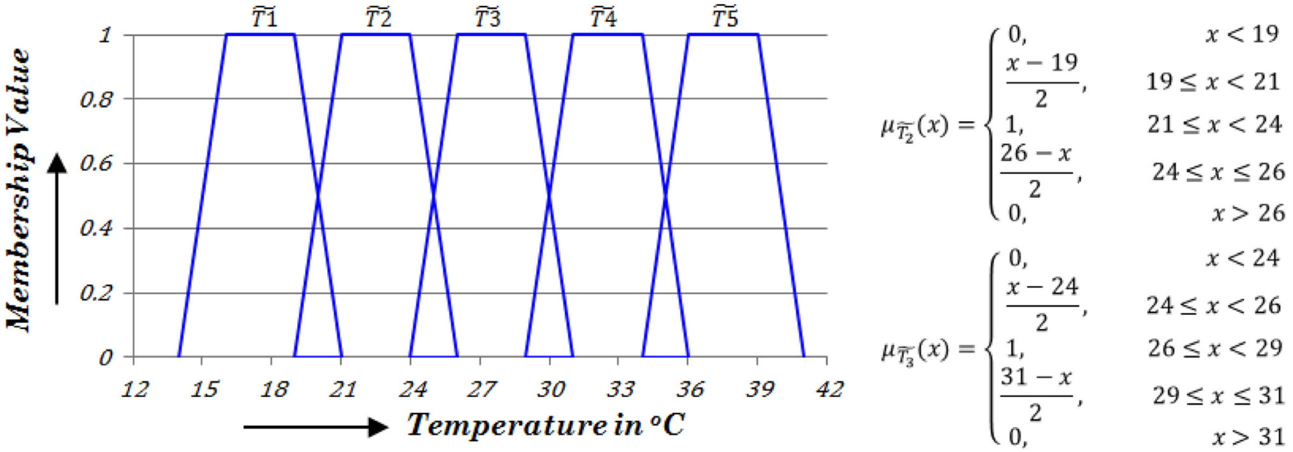
If, for $d = 365$ the $shortVar(d)$ has the minimum value and also tends to 0, then the series is said to have *yearly* variation. Similarly,

Table 1
Symbols and notations used in the present paper.

Notation	Meaning
$\mu_{\tilde{A}}(x)$	Fuzzy membership of the value x in the fuzzy set $\tilde{A}$
$\tilde{A}$	Fuzzy set corresponding to an event A
CA	Set of observed climate variables
$D(q)$	q-order generalized multifractal dimension
d_i	Temporal distance of year y_i from the prediction year
$Fluct_q(l)$	q-order fluctuation over series segment of length l
$h(q)$	q-order generalized Hurst exponents
I	Total number of interval or discretized range value
$I_s(v_i)$	Size of interval (discretized range) for a variable v_i
$[l_j, u_j]$	jth range/discretized value (lower limit: l_j and upper limit: u_j)
L_l	Total count of series segments each having length l
Loc	Set of spatial locations with known historical data
$max(v_i)$	Maximum observed value for the variable $v_i \in V$
$mean(s, l)$	Mean or average value of a series segment s of length l
$min(v_i)$	Minimum observed value for the variable $v_i \in V$
n	Total number of climatological variables considered
$P(A)$	Marginal probability of occurrence of the event A
$P(A\|B)$	Conditional probability of occurrence of the event A, given evidence B
P_f	Probability estimate corresponding to the final year
$Poly_s^m$	mth order fitting polynomial over series segment s
P_{y_i}	Probability estimate corresponding to the year y_i
$R_j^{v_i}$	jth fuzzified range corresponding to the variable v_i
SA	Set of spatial attributes (e.g. latitude, elevation etc.)
SA_i^x	ith spatial attribute, associated with location x, which belong to SA
$series_x$	Time series corresponding to a variable $x \in V$
$\|series_x\|$	Length of time series corresponding to variable $x \in V$
$\overline{series_x}$	Mean of the time series corresponding to variable $x \in V$
$s\_profile$	Profile of the series
V	Set of climatological variables or the representative nodes
$var(s, l)$	Variance within a series segment s of length l
v_i	ith variable $\in V$
y_i	ith training year
Z	Set of new spatial locations with unknown data in the historical years

Table 2
Discretized range of temperature (T).

Ranges	T_1	T_2	T_3	T_4	T_5
Temperature (°C)	$15 \le T < 20$	$20 \le T < 25$	$25 \le T < 30$	$30 \le T < 35$	$35 \le T \le 40$



$$\mu_{\widetilde{T_2}}(x) = \begin{cases} 0, & x < 19 \\ \dfrac{x-19}{2}, & 19 \le x < 21 \\ 1, & 21 \le x < 24 \\ \dfrac{26-x}{2}, & 24 \le x \le 26 \\ 0, & x > 26 \end{cases}$$

$$\mu_{\widetilde{T_3}}(x) = \begin{cases} 0, & x < 24 \\ \dfrac{x-24}{2}, & 24 \le x < 26 \\ 1, & 26 \le x < 29 \\ \dfrac{31-x}{2}, & 29 \le x \le 31 \\ 0, & x > 31 \end{cases}$$

Fig. 3. Example for fuzzification of discretized ranges of temperature variable (T).

if, for $d = 30$ the *shortVar*(d) is the minimum and tends to 0, then the series is said to have *monthly* variation. If, for $d = 7$ the *shortVar*(d) is the minimum and tends to 0, then the series is said to have *weekly* variation. Else, the series is considered to have daily variation. Based on detected variation, new datasets corresponding to each training year are prepared in a manner as described in Algorithm 1.

These newly prepared datasets are used in the next step to train the fuzzy Bayesian network for capturing and modeling inter-variable spatio-temporal relationships for the corresponding training years.

2.1.2. Learning spatio-temporal relationship

In order to learn the spatio-temporal inter-relationships among climate variables, an extension of *new fuzzy Bayesian net-*

Algorithm 1: Dataset preparation.

```
/* This algorithm prepares new datasets for each training year based
on short-term climatic variations of the variables. */
Input : Historical dataset $H = \{H_{y_1}, H_{y_2}, \cdots, H_{y_t}\}$ of past $t$ years,
        Directed Acyclic Graph $G$ of the Bayesian network, Set of
        climate variables of interest $V$, and Prediction day $d$
Output: New dataset $TD_{y_i}$ corresponding to each training year
        $y_i (1 \le i \le t)$
$TD_{y_i} \leftarrow \phi$
for each variable $v \in V$ do
    $Ch_v$ = Set of child climate variables of $v$ in $G$
    if $v$ shows daily variation, or there exists $c \in Ch_v$ such that $c$ shows
    daily variation then
        for each training year $y_i (1 \le i \le t)$ do
            $TD_{y_i} \leftarrow TD_{y_i} \cup$ (whole-year data from $H_{y_i}$ for variable $v$)
        end
    end
    else if $v$ shows monthly variation, or there exists $c \in Ch_v$ such that $c$
    shows monthly variation then
        for each training year $y_i (1 \le i \le t)$ do
            $TD_{y_i} \leftarrow TD_{y_i} \cup$ (data from $H_{y_i}$ corresponding to the month
            of $d$ for variable $v$)
        end
    end
    else if $v$ shows weekly variation, or there exists $c \in Ch_v$ such that $c$
    shows weekly variation then
        for each training year $y_i (1 \le i \le t)$ do
            $TD_{y_i} \leftarrow TD_{y_i} \cup$ (data from $H_{y_i}$ corresponding to the week of
            $d$ for variable $v$)
        end
    end
    else if $v$ shows yearly variation then
        for each training year $y_i (1 \le i \le t)$ do
            $TD_{y_i} \leftarrow TD_{y_i} \cup$ (data from $H_{y_i}$ corresponding to the day $d$
            for variable $v$)
        end
    end
end
```

work (NFBN) learning, as proposed in our earlier work (Das & Ghosh, 2014a), has been utilized here. NFBN (Das & Ghosh, 2014a) is a variant of FBN (Tang & Liu, 2007). However, it is more precise and computationally more efficient than FBN (Tang & Liu, 2007).

New Fuzzy Bayesian Network (NFBN)

The working principle of NFBN (Das & Ghosh, 2014a) is as follows:

Let $A = \{A_1, A_2, \cdots, A_m\}$ and $B = \{B_1, B_2, \cdots, B_n\}$ be two sets of events corresponding to the variables x and y respectively —where, $A_1, \cdots, A_m$ and $B_1, \cdots, B_n$ are in the form of range of values achieved by x and y. Also let $\tilde{A}$ and $\tilde{B}$ be two corresponding fuzzy events.

Then according to *NFBN*,

$$P(\tilde{B}/\tilde{A}) = \frac{|\{m_i | \mu_{\tilde{B}}(y_{m_i}) > 0, \quad \mu_{\tilde{A}}(x_{m_i}) > 0\}|}{P(\tilde{A})}, \tag{4}$$

where, $m_i \in \{m_1, m_2, \cdots, m_M\}$, a set of all the observations for the variable x and y; M is the total number of such observations; x_{m_i}= Value of the variable x in the ith observation (m_i); y_{m_i}= Value of the variable y in the ith observation (m_i); $\mu_{\tilde{A}}(x_{m_i})$ = Membership of the value x_{m_i} in the fuzzy set $\tilde{A}$; and $\mu_{\tilde{B}}(y_{m_i})$ = Membership of the value y_{m_i} in the fuzzy set $\tilde{B}$.

Here, in *NFBN*, the fuzzy marginal probability $P(\tilde{A})$ is defined as:

$$P(\tilde{A}) = \frac{|\{n_i | \mu_{\tilde{A}}(x_{n_i}) > 0, n_i \in \{n_1, n_2, \cdots, n_N\}\}|}{N} \tag{5}$$

where, $\{n_1, \cdots, n_N\}$ is a set of all observations for the variable x; N is the total number of observations for x; and $\mu_{\tilde{A}}(x_{n_i})$ = Membership of the value x_{n_i} in the fuzzy set $\tilde{A}$.

The present work proposes an *extension* of spatio-temporal inter-relationship learning which is based on the principle of NFBN with *explicitly incorporated spatial information*. As shown in Fig. 4, the network (also called causal dependency graph) in the proposed learning framework not only consists of the climatological variables, but also explicitly includes the *spatial attributes* (*SA*s) for incorporating *spatial information*, like *land elevation, latitude, land-cover category* etc. depending on which the climatological variables show variant behavior. The incorporation of these spatial information in the network helps in modeling the spatio-temporal dependency among the climatological variables in a *more exhaustive* manner rather than considering their implicit influence as used by Das and Ghosh (2014a). Therefore, given the spatial attributes, like latitude, land elevation, land use land cover (LULC) type etc., the proposed framework is capable of forecasting climatological time series for any location outside the study/training region as well. However, better accuracy can be achieved by training the model with the historical time series data of a large set of locations with varying spatial attribute combinations.

Utilizing the Eqs. (4) and (5), the network is trained with the given data for each training year (y_1, y_2, $\cdots$, y_t, t= total number of available training years) separately, to learn the corresponding spatio-temporal relationships among the variables, in terms of probability estimates. In the Fig. 4, the network, separately trained for each training year, has been denoted by BN_{y_1}, $BN_{y_2}, \cdots$, BN_{y_t} respectively. At the end of training for each year, the fuzzy probabilities obtained for each considered variable are averaged to get the corresponding fuzzy probabilities for the prediction year. The averaging is performed in an weighted manner, with consideration to *temporal auto-correlation* among the historical years (Das et al., 2017). Temporal autocorrelation occurs when the course of a time series is influenced by its recent past. For example, the weather condition of a day in one year is more similar to that in its previous year than that in longer past. Therefore, based on this concept, the weighted average of the estimated probability values has been performed by assigning higher weights to the captured probabilities corresponding to a year which is nearer to the prediction year. For any training year y_i, if d_i is its temporal distance from the prediction year, then the final probability P_f is estimated as follows:

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

where, t is the total number of years considered for training; and P_{y_i} is the estimated marginal/ conditional probability of any variable, for the year y_i.

2.2. Module-2: measuring intrinsic regularity

The objective of the second module (refer Module-2 in Fig. 2) is to model the *intrinsic chaos*, or in other sense, the *intrinsic regularity* in each of the considered climatic time series. The climate system is governed by a variety of physical processes and exhibits a great deal of fluctuations especially at various temporal scales. It has been observed by research communities that these fluctuations or changes in climate system show fractal phenomenon having asymptotic power-law scaling for several long records (Lin & Fu, 2008). Moreover, the recent researches indicate that only a single scaling exponent is not sufficient to fully characterize the complex dynamics of any climatological time series. Therefore, the *multifractal analysis*, which can identify and quantify the multiple scaling exponents in the data, is more appropriate in this regard.

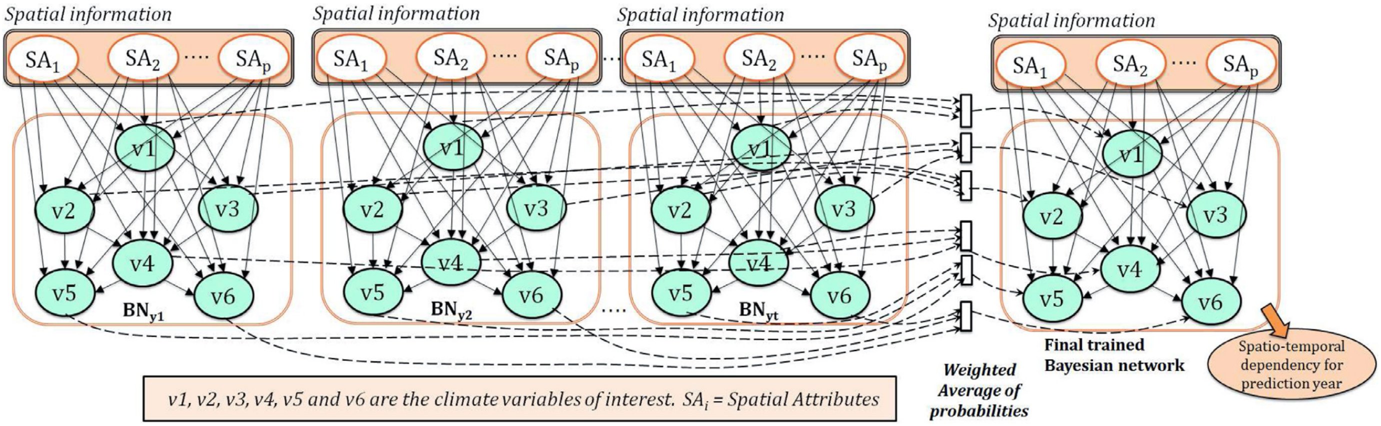

Spatial information
SA1
SA2
SAp
v1
v2
v3
v4
v5
v6
BNy1
BNy2
BNyt
Weighted Average of probabilities
Final trained Bayesian network
Spatio-temporal dependency for prediction year
v1, v2, v3, v4, v5 and v6 are the climate variables of interest. SAi = Spatial Attributes

Fig. 4. Proposed spatio-temporal relationship learning framework based on new fuzzy Bayesian network (NFBN) with incorporated *spatial information*.

In our proposed framework, the Module-2, introduced for capturing intrinsic regularity in climatic time series, is based on the *multifractal detrended fluctuation analysis (MF-DFA)* technique (Kantelhardt et al., 2002). As shown in Fig. 2, the module takes the data series of past years as input, and finally captures the regularity information in the form of *multifractal dimensions* of each series, which are then fed to the next module for final prediction. The whole module is comprised of two major steps: (a) Capturing data trend, and (b) Measuring intrinsic regularity.

2.2.1. Capturing data trend

In this step we apply MF-DFA technique to analyze the logarithmic plots of series fluctuations versus different lengths of time scale (refer Fig. 5(a)). The characteristics of these plots help in determining the actual trend in the data. The overall procedure is described below.

Let, the time series data be associated with a variable $v \in V$. So, as per the principles of MF-DFA, for each particular length of time scale (l), the profile of the $series_v$ is first divided into $L_l = \lfloor |series_v|/l \rfloor$ number of segments (s) starting from each end separately. Then, the qth order fluctuation in $series_v$ is estimated as follows:

$$Fluct_q(l) = \left\{ \frac{1}{2L_l} \sum_{s=1}^{2L_l} [F^2(l,s)]^{q/2} \right\}^{1/q} \tag{7}$$

where, $q \neq 0$, $s \in 1, 2, 3, \cdots, 2L_l$, and $F^2(l, s)$ is the local variance at the segment s. In case $q = 0$, the qth order fluctuation in $series_v$, for a particular length of time scale (l) is measured using logarithmic averaging procedure in following manner:

$$Fluct_q(l) = \exp\left\{ \frac{1}{4L_l} \sum_{s=1}^{2L_l} ln[F^2(l,s)] \right\} \tag{8}$$

Now, if $Poly_s^m$ is the mth order fitting polynomial for a segment s, and $s\_profile$ is the series profile, obtained by performing cumulative sum of series deviation from the series mean, then the value of $F^2(l, s)$ is calculated as follows:

$$F^2(l,s) = \frac{1}{l} \sum_{i=1}^{l} \{s\_profile[(s-1)*l+i] - Poly_s^m(i)\}^2 \tag{9}$$

when $s = 1, 2, \cdots, L_l$ and

$$F^2(l,s) = \frac{1}{l} \sum_{i=1}^{l} \{s\_profile[L-(s-L_l)*l+i] - Poly_s^m(i)\}^2 \tag{10}$$

when $s = L_l+1, L_l+2, \cdots, 2L_l$.

If the considered m is too small, then in the $log - log$ plot of $Fluct_q(l)$ vs l the $Fluct_2(l)$ shows a prominent crossover to a regime with larger slope for large scales l which disappears gradually with the increasing value of m, as shown in Fig. 5(a). Once the properly fitting polynomial degree m is finalized, the *trend in original data series* is estimated as $(m-1)$.

2.2.2. Measuring intrinsic regularity

In this step, we measure the intrinsic regularity in each climatic time series by estimating its generalized multifractal dimensions using MF-DFA technique. As mentioned earlier, MF-DFA technique primarily analyzes the *log-log* plots of series fluctuations versus different lengths of time scale. The plots show different slopes for different orders of fluctuation in case the series is multifractal. All these slopes collectively provide the *generalized Hurst exponents* (refer Fig. 5(b)), which are further utilized to determine the *multifractal dimensions* of the series.

Once the actual data trend m is captured for a particular series, the corresponding *generalized Hurst exponents* i.e. $h(q)$-values are estimated by solving the power law equation as follows:

$$Fluct_q(l) \propto l^{h(q)} \tag{11}$$

$$\log Fluct_q(l) = h(q)\log l + \log C \tag{12}$$

$$h(q) = \frac{\log Fluct_q(l)}{\log l} - \frac{\log C}{\log l} \tag{13}$$

$$h(q) = \frac{\log Fluct_q(l)}{\log l} + C' \tag{14}$$

where, C' is a constant.

The $h(q)$ is now used to calculate the *generalized multifractal dimensions*, denoted by $D(q)$ (refer Fig. 5(c)), in following manner:

$$D(q) = (qh(q)-1)/(q-1) \tag{15}$$

The $D(q)$ values basically represent the *chaotic nature* of the concerned series in the form of a set of non-integer dimensions, and are fed along with the estimated data trend information (m) to the next module to aid in final prediction.

2.3. Module-3: Incorporating natural regularities in multivariate prediction

The objective of the third module (*Module-3*, refer Fig. 2) in our proposed framework is to incorporate the natural regularities in multivariate prediction. The module basically tunes the inferred

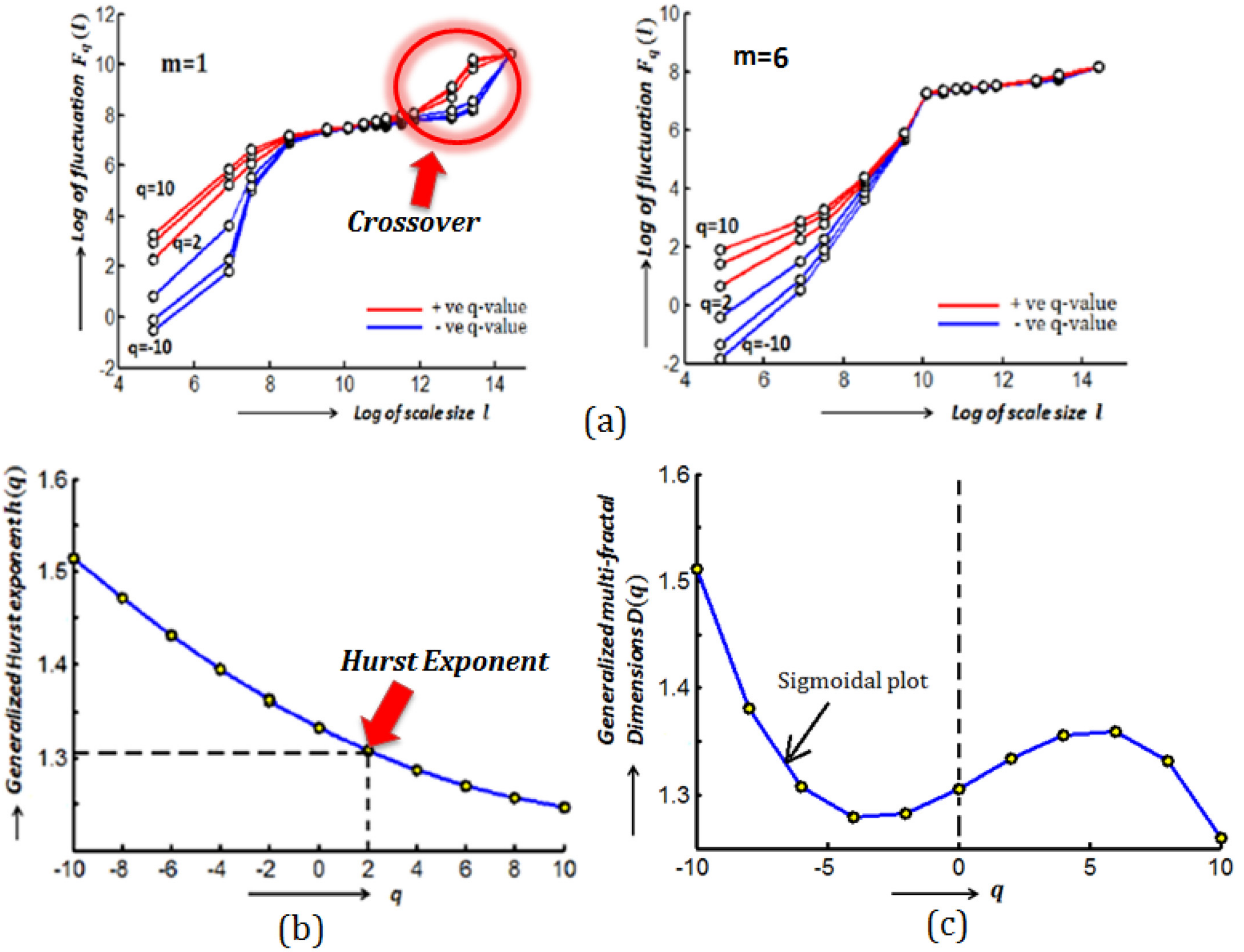


Fig. 5. Multifractal analysis: (a) *Crossover*, (b) *Generalized Hurst exponents h(q) vs. q*, (c) *Fitted sigmoidal curve* for *Generalized Multifractal Dimensions D(q) vs. q*.

value of prediction variable, as obtained using probability distributions from the first module, into some value which keeps the intrinsic regularity within the respective time series almost unaltered.

The proposed data tuning process is an upgraded version of the earlier work of Das and Ghosh (2014b). For each prediction day, the data tuning process in our earlier work starts by considering the predicted value of the previous day as seed value. As a result, the error of prediction, encountered in the previous day, cumulatively increases with the prediction for the next days, and this event restricts the approach (Das & Ghosh, 2014b) applicable for a short-term prediction only. In contrast, the present approach first infers the value of the concerned variable by utilizing the spatio-temporal inter-relationships as learnt in first module. Then this inferred value is treated as the seed value and is tuned further to get the final value of prediction, which conforms to the intrinsic regularity within the observed series. Therefore, our present data tuning process becomes independent of the prediction value of previous day and thereby overcomes the cumulative effect of prediction error, leading to efficient performance in long-term prediction as well. The process is accomplished by the *Module-3*, illustrated below.

As shown in Fig. 2, the *Module-3* takes as input the *probabilistic information* of the spatio-temporal dependency among the concerned variables and the *captured regularity* present in each of the climatic data series, as generated by the *Module-1* and *Module-2*, respectively. Finally, the module generates the forecast result in terms of the future states/values of the considered climatological variables. The forecasting is performed based on following two assumptions:

- The higher the inferred fuzzy probability of a particular state/value of a variable, the more the tendency of occurrence of that state/value for it.
- The future series corresponding to each climatic variable must be consistent with the regularity in the past series, expressed through generalized multifractal dimensions.

During the multivariate prediction in *Module-3* (refer Fig. 2), each climatic variable v_i is inferred from the given spatial information, by using the fuzzy Bayesian inference technique. The range/discretized values corresponding to the first and second highest fuzzy probabilities, say pr_1: $[l_1, u_1]$ and pr_2: $[l_2, u_2]$ respectively, are considered to finalize the inferred range ($[l_f, u_f]$) of v_i in a manner as described below:

Let, $R_j^{v_i}$ be the jth fuzzified range value corresponding to the variable v_i. $SA_1, \ldots, SA_p$ are the observed spatial attributes corresponding to the prediction location, and CA is the set of observed climate variables, i.e. $CA \subseteq V$ (as per the problem definition). Then, the first and second highest fuzzy probabilities i.e. pr_1 and pr_2 are estimated as follows:

$$pr_1 = P([l_1, u_1]/CA, SA_1, \dots, SA_p) \tag{16}$$

$$= \underbrace{max}_{\forall j} \left\{ P(R_j^{v_i}/CA, SA_1, \dots, SA_p) \right\} \tag{17}$$

and

$$pr_2 = P([l_2, u_2]/CA, SA_1, \dots, SA_p) \tag{18}$$

$$= \underbrace{second\_max}_{\forall j} \left\{ P(R_j^{v_i}/CA, SA_1, \dots, SA_p) \right\} \tag{19}$$

By utilizing these fuzzy probabilities (i.e. pr_1 and pr_2), we calculate the final inferred range as follows:

$$[l_f, u_f] = \left[B_{val} - \left(\frac{I_s(v_i) * pr_1}{pr_1 + pr_2} \right), B_{val} + \left(\frac{I_s(v_i) * pr_2}{pr_1 + pr_2} \right) \right] \tag{20}$$

where, B_{val} is the boundary value between $[l_1,\ u_1]$ and $[l_2,\ u_2]$, $l_1 < l_2$. i.e. $B_{val} = u_1 = l_2$ (as the discretized intervals are non-overlapping), and $I_s(v_i)$= size (or length) of interval/range for v_i. If $l_1 > l_2$, then the final value of inferred range can be obtained by exchanging pr_1 and pr_2 in Eq. (20).

Now, to *incorporate* the *natural regularity* as captured for each prediction variable, the central value of the inferred range is taken as the seed value, and various amounts of fluctuation is added to this seed value to generate a set of candidate prediction values. Then, each of these candidate values is separately appended at the end of the already obtained series and the newly formed series is checked for conformity with the original series, in terms of deviation from multifractal dimensions, using Eq. (21).

$$deviation = \sqrt{\frac{1}{(q_b - q_a + 1)} \sum_{q=q_a}^{q_b} [(D(q) - D_{new}(q)]^2} \tag{21}$$

where, $D_{new}(q)$ are the multifractal dimensions for the new series including the candidate value of prediction; $D(q)$ are the multifractal dimensions for the original series; $[q_a,\ q_b]$ is the sub-range of considered q-values (refer to the step of *Data trend capture*).

Now, the fluctuation amount $fluctn_{bestIndex}$, for which the adjusted forecast-value gives the least deviation from the original multifractal dimensions, is considered finally, and the final predicted value of v_i becomes:

$$best\_candidate\_val = \left\{ \frac{(l_f + u_f)}{2} + fluctn_{bestIndex} \right\} \tag{22}$$

The various steps of incorporating the natural regularity in prediction process have been presented through Algorithm 2 .

Algorithm 2: Incorporating natural regularity.

```
/* The algorithm incorporates natural regularity in prediction of a
climatic variable $X$ for future $k$ days. The series corresponding to $X$ is
denoted by $x(t)$, and the corresponding predicted series has been
denoted by $ps(t)$*/
Input  : Central values of the inferred range $v^i_{mid}$ $(1 \le i \le k)$, Degree of
         overall data trend $(m-1)$, and the generalized multifractal
         dimensions $D(q)$ for the original historical data series
         $x(t) = \{x_1, x_2, \cdots, x_d\}$.
Output : Predicted series $ps(t) = \{ps_1, ps_2, \cdots, ps_k\}$, for the next $k$
         days.
k=Number of prediction days.
d=Total number of observations in input series $x(t)$.
fluctn=Set of $g$ number of fluctuation values considered for
  adjustment purpose.
add(S, val)=A function that add/ include an observation val at the end
  of series S.
del_first(S)=A function that delete the first observation from the
  beginning of series S.
del_last(S)=A function that delete the last observation from the end of
  series S.
$TS \leftarrow$ { last $(d-1)$ entries from $x(t)$} $= \{x_2, x_3, \cdots, x_d\}$;
for each prediction day $i (1 \le i \le k)$ do
    for each considered fluctuation amount $fluctn_j (1 \le j \le g)$ do
        $candidate\_val \leftarrow (v^i_{mid} + fluctn_j)$;/* candidate_val is a
          candidate prediction value. */
        $TS \leftarrow add(TS, candidate\_val)$;
        Apply MF-DFA$_m$ to calculate the generalized Hurst exponents
          $h_{new}(q)$ for the new series TS;
        Calculate the multifractal dimensions
          $D_{new}(q) \leftarrow (qh_{new}(q) - 1)/(q-1)$ for TS;
        $deviation_j \leftarrow \sqrt{\frac{1}{(q_b-q_a+1)} \sum_{q=q_a}^{q_b} [D(q) - D_{new}(q)]^2}$ ;/*$[q_a, q_b]$ is
          the range of considered q-values for which $\alpha$ has a
          corresponding positive $f(\alpha)$ in the multifractal singularity
          spectrum $f(\alpha)$ vs. $\alpha$.*/
        $TS \leftarrow del\_last(TS)$;
    end
    $minDeviation \leftarrow minimum(deviation_1, \cdots, deviation_g)$;
    bestIndex ← value of $j$ $(1 \le j \le g)$ for which $deviation_j$ is equal to
      minDeviation;
    $best\_candidate\_val \leftarrow (v^i_{mid} + fluctn_{bestIndex})$;
    $ps_i \leftarrow best\_candidate\_val$;
    $TS \leftarrow add(TS, best\_candidate\_val)$;
    $TS \leftarrow del\_first(TS)$;
end
Print $ps_i (1 \le i \le k)$ as the predicted values of the prediction variable X
  for ith prediction day.
```

3. Experimentation

This section describes the dataset, experimental set up, and the various outcomes of our experimentation. The overall results are found to be encouraging.

3.1. Data

The experimentation has been carried out with a collection of sixteen-year (2001–2016) data, corresponding to three different training locations in *India*, namely *Kolkata* [22.58°N, 88.36°E], *Lucknow* [26.85°N, 80.91°E], and *Raipur* [21.25°N, 81.63°E] (refer Fig. 6). The location *Kolkata* is in *eastern India* and belongs to tropical climate zone, whereas the locations *Raipur*, and *Lucknow* belong to the temperate climate zone in *central India* and *north India* respectively. The experimental data are over four major climatological variables, namely, *Temperature, Relative humidity, Precipitation rate*, and *Soil moisture*, which have been collected from the Fetch-Climate Explorer (Microsoft-Research, 2015). Once the proposed model is trained, the testing on spatio-temporal extrapolation has been made for all the locations in the training set (*Kolkata, Raipur, and Lucknow*), and two more locations outside the set, namely, *Baleshwar, India (21.49°N, 86.93°E)*, and *Kharagpur, India (22.33°N, 87.24°E)* as well.

3.2. Experimental results

The performance of prediction using FB-STEP has been expressed in terms of prediction error (RMSE: *Root Mean Square Error* and MAE: *Mean Absolute Error* Wang, Xu, Tang, Yuan, & Wang, 2017) for two test years: 2015 and 2016, along with comparison to other existing methods, including exponential smoothing with Holt-Winters Approach (Holt, 2004), Automated ARIMA (R-Tool 3.1.1), Vector Auto-Regressive Moving Average or VARMA (De Gooijer & Hyndman, 2006; Tsay, 2013), Neural Network (NNTool, MATLAB R2011a), Hierarchical Bayesian Auto-Regressive model or HBAR (Sahu & Bakar, 2012), standard BN, and FBN

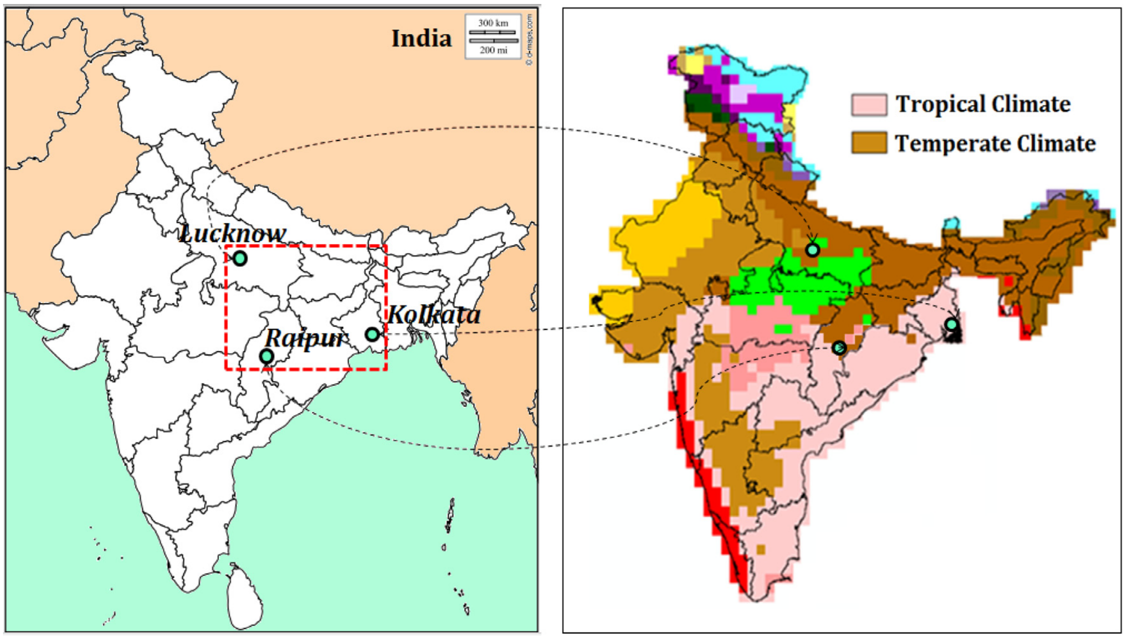


Fig. 6. Study area containing *three* training locations: *Kolkata, Raipur,* and *Lucknow*.

(Mrad, Delcroix, Maalej, Piechowiak, & Abid, 2012; Ryhajlo, Sturlaugson, & Sheppard, 2013; Tang & Liu, 2007). In order to make prediction for the year 2015, the datasets of 2001–2014 have been used as training dataset, whereas, the prediction for 2016 has been made based on the training datasets of 2001–2015.

Fig. 7(a)–(d) present the comparative results of predicting climatic condition in Kolkata, Lucknow, Raipur, Baleshwar, and Kharagpur, for the target year 2015, in terms of *Temperature, Relative humidity, Precipitation rate* and *Soil moisture* respectively. Similarly, Fig. 8(a)–(d) present the same for the prediction year 2016. The results of prediction for Baleshwar and Kharagpur, as depicted in Figs. 7 and 8, have been compared only with HBAR, BN and FBN. It is because the other models for comparison do not explicitly considers the spatial properties of test locations, and therefore are not fit for spatio-temporal extrapolation for any location *outside the training set*. Both for BN and FBN, the same causal dependency graph, as used in our approach, has been used for incorporating spatial information.

Moreover, in order to provide the *quantification for uncertainty* in prediction for each variable, we have determined the *Dawid-Sebastiani score* (Gneiting & Katzfuss, 2014) corresponding to our proposed FB-STEP approach and that for the other forecasting techniques as well. The score (*DSS*) is measured as follows:

$$DSS(F, y) = \left(\frac{y - \mu_F}{\sigma_F}\right)^2 + 2\ log\ \sigma_F \tag{23}$$

where, y is the observed value, F is the forecast time series, μ_F is the mean forecast value, and σ_F^2 is the variance of the forecast time series.

The *DSS* for prediction of *Temperature, Humidity, Precipitation* and *Soil moisture* have been tabulated in Table 3 to Table 6 respectively. Since the Holt-Winters Approach, ARIMA, VARMA, and the considered Neural Network model cannot extrapolate the time series for the location Kharagpur and Baleshwar, the tables shows the *DSS* values for three locations (Kolkata, Raipur, and Lucknow) only.

3.2.1. Discussion

From the experimental results (refer Figs. 7 and 8, and Tables 3–6) the following inferences can be drawn:

- As depicted in Figs. 7(a)–(d) and 8(a)–(d), the result of prediction using the proposed FB-STEP is far better than that of the other approaches with respect to both RMSE and MAE. It not only proves the worth of considering natural regularity (obtained using *multifractal analysis*) during prediction, but also establishes the effectiveness of our *extended* NFBN-based *learning* that considers the *spatial information* in an explicit manner.
- It is also evident from the Figs. 7 and 8 that with the increase in training data, the prediction error decreases, i.e. the prediction accuracy improves. This ensures the *consistency* of FB-STEP in multivariate prediction.
- Moreover, the proposed approach also shows the best performance in accomplishing spatio-temporal extrapolation, as depicted in the Figs. 7(a)–(d) and 8(a)–(d), corresponding to the two new locations, i.e. Baleshwar and Kharagpur respectively.
- It is evident from the Table 3–6 that the Dawid-Sebastiani scores for the proposed FB-STEP-based predictions are significantly less in most of the cases of prediction. The scores for FB-STEP are also very close to the *ideal scenario*, in which the predicted time series is same as that of the observed time series. This ensures that the prediction uncertainty in case of FB-STEP is also lesser than that of the other forecasting models, used in the comparative study.

Additionally, in the Figs. 9 and 10, we have plotted the predicted values of the considered climatological variables over two sample locations, namely *Raipur* and *Baleshwar*, for nine *randomly* selected days in the year 2015 and 2016 respectively. *Raipur* has been chosen as a representative of our training locations and *Baleshwar* has been chosen as a representative of the locations outside the training set. For each representative location, the prediction days have been randomly selected from three major observable seasons in the associated region: *pre-monsoon, monsoon*, and *post-monsoon*. It is apparent from the figures that the predictions made by proposed FB-STEP are more towards the actual ones, compared to the others.

Overall, the proposed FB-STEP produces least prediction error in most of the cases and delivers superior prediction performance. Since our proposed approach pre-processes the training data to capture the short-term climatic variation as described in

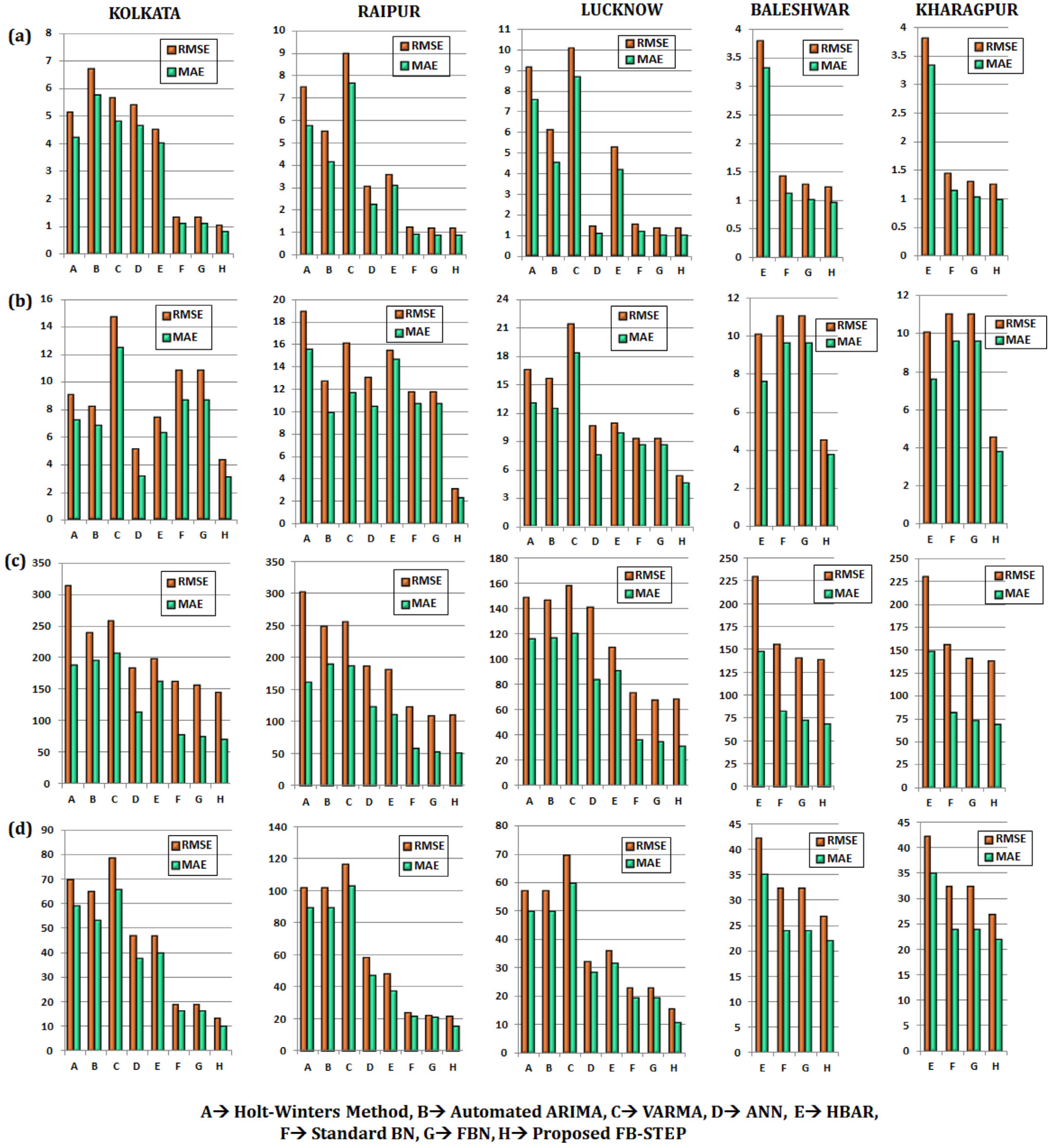


Fig. 7. Comparative study of proposed approach (FB-STEP) with existing prediction techniques considering various climatic variables for the prediction year 2015: (a) Temperature, (b) Humidity, (c) Precipitation rate (d) Soil Moisture.

Section 2.1.1(refer Algorithm 1), the fuzzy Bayesian network based training in case of the proposed approach becomes more effective than that in case of the other benchmark and state-of-the-art forecasting techniques considered. The fuzzification of the discretized data also helps to reduce the uncertainty in prediction. The other reason behind superior performance of FB-STEP is the refinement of the inferred value of the prediction variable by using the multifractal analysis. The multifractal analysis, as described in the Section 2.2, measures the intrinsic regularity in each of the time series under consideration and finally utilizes this inherent property to tune the value inferred by the trained fuzzy Bayesian network.

The Fig. 11 shows the *maximum percentage improvement* in prediction with incorporated regularity information. The improvement has been averaged over all the prediction locations.

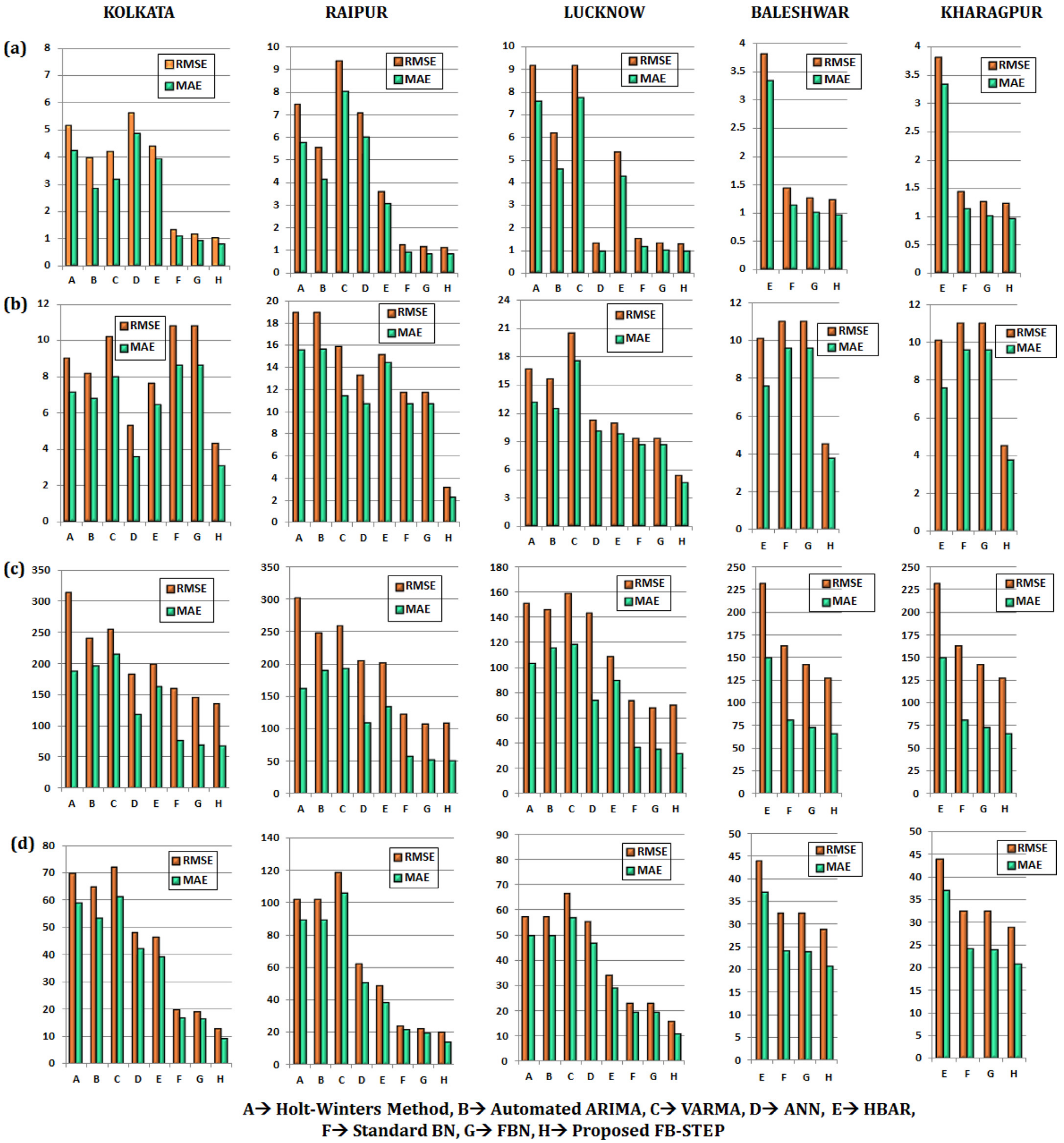


Fig. 8. Comparative study of proposed approach (FB-STEP) with existing prediction techniques considering various climatic variables for the prediction year 2016: (a) Temperature, (b) Humidity, (c) Precipitation rate (d) Soil Moisture.

The gain of this extra step of capturing and incorporating intrinsic regularity has also been quantified in terms of *99% confidence intervals of absolute prediction error*, as shown in Table 7. The upper and lower bounds of confidence intervals have been determined considering all the prediction locations used in the experimentation. It is evident from the tabulated values that the extra step of capturing and incorporating intrinsic regularity leads to improved performance with reduced prediction error. Considering training set and prediction locations from the same climate zone may yield even better performance.

3.2.1.1. Major implications. The overall experimental study clearly shows the utility of incorporating spatial information in climatological time series prediction. The study reveals that the climatic information from the neighboring locations with similar spatial properties can effectively help to determine the climatic condition

Table 3
Dawid-Sebastiani score (*DSS*) in prediction of temperature.

Forecasting techniques	Prediction year					
	2015			2016		
	Kolkata	Raipur	Lucknow	Kolkata	Raipur	Lucknow
Holt-Winters method	4.268	5.575	7.216	4.268	5.575	7.215
Automated ARIMA	5.130	4.402	4.715	3.863	4.423	4.752
VARMA	4.502	6.890	8.325	3.923	7.294	7.207
ANN	4.301	3.716	**3.311**	4.975	5.129	**3.290**
HBAR	4.037	3.751	4.323	3.998	3.490	4.345
Standard BN	3.372	3.351	3.346	3.372	3.351	3.346
FBN	3.372	3.353	3.336	3.355	3.350	3.332
Proposed FB-STEP	**3.347**	**3.348**	3.342	**3.346**	**3.349**	3.331
Ideal scenario	3.308	3.307	3.269	3.308	3.307	3.269

Table 4
Dawid-Sebastiani score (*DSS*) in prediction of humidity.

Forecasting techniques	Prediction year					
	2015			2016		
	Kolkata	Raipur	Lucknow	Kolkata	Raipur	Lucknow
Holt-Winters method	5.409	10.231	8.348	5.409	10.231	8.348
Automated ARIMA	5.188	6.851	8.111	5.188	10.323	8.110
VARMA	7.704	17.988	14.256	5.754	18.856	13.118
ANN	4.623	7.205	6.008	4.624	7.211	6.837
HBAR	5.043	21.824	6.184	5.113	19.541	6.170
Standard BN	5.800	7.103	5.763	5.800	7.103	5.763
FBN	**4.521**	7.103	5.763	5.800	7.103	5.763
Proposed FB-STEP	**4.521**	**4.162**	**4.521**	**4.521**	**4.161**	**4.520**
Ideal scenario	4.254	3.972	4.027	4.254	3.972	4.027

Table 5
Dawid-Sebastiani score (*DSS*) in prediction of precipitation.

Forecasting techniques	Prediction year					
	2015			2016		
	Kolkata	Raipur	Lucknow	Kolkata	Raipur	Lucknow
Holt-Winters method	3.3e + 08	2.9e + 08	2.7e + 08	3.3e + 08	2.9e + 08	4.0e + 08
Automated ARIMA	339.642	437.363	233.484	339.603	434.852	236.796
VARMA	507.254	570.512	375.89	411.633	587.186	401.004
ANN	3.2e + 05	5.1e + 05	7.2e + 04	2.5e + 05	3.4e + 05	6.3e + 05
HBAR	**167.559**	331.713	97.163	**167.924**	328.53	96.817
Standard BN	1323.648	949.661	109.699	1323.188	841.313	109.699
FBN	871.513	764.385	93.264	760.776	760.154	143.248
Proposed FB-STEP	587.771	**220.128**	**65.001**	723.606	**220.472**	**65.029**
Ideal scenario	−0.996	−2.267	−3.443	−0.996	−2.267	−3.443

Table 6
Dawid-Sebastiani score (*DSS*) in prediction of soil moisture.

Forecasting techniques	Prediction year					
	2015			2016		
	Kolkata	Raipur	Lucknow	Kolkata	Raipur	Lucknow
Holt-Winters method	27.365	46.395	25.105	27.365	46.395	25.105
Automated ARIMA	23.108	46.395	25.105	23.227	46.395	25.105
VARMA	35.798	84.882	39.357	28.959	88.571	35.437
ANN	16.52	29.116	11.083	16.029	28.954	24.148
HBAR	14.204	16.069	12.312	14.204	16.306	11.63
Standard BN	7.196	8.522	9.383	7.427	8.522	9.383
FBN	7.196	**8.022**	9.383	7.196	8.022	9.383
Proposed FB-STEP	**6.168**	8.035	**6.600**	**6.120**	**7.798**	**6.658**
Ideal scenario	5.402	5.295	5.166	5.402	5.295	5.166

of a very new spatial location for which no observed data from past years is available. The experimental outcomes also demonstrate the effectiveness of modeling intrinsic chaos within the associated time series, in a climatological prediction framework. Similar approach may be successfully employed for the prediction of various other natural as well as artificial time series, including hydrological and atmospheric time series, human heartbeat, respiratory excursions, and financial time series etc., which are inherently chaotic in nature.

Nonetheless, the proposed FB-STEP framework has its own limitations as well. First of all, the framework, being based on Bayesian network model, may require exponential time and space during

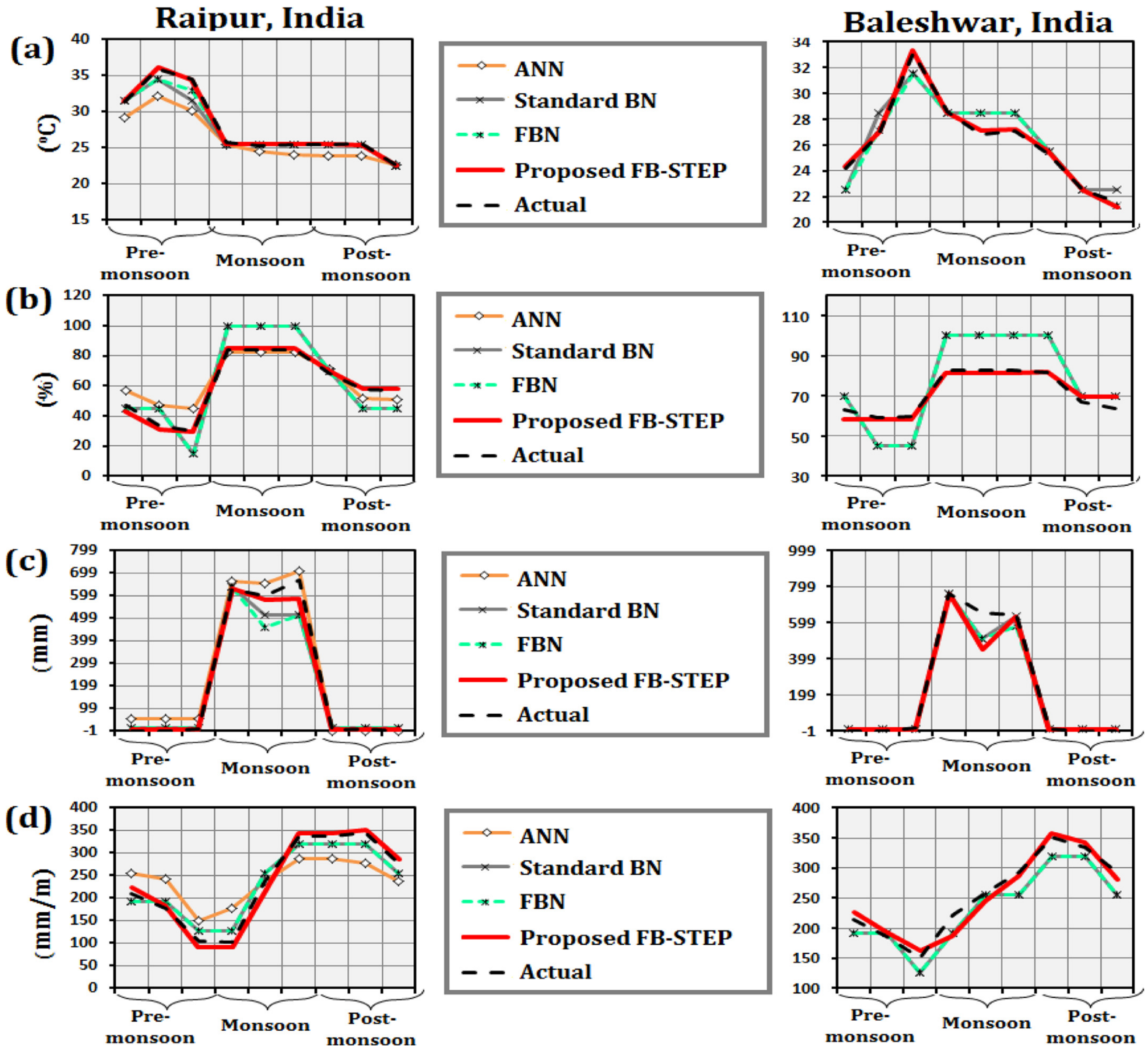


Fig. 9. Comparative study of prediction for the year 2015 in two sample locations (*Raipur* and *Baleshwar*): (a) Temperature, (b) Humidity, (c) Precipitation rate, (d) Soil moisture.

Table 7
99% confidence interval of absolute error in prediction for 2015 and 2016.

Prediction	Cases	2015		2016	
variable		Lower bound	Upper bound	Lower bound	Upper bound
Temperature	Case-1	00.791	00.996	00.816	01.025
(°C)	**Case-2**	**00.782**	**00.988**	**00.761**	**00.960**
Humidity	Case-1	06.895	07.908	08.587	09.913
(%)	**Case-2**	**02.913**	**03.617**	**02.912**	**03.615**
Precipitation	Case-1	42.177	68.086	39.623	64.877
(mm)	**Case-2**	**38.118**	**63.723**	**36.897**	**61.767**
Soil Moisture	Case-1	16.034	18.653	17.516	20.188
(mm/m)	**Case-2**	**10.312**	**13.507**	**09.530**	**12.731**

Case-1: Without considering extra step of capturing and incorporating intrinsic regularity;
Case-2: Considering extra step of capturing and incorporating intrinsic regularity;

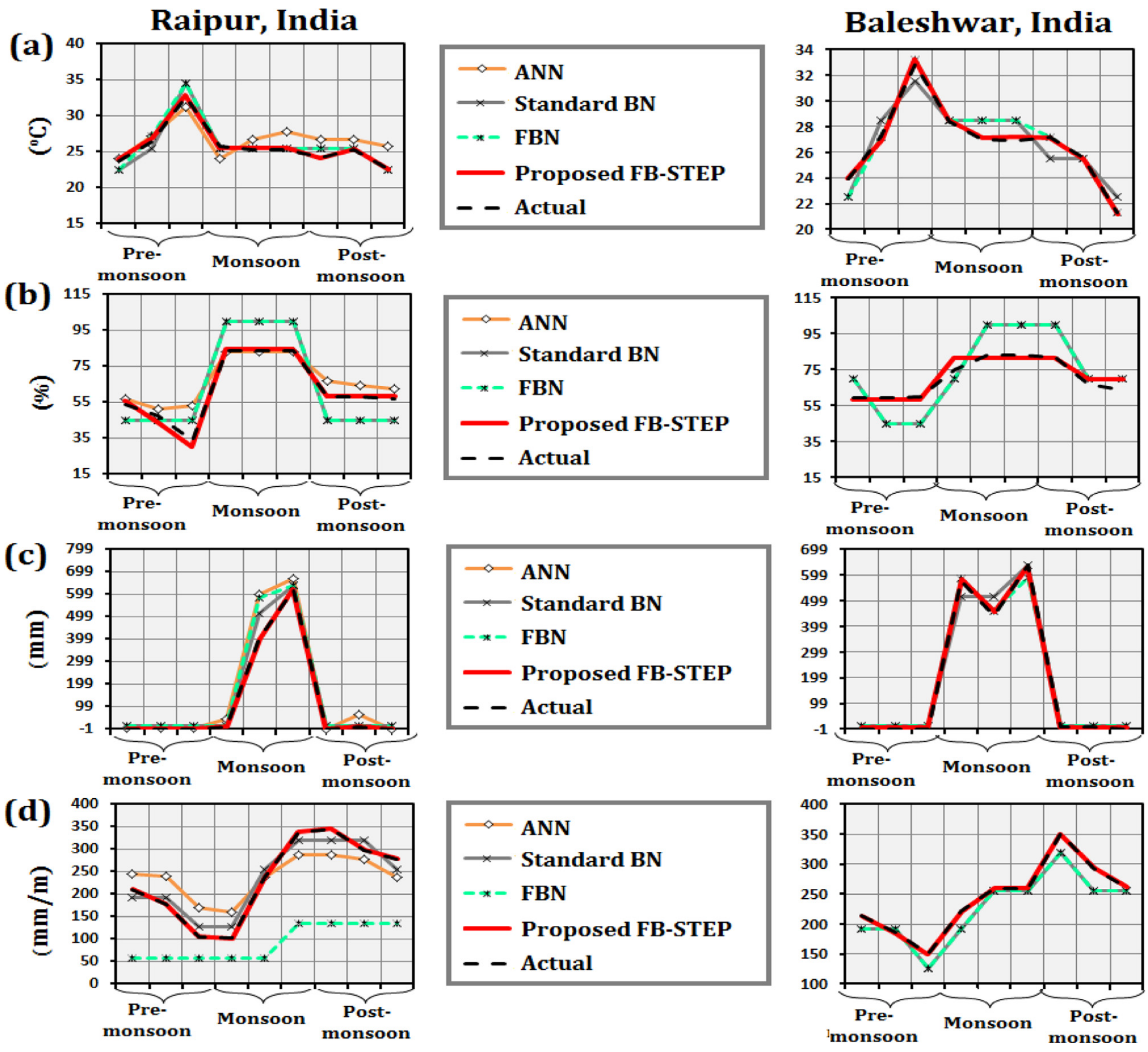


Fig. 10. Comparative study of prediction for the year 2016 in two sample locations (*Raipur* and *Baleshwar*): (a) Temperature, (b) Humidity, (c) Precipitation rate, (d) Soil moisture.

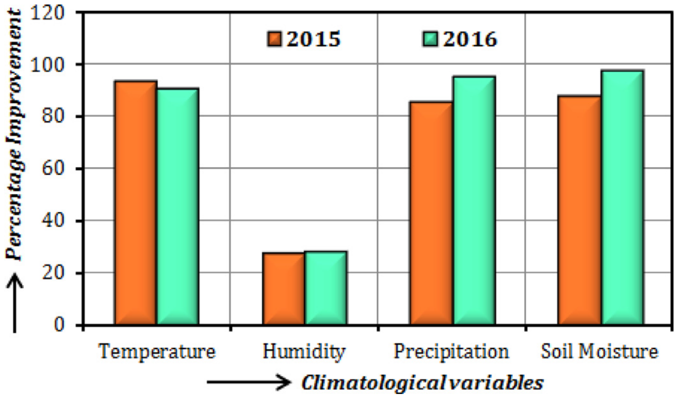
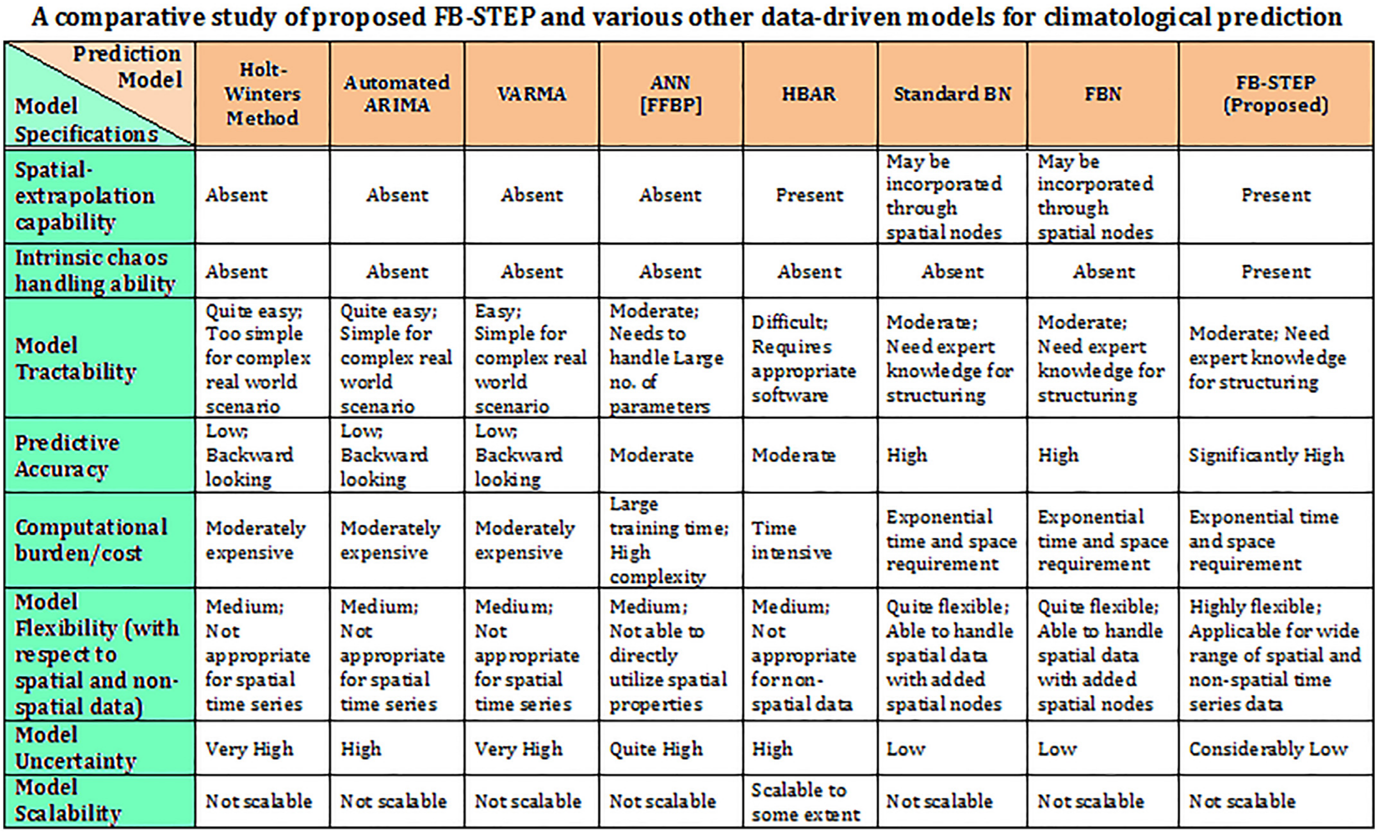


Fig. 11. Percentage improvement in prediction with consideration to the step of capturing and incorporating intrinsic regularity.

spatial dependency modeling. Secondly, since FB-STEP uses multifractal analysis, it becomes necessary to supply sufficiently long time series so that the intrinsic regularity within the time series is properly captured. Therefore, the framework is not very suitable when the length of training time series is considerably low. Further, like almost all the above-discussed prediction approaches, the FB-STEP framework is also not able to capture variability at varying spatial and temporal scales. Finally, the framework largely depends on the domain expertise, to determine the structure of the causal dependency graph and also to set the appropriate membership function for fuzzification. Consequently, this diminishes the tractability of the proposed framework. A summary of the strengths and the weaknesses of the proposed FB-STEP is presented through Fig. 12, in comparison with the other considered data-driven techniques.

4. Conclusions

Climatological events are typically non-linear and chaotic in nature, and thus it is extremely difficult to predict them accurately even with the help of state-of-the-art data-driven approaches. The present work proposes FB-STEP, a hybrid CI-based improved spatio-temporal analysis framework for multivariate prediction of climatological time series data. The primary objective is to ex-



A comparative study of proposed FB-STEP and various other data-driven models for climatological prediction

Prediction Model Model Specifications	Holt-Winters Method	Automated ARIMA	VARMA	ANN [FFBP]	HBAR	Standard BN	FBN	FB-STEP (Proposed)
Spatial-extrapolation capability	Absent	Absent	Absent	Absent	Present	May be incorporated through spatial nodes	May be incorporated through spatial nodes	Present
Intrinsic chaos handling ability	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present
Model Tractability	Quite easy; Too simple for complex real world scenario	Quite easy; Simple for complex real world scenario	Easy; Simple for complex real world scenario	Moderate; Needs to handle Large no. of parameters	Difficult; Requires appropriate software	Moderate; Need expert knowledge for structuring	Moderate; Need expert knowledge for structuring	Moderate; Need expert knowledge for structuring
Predictive Accuracy	Low; Backward looking	Low; Backward looking	Low; Backward looking	Moderate	Moderate	High	High	Significantly High
Computational burden/cost	Moderately expensive	Moderately expensive	Moderately expensive	Large training time; High complexity	Time intensive	Exponential time and space requirement	Exponential time and space requirement	Exponential time and space requirement
Model Flexibility (with respect to spatial and non-spatial data)	Medium; Not appropriate for spatial time series	Medium; Not appropriate for spatial time series	Medium; Not appropriate for spatial time series	Medium; Not able to directly utilize spatial properties	Medium; Not appropriate for non-spatial data	Quite flexible; Able to handle spatial data with added spatial nodes	Quite flexible; Able to handle spatial data with added spatial nodes	Highly flexible; Applicable for wide range of spatial and non-spatial time series data
Model Uncertainty	Very High	High	Very High	Quite High	High	Low	Low	Considerably Low
Model Scalability	Not scalable	Not scalable	Not scalable	Not scalable	Scalable to some extent	Not scalable	Not scalable	Not scalable

Fig. 12. Study of proposed FB-STEP in comparison with the other data-driven prediction models.

ploit the power of computational intelligence (CI) for dealing with the various challenges in climatological prediction and thereby to enhance the potentials of data-driven approaches as the complements for the physics-driven forecast models.

FB-STEP is based on the principles of a *fuzzy Bayesian network (FBN*), and *multifractal detrended fluctuation analysis (MF-DFA)*. The fuzzy Bayesian network helps in capturing the spatio-temporal inter-relationships among the climatological variables and also reduces the epistemic uncertainty in prediction process, whereas, the MF-DFA technique captures the natural regularities present in climatic time series, and incorporates these during the spatio-temporal prediction. A comparative study has been carried out with the traditional statistical methods, vector process-based models, classical space-time models, and other non-linear approaches to forecast climatic conditions of five major cities in India (*Kolkata, Raipur, Lucknow, Kharagpur* and *Baleshwar*), using the historic data on *temperature, humidity, precipitation rate*, and *soil moisture*. From the predicted results, it is observed that the proposed FB-STEP outperforms the other state-of-the-art and benchmark techniques by producing the minimum error and least uncertainty in prediction for each considered variable.

4.1. Future research directions

In future, the work can be extended to incorporate the spatio-temporal climate change pattern (Das & Ghosh, 2015) in the proposed framework to improve the prediction accuracy. Incorporation of the spatial semantics (Das & Ghosh, 2017) in the proposed spatio-temporal prediction technique may also be explored. Moreover, the presently proposed FB-STEP is a generic prediction framework which is applicable not only for climatological time series but also for spatio-temporal data from diverse domains of application. For example, let us consider the real estate price or housing price from the domain of finance and economy. It can be noted that the real estate price is significantly affected by the recent selling prices of the nearby real estates/houses, and therefore, prominently shows spatio-temporal dependencies among such prices. The financial time series is also proved to show multifractal characteristics (Thompson & Wilson, 2016). Therefore, the FB-STEP framework may successfully be applied for predicting such data. Similarly, huge scopes also remain in exploring the FB-STEP framework to predict spatio-temporal data from other application domains, including atmospheric research, hydrology, medicine, and so on. In the subsequent part of this section, we discuss a few more open problems (enumerated below) for future research, which are mainly centered around the limitations of our proposed framework and can be envisaged as important directions to explore in the generic field of spatio-temporal prediction through data-driven approaches.

i) Extending the proposed framework with continuous BN analysis: In our research we have made an attempt to extend discrete Bayesian analysis for the spatio-temporal relationship learning purpose. Consequently, ample scope remains in refining the proposed prediction framework in terms of using continuous BN analysis in the module-1. This will eventually make the prediction model more flexible for application in diverse domains where the relevant time series are by nature continuous and it often becomes difficult to acquire appropriate domain knowledge to decide the discretized range boundaries.
ii) Employing hierarchical extensions of BN model: Not only the climatic time series, but also the majority of the spatio-temporal data often contain variability at several spatial and temporal scales. The space-time variability is further complicated due to different spatial behaviors at different time instants and vice-versa. So, defining a more flexible version for

the proposed ST prediction framework is necessary. Employing space-time dynamic hierarchical extensions of the proposed fuzzy BN modeling may be an effective solution in this respect.

iii) Dealing with unknown structure of the causal dependency graph: In this research, we have assumed that the network structure (causal dependency graph) of the BN model is expert-determined or known a-priory. The extension has been made in terms of *incorporating fuzziness and spatial information* during *parameter learning* and *inference generation mechanism*. Therefore, huge scope remains in dealing with unknown structures of the causal dependency graphs, by developing appropriate structure learning algorithms.

iv) Increasing scalability of the proposed framework: Since our proposed FB-STEP framework uses multifractal analysis to capture the intrinsic regularity in the time series, it needs historical dataset over long duration in past. Prediction with time series of a very short duration may not reflect the expected performance. Hence, the future scope also remains in increasing the scalability of the proposed prediction model.

v) Extending the proposed framework to deal with external impacts: While employing the multifractal analysis, our proposed FB-STEP framework assumes that the fluctuations within the concerned time series are natural. It does not take into account the external effects, like those arise due to anthropogenic activities. Therefore, in future, the framework can be upgraded to deal with the impacts from artificial factors as well.

vi) Developing software tool/package for FB-STEP: Huge scope also remains in developing software tool for our proposed FB-STEP framework and commercially deploying the same, for easy access to wide range of users. Separate package may also be built for FB-STEP, so that it can be integrated with existing mathematical computing software, like MATLAB, R-tool etc.

vii) Synergism between BN and deep learning architecture: Though the BN models are intrinsically capable of reasoning under uncertainty, these may not work efficiently when the input data is very large and complex. On the other hand, though the deep learning frameworks can model complex processes by exploiting their hierarchical representation power, these are, in general, not able to understand and model their uncertainty. Hence, there remains ample scope of further enhancing the proposed FB-STEP framework by employing Bayesian deep learning for ST relationship modeling purpose.

viii) Combining data-driven and physics-driven approaches for improved prediction: Finally, employing only data-driven analysis and ignoring the basic physical laws underneath a system cannot be a complete approach to extract all the insights. Hence, efficiently combining both the physical and the data-driven approaches, for developing theory-guided data-driven models for climatological and other spatio-temporal prediction, can be an interesting as well as useful research topic in future.

References

Abhishek, K., Kumar, A., Ranjan, R., & Kumar, S. (2012). A rainfall prediction model using artificial neural network. In *Control and system graduate research colloquium (ICSGRC), 2012 IEEE* (pp. 82–87). IEEE.

Aguilera, P., Fernández, A., Fernández, R., Rumí, R., & Salmerón, A. (2011). Bayesian networks in environmental modelling. *Environmental Modelling & Software, 26*(12), 1376–1388.

Awan, M. S. K., & Awais, M. M. (2011). Predicting weather events using fuzzy rule based system. *Applied Soft Computing, 11*(1), 56–63.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control*. John Wiley & Sons.

Chatfield, C. (2013). *The analysis of time series: An introduction*. CRC press.

Cofıno, A. S., Cano, R., Sordo, C., & Gutierrez, J. M. (2002). Bayesian networks for probabilistic weather prediction. In *15th Eureopean conference on artificial intelligence, ECAI* (pp. 695–700). IOS Press.

Das, M., & Ghosh, S. (2015). Spatio-temporal pattern analysis for regional climate change using mathematical morphology. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2*(4), 185–192.

Das, M., Ghosh, S., Gupta, P., Chowdary, V., Nagaraja, R., & Dadhwal, V. (2017). Forward: A model for forecasting reservoir water dynamics using spatial Bayesian network (spabn). *IEEE Transactions on Knowledge and Data Engineering, 29*(4), 242–855. doi:10.1109/TKDE.2016.2647240.

Das, M., & Ghosh, S. K. (2014a). A probabilistic approach for weather forecast using spatio-temporal inter-relationships among climate variables. In *Industrial and information systems (ICIIS), 2014 9th international conference on* (pp. 1–6). IEEE.

Das, M., & Ghosh, S. K. (2014b). Short-term prediction of land surface temperature using multifractal detrended fluctuation analysis. In *India conference (indicon), 2014 annual IEEE* (pp. 1–6). IEEE.

Das, M., & Ghosh, S. K. (2017). Sembnet: A semantic Bayesian network for multivariate prediction of meteorological time series data. *Pattern Recognition Letters, 93*, 192–201. doi:10.1016/j.patrec.2017.01.002.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 Years of time series forecasting. *International Journal of Forecasting, 22*(3), 443–473.

Drignei, D., Forest, C. E., Nychka, D., et al. (2008). Parameter estimation for computationally intensive nonlinear regression with an application to climate modeling. *The Annals of Applied Statistics, 2*(4), 1217–1230.

Faghmous, J. H., & Kumar, V. (2014). Spatio-temporal data mining for climate data: Advances, challenges, and opportunities. In *Data mining and knowledge discovery for big data* (pp. 83–116). Springer.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application, 1*, 125–151.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting, 20*(1), 5–10.

Jun, K.-S., Chung, E.-S., Kim, Y.-G., & Kim, Y. (2013). A fuzzy multi-criteria approach to flood risk vulnerability in South Korea by considering climate change impacts. *Expert Systems with Applications, 40*(4), 1003–1013.

Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., Havlin, S., Bunde, A., & Stanley, H. E. (2002). Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and Its Applications, 316*(1), 87–114.

Kirtman, B. P., Bitz, C., Bryan, F., Collins, W., Dennis, J., Hearn, N., et al. (2012). Impact of ocean model resolution on CCSM climate simulations. *Climate Dynamics, 39*(6), 1303–1328.

Lin, G., & Fu, Z. (2008). A universal model to characterize different multi-fractal behaviors of daily temperature records over china. *Physica A: Statistical Mechanics and Its Applications, 387*(2), 573–579.

Madadgar, S., & Moradkhani, H. (2013). A Bayesian framework for probabilistic seasonal drought forecasting. *Journal of Hydrometeorology, 14*(6), 1685–1705.

Madadgar, S., & Moradkhani, H. (2014). Spatio-temporal drought forecasting within Bayesian networks. *Journal of Hydrology, 512*, 134–146.

Microsoft-Research (2015). FetchClimate. http://research.microsoft.com/en-us/um/camb ridge/projects/fetchclimate, [Online; Jan-2015].

Mrad, A. B., Delcroix, V., Maalej, M. A., Piechowiak, S., & Abid, M. (2012). Uncertain evidence in Bayesian networks: Presentation and comparison on a simple example. In *Advances in computational intelligence* (pp. 39–48). Springer.

Nandar, A. (2009). Bayesian network probability model for weather prediction. In *Current trends in information technology (CTIT), 2009 international conference on the* (pp. 1–5). IEEE.

Nayak, R., Patheja, P., & Waoo, A. (2012). An enhanced approach for weather forecasting using neural network. In *Proceedings of the international conference on soft computing for problem solving (SOCPROS 2011) december 20–22, 2011* (pp. 833–839). Springer.

NIPCC (2014). Climate Change Reconsidered. http://www.nipccreport.org/reports/2011/pdf/01ClimateModels.pdf, [Online; Dec-2014].

Nourani, V., Mogaddam, A. A., & Nadiri, A. O. (2008). An ann-based model for spatiotemporal groundwater level forecasting. *Hydrological Processes, 22*(26), 5054–5066.

Riahy, G., & Abedi, M. (2008). Short term wind speed forecasting for wind turbine applications using linear prediction method. *Renewable Energy, 33*(1), 35–41.

Ryhajlo, N., Sturlaugson, L., & Sheppard, J. W. (2013). Diagnostic Bayesian networks with fuzzy evidence. In *Autotestcon, 2013 IEEE* (pp. 1–8). IEEE.

Sahu, S. K., & Bakar, K. S. (2012). Hierarchical Bayesian autoregressive models for large space–time data with applications to ozone concentration modelling. *Applied Stochastic Models in Business and Industry, 28*(5), 395–415.

Tang, H., & Liu, S. (2007). Basic theory of fuzzy Bayesian networks and its application in machinery fault diagnosis. In *Fourth international conference on fuzzy systems and knowledge discovery, 2007. FSKD 2007: Vol. 4* (pp. 132–137). IEEE.

Thompson, J. R., & Wilson, J. R. (2016). Multifractal detrended fluctuation analysis: Practical applications to financial time series. *Mathematics and Computers in Simulation, 126*, 63–88.

Tsay, R. S. (2013). *Multivariate time series analysis: With R and financial applications*. John Wiley & Sons.

Venkadesh, S., Hoogenboom, G., Potter, W., & McClendon, R. (2013). A genetic algorithm to refine input data selection for air temperature prediction using artificial neural networks. *Applied Soft Computing, 13*(5), 2253–2260.

Wang, G., Xu, T., Tang, T., Yuan, T., & Wang, H. (2017). A Bayesian network model for prediction of weather-related failures in railway turnout systems. *Expert Systems with Applications, 69*, 247–256.