ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model

Moumita Saha [a,b,d,*], Anirban Santara [b], Pabitra Mitra [b], Arun Chakraborty [c], Ravi S. Nanjundiah [d,e,f]

[a] Department of Computer Science, University of Colorado Boulder, USA
[b] Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India
[c] Centre for Oceans, Rivers, Atmosphere and Land Sciences, Indian Institute of Technology Kharagpur, India
[d] Centre for Atmospheric and Oceanic Sciences, Indian Institute of Science, Bangalore, India
[e] Divecha Centre for Climate Change, Indian Institute of Science, Bangalore, India
[f] Indian Institute of Tropical Meteorology, Pune, India

ARTICLE INFO

Keywords:
Stacked autoencoder
Automated feature learning
Predictor identification
Monsoon prediction
Ensemble regression model
Indian summer monsoon

ABSTRACT

The study of climatic variables that govern the Indian summer monsoon has been widely explored. In this work, we use a non-linear deep learning-based feature reduction scheme for the discovery of skilful predictors for monsoon rainfall with climatic variables from various regions of the globe. We use a stacked autoencoder network along with two advanced machine learning techniques to forecast the Indian summer monsoon. We show that the predictors such as the sea surface temperature and zonal wind can predict the Indian summer monsoon one month ahead, whereas the sea level pressure can predict ten months before the season. Further, we also show that the predictors derived from a combination of climatic variables can outperform the predictors derived from an individual variable. The stacked autoencoder model with combined predictors of sea surface temperature and sea level pressure can predict the monsoon (June-September) two months ahead with a 2.8% error. The accuracy of the identified predictors is found to be superior to the state-of-the-art predictions of the Indian monsoon.

1. Introduction

The Indian summer monsoon is a complex climatic phenomenon. It refers to the rainfall occurring between June and September, which contributes to more than 75% of the aggregate annual rainfall of the country. The India Meteorological Department defines the monsoon process as the seasonal reversal of the wind direction down the shores of the Indian Ocean; blowing from the southwest direction. The term 'monsoon' is derived from the Arabic word 'mausim,' meaning season. The geographical features of the subcontinent, atmospheric conditions, oceanic components, and geophysical components are significant factors influencing the meteorological event under study. The distinctive heating of landmass and the sea surface triggers the monsoon in its initial stages. This phenomenon is then sustained by tropospheric temperature gradients. These tropospheric temperature gradients, in turn, cause pressure gradients, which lead to monsoonal circulation. The pressure gradient advects moisture into the Indian landmass; most of this moisture comes from the Arabian Sea. Lows and depressions form over the Bay of Bengal and move inland, bringing copious amounts of rain over regions of India. The moisture-laden winds rise-up in altitude, resulting in orographic rainfall due to the presence of geographical reliefs, such as the Western Ghats in the western coast and the Himalayas along the north-eastern boundary of India.

* Corresponding author at: Department of Computer Science, University of Colorado Boulder, USA.
E-mail address: moumita.saha2012@gmail.com (M. Saha).

https://doi.org/10.1016/j.ijforecast.2020.03.001
0169-2070/

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

The distribution of rainfall over the country is highly variable, which adds more difficulty to its understanding and prediction. The monsoon has a significant impact on agriculture, growth of flora-fauna, hydro-electricity production, freshwater renewal, and on the overall economic growth of the country. The India Meteorological Department (IMD) continually assesses different climatic variables and updates its models to forecast the Indian monsoon with improved accuracy (Rajeevan, Pai, Dikshit, & Kelkar, 2004; Rajeevan, Pai, Kumar, & Lal, 2007).

Climate change is predominant and it affects the Indian summer monsoon in terms of its onset, intensity, distribution, and span. Though climate change and its impact on the monsoon are not the aim of this article, we briefly discuss some of the related research. Koll and Chaithra (2018) discussed the increase in variability of the Indian monsoon process, which is symbolized by a higher count of prolonged dry periods with no or deficient rainfall, and short, intense spells. The large-scale changes in precipitation are due to an increase in global emissions of greenhouse gasses and local changes through deforestation, land-use changes, and urbanization. Loo, Billa, and Singh (2015) showed the presence of a varying degree of fuzziness between seasonal atmospheric flow and worldwide warming in the monsoon period. An increase in anomalous behavior of the Southeast Asia monsoon, along with delayed onset, is observed as corresponding to the rise in global temperature anomalies. Annamalai, Hafner, Sooraj, and Pillai (2013) showed an east–west shift in monsoon for rising sea surface temperature over the Indo-Pacific region with an additional increase in greenhouse gasses. The east–west shift results in decreased precipitation over South Asia and corresponding increased rainfall in the tropical western Pacific region. This change is also shown with the simulation of a coupled model, which includes the current increased greenhouse gas concentration. Turner and Annamalai (2012) focused on the importance of understanding the process driving the monsoon, the modes of variability of monsoon rainfall, and the seasonal cycles, which can assist in building improved models to reduce the uncertainty in prediction of monsoon rainfall. Many studies have been conducted to understand the effect of anthropogenic climate change on the Indian Monsoon (Krishnan et al., 2016; Priya, Krishnan, Mujumdar, & Houze, 2017). Patil, Venkataraman, Muduchuru, Ghosh, and Mondal (2019) disentangled the effects of changes in the gradient of sea surface temperature from those by anthropogenic aerosol emissions (as simulated in an atmospheric general circulation model) on the South Asian monsoon rainfall.

Multiple climatic variables influence the Indian monsoon process. Predictors of various climatic variables like zonal wind velocity, sea level pressure, air temperature, sea surface temperature, etc. have an impact on the event at multiple lead times. The monsoon has a large amount of variability. Additionally, the influencing predictors also change with time (Saha, Chakraborty, & Mitra, 2016). Consequently, meteorological experts evolve and update the monsoon predictors (Rajeevan et al., 2004, 2007) to improve the forecast. DelSole and Shukla (2002) have chosen predictors by measuring the errors of models with various sets by neglecting the models with low accuracy. Saha and Mitra (2016) have utilized recurrent neural networks to understand and forecast the Indian monsoon. Clustering techniques are also studied to predict the precipitation of India (Saha & Mitra, 2019; Saha, Mitra, & Chakraborty, 2015). DelSole and Shukla (2012) have also highlighted the use of sea surface temperatures in monsoon prediction with coupled ocean-atmosphere models, because the steadily changing values of sea surface temperatures contribute as a source of predictability.

We selected the climatic predictors influencing the monsoon by studying physical processes like sea-atmospheric interactions and wind flow patterns. However, these methods have the following pitfalls: (i) climatic predictors are derived from variables within localized geographical regions; however, monsoon predictors can be combinations of variables located at different locations (e.g. pressure gradients between ocean and land advect the moist air towards landmasses, and are responsible for rainfall), (ii) the method does not examine all areas over the world; only some geographical regions are explored and analyzed from a meteorological point of view. However, predictors influencing a phenomenon change with time, and therefore it is necessary to explore new locations to include novel influencing predictors of the monsoon. The motivation of our proposed work is to provide provisions to the pitfalls of past approaches. We propose a method for automated identification of monsoon predictors using machine learning techniques to predict the Indian summer monsoon.

Advanced non-linear techniques are required to identify predictors from climatic variables. Autoencoders are extensively used in dimensionality reduction (Hinton & Salakhutdinov, 2006) and are efficient in their ability to handle enormous amounts of data. Liu, Hu, He, Chan, and Lai (2015) have provided weather forecasts utilizing an autoencoder network. The study has derived useful features from hourly wind velocity and temperature data, and used them for forecasting the temperature. Autoencoder architecture extracts non-linear relationships from the input variables and converts the complex form of data into a format that is stable and efficient (Song, Liu, Huang, Wang, & Tan, 2013).

A stacked autoencoder is a multilayer network built with a sparse autoencoder at each layer. It assists in identifying more complex and proficient monsoon predictors from climatic variables than a shallow single-layer autoencoder. Single-layer autoencoders have been used to identify predictors of the Indian monsoon (Saha, Mitra, & Nanjundiah, 2016). However, the proposed approach using stacked autoencoders produces highly sophisticated predictors with more information content. The proposed model performs better than the predictors obtained from single-layer autoencoders for monsoon prediction (Saha et al., 2016). Saha, Mitra, and Nanjundiah (2016) used stacked-autoencoders for the forecast of phases (late or early) of the Indian monsoon. Stacked autoencoder network also showed their imprint on the prediction of monsoon for regional parts of India (Saha, Mitra, & Nanjundiah, 2017). The proposed deep-learning-based approach is focused on the identification of monsoon predictors and

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

the prediction of the annual aggregate Indian summer monsoon, exploring more climatic variables and modeling specifically for the aggregate Indian summer monsoon. For the autoencoder architecture, the number of parameters are kept the same in both the shallow and deep autoencoder; the stacked (deep) autoencoders giving lower reconstruction errors compared to the shallower ones (Hinton & Salakhutdinov, 2006). The deeper the level of stacked autoencoders, the more complex the identified predictors; predictors at each layer are formed from a combination of features produced at the previous layer. The learning algorithm applied over the stacked autoencoder, which adapts by layer-wise training, is very useful for learning the weights of a deep-stacked autoencoder (Hinton, Osindero, & Teh, 2006). All the layers of features capture strong correlations between the units in the previous layer. The stacked encoder with layer-wise pre-training presents an effective way to disclose non-linear features of the input progressively and, thus, help to identify new monsoon predictors.

The proposed approach has two significant contribution directions: (i) identification of new predictors of monsoons (which could be non-linear), from variables across the world by utilizing stacked autoencoders; (ii) forecasting the Indian summer monsoon with improved accuracy using newly-identified predictors with advanced machine learning models.

The approach initiates by dividing the globe into individual grid boxes. The spatially averaged (over the grid box) time series now represents the parameter of the corresponding grid box. The climatic variables corresponding to different grid boxes are the inputs for the architecture of stacked autoencoders. After developing and optimizing the stacked autoencoder, we collected the new predictors from its internal layers. Two advanced machine learning prediction models, namely, an ensemble regression tree with a bagging method and an ensemble decision tree, are utilized to forecast the monsoons of India.

Section 2 gives a descriptive study of the climatic variables used for the discovery of new predictors of monsoons, and also describes the precipitation data. Section 3 explains the stacked autoencoder design, architecture, and its working. Section 4 describes the predictor identification method using a stacked autoencoder network. The monsoon prediction models are elaborated in Section 5. Section 6 explains the experimental outcomes and a detailed analysis of identified predictors for their forecasting skills. A discussion of identified climatic predictors is presented in Section 7. Finally, the article concludes in Section 8.

2. Climatic variables

The climatic variables examined for the proposed approach of identifying new predictors of monsoons, and thereby predicting the Indian summer monsoon are sea surface temperature (SST), zonal u-wind (UWND) at a 200 hPa pressure level, and sea level pressure (SLP). The gradient of SLP between the sea and land surface advects the moisture-laden winds towards the landmass, causing rainfall (Rajeevan et al., 2007). The Arabian Sea, Bay of Bengal, and Indian Ocean surround the Indian peninsula, and the zonal winds flowing from the ocean to the land carry moisture, which influences the quantity and quality of the monsoon. Additionally, the Himalayan range and the Tibetan plateau surround India on the north-eastern side and the Western Ghats run down the western coast of the mainland. These geographical features make the winds in the lower atmosphere ineffectual to India. Thus, u-winds (zonal winds) at the 200 hPa pressure level, which can have an impact on the Indian monsoon, is considered for this study. Swapna, Krishnan, and Wallace (2014) showed a relationship between trends of Indian summer rainfall and variation in the meridional gradient of the monsoonal zonal wind. The reverse link of influence on interannual variations of Indian rainfall on the strength of the zonal wind cycle in the equatorial Indian Ocean is also explored (Kelly & Mapes, 2011; Ogata & Xie, 2011).

Both the sea level pressure and zonal wind data are acquired from reanalysis data by NCEP (*www.esrl.noaa.gov/psd/*), provided at a 2.5° × 2.5° grid location (Kalnay et al., 1996). We observed a high correlation between the anomaly of SST and the summer monsoon of the country (Rajeevan et al., 2004, 2007). The SST data are collected from version 2 of NOAA Optimum Interpolation data (*NOAA_OI_SST_V2*), present at a 2° × 2° resolution (Reynolds, Rayner, Smith, Stokes, & Wang, 2002). Data are considered for each month from 1948 through 2014.

We segregated sixty-seven years (1948–2014) into the training and testing sets. The training set is utilized to discover the monsoon predictors using a stacked autoencoder and develop a prediction model for the monsoon. The test set is used to evaluate the performance of new predictors. The training period is 1948–2000, and the test period is 2001–2014.

The monthly anomaly series (denoted as $processData_m^y$ for the month m of the y^{th} year) of the input variables are evaluated with the subtraction of the monthly climatology from the data values, as shown in Eq. (1).

$$processData_m^y = X_m^y - mean(X_m), \tag{1}$$

where X_m^y is the variable for the m^{th} month of the y^{th} year. The mean (X_m) denotes the average of the m^{th} month of the input variable for the years being studied.

We concentrate on forecasting the India-wide summer monsoon prevailing for June-September. Rainfall data is obtained from the premier India Meteorological Department (*www.imdpune.gov.in*) for 1948–2014. The monsoon rainfall is presented in a millimeter scale (mm). The long-period average (LPA) of the monsoon is calculated by considering a period of 30 years from 1961 to 1990. We expressed all the rainfall values as a percentage of the LPA rainfall. The LPA rainfall is 887.9 mm, with an overall variation of 10% (88 mm). The distribution of rainfall over 1948–2014 is presented in Fig. 1. We categorize monsoon years into three classes: (i) drought - years having rainfall as 10% less than the LPA rainfall (red bars); (ii) excess - years having monsoon as 10% more than the LPA rainfall (blue bars); and (iii) normal - years having precipitation within 10% boundary from the LPA on both sides (green

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

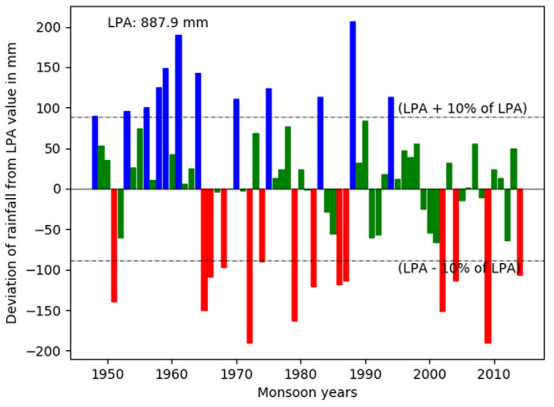

Fig. 1. Indian summer monsoon distribution as the deviation from the long-period average rainfall value for the study period 1948–2014 (red bars are drought years, blue bars are excess rainfall years, green bars are normal monsoon years). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

bars). We have also performed some statistical analysis of the rainfall. The mean rainfall over the period is 890.19 mm, with a standard deviation of 90.47 mm. The highest rainfall year is 1988 with rainfall of 1094.1 mm, and the lowest rainfall year is 1972, with 697.4 mm of rain.

3. Stacked autoencoders

An autoencoder is an architecture belonging to the class of artificial neural networks. It has an input, an internal layer, and an output layer. We used this architecture to acquire the intricate features of data, and utilized it for the reduction in dimensionality of the data. The autoencoder sets the output values to be the same as the input values, and attempts to recreate the input information from the representation learned in the internal layer. The architecture is capable of learning a non-linear function by using the process of iterative training of the model. The model aims at reducing the re-construction errors in re-building the output from the representation discovered in the internal layer.

Several single-layered autoencoders are heaped to generate a deep neural model. The outcome of the first autoencoder acts as an input to the second, and this continues to the desired depth of the architecture. We show the structure of the stacked autoencoder and its training method in Fig. 2.

The training of the stacked autoencoder is unsupervised, which does not require any information about the predictand variable. The pre-training is performed, taking into consideration a single layer at a time. The layers of the stacked autoencoder are trained with the motivation of minimization of errors in input reconstruction individually for the layers (shown in the top portion of Fig. 2). The pre-training assists in generalization as it guarantees that the information learned is from the input without any output label. After pre-training all the layers, the whole stacked network is fine-tuned with a run of

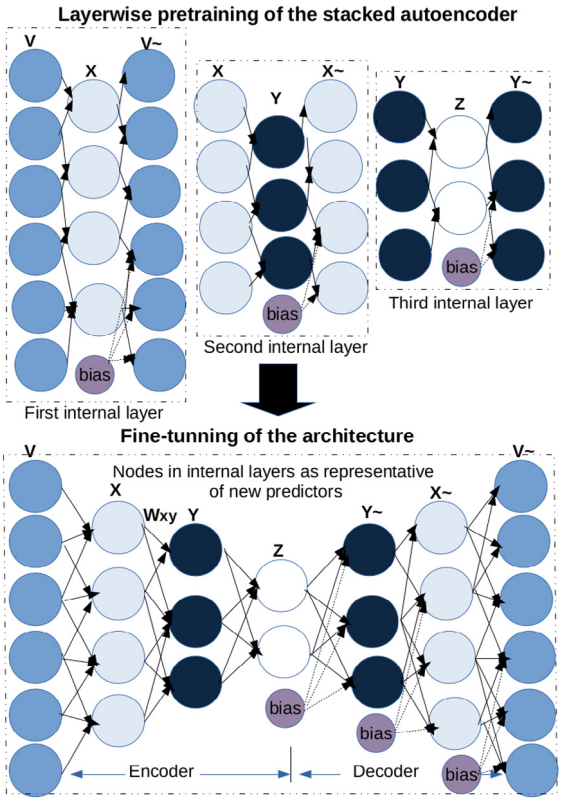

Fig. 2. Stacked autoencoders showing individual layer-wise pre-training, followed by fine-tuning of the total network.

gradient descent, and the weights assigned to each layer are adjusted (shown in the bottom portion of Fig. 2). We learned novel complex features in the internal layers of the deep autoencoder model.

In formal notation, $x \in R^n$ is an input. The activation of each node at the internal layers is denoted by h_i^t, where i varies as $i = 1, \ldots, p$, and p denotes the count of nodes in the t^{th} internal layer Eq. (2).

$$h^t(x) = f\left(W^t x + b^t\right), \tag{2}$$

where $f(z) = \frac{e^{2z}-1}{e^{2z}+1}$ is the hyperbolic tangent activation function, $h^t(x) \in R^p$ is the node activation for the t^{th} internal layer, the dimension of W^t is $(p \times n)$ (representing a weight matrix from the antecedent $((t-1)^{th})$ to the present (t^{th}) internal layer), $b^t \in R^p$ is a bias of the present layer, and p and n are the node counts for the present and antecedent internal layer. When $t = 1$, that is for the first internal layer, the previous layer is the input layer.

We applied the hyperbolic tangent function in transition between a layer and its subsequent layer, except for in the case of the change from the last internal layer to the output layers. Eq. (3) shows the transition function applied the output layer. Random weight values are used to initialize the weight matrix. The weight matrix gets updated at every epoch with back-propagation of error calculated between the predicted and actual output. The matrix is finally populated with revised values to capture

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

the non-linearity of the input data by minimizing the errors.

$$\widehat{x} = g\left(W^r h^r(x) + b^r\right), \tag{3}$$

where $g(z) = ((a * z) + b)$ represents a linear transition, a, b are constants, $\widehat{x} \in R^n$ is a vector of output nodes, W^r is the weight matrix, and $b^r \in R^n$ represents the bias vector corresponding to the output layer.

The proposed method considers climatic variables like SLP, SST, or UWND, or a combination of these variables (described in Section 4.2.1) as the inputs for the stacked model. We acquired the potential predictors from internal levels of the architecture as non-linear combinations of variables in various distant regions. Therefore, the inputs and outputs of the model are the variables and non-linear combinations of the input variables, respectively. In conclusion, the autoencoder helps to reduce the dimension or new composite feature identification.

Formally, provided an s input set denoted by x_i, where $i = 1, \ldots, s$, the bias vector b^t and the weight matrix W^t corresponding the t^{th} layer are learned using a gradient descent method to reduce the input recreation error $\sum_{i=1}^{s} \|x_i - \widehat{x}\|^2$.

4. Automated monsoon predictors' identification using a stacked autoencoder model

The proposed method of automated identification of predictors for monsoons using a stacked autoencoder and, thereby, the prediction of the Indian summer monsoon is shown in Fig. 3.

4.1. Climatic variables as an input to the autoencoders

We partitioned the whole globe into spatial rectangular boxes of dimension 10° latitude × 20° longitude, which adds up to 324 rectangular grid boxes. We considered the grid boxes to have more than 80% of their values as non-Null further (e.g. neglected grids near the poles which have mostly Null values). The time series of the input variables in the grid box are averaged to acquire one series that represents the box. The inputs of the first autoencoder layer are the series of the chosen grids. The count of nodes to the input level (layer) of the stacked autoencoder built for variables sea level pressure (Stk_SLP) and zonal wind (Stk_UWND) are 324, individually. The model designed for sea surface temperature (Stk_SST) has 240 input nodes.

4.2. Identifying the predictors of the monsoon

The predictor identification process initiates with unsupervised training and fine-tuning of a stacked autoencoder, followed by a threshold over the acquired weights. Finally, the required predictors are filtered from their correlation study with the Indian summer monsoon.

4.2.1. The structure of stacked autoencoders

We identified two types of predictors using a stacked autoencoder, the first from individual climatic variables and the second from a combination of variables.

The study uses triple-layered stacked autoencoders. The input for the first autoencoder of the stacked model comprises input variables of the selected boxes. The ratio of nodes between two consecutive layers is considered to be 15:1. The nodes learned in the inner layer are representative of composite learned features. These are the new probable monsoon predictors. The innermost layer of the first autoencoder is the input for the next autoencoder. A similar trend is repeated for the last autoencoder.

The model built for sea surface temperature (Stk_SST) has architecture as [240 72 21 6 21 72 240], where the count of input variables is 240. The count of nodes in the internal layer of the first autoencoder is 72, the next is 21, and for the last autoencoder it is 6. The following values represent the mirror-image of the previous layers, which helps in the reconstruction of the input. Similarly, the structure of Stk_SLP and Stk_UWND are [324 97 29 9 29 97 324].

Apart from these three designs, we designed fourth and fifth stacked autoencoders with an input of the combined features of SST and SLP (Stk_SLP_SST) and that of UWND and SLP (Stk_SLP_UWND). Stacked autoencoder Stk_SLP_SST has the layers [564 169 50 15 50 169 564], where 564 nodes comprise 240 SST input nodes and 324 SLP input nodes. Stk_SLP_UWND has structure [648 194 58 17 58 194 648], where both SLP and UWND have 324 input nodes each.

4.2.2. Pre-training of stacked autoencoders for monsoon predictor identification

We examined input data for predictor identification from 1948–2000 for developing the stacked autoencoder architecture. The input corresponds to the data at a monthly scale, which adds to 636 (53 years × 12 months) instances. Unsupervised pre-training of the stacked autoencoder is performed considering a single layer at a time. We trained the layers by reducing the error in reconstructing the input following the principle of the autoencoder. All the autoencoders are pre-trained at first, and then the complete architecture is fine-tuned using a gradient descent back-propagation algorithm. The nodes learned in the internal levels of the stacked autoencoder present the composite and sophisticated features.

4.2.3. Post-treatment by thresholding of weights

We acquired three sets of predictors from the three internal levels of the stacked architecture. The threshold is adapted to consider the input nodes that actively influence the node in the inner layer while discarding the rest. We put a limit on the weight matrix by considering the weights with a value higher than twice the standard deviation over the mean learned in the weight matrix. The threshold is calculated empirically, such that it engrosses more than 10% of the nodes in the input layer. This contributes to an evaluation of the node in the internal layer, which is the seed for a new potential monsoon predictor.

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

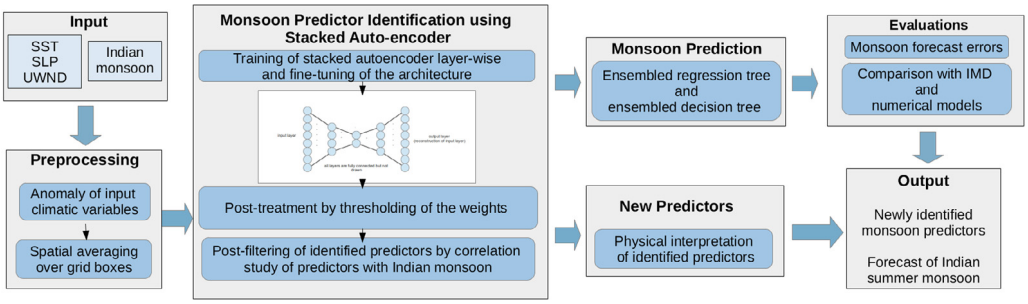

Fig. 3. Identification of Indian monsoon predictors utilizing stacked autoencoders.

New climatic predictors are evaluated as the weighted sum of the nodes in the input layer that are being selected after the threshold method.

In formal notation, x_j denotes the input node, where j varies as $j = 1, \ldots, n$, and the variable h_k denotes an internal node, where k varies as $k = 1, \ldots, p$. Finally, W_{jk} denotes the learned weight of input node x_j for the internal node h_k. The predictor obtained from internal node h_k is calculated as follows Eq. (4).

$$h_k = \sum_{j=1}^{n} W_{jk} x_j, \quad \forall j, \ | \ W_{jk} > thres_k, \tag{4}$$

where $thres_k$ denotes the threshold value considered for the internal node h_k of the specific layer.

4.2.4. Post-filtering of identified monsoon predictors

The newly-identified predictors of each layer are ranked by studying the predictors' (corresponding to the nodes of the internal layers) correlation with the Indian summer monsoon. This considers a lead of one to twelve months for finding the highest correlated month for all the identified predictors using the Pearson correlation coefficient (γ) Eq. (5). We considered the best lead month of the identified predictors for further evaluation. Thus, for all three internal layers of the stacked autoencoder, a sorted list of identified monsoon predictors is obtained by considering the correlation in descending order. We observed high correlations between some of the obtained predictors and rainfall. The predictors at deep layers outperform predictors obtained in the shallow layer (explored in Section 6.3.2). Higher predictability of predictors from a deeper layer highlights the advantages of using a stacked autoencoder over a single layer autoencoder. The spatial location of the identified monsoon predictors is shown in Section 6.2.

$$\gamma = \frac{\Sigma_{i=1}^{n} (p_i - \overline{p})(r_i - \overline{r})}{\sqrt{\Sigma_{i=1}^{n} (p_i - \overline{p})^2} \sqrt{\Sigma_{i=1}^{n} (r_i - \overline{r})^2}}, \tag{5}$$

where p_i and r_i denote the identified predictor's values and Indian summer monsoon at the i^{th} year, $\overline{p}$ and $\overline{r}$ are the means for the study period, and n is the number of years under study.

5. Models for forecasting the Indian summer monsoon

The learned predictors are assembled to build the predictor sets for predicting the rainfall of the country. We consider: (i) a set of learned predictors for variables SST, SLP, and UWND, individually; and (ii) a set containing derived predictors obtained from a combination of climatic variables SLP+SST and SLP+UWND. Advanced machine learning models (described in the following section) are designed to forecast the rainfall. The training period is from 1948 to 2000, and the test period is from 2001 to 2014. We have used an ensemble regression tree with a bagging algorithm and an ensemble decision tree for designing the prediction models. The models are selected for the following reasons: (i) the bagging is a bootstrap aggregating method used for improving the approximation; (ii) bagging helps in enhancing the forecasting skill of the underlying built-in regression trees; and (iii) the ensembles are capable of dealing with non-linear feature sets, as well as being able to deal with high-dimensional data.

5.1. Ensemble regression tree with a bagging method (RegTree)

A fitted ensemble merges a set of weak learners that learn using the regression tree and the training data (Loh, 2008). The ensemble model can predict the response for new data by summing the forecasts from its subsequent weak learner models. We used the bagging technique for training the underlying learners using regression trees. The count of weak learner models in an ensemble is selected empirically, such that it maintains a balance between the speed of the algorithm and its accuracy performance. In our case, five weak learner models are the ensemble to forecast the monsoon rainfall. The model takes the set of predictors as inputs, trains the weak, vulnerable learners using bagging algorithms based on tree learners (MATLAB, 2012), and predicts the Indian summer monsoon rainfall.

5.2. Ensemble decision tree (DecTree)

We considered an ensemble decision tree model for regulating regression-based models (Liaw & Wiener, 2002). All the trees are built on an independently selected bootstrap set of input instances. The final forecast by this proposed model for a new test input is computed by averaging the predictions provided by individual trees. The different trees are built using the functionality of the regression tree. The model creates an ensemble model of

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

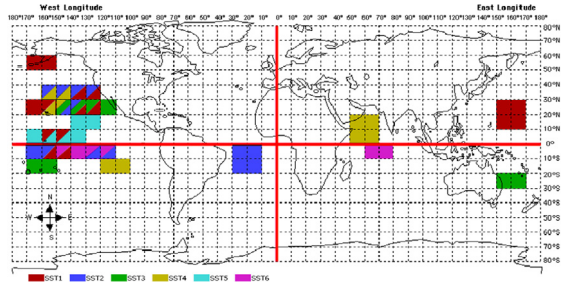

Fig. 4. Geographical areas of SST-based identified monsoon predictors.

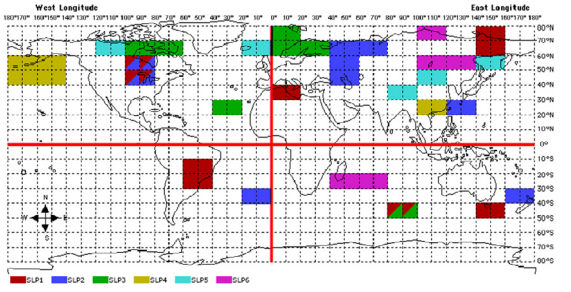

Fig. 5. Geographical areas of SLP-based identified monsoon predictors.

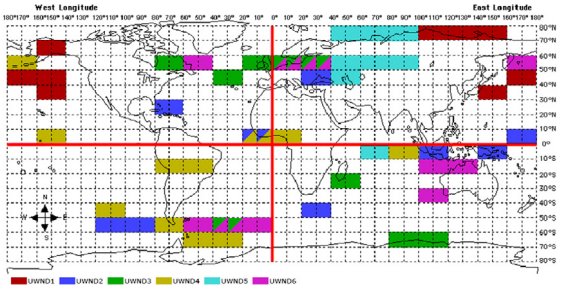

Fig. 6. Geographical areas of UWND-based identified monsoon predictors.

the number of decision trees for predicting rainfall (MATLAB, 2012). A weighted mean over forecasts from every single tree is calculated to obtain the final prediction from the fitted ensemble decision tree model as shown in Eq. (6).

$$\widehat{y}_{bag} = \frac{1}{\sum_{m=1}^{M} a_m I(m \in S)} \sum_{m=1}^{M} a_m \widehat{y}_m I(m \in S), \tag{6}$$

where $\widehat{y}_m$ is the forecast provided by the tree m of the ensemble. S is the predictor set of the selected trees which participate in forecasting, and $I(m \in S)$ is ascertained as 1 when m is present in the set S, otherwise, 0. a_m denotes the weight assigned to the m^{th} tree, and the total number of trees is M in the ensemble model.

6. Experimental results

6.1. Identified predictors at multiple levels of a stacked autoencoder

We assessed the new potential predictors of monsoon from the nodes of the internal layers of a stacked autoencoder. The weighted sum of nodes attaining the threshold gives the predictors. These predictors are filtered and sorted in a decreasing trend of their correlation coefficient value, with the Indian monsoon as being significant for the monsoon phenomenon.

6.2. Geographical regions of identified predictors

The geographical coverage of monsoon predictors obtained from the first internal layer of a stacked autoencoder for climatic variables sea surface temperature (SST), sea level pressure (SLP), and u-wind (UWND) are shown in Figs. 4, 5, and 6, respectively. The figures show, for all the input climatic variables, the highest six correlated predictors with the Indian monsoon from the sorted list. The new monsoon predictors are a combination of climatic variables from various distant areas across the world. Thus, the predictors are not geographically localized, they are from variable changes of different spatial locations.

Each color in the figure denotes the geographical locations of the variables that are combined to generate a potential predictor of monsoon. Different regions are combined non-linearly as the layers of the stacked autoencoder are developed and optimized, utilizing a non-linear tan-hyperbolic function. To each area, we assign a weight at the end of the training of the encoders. The weighted average of the predictor from different regions is amalgamated to produce a new predictor. The identified predictors are sorted by their correlation coefficient value with the monsoon, and provided in the same sorted fashion in the figures (e.g. SST1 and SST6 have the highest and lowest correlation with the Indian summer monsoon among the presented SST-based predictors).

Various colors in the same location over the globe represent the engagement of that area in all those monsoon predictors marked by respective colors. Some new predictors correspond to the areas which are known to influence the monsoon process; present a validation of our method for identifying the predictors of the Indian monsoon phenomenon.

6.3. Prediction skills of monsoon predictors

We focus on predicting the aggregate Indian summer monsoon. We will present the prediction accuracy using different predictors and we will also compare the forecasting skills of the proposed stacked autoencoder-based method with existing monsoon prediction models.

6.3.1. Monsoon predictor sets

We built the monsoon predictor sets with predictors from input variables at different layers of stacked autoencoders. The identified predictors at each layer are sorted in descending order according to their correlations with the Indian summer monsoon individually. This was done for all input variables and their combinations.

Predictors derived from the individual layers of the autoencoder are assembled to construct the predictor sets

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

Table 1
Prediction of the Indian monsoon with SST-based predictors derived from three layers of stacked autoencoder architecture for 2001 to 2014 (me: mean absolute error in %, re: root mean square error in % and pred. mon.: prediction month).

	D1		D2		D3		D4		D5	
	me	re	me	re	me	re	me	re	me	re
Identified predictors of first layer pred. mon. D1- August, D2- September, (D3, D4, D5)- April										
RegTree	6.6	8.2	6.4	7.5	5.9	8.6	6.2	7.9	**5.4**	7.4
DecTree	6.7	8.2	5.9	7.8	6.7	8.3	5.9	7.7	5.6	7.0
Identified predictors of second layer pred. mon. (D1, D2, D3, D4, D5)- May										
RegTree	5.5	6.6	**4.3**	5.7	5.1	7.7	5.3	8.1	5.1	7.5
DecTree	5.1	7.4	5.1	7.1	4.9	6.8	5.4	8.3	5.4	7.8
Identified predictors of third layer pred. mon. (D1, D2, D3)- May										
RegTree	5.0	7.2	4.8	6.7	4.9	6.9	–	–	–	–
DecTree	**4.4**	6.3	4.6	7.5	4.7	6.8	–	–	–	–

Table 2
Prediction of the Indian monsoon with SLP-based predictors derived from three layers of stacked autoencoder architecture for 2001 to 2014 (me: mean absolute error in %, re: root mean square error in % and pred. mon.: prediction month)

	D1		D2		D3		D4		D5	
	me	re	me	re	me	re	me	re	me	re
Identified predictors of first layer pred. mon. (D1, D2, D3, D4, D5)- November										
RegTree	5.9	7.7	4.8	6.5	5.3	6.7	5.4	6.9	5.4	7.5
DecTree	**4.4**	6.0	5.0	6.7	5.6	7.0	5.3	7.6	5.5	7.2
Identified predictors of second layer pred. mon. (D1, D2)-August, D3- October, (D4, D5)- January										
RegTree	4.9	6.5	5.0	6.6	5.0	6.4	4.8	6.7	4.6	5.7
DecTree	5.2	6.8	4.9	5.5	4.9	5.9	**4.5**	5.7	**4.5**	5.8
Identified predictors of third layer pred. mon. (D1, D2, D3, D4)- August										
RegTree	4.9	6.0	4.8	6.3	5.2	5.9	5.0	6.2	–	–
DecTree	**4.0**	4.9	4.8	6.5	4.8	6.7	5.2	6.1	–	–

Table 3
Prediction of the Indian monsoon with UWND-based predictors derived from three layers of stacked autoencoder architecture for 2001 to 2014 (me: mean absolute error in %, re: root mean square error in % and pred. mon.: prediction month).

	D1		D2		D3		D4		D5	
	me	re	me	re	me	re	me	re	me	re
Identified predictors of first layer pred. mon. (D1, D2, D3)- October, (D4, D5)- May										
RegTree	5.3	6.7	4.7	5.6	5.2	6.8	**4.2**	4.8	4.6	6.9
DecTree	5.7	7.1	5.5	6.8	5.0	6.1	4.7	5.8	**4.2**	5.7
Identified predictors of second layer pred. mon. D1- August, (D2, D3, D4, D5)- May										
RegTree	4.9	7.0	5.2	7.0	4.4	5.9	**4.3**	5.7	4.7	6.8
DecTree	4.3	5.9	5.0	6.6	5.0	7.0	**4.3**	6.1	4.9	6.6
Identified predictors of third layer pred. mon. D1- January, (D2, D3, D4)- February										
RegTree	5.0	8.2	5.1	7.4	4.6	6.0	4.3	6.1	–	–
DecTree	5.2	8.2	5.0	7.1	**4.1**	5.9	5.1	7.4	–	–

separately for each level. The highest correlated four, five, six, eight, and ten identified predictors from the sorted list of each layer built the D1, D2, D3, D4, and D5 predictor sets, respectively. The predictors of the first and second layers of the stacked autoencoder for all climatic variables built five predictor sets, individually. For the third layer, four predictor sets for SLP and UWND, and three for SST corrsepond to the count of identified potential predictors at the innermost level (the number of nodes are six for SST, and nine for SLP and UWND, for the designed autoencoder architecture; refer to Section 4.2.1).

6.3.2. Prediction by predictors derived from individual climatic variables

We presented the performance of the monsoon predictors by the mean absolute error (me), root mean square error (re), Pearson correlation, and Spearman correlation measures for the test period 2001–2014 by training the proposed forecasting models from 1948 to 2000. Table 1 shows the prediction of the monsoon by identified predictors of SST. Prediction accuracies by predictors of UWND and SLP are shown in Tables 2 and 3.

The forecasting errors in deeper layers are lower in comparison to the errors obtained in shallower layers. Considering the lead month of every climatic predictor in the predictor set, the sea surface temperature predictors derived from the third layer provide 4.3% error in May (as the minimum lead of the predictor is one). Interestingly, we find that using predictors based on SLP from August of the previous year (a lead of ten months) shows an error of 4.0% while predictors based on UWND from February (a lead of four months) have a mean absolute error of 4.1%. For all the three climatic predictors, prediction accuracy increases with more compressed and composite features at the deeper (innermost) layers. The accuracy in prediction improves with identified monsoon predictors in the first layer to subsequent deeper layers, with the best accuracy given by predictors evaluated at the innermost (third) layer of the stacked autoencoder. Pearson correlation coefficients (γ) of 0.74, 0.82, and 0.87 with p-values$<$.01 are observed between actual and predicted rainfalls by SST, SLP, and UWND-based monsoon predictors, respectively. We are also adding a non-linear Spearman correlation coefficient to compare the actual and predicted precipitation for more detailed analysis. We observed Spearman correlation coefficients of 0.78, 0.86, and 0.90 with p-values$<$.01 between the predicted and actual monsoon for prediction provided by the learned predictors of SST, SLP, and UWND variables, respectively. Fig. 7 shows the predicted and actual rainfall in terms of departure from the long period average (LPA) monsoon by the SST, SLP, and UWND predictors during 2001 to 2014 (the testing period).

The SST-, SLP-, and UWND-based predictors forecast the Indian summer monsoon at leads of one, ten, and four months, respectively. Different symbols show the forecast rainfall by identified predictors, and the bar represents the actual rainfall. The negative or positive departure from the long period average value is predicted correctly, as observed for most years in the test period. The extremes of the monsoon are also predicted well by the proposed

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

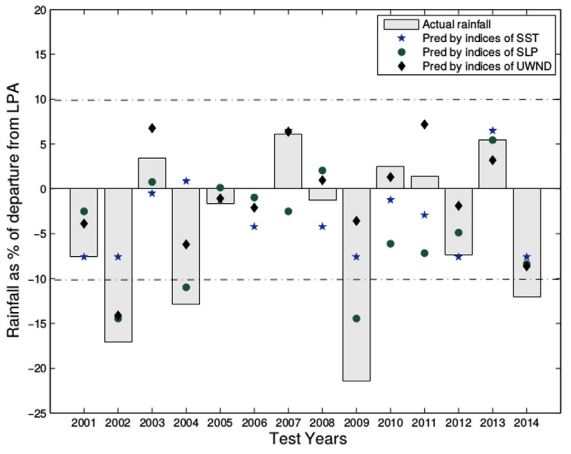

Fig. 7. Performance of monsoon predictors derived from climatic variables SST, SLP, and UWND using stacked autoencoders for predicting the Indian summer monsoon between 2001 and 2014.

method. All four drought years of 2002, 2004, 2009, and 2014 are predicted correctly with a negative sign of departure from the long period average by SST-, SLP-, and UWND-based predictors (excluding 2004 with the SST-based predictor). The identified SLP-based monsoon predictors capture the drought (or negative-anomaly) years efficiently.

6.3.3. Prediction by predictors derived from combined climatic variables

Using two climatic variables as inputs for the stacked autoencoder will lead to feature identification, which is a combination of the respective climatic variables. We have considered two combinations: (i) a combination of SLP and SST as both variables play essential roles in the monsoon process; and (ii) a combination of SLP and UWND, as differences in the pressure zone could modify the monsoonal winds. Thus, the pressure at sea level, as well as the zonal wind at 200 hPa, are considered owing to the Great Himalayas and the Tibetan plateau surrounding the Indian subcontinent. This helps with examining both the upper- and lower-level circulations. Predictions by the combined predictors are shown in Table 4.

The identified predictors of combined variables perform better than the individual variables, as explained in the previous section. However, for the combined predictors, the identified features at the first internal layer of the stacked autoencoder perform the best, and the accuracy degrades with combined features at the deeper levels of the stacked architecture. The first layer's features learned from a combination of two climatic variables retain the essential characteristics of both of the variables and, thus, predict monsoons with reasonable accuracy. However, with deeper layers that produce predictors as a compression of two already combined variable-based predictors, the prediction precision is observed to be lower for a loss of information.

The identified combined SLP+SST-based monsoon predictors predict the Indian monsoon with a mean absolute error of 2.8% in April (a lead of two months). This error

Table 4
Prediction of the Indian monsoon with combined variables of SLP+SST and SLP+UWND from three layers of stacked autoencoder architecture for 2001-2014 (me: mean absolute error in %, re: root mean square error in % and pred. mon.: prediction month).

	D1		D2		D3		D4		D5	
SLP+SST										
Identified predictors of first layer pred. mon. (D1, D2, D3, D4, D5)- April										
	me	re	me	re	me	re	me	re	me	re
RegTree	3.8	5.5	4.3	5.7	3.9	5.1	4.4	5.9	4.1	5.2
DecTree	3.1	4.6	**2.8**	4.5	4.3	5.5	4.4	5.0	4.5	5.3
Identified predictors of second layer pred. mon. (D1, D2, D3, D4, D5)- October										
RegTree	5.5	7.6	5.4	7.4	5.2	6.9	5.3	7.4	5.4	7.5
DecTree	**4.9**	6.7	5.2	7.0	5.4	7.6	5.4	6.8	5.7	7.9
Identified predictors of third layer pred. mon. (D1, D2, D3, D4, D5)- February										
RegTree	5.2	6.5	6.1	7.8	5.9	8.0	**5.0**	6.6	5.7	6.8
DecTree	5.1	6.3	6.0	8.3	5.2	7.2	5.5	6.8	5.4	7.2
SLP+UWND										
Identified predictors of first layer pred. mon. (D1, D2, D3, D4, D5)- April										
	me	re	me	re	me	re	me	re	me	re
RegTree	4.3	5.8	3.9	4.8	**3.7**	5.9	4.0	5.2	3.8	4.9
DecTree	4.5	6.1	4.2	5.4	4.0	5.1	4.1	6.0	3.8	5.1
Identified predictors of second layer pred. mon. (D1, D2, D3, D4, D5)- January										
RegTree	4.5	5.9	**4.2**	5.0	**4.2**	5.1	**4.2**	5.2	4.8	5.8
DecTree	4.4	5.2	4.7	5.8	4.6	5.5	**4.2**	5.9	4.5	6.4
Identified predictors of third layer pred. mon. (D1, D2, D3, D4, D5)- January										
RegTree	4.8	6.1	4.9	7.2	4.8	6.5	5.2	6.7	5.0	6.1
DecTree	5.0	7.1	4.8	6.9	**4.7**	6.2	5.3	6.5	5.2	6.8

is quite low for a complex system such as the monsoon (and much lower than that of existing state-of-the-art coupled models or empirical models used for monsoon forecasts). Lastly, the combined predictors of SLP+UWND provide a 3.7% mean absolute error in the Indian monsoon forecast in April (two month lead). The Pearson correlation coefficients (γ) between the predicted and actual rainfall by SLP+SST- and SLP+UWND-based predictors are 0.85 and 0.89 with p-values$<$.01, respectively. The nonlinear Spearman correlation coefficients are 0.89 and 0.88 with p-values$<$.01, respectively.

The actual and predicted rainfall by the combined SLP+SST- and SLP+UWND-based predictors are shown in the scatter plot in Fig. 8. The points are closer to the line representing 45°, which indicates the closeness of the predicted and observed rainfall. We have also tried to estimate predictors from the combinations of all three variables (SST+SLP+UWND). The performance of the combined SST+SLP+UWND indices is lower than that of individual predictors or the previously described combined predictors. An explanation for their poor performance may be because some critical characteristics are lost, or because some irrelevant features, formed from the combination of these three variables, are incapable of capturing and predicting the monsoon.

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

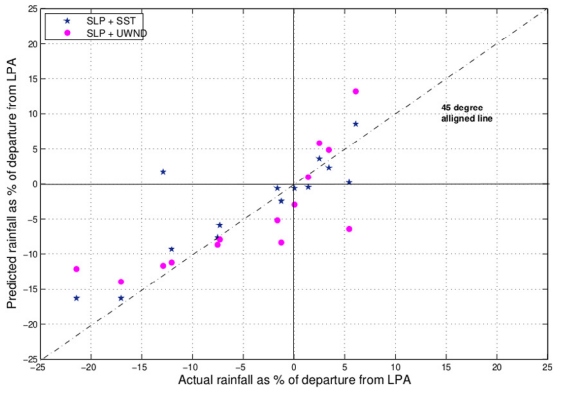

Fig. 8. Scatter plot of predicted versus actual rainfall using combined SLP+SST- and SLP+UWND-based predictors for 2001 to 2014.

6.3.4. Performance of combined monsoon predictors over individual monsoon predictors

The predictors identified from a combination of climatic variables outperform the predictors derived from a single climatic variable. If we consider the best-performing predictor sets for each case, the identified SLP- and SST-based monsoon predictors predict with 4.0% and 4.3% errors, individually. In contrast, the combined identified SLP+SST-based predictors forecast the Indian monsoon with a 2.8% error. The same trend is seen for the combined identified predictors of SLP+UWND with a mean absolute error of 3.7%. The individual counterparts, SLP- and UWND-based predictors, forecast with 4.0% and 4.1% errors, respectively. The combined predictors also capture extremes to a greater extent than individually identified predictors.

Figs. 9 and 10 show the predictions of the combined predictors over the individual predictors for SLP+SST and SLP+UWND, respectively. All the drought years (2002, 2004, 2009, and 2014) during this period are captured by the combined predictors of SLP+UWND, and they are also well-captured by the combined predictors of SLP+SST (except for in the year 2004).

6.3.5. Comparison of the model using identified monsoon predictors with the India Meteorological Department's monsoon prediction

We compared the proposed monsoon model with the India Meteorological Department (IMD) monsoon prediction models (Rajeevan et al., 2004, 2007). We compared the suggested method with the power regression model from the IMD (Rajeevan et al., 2004) and the present model based on the pursuit projection regression (PPR) (Rajeevan et al., 2007). The predictions provided by our stacked autoencoder-based identified predictors are compared for 2003–2014 following the availability of the IMD monsoon forecast. The IMD operational model predicts rainfall with a 7.5% mean absolute error. The PPR model from IMD predicts the monsoon at two leads, once in April (lead of two months, LRF1) and the succeeding in June (at the start of the season, LRF2). The Indian summer monsoon is forecast with 7.1% and 6.5% errors by the LRF1 and LRF2 models, respectively.

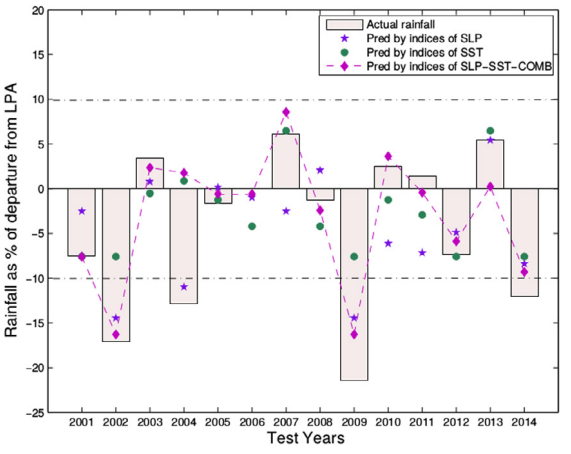

Fig. 9. Monsoon prediction by identified combined SLP+SST-based predictors versus individual monsoon predictors during the test-period of 2001 to 2014.

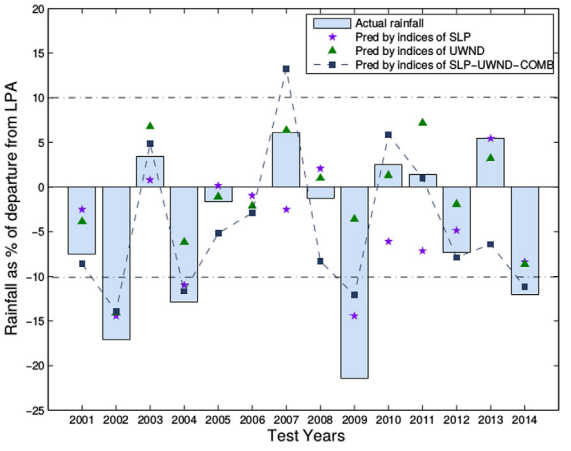

Fig. 10. Prediction by combined SLP+UWND-based predictors versus individual monsoon predictors during the test-period of 2001 to 2014.

However, the identified SST- and UWND-based predictor sets from our proposed method forecast monsoons with 4.7% and 4.2% errors in May of the present year, and SLP-based predictors forecast with a 4.1% error in August of the preceding year. The SLP+SST- and SLP+UWND-based predictors also produce 3.2% and 3.9% errors in forecasting the Indian monsoon with a two month lead in April. Therefore, the predictions by our proposed approach have high future potential. Table 5 shows detailed measures of prediction for the proposed model and IMD models. Finally, Fig. 11 shows the mean absolute errors in monsoon predictions by our proposed method and the IMD models.

Fig. 12 shows comprehensive forecasts by the identified combined SLP+SST-based predictors against the predictions provided by the IMD models for 2003 to 2014. The estimates by our approach are closer to the observed rainfall than the forecasts provided by the India Meteorological Department models.

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

Table 5
Comparison of the proposed stacked autoencoder approach with IMD monsoon prediction model for 2003-2014 (me: mean absolute error in %, re: root mean square error in %, pear.: Pearson correlation, spear.: Spearman correlation, pred. mon.: month of rainfall prediction.

Stacked autoencoder model					
Predictor	me	re	pear.	spear.	pred. mon.
SST	4.7	6.1	0.70	0.77	May
SLP	4.1	5.1	0.80	0.63	August (prev. yr.)
UWND	4.2	4.7	0.82	0.66	May
SLP+SST	3.2	4.9	0.82	0.74	April
SLP+UWND	3.9	5.0	0.80	0.80	April
IMD monsoon prediction model					
Operational	7.5	9.3	0.01	−0.05	June
LRF1	7.1	8.5	−0.06	−0.24	May
LRF2	6.5	8.1	0.21	0.06	June

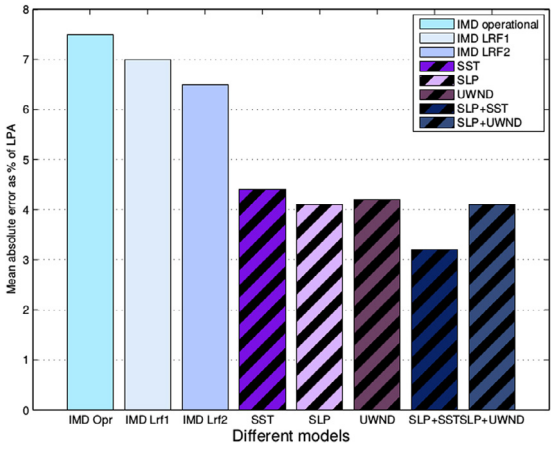

Fig. 11. Monsoon prediction by identified individual SST-, SLP-, UWND-based predictors, and identified combined SLP+SST-, SLP+UWND-based predictors and IMD prediction models for the period 2003 to 2014 (Rajeevan et al., 2004, 2007).

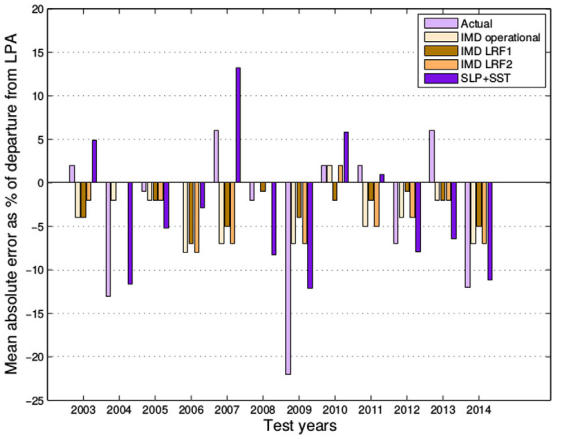

Fig. 12. Monsoon prediction by combined SLP+SST-based predictors and IMD prediction models (Rajeevan et al., 2004, 2007) for 2003 to 2014.

Table 6
Indian monsoon prediction using identified monsoon predictors with the proposed stacked autoencoder-based approach and single-layer autoencoder-based approach (me: mean absolute error in %, re: root mean square error in %).

Stacked autoencoder			Single-layer autoencoder		
Predictor	me	re	Predictor	me	re
SLP	4.0	4.9	SLP	5.5	8.0
SST	4.3	5.7			
UWND	4.2	4.8	SST	5.6	7.8
SLP+SST	2.8	4.5			
SLP+UWND	3.9	4.8	AT	5.6	7.8

Table 7
Pearson correlation (pear.) and Spearman correlation (spear.) between actual and predicted rainfall by predictors obtained from the proposed stacked autoencoder-based method and single-layer autoencoder-based method.

Stacked autoencoder			Single-layer autoencoder		
Predictor	pear.	spear.	Predictor	pear.	spear.
SLP	0.82	0.86	SLP	0.49	0.45
SST	0.74	0.78			
UWND	0.87	0.90	SST	0.51	0.56
SLP+SST	0.85	0.89			
SLP+UWND	0.89	0.88	AT	0.69	0.69

6.3.6. Comparison of the identified monsoon predictors using a stacked autoencoder with predictors derived from a single-layer autoencoder

We compared the Indian summer monsoon prediction with the predictors obtained from our proposed stacked autoencoder-based approach with the forecast provided by a single-layer autoencoder (Saha et al., 2016). Saha et al. (2016) utilize the single-layer autoencoder to identify new monsoon predictors from sea level pressure (SLP), sea surface temperature (SST), and air temperature (AT). We have considered the same test-period (2001–2014) for comparing the performance of identified predictors with predictors obtained from a single-layer autoencoder. The stacked autoencoder-based identified predictor is observed to be superior in predicting the aggregate Indian monsoon. The mean absolute errors in monsoon prediction by predictors identified by our approach and the single-layer autoencoder method are elaborated in Table 6. The identified SLP+SST predictors forecast the monsoon with a 2.8% error, whereas the SLP predictors using the single-layer autoencoder forecast the monsoon rainfall with a mean absolute error of 5.5%.

Table 7 shows the correlation between the predicted and actual monsoon by both methods. The stacked autoencoder-based predictors forecast the rainfall with a high correlation of 0.85 against 0.69 for the predictors of a single-layer autoencoder.

Fig. 13 shows the prediction by SLP+SST predictors obtained using stacked autoencoders and SLP predictors identified by single-layer autoencoders. The stacked autoencoder-based forecasts are closer to the observed monsoon. Stacked autoencoder-based predictors correctly forecast the drought years of 2002, 2009, and 2014. The single-layer autoencoder approach is observed to forecast rainfall above the long period average for the drought year of 2002. It even predicts the erroneous sign of anomaly

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

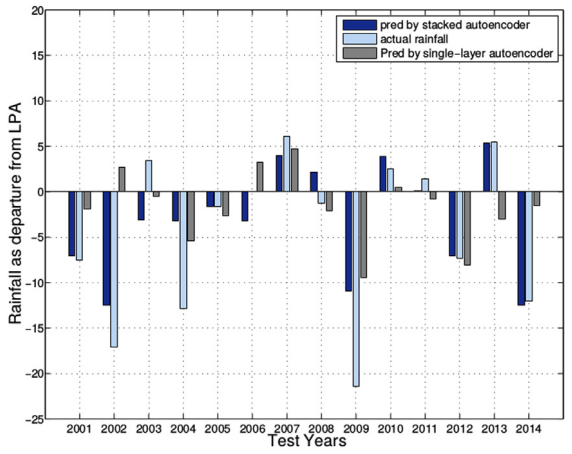

Fig. 13. Comparison of prediction by staked autoencoder-based identified SLP+SST predictors and single-layer autoencoder-based identified SLP predictors during 2001 to 2014.

of rainfall for 2002, 2003, 2006, 2011, and 2013. Thus, all the outcomes highlight the efficacy of the stacked autoencoder over the single-layer autoencoder.

6.3.7. Comparison of the stacked autoencoder-based monsoon prediction with dynamical model forecasts

The Indian monsoon rainfall is simulated using hindcasts of the National Centers for Environmental Prediction Climate Forecast System version 2 (CFSv2) (Ramu et al., 2016). We evaluated the forecast accuracy of the rainfall at two horizontal resolutions, namely T-126 (∼100 km) and T-382 (∼38 km). Ramu et al. (2016) reported Pearson correlation coefficients of 0.49 and 0.55 between the actual and forecast precipitation for T126 and T382 resolutions. The proposed approach shows Pearson correlation coefficients of 0.74, 0.82, 0.87, 0.85, and 0.89 with p-values$<$0.01 between the predicted and actual rainfall by predictors of SST, SLP, UWND, SLP+SST, and SLP+UWND, respectively. The predictions by the stacked autoencoder-based approaches outperform the coupled model predictions at both low and high resolutions.

We compared the prediction by the proposed approach with an ensemble of multiple model forecasts (Rajeevan, Unnikrishnan, & Preethi, 2012). Six ocean-atmosphere coupled models with nine ensemble members are each considered for developing a new ensemble multi-model (ENSEMBLE MME). The six coupled models considered for this purpose were ECMWF, UKMO, Meteo France, CMCC-INGV, IFM-GEOMAR, and DePreSys, UKMO. The successs of a new ensemble multi-model in the prediction of the monsoon is showed to have improved over the DEMETER MME, a multi-model ensemble from a previous intercomparison project. The six mentioned coupled models, DEMETER MME, and ENSEMBLE MME, show Pearson correlation coefficients of 0.37, 0.34, 0.34, 0.39, 0.39, 0.27, 0.28, and 0.45, respectively, between the predicted and actual monsoon rainfall during 1960–2005. The prediction of the monsoon by identified predictors (SLP+SST) using our proposed approach reports a correlation of 0.85 during 2001–2014.

Table 8
Observed and prediction of the Indian summer monsoon (as a percentage of LPA) using the stacked autoencoder model for 2015 to 2018.

Year	Observed	Predicted
2015	86%	93.0% ± 3.0%
2016	97%	95.7% ± 2.3%
2017	95%	99.7% ± 1.7%
2018	91%	97.0% ± 1.7%

Nanjundiah, Francis, Ved, and Gadgil (2013) focused on analyzing the inter-annual variation of the monsoon phenomenon and predicting the extremes using seven coupled ocean-atmospheric models from predicting centers in the USA and Europe. Five models are identical to the ones considered by Rajeevan et al. (2012) except DePreSys, UKMO. The last two models are a coupled forecast system, CFS1 and CFS2 of NCEP. CFS1 and CFS2 models predict the Indian summer monsoon with Pearson correlation coefficients of 0.14 and 0.41, respectively, between the actual and forecast rainfall during 1982–2009. The predictors derived from the climatic variables using our method show superior Pearson correlation coefficients (0.82, 0.74, 0.87, 0.85, and 0.89) between the observed and predicted monsoons during the test-period (2001–2014). Even if we consider the same period (1982–2009) as the coupled forecast system model, a part of the period from 1982 to 2000 lies in our training phase and the remaining period from 2001 to 2009 in the testing phase. Even then the correlation between the actual and forecast rainfall is high (0.88 for SLP+SST predictors during 1982–2009). All the results establish the efficacy of the proposed stacked autoencoder-based approach in the forecasting of the Indian monsoon over global-coupled models. We, however, would like to reiterate that a dynamical model generates not only seasonal forecasts but many other useful products, and in recent years, it has shown significant improvements in skill related to the improved incorporation of physical processes, improved resolution, etc.

6.3.8. Predictions for the years 2015 to 2018

The monsoon predictors are used to forecast the Indian monsoon for the years from 2015 to 2018 (Table 8). The selected predictor set for performing the forecast is the one that has the least errors during the test-period 2001–2014.

7. Discussion of identified predictors

Identified predictors are a combination of variables from various geographical regions; some areas are well-known areas influencing the Indian monsoon, while some new regions also participate in the formation of monsoon predictors.

Predictors identified from sea surface temperature show a strong resemblance to the areas of the Equatorial Pacific Ocean; the region of the El-Niño event that has a high impact on the intensity of the Indian rainfall (Joseph, Eischeid, & Pyle, 1994). It also encapsulates the area of the

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

Equatorial and South-Western Indian Ocean that affects the flow of moisture into the Indian landmass.

Predictors derived from sea level pressure are over the regions of the Tibetan Plateau and the sea near Madagascar (in the vicinity of the Mascarenes High, which has an impact on precipitation of the Indian summer monsoon) (Krishnamurti & Bhalme, 1976). The pressure gradient between the two regions can influence circulation and wind patterns and, hence, the strength of the monsoon rainfall. Other regions include the North-Western part of North and South America, Northern Europe, and the Siberian regions. The sea level pressure of these regions is found to participate in identified predictors that are significant for the Indian summer monsoon.

We observed the zonal wind-based predictors sparsely located over the globe. Winds flowing over the regions of Western Europe and the North Pacific Ocean participate in new predictors. Another major area obtained is the Equatorial Indian Ocean. The zonal winds blowing in this part carry the moisture from the sea towards the landmass. Lastly, the wind over some portions of the South Atlantic and South Pacific Oceans also engages in UWND-based identified predictors that influence the monsoon over the Indian region. A more detailed analysis of these identified predictors needs to be conducted to understand their role in modulating the Indian monsoon.

8. Conclusions

A stacked autoencoder is utilized to identify non-linear climatic predictors as an amalgamation of variables from various geographical regions. We captured the predictors from the learned nodes at the internal layers of the architecture. The first inner layer mainly assists in the formation of new climatic predictors, whereas the subsequent layers compress the feature formed in the initial layer. Prediction models show their skill in forecasting the monsoon using the identified predictors. Both the models (RegTree and DecTree) perform proficiently. But, DecTree is observed to predict the monsoon with a higher accuracy using most of the identified predictors (both for predictors derived from an individual variable and a combination of variables). Predictions by combined monsoon predictors show a better performance than individual predictors. Identified predictors also capture extremes, and they are comparable to the current IMD monsoon prediction models and numerical models. Lastly, the significance of newly-identified predictors in the meteorological domain is also explored.

Future scope may involves the use of more complex deep architecture, namely, convolution neural networks, long-short term memory models, etc. These models will assist in the identification of sophisticated predictors from a combination of multiple input variables from various collections of the network models. Moreover, the predictors can be analyzed more crucially over geographical domains, and can be explored to understand their effects on climatic phenomena.

References

Annamalai, H., Hafner, J., Sooraj, K. P., & Pillai, P. (2013). Global warming shifts the monsoon circulation, drying South Asia. *Journal of Climate*, *26*(9), 2701–2718.

DelSole, T., & Shukla, J. (2002). Linear prediction of Indian monsoon rainfall. *Journal of Climate*, *15*, 3645–3658.

DelSole, T., & Shukla, J. (2012). Climate models produce skillful predictions of Indian summer monsoon rainfall. *Geophysical Research Letters*, *39*(9).

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527–1554.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Joseph, P. V., Eischeid, J. K., & Pyle, R. J. (1994). Interannual variability of the onset of the Indian summer monsoon and its association with atmospheric features, El Nino, and sea surface temperature anomalies. *Journal of Climate*, *7*(1), 81–105.

Kalnay, E., et al. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, *77*(3), 437–471.

Kelly, P., & Mapes, B. (2011). Zonal mean wind, the Indian monsoon, and july drying in the western atlantic subtropics. *Journal of Geophysical Research: Atmospheres*, *116*(D21).

Koll, R. M., & Chaithra, S. T. (2018). *Impacts of climate change on the Indian summer monsoon*. Ministry of Environment, Forest and Climate Change, Government of India.

Krishnamurti, T. N., & Bhalme, H. (1976). Oscillations of a monsoon system. Part I. Observational aspects. *Journal of the Atmospheric Sciences*, *33*(10), 1937–1954.

Krishnan, R., Sabin, T. P., Vellore, R., Mujumdar, M., Sanjay, J., Goswami, B. N., et al. (2016). Deciphering the desiccation trend of the South Asian monsoon hydroclimate in a warming world. *Climate Dynamics*, *47*(3–4), 1007–1027.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22.

Liu, J. N. K., Hu, Y., He, Y., Chan, P. W., & Lai, L. (2015). Deep neural network modeling for big data weather forecasting. In *Information granularity, big data, and computational intelligence* (pp. 389–408). Springer.

Loh, W. Y. (2008). Classification and regression tree methods. *Encyclopedia of Statistics in Quality and Reliability*, 315–323.

Loo, Y., Billa, L., & Singh, A. (2015). Effect of climate change on seasonal monsoon in Asia and its impact on the variability of monsoon rainfall in Southeast Asia. *Geoscience Frontiers*, *6*(6), 817–823.

MATLAB (2012). *Statistics and machine learning toolbox*. Natick, Massachusetts, United States: MATLAB version 2012b, The MathWorks Inc.

Nanjundiah, R. S., Francis, P. A., Ved, M., & Gadgil, S. (2013). Predicting the extremes of Indian summer monsoon rainfall with coupled ocean-atmosphere models. *Current Science*, *104*(10), 1380–1393.

Ogata, T., & Xie, S. (2011). Semiannual cycle in zonal wind over the Equatorial Indian ocean. *Journal of Climate*, *24*(24), 6471–6485.

Patil, N., Venkataraman, C., Muduchuru, K., Ghosh, S., & Mondal, A. (2019). Disentangling sea-surface temperature and anthropogenic aerosol influences on recent trends in South Asian monsoon rainfall. *Climate Dynamics*, *52*(3–4), 2287–2302.

Priya, P., Krishnan, R., Mujumdar, M., & Houze, R. A. (2017). Changing monsoon and midlatitude circulation interactions over the Western Himalayas and possible links to occurrences of extreme precipitation. *Climate Dynamics*, *49*(7–8), 2351–2364.

Rajeevan, M., Pai, D. S., Dikshit, S. K., & Kelkar, R. R. (2004). IMD's new operational models for long-range forecast of southwest monsoon rainfall over India and their verification for 2003. *Current Science*, *86*(3), 422–431.

Rajeevan, M., Pai, D. S., Kumar, R. A., & Lal, B. (2007). New statistical models for long-range forecasting of southwest monsoon rainfall over India. *Climate Dynamics*, *28*(7–8), 813–828.

Rajeevan, M., Unnikrishnan, C. K., & Preethi, B. (2012). Evaluation of the ENSEMBLES multi-model seasonal forecasts of Indian summer monsoon variability. *Climate Dynamics*, *38*(11–12), 2257–2274.

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.

Ramu, D. A., Sabeerali, C. T., Chattopadhyay, R., Rao, D. N., George, G., Dhakate, A. R., et al. (2016). Indian Summer monsoon rainfall simulation and prediction skill in the CFSv2 coupled model: Impact of atmospheric horizontal resolution. *Journal of Geophysical Research: Atmospheres*, *121*, 2205–2221.

Reynolds, R. W., Rayner, N. A., Smith, T. M., Stokes, D. C., & Wang, W. (2002). An improved in situ and satellite SST analysis for climate. *Journal of Climate*, *15*, 1609–1625.

Saha, M., Chakraborty, A., & Mitra, P. (2016). Predictor-year subspace clustering based ensemble prediction of Indian summer monsoon. *Advances in Meteorology*, *2016*(9031625), 1–12.

Saha, M., & Mitra, P. (2016). Recurrent neural network based prediction of Indian summer monsoon using global climatic predictors. In *2016 international joint conference on neural networks* (pp. 1523–1529).

Saha, M., & Mitra, P. (2019). Identification of Indian monsoon predictors using climate network and density-based spatial clustering. *Meteorology and Atmospheric Physics*, *131*(5), 1301–1314.

Saha, M., Mitra, P., & Chakraborty, A. (2015). Fuzzy clustering-based ensemble approach to predicting Indian monsoon. *Advances in Meteorology*, *2015*(32835), 1–12.

Saha, M., Mitra, P., & Nanjundiah, R. S. (2016). Autoencoder-based identification of predictors of Indian monsoon. *Meteorology and Atmospheric Physics*, *128*(5), 613–628.

Saha, M., Mitra, P., & Nanjundiah, R. (2016). Predictor discovery for early-late Indian summer monsoon using stacked autoencoder. *Procedia Computer Science*, [ISSN: 1877-0509] *80*, 565–576, (ICCS 2016).

Saha, M., Mitra, P., & Nanjundiah, R. (2017). Deep learning for predicting the monsoon over the homogeneous regions of india. *Journal of Earth System Science*, *126*(4), 1–18.

Song, C., Liu, F., Huang, Y., Wang, L., & Tan, T. (2013). Auto-encoder based data clustering. In *Progress in pattern recognition, image analysis, computer vision, and applications* (pp. 117–124). Springer.

Swapna, P., Krishnan, R., & Wallace, J. M. (2014). Indian OCean and monsoon coupled interactions in a warming environment. *Climate Dynamics*, *42*(9–10), 2439–2454.

Turner, A. G., & Annamalai, H. (2012). Climate change and the South Asian summer monsoon. *Nature Climate Change*, *2*(8), 587–595.

Please cite this article as: M. Saha, A. Santara, P. Mitra et al., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.03.001.