***Original Research Article***

# Multi-variate time-series analysis using VECM identifies the best set of exogenous predictors for rainfall and temperature in India: a data analytical approach




## ABSTRACT

The complex nature of climate change with multitude of underlying factors poses a major hindrance in data analysis and decision making by policy makers. Here, we utilize data analytic techniques to identify the impact of different climate change indicators on precipitation and temperature data from India. The observed values of important climatic parameters namely, rain, maximum temperature, minimum temperature and mean temperature in India were analyzed along with the observed values of selected socio-environmental indicators featured by WHO for climate change as exogenous variables for a period of 61 years. Data were pre-processed to identify ten exogenous indicators which were then modelled using Vector Error Correction Model (VECM). 1024 VECM models were built and evaluated for the prediction of the four endogenous variables using all possible combinations of the selected indicators. Seven exogenous variables were determined as the best set of indicators based on the AIC of the different models. The model built using the identified variables was compared to others to illustrate the probable impact of this combination of variables. The study thus demonstrates a simple but rational data-driven approach for use in decision making.


*Keywords: Climate change, time-series, data analytics, VECM, AIC*

## 1. INTRODUCTION

The rate of climate change is at an all-time high since the 20$^{th}$ century. Research seeking solutions to global and local climate changes are underway and is the need of the hour [12]. Efforts are underway at a global scale by the Governments of different countries to tackle climate change [2]. The World Bank enlists a set of 76 climate change indicators pertaining to economic, social, environmental, health, education, development, and energy domains. The Climate Change Knowledge Portal (CCKP) was established to provide access to the global, regional, and national data and reports consolidated from the World Bank as a learning resource [15]. Currently, the portal presents 34 indicators as featured climate change indicators for any nation and provides the available observed data for countries including India.

Climate change data recorded as yearly series is an invaluable source of knowledge. In the data-driven world, scientists use analytic procedures to develop solutions to critical challenges spanning across different domains including business management, economics, law enforcement and medicine [4, 6, 10, 24]. Methods and processes involved in data analytics have helped to generate knowledge for decision making in issues pertaining to climate change also [9, 23]. Several works have been conducted in India that dealt with forecasting of weather conditions such as precipitation and temperature by predictive analytic approaches. Several techniques of data analytics based on statistical or machine learning principles such as Artificial Neural Networks (ANN) [3, 16], Support Vector Machines (SVM)

[20], regression analysis and time-series modelling have been utilized. Wamanse and Patil (2022) utilized regression analysis to study the impact of different featured indicators in climate change [23]. Shivhare et al., 2019 utilized auto regressive integrated moving average (ARIMA), time-series specific algorithm for univariate analysis to develop a daily weather forecast tool for Varanasi. The work by Dimri et al., 2020 is commendable for the ARIMA analysis incorporating the seasonality of the variables. The work also reviews in details previous efforts carried out using such techniques in India. Time-series modelling by autoregression is an interesting technique used to forecast variables of climate change using past values and trends [7, 11]. Majority of these studies utilized the ARIMA model to forecast a parameter based on historical values. However, ARIMA can be applied only for univariate analysis, and it is important that multiple parameters are considered for prediction of variables pertaining to a multi-layered issue like climate change. Kaur et al., 2021 compared different types of models for multi-variate analysis including machine learning techniques and time-series analysis to study the changes in water levels and rate of flow in water bodies. When considering more than one variable at a time, multivariate approaches for regression such as Vector Auto Regression (VAR) or Vector Error Correction Model (VECM) are preferred over ARIMA models to investigate the relationships between them [7]. While VAR is implemented on stationary variables, non-stationary, co-integrated variables are examined using VECM by incorporating error correction.

A generalized VECM with $p$ lags and co-integration rank $r \leq k$ is expressed as

$$\Delta y_t = \prod y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + D_t + \phi_t,$$

where:
$\Delta$ is the operator differencing with $\Delta y_t = y_t - y_{t-1}$;
$y_{t-1}$ = vector variable endogenous with lag 1;
$\phi_t = k \times 1$ vector residuals;
$D_t = k \times 1$ vector constant;
$\Pi$ is the matrix coefficient of cointegration = $\alpha\beta^t$;
$\alpha$ = vector adjustment, $k \times r$ matrix and
$\beta$ = matrix cointegration (long-run parameter) $(k \times r)$;
$\Gamma_i = k \times k$ matrix coefficient for the $i^{th}$ variable endogenous;
and $k$ is the size of vectors considered.

The classical forms of VAR consider all variables as endogenous and is modelled as a linear function of their lags. However, in complex problems like climate change, it is possible to identify other independent variables which influence the value of endogenous variables. In such cases, they are incorporated as exogenous variables and their current/lagged values are also considered for prediction of endogenous variables. Thus, depending upon the influence of the exogenous variables, these models can perform better than conventional univariate models by incorporating the impact of these variables. The model nevertheless does not include modelling of the exogenous variables. VECM models have been used in various complex forecast problems including market prices of products, agricultural output research, economic issues, stock prices and farmers trade [1, 17, 19, 22].

In the current work, we aim to explore the utility of the WHO featured climatic change indicators as exogenous variables in forecasting the future values of rain and temperature as endogenous variables. The rationale of the work was to identify the best combination of exogenous variables for the prediction of rainfall and temperature by comparing the performances of various VECM models based on the currently available annual data from India.

## 2. MATERIALS AND METHOD

The present study discovers the impact of ten featured climate change indicators in forecasting the values of rain, mean temperature, maximum temperature and minimum temperature.

## 2.1 Dataset selection

A consolidated dataset of available observed annual values of 65 Economic, Social, Environmental, Health, Education, Development and Energy indicators of India indicators in India between the period of 1960 and 2020 were sourced from the World Bank's data portal at (https://data.humdata.org/dataset/world-bank-climate-change-indicators-for-india). As the primary target variables, the recorded annual values of four climatic parameters namely rain/precipitation, mean temperature, maximum temperature and minimum temperature were retrieved from the Climate Knowledge Portal of the World Bank (https://climateknowledgeportal.worldbank.org/).

## 2.2 Data pre-processing

The WHO lists 34 featured indicators of climate change. The raw data of socio-environmental indicators were filtered to extract only the recorded values of these featured indicators from 1960 to 2020. The dataset was available in the form of an edgelist matrix that was converted to an adjacency matrix datafile using an in-house coded R script. The converted matrix consisted of the various indicators as columns of time-series values and the year of record as the rows. Any series with more than 20% missing values were removed from the analysis to reduce bias and ensure that the lack of variation in these series does not affect the model. The threshold value for removal of variables based on missing data was set after analyzing the role of these factors in climate change manually. The observed annual values of precipitation and temperature for the respective years were then appended to the dataset. The data consisted of observed values of different indicators that were spread across different ranges. To achieve linearity and stabilization of variances, the values were log-transformed and scaled to a similar range of values for further calculation purposes. The multi-variate time-series data was then split into training and test sets with the final ten years categorized for testing of the models.

## 2.3 Multi-variate time-series analysis

The primary aim of the study was to predict the utility of the socio-environmental indicators in forecasting climatic parameters. For time-series forecast analysis, the target variables, whose future values are to be predicted, were considered as endogenous variables while the ten indicators that serve as probable additional predictors, but will not be predicted, were considered as exogenous variables.

Once the multi-variate time-series problem was finalized, the next step was to identify stationarity and co-integration rank of the variables for appropriate selection of time-series analysis algorithm. The VARselect package in R was utilized first to identify the optimal lag. Stationarity and co-integration rank were tested using the augmented Dickey–Fuller (ADF) test and the Johansens test. Based on the results, the dataset was then modelled using a modified version of VAR with error correction known as the VECM.

## 2.4 Model building and evaluation

To study the impact of different combinations of the exogenous variables and to identify the best combination, all possible models were built using the VECM() function in the R package tsDyn. All the four endogenous variables were provided as the input for all the models while the exogenous variables were provided differently, in all possible combinations.The Akaike Information Criterion (AIC) of the models were then analyzed to select the best model. The models were then examined for their fit, residuals and accuracy in predicting the test data set. Predictions were carried out using the models built using (i) the selected model with the apparently best combination of exogenous variables, (ii) the model built with all the exogenous variables and (iii) the model built without any exogenous variable.

## RESULTS

### 3.1 Data collection and analysis

After pre-processing, a data matrix of 14 parameters for 61 years were obtained to be used as a multi-variate time-series dataset. Among the retrieved list of 34 featured socio-environmental indicators, data was available for 31 indicators while the data for three indicators namely, Electric power consumption (kWh per capita), Energy use (kg of oil equivalent per capita) and Renewable electricity output (% of total electricity output) were missing. After removal of indicators with more than 20% missing values, 10 indicators were retained along with the recordings of four other variables – rain, mean temperature, maximum temperature and minimum temperature. The values of four indicators were missing for the year 1960 which was then imputed using standard techniques. Among the time-series data spanning over a period of 61 years, the first 51-year data (1960-2010) was used as the training set and the last ten-year data from 2011 to 2020 was utilized as testing data for evaluation purposes.

The forecast problem was presented as a multi-variate time series analysis with four endogenous (rain, mean temperature, maximum temperature and minimum temperature) and ten exogenous variables (Table 1). The selection of endogenous/exogenous variables were done based on the respective definitions of variables that shall be forecasted. Endogenous variables were the ones that will be predicted into the future while values of exogenous variable will only be used for the prediction of endogenous variables. Figures 1 and 2 presents the time-series plots of these indicators.

**Table 1: List of exogenous variables used in the study and the indicator codes**

| | |
|---|---|
| Agricultural land (% of land area) | AG.LND.AGRI.ZS |
| Arable land (% of land area) | AG.LND.ARBL.ZS |
| Cereal yield (kg per hectare) | AG.YLD.CREL.KG |
| Population in urban agglomerations of more than 1 million (% of total population) | EN.URB.MCTY.TL.ZS |
| Agriculture, forestry, and fishing, value added (% of GDP) | NV.AGR.TOTL.ZS |
| Mortality rate, under-5 (per 1,000 live births) | SH.DYN.MORT |
| Population growth (annual %) | SP.POP.GROW |
| Population, total | SP.POP.TOTL |
| Urban population | SP.URB.TOTL |
| Urban population (% of total population) | SP.URB.TOTL.IN.ZS |

### 3.2 Selection of time-series analysis algorithm

The optimal lag number for the selected endogenous and exogenous variables for multi-variate time-series analysis was identified as K=2 according to the AIC criterium. The p-values of ADF tests (as well as the time-series plots in Figure 1) indicated that all variables except rain were non-stationary ($p>0.05$). Further, Johansens test was carried out to identify the co-integration rank of the endogenous variables as r=1. As the problem was identified as non-stationary and cointegrated, a modified version of VAR with error correction known as VECM was utilized accordingly.

### 3.3 Model building and evaluation

Models were built using the four predictor variables as endogenous. The set of exogenous variables were shuffled to obtain all possible combinations of the ten indicators. An exhaustive set of all possible

combinations of the ten endogenous variables with number of variables n=1 to 10; ranging from each variable separately (n=1, 10 possible combinations) to all the ten together (n=10, 1 possible combination) were considered.

Thus, a total of 1023 VECM models were built using all possible combinations of the ten exogenous variables and four endogenous variables with parameters set as K=1 and r=2. An additional model with no exogenous variable was also built for comparison. The model with the least AIC value was then selected to identify the best fit model. The results demonstrate that a combination of seven exogenous variables, namely Agricultural land (% of land area), Arable land (% of land area), Cereal yield (kg per hectare), Population in urban agglomerations of more than 1 million (% of total population), Agriculture, forestry, and fishing, value added (% of GDP), Population growth (annual %), and Urban population (% of total population) provided the best model with the minimum AIC value of -2044.335.

### 3.4 Prediction and performance evaluation of the best model

The selected model with the least AIC, built using the above combination of exogenous variables (hereafter referred to as M1) were compared with two other models for performance evaluation: the model built using all the ten exogenous variables (M2) and the model built without any of the exogenous variables (M3) The values of the endogenous variable for subsequent ten years, i.e from 2011 to 2020, were predicted using these models. The predicted values of the four variables were then compared with their actual values from test data set to estimate the accuracy of each of these models (Table 2). The forecast plots as well as the fitted vs residual plots of the model M1 are provided in Figures 3 and 4.

## 3. DISCUSSION

Time-series forecasting throw light into the patterns underlying the metrics being observed. Conducting multivariate time-series analysis aids in observing a group of variables over a specific duration. Compared to machine learning models, these models allow for investigation of causal relationships between the variables with least bias and not posing problems like data imbalance.

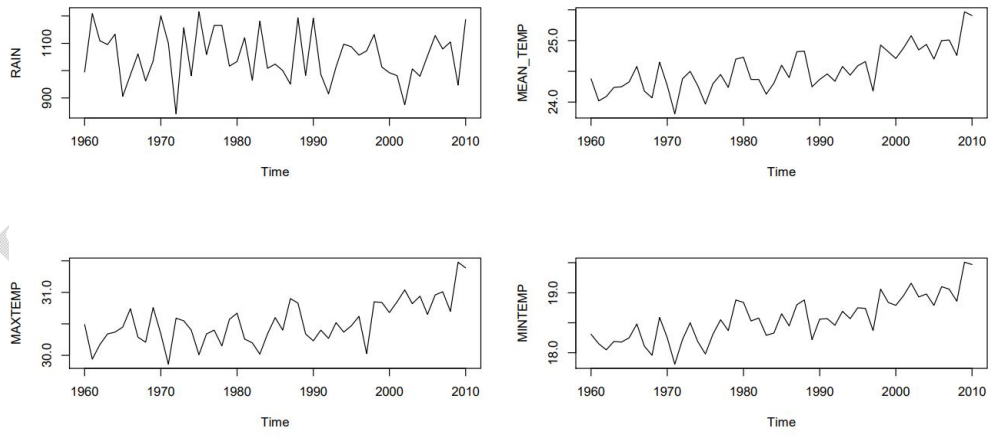


**Fig 1: The time-series plots of endogenous variables used in the study**

Non-parametric and data-driven analysis by time-series interrogation is hence known to improve such prediction tasks [7]. To identify the best set of exogenous variables that influence the values of the endogenous variables, all possible combinations of the available ten exogenous variables were first formulated. A three-step analysis was undertaken to investigate the possible impact of each combination. Stationarity was checked using ADF test and the integration of the variables were verified using Johansen co-integration test. Error correction models for multi-variate series were built with all other parameters same but with different set of exogenous variables.

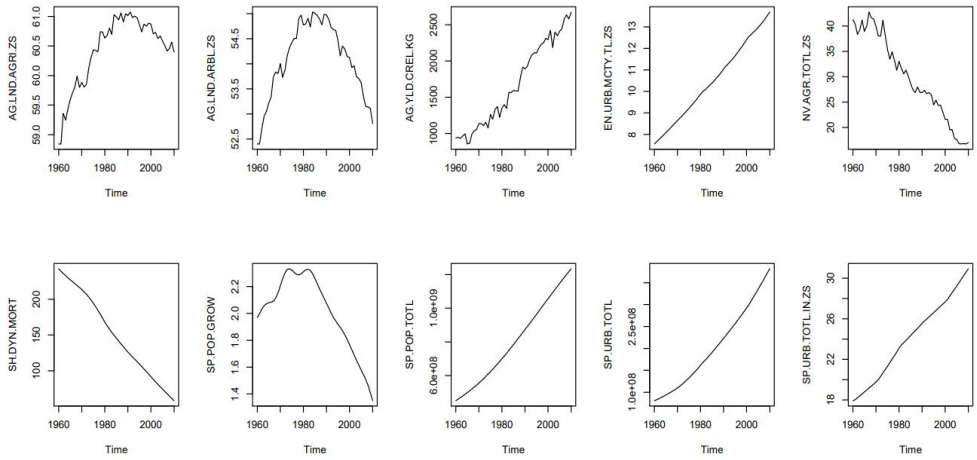


**Fig 2: The time-series plot of exogenous variables used in the study**

The best model for prediction was then selected based on AIC values as it has been previously established [25]. Rationally, this model should also predict the best values of endogenous variables. In order to test the prediction, the model predictions were then compared to the predictions by model M2 and M3 in terms of accuracy statistics estimated by the RMSE. Substantiating the study approach, the model M1 was found to perform better for all the four variables (Table 2). The approach used in this work thus presents an innovative turn-around of the VECM technique to identify the most impactful variables for any study.

**Table 2: Model performance evaluation**

| Model | AIC | RMSE-Rain | RMSE-MeanTemp | RMSE-MaxTemp | RMSE-MinTemp |
|---|---|---|---|---|---|
| M1 | -2044.335. | 0.093866 | 0.016521 | 0.014657 | 0.01954 |
| M2 | -2002.907 | 0.749467 | 0.027782 | 0.035837 | 0.022416 |
| M3 | -2037.593 | 0.711903 | 0.023055 | 0.0136 | 0.040782 |

**See text for details on M1, M2 and M3*

As of date, WHO lists 76 variables as indicators of climate change in India among which 34 has been listed in featured indicators. Previous research conducted in climate change though attempted to forecast future values based on past values, do not look into the complex relationships between the explicit

parameters such as precipitation, temperature, sea level etc. and implicit indicators such as the ones described by WHO. These implicit indicators could be hence considered as exogenous or independent variables and it is important to include them while studying the endogenous variables. Such studies have been conducted in econometric and business management domains, but have not been found exploited in environmental sciences or climatology.

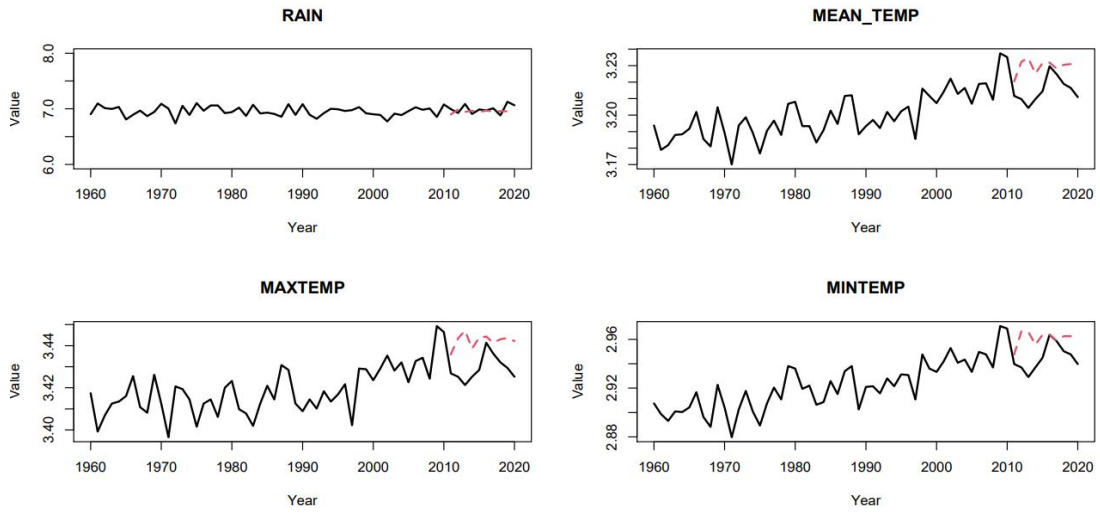


**Fig 3: The forecast plots comparing the predicted and actual values built using the minimum AIC model M1. The black solid line indicates the actual values while the red dashed ones are the predicted values**

Some of the previous works carried out in India with reference to climate change utilized ARIMA or the seasonal ARIMA (SARIMA) models for time-series analysis [5, 8, 13, 21, 18]. Though all these models have been successful in varying levels to predict temperature and rainfall in the respective areas, they have all been essentially univariate studies, in which the influences of one or more variables over the others were not evaluated. This is a major drawback when investigating complex multi-layer issues like climate change. Hence, in the current work we have adopted the utility of climate change related indicators to forecast rain and temperature values. This overcomes the disadvantage of relying only on the historical values of the same variable for prediction. The primary contributions of this study thus include the prediction of climatic variables by considering not only the historical values of that variable but also other variables after suitable identification of the most influential variables.

To the best of our knowledge, no study reports the relevance or ranking of WHO-featured climatic change indicators with respect to the scenario in Indian subcontinent. As a pioneer effort in this direction, we gathered the available data of different indicators for 50 years to train error correction models for multi-variate time series. Though recordings are available for numerous other variables too, it was found to be incomplete; comprising of a large amount of missing data. Eyeballing the raw data provided another possible option for data pre-processing; one in which a greater number of variables could be included for study, but at the cost of reduced number of years.

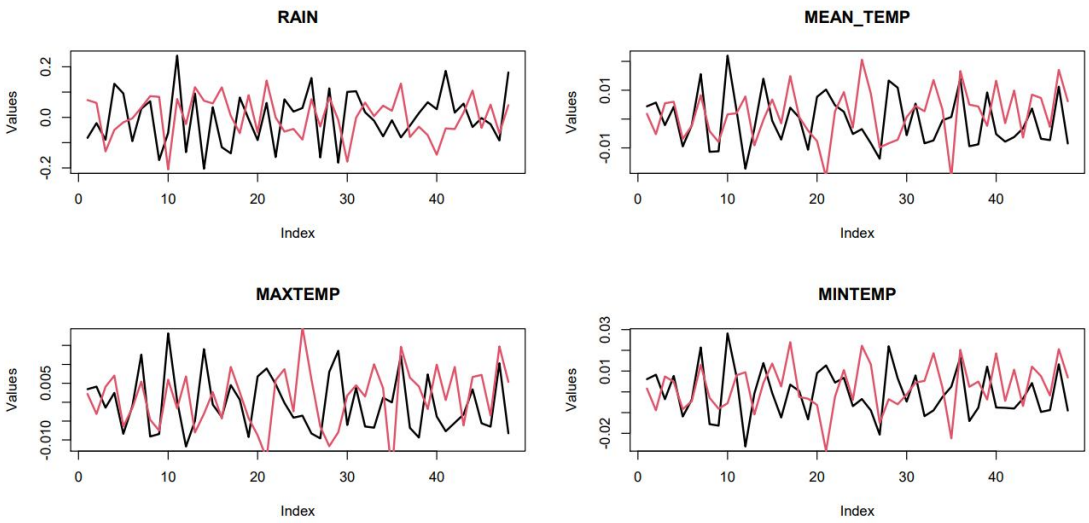


**Fig 4: The fitted vs residual plots of the minimum AIC model M1. The black solid line indicates the fit values while the red dashed ones are the residual values**

The limited number of data points in terms of the temporal variable can affect the performance of a time-series model as used in this study. The study by Wamanse and Patil (2022) used such a scenario and used machine learning techniques for regression and forecasting of the featured indicators only but without any attempt to look into any relationship investigation or relevance study of the variables. They did not consider any of the explicit parameters such as precipitation or temperature too. In contrast, the work described in the current text, utilizes an efficient model of multi-variate time series algorithm to investigate the relationship of available variables in determining the values of explicit parameters for climatic change such as rain-fall and temperature.

## 5. CONCLUSION

Time series analysis is an efficient technique to forecast complex time-based parameters based on the past values. However, regular time-series analysis predicting the future values solely based on its own past values is not efficient when it comes to complex variables with different predictors such as those involved in climate change. In the current study, we use the error correction model for vector autoregression to predict the future values of different climatic parameters not only by analyzing their past values but also taking into consideration some of the other available indicators. The study utilized VECM algorithm to identify the best combination of these variables and presented the most efficient model based on the available data. The study thus endorses the detection of important variables to be monitored for creating an impact in tackling climate change issue. However, this study is not without limitations. Primarily, the study utilized only a limited dataset of both endogenous and exogenous variables for prediction. Climate change is a major phenomenon which deals with various other parameters that can be investigated similarly. Secondly, the models built in the study are linear predictors. This is expected to produce a lower bias but at the cost of a higher RMSE. Nevertheless, such rational analyses as described in this study could be further utilized for other possible datasets with selected interesting endogenous and exogenous variables to explore their relationship. In cases such as climate change,

where multiple factors underlie the apparent measurements of rainfall, temperature, sea levels and snowfall, it is important that an integrated view is considered for effective remedial actions.

## REFERENCES


[1] Adinew, Melaku, and Gebrekirstos Gebresilasie. "Effect of Climate Change on Agricultural Output Growth in Ethiopia: Co-Integration and Vector Error Correction Model Analysis." *Budapest International Research in Exact Sciences (BirEx) Journal* (2019): 132-143.

[2] Beg, Noreen, Jan Corfee Morlot, Ogunlade Davidson, Yaw Afrane-Okesse, Lwazikazi Tyani, Fatma Denton, Youba Sokona et al. "Linkages between climate change and sustainable development." *Climate policy* 2, no. 2-3 (2002): 129-144.

[3] Chithra, N. R., Santosh G. Thampi, Sujith Surapaneni, Revanth Nannapaneni, A. Reddy, and J. Dinesh Kumar. "Prediction of the likely impact of climate change on monthly mean maximum and minimum temperature in the Chaliyar river basin, India, using ANN-based models." *Theoretical and Applied Climatology* 121, no. 3 (2015): 581-590.

[4] de Medeiros, Mauricius Munhoz, Norberto Hoppen, and Antonio Carlos Gastaud Maçada. "Data science for business: benefits, challenges and opportunities." *The Bottom Line* (2020).

[5] Dwivedi, D. K., G. R. Sharma, and S. S. Wandre. "Forecasting mean temperature using SARIMA Model for Junagadh City of Gujarat." IJASR 7, no. 4 (2017): 183-194.

[6] Edmondson, Marquay, Walter R. McCollum, Mary-Margaret Chantre, and Gregory Campbell. "Exploring critical success factors for data integration and decision-making in law enforcement." *International Journal of Applied Management and Technology* 18, no. 1 (2019): 4.

[7] Kaur, Harleen, Mohammad Afshar Alam, Saleha Mariyam, Bhavya Alankar, Ritu Chauhan, Rana Muhammad Adnan, and Ozgur Kisi. "Predicting Water Availability in Water Bodies under the Influence of Precipitation and Water Management Actions Using VAR/VECM/LSTM." *Climate* 9, no. 9 (2021): 144.

[8] Kaushik, Inderjeet, and Sabita Madhvi Singh. "Seasonal ARIMA model for forecasting of monthly rainfall and temperature." Journal of Environmental Research and Development 3, no. 2 (2008): 506-514.

[9] Kneier, Fabian, Carina Zang, Dirk Schwanenberg, Stephan Dietrich, Harald Köthe, and Petra Döll. "Co-Developing a Knowledge Portal for the Presentation and Analysis of Uncertain Global Multi-Model Based Information on Freshwater-Related Hazards of Climate Change." In *Geophysical Research Abstracts*, vol. 21. 2019.

[10] May, Peter, Charles Normand, J. Brian Cassel, Egidio Del Fabbro, Robert L. Fine, Reagan Menz, Corey A. Morrison, Joan D. Penrod, Chessie Robinson, and R. Sean Morrison. "Economics of palliative care for hospitalized adults with serious illness: a meta-analysis." *JAMA internal medicine* 178, no. 6 (2018): 820-829.

[11] Mukadi, Pitshu Mulomba, and Concepción González-García. "Time Series Analysis of Climatic Variables in Peninsular Spain. Trends and Forecasting Models for Data between 20th and 21st Centuries." *Climate* 9, no. 7 (2021): 119.

[12] Richardson, Katherine, Will Steffen, and Hans Joachim Schellnhuber. "Climate change, global risks, challenges and decisions. Synthesis report." (2009).

[13] Sarraf, Amirpouya, Seyed Farnood Vahdat, and Azita Behbahaninia. "Relative humidity and mean monthly temperature forecasts in Ahwaz Station with ARIMA model in time series analysis." In International Conference on Environment and Industrial Innovation IPCBEE, Singapore, vol. 12. IACSIT Press Singapore, 2011.

[14] Shivhare, Nikita, Atul Kumar Rahul, Shyam Bihari Dwivedi, and PRABHAT KUMAR SINGH Dikshit. "ARIMA based daily weather forecasting tool: A case study for Varanasi." *Mausam* 70, no. 1 (2019): 133-140.

[15] Showstack, Randy. "Climate change portal established." (2011): 455-455.

[16] Shrivastava, Gyanesh, Sanjeev Karmakar, Manoj Kumar Kowar, and Pulak Guhathakurta. "Application of artificial neural networks in weather forecasting: a comprehensive literature review." *International Journal of Computer Applications* 51, no. 18 (2012).
[17] Suharsono, Agus, Auliya Aziza, and Wara Pramesti. "Comparison of vector autoregressive (VAR) and vector error correction models (VECM) for index of ASEAN stock price." In *AIP Conference Proceedings*, vol. 1913, no. 1, p. 020032. AIP Publishing LLC, 2017.
[18] Tanusree, D. R., and K. D. Kishore. "Modeling of mean temperature of four stations in Assam." Int. J. Advanced Res 4, no. 12 (2016): 366-370.
[19] Thierry, Belinga, Zhou Jun, Doumbe Doumbe Eric, Gahe Zimy Samuel Yannick, and Koffi Yao Stéphane Landry. "Causality relationship between bank credit and economic growth: Evidence from a time series analysis on a vector error correction model in Cameroon." *Procedia-Social and Behavioral Sciences* 235 (2016): 664-671.
[20] Tripathi, Shivam, V. V. Srinivas, and Ravi S. Nanjundiah. "Downscaling of precipitation for climate change scenarios: a support vector machine approach." *Journal of hydrology* 330, no. 3-4 (2006): 621-640.
[21] Tularam, Gurudeo Anand, and Mahbub Ilahee. "Time series analysis of rainfall and temperature interactions in coastal catchments." Journal of Mathematics and Statistics 6, no. 3 (2010): 372-380.
[22] Usman, Mustofa, Dhia Fadhilah Fatin, M. Yusuf S. Barusman, and Faiz AM Elfaki. "Application of Vector Error Correction Model (VECM) and impulse response function for analysis data index of farmers' terms of trade." *Indian Journal of Science and Technology* 10, no. 19 (2017).
[23] Wamanse, Rutvij, and Tushuli Patil. "Analysis of various climate change parameters in India using machine learning." *arXiv preprint arXiv:2201.10123* (2022).
[24] Xu, Lizhi, Shouyang Wang, Jingjing Li, Ling Tang, and Yanmin Shao. "Modelling international tourism flows to China: A panel data analysis with the gravity model." *Tourism Economics* 25, no. 7 (2019): 1047-1069.
[25] Zhu, Lixin, Lifang Li, and Zhenlin Liang. "Comparison of six statistical approaches in the selection of appropriate fish growth models." *Chinese Journal of Oceanology and Limnology* 27, no. 3 (2009): 457-467.