A Case Study on the Prediction of Heatwave Days Using Machine Learning Algorithms over Telangana

Chapter · October 2023

DOI: 10.1007/978-981-99-4811-6_5

Srikanth Bhoopathi
National Institute of Technology Warangal
5 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Manali Pal
National Institute of Technology Warangal
21 PUBLICATIONS **193** CITATIONS
SEE PROFILE

Chapter 5
A Case Study on the Prediction of Heatwave Days Using Machine Learning Algorithms over Telangana

B. Srikanth and Manali Pal

Abstract This study aims to develop a heatwave prediction models for 7-, 15-, and 30-day lead times using machine learning algorithms, that is, support vector regression (SVR) and random forest (RF) for Telangana, a semiarid region vulnerable to heatwaves. The study uses five meteorological variables, namely, geopotential height, U-wind, V-wind, air temperature, and relative humidity for four atmospheric pressure levels (1000, 925, 850, and 700 hPa) from 1990 to 2019 as predictors. The input data is obtained from ECMWF Reanalysis Version 5 (ERA5), and the spatially averaged daily maximum temperatures of summer (i.e., for the months of April, May, and June) are obtained from the India Meteorological Department (IMD) as predictand. The Principal Component Analysis is used on spatially averaged meteorological variables to reduce the number of closely related variables for different pressure levels. The study shows significant accuracy in predicting the total number of annual heatwave days (HWDs) for Telangana for seven-day lead time, but model performances decrease with increasing lead time. Despite spatiotemporal variations in the connections between heatwaves and predictors, the models are satisfactory, and SVR outperforms RF in predicting HWDs. The study's findings indicate that the spatiotemporal dynamics of meteorological variables could be used for long-term heatwave prediction, and both SVR and RF models have the potential for reliable usage in this context.

Keywords Heatwave days · Machine learning algorithm · Random forest · Support vector regression

B. Srikanth · M. Pal (✉)
Department of Civil Engineering, NIT, Warangal, Telangana, India
e-mail: sb712007@student.nitw.ac.in

M. Pandey et al. (eds.), *River, Sediment and Hydrological Extremes: Causes, Impacts and Management*, Disaster Resilience and Green Growth,
https://doi.org/10.1007/978-981-99-4811-6_5

5.1 Introduction

Heatwaves (HWs) are days for a particular place experiencing abnormally high temperature that exceeds its long-term normal value that is calculated based on the maximum average temperature for a given base period of 30 years. These extreme heat events have serious and deadly consequences for various systems such as human health, agriculture, energy demand, and forest ecosystems and also significant impact on hydrological extremes, particularly in regions with already limited water resources. During a heatwave, high temperatures can increase evaporation rates, leading to decreased soil moisture and lower water levels in rivers, lakes, and reservoirs. This can exacerbate existing drought conditions, leading to a greater risk of water scarcity and water-related conflicts (Afroz et al. 2022). Additionally, high temperatures can increase the likelihood of intense rainfall events, leading to flash floods and landslides in areas with insufficient drainage capacity. These hydrological extremes can have significant economic, social, and environmental consequences, including damage to infrastructure and buildings, loss of crops, and increased risk of waterborne diseases. Therefore, it is essential to consider the potential impacts of heatwaves on hydrological extremes in planning for future water management and adaptation strategies. In recent years, heatwaves have become more common globally and have resulted in many fatalities. For instance, in 2003, a severe heatwave in Western Europe led to the death of over 70,000 people (Coumou and Rahmstorf 2012). Similarly, the 2010 heatwave in Russia, which lasted for a month (Dole et al. 2011; McMichael and Lindgren 2011) and the 2009 heatwave in Southeastern Australia caused the death of 54,000 and 374 people, respectively. In India, heatwave is a significant concern, particularly in the northern and central regions (Das and Umamahesh 2022; Das et al. 2022). The country experiences heatwaves every year, and they have become more frequent and severe in recent years, with temperatures regularly exceeding 45 °C in many regions. This extreme heat can cause heatstroke, dehydration, and other health problems, especially among vulnerable populations such as the elderly, children, and those with pre-existing medical conditions and who also experienced many fatalities due to various occurrences of these HWs. For example, the HW occurred in 1988 killed approximately 1300 people (De and Mukhopadhyay 1998) and in 1998 and 2003 killed approximately 2042 (Jenamani 2012) and 3054 people, respectively; and the toll was 2248 across different parts of the country, in the HWs that occurred in 2015 (Guha et al. 2016). Heatwaves in India also have a significant impact on the agricultural sector, as they can cause crop failures, loss of livestock, and damage to infrastructure. This, in turn, can lead to food shortages and price hikes, affecting the livelihoods of many farmers and their families. Thus, it can have a negative impact on the overall economy, as they can disrupt transportation and energy systems, and reduce productivity in various sectors. In addition, the increased demand for electricity during heatwaves can put a strain on the power grid, leading to power outages and blackouts.

Based on scenario-based projection studies, it is expected that global temperatures will increase by 1.4–5.8 °C (De Perez et al. 2018), which could result in a significant rise in the number of heat-related deaths and illnesses (Meehl et al. 2009; Zhang et al. 2017). The Intergovernmental Panel on Climate Change (IPCC) assessment suggests that the frequency and length of warm days have been increasing since the 1950s and that most of Asia will experience more temperature extremes. Over recent decades, areas with high population density, particularly urban regions, have been more severely affected by these extreme temperatures (Christidis et al. 2015; Mishra et al. 2015). Heatwaves in India typically occur from March to May, which is the pre-monsoon season, and have varying levels of intensity, duration, and negative impacts on different parts of the country (Pai et al. 2013; Basha et al. 2017). A study by Rohini et al. (2016) and Das et al. (2020) used a gridded dataset to investigate the "excessive heat factor," which included both the "excess heat index" and "heat stress index" from 1961 to 2013. The research found that while the frequency and duration of heatwaves are increasing in some parts of central and northwestern India, there were no significant trends in the rest of the country. Khan et al. (2019) suggest that heatwaves could cause even more harm in the coming decades, not just in Asia but also in neighboring regions. To address the challenges posed by heatwaves in India, it is crucial to take steps to mitigate their effects. Therefore in implementing early warning systems, there is a need for the development of a robust model for forecasting heatwaves as a potential climate change mitigation measure.

In general, there are two main approaches commonly used for predicting heatwaves: (a) dynamic climate models and (b) statistical models. Dodla et al. (2017); Ahsan Khan et al. (2020); Naveena et al. (2021) employed the Weather Research and Forecasting (WRF) model, a type of dynamic climate model, to predict heatwaves up to 72 h in advance. Their study found that the root mean square error (RMSE) values ranged from 0.8 to 2.24 K. Similarly, Amna et al. (2013); Mandal et al. (2019) used a multi-model dynamical ensemble prediction system for heatwave prediction but found that the system's accuracy decreased for extreme forecast probabilities beyond a one-week lead time. Although dynamic climate models can capture complex interactions between the atmosphere, land, and ocean due to their physical basis, they are computationally demanding, requiring significant investment in data assimilation and a longer time for model building and parameterizations. Recently, a few studies have started to use machine learning (ML) algorithms for the same (Das and Nanduri 2018; Pandey et al. 2020, 2022). Khan et al. (2019) applied quantile regression forest (QRF) and random forest (RF) models in Pakistan, using synoptic climate variables to predict heatwaves at different time lags. The QRF model demonstrated accuracy in predicting heatwave triggering and departure dates for lead times of 1–10 days. Similarly, Khan et al. (2021) developed a climate change resilient heatwave prediction model using support vector regression (SVR), RF, and artificial neural network (ANN). The study found that SVR performed better than RF and ANN in predicting heatwave days and has the potential to provide accurate forecasting in the context of climate change (Sharma and Goyal 2017; Das and Nanduri 2018; Pandey and Md Azamathulla 2021; Singh et al. 2022). Further, in a study by Jacques-Dumas et al. (2022), a convolutional neural network was trained using 1000 years of climate model outputs to forecast extreme heatwave occurrences. The model demonstrated the ability to predict heatwaves at three different levels of intensity as early as 15 days ahead of the event. Asadollah et al. (2021) utilized decision tree (DT), random forest (RF), and AdaBoost regression and decision tree (ABR-DT) to predict heatwaves and found that ABR-DT outperformed the other models even when one or multiple variables were removed. Imran Khan and Maity (2022) used a combination of one-dimensional neural network (Conv 1D) and long short-term memory (LSTM) neural network and found that the model's efficiency decreased to 50% when predicting more than five days ahead. These machine learning algorithms are efficient in recognizing highly nonlinear relationships between predictors and predictands and have been used extensively to predict various climate variables, such as wind, evapotranspiration, and extreme events. Therefore, considering the benefits of using these ML algorithms, along with the limited number of studies performed using them, the present study aims to predict heatwaves using support vector regression (SVR) and random forest (RF) for the state of Telangana and evaluate the models' performance for the same.

5.2 Materials and Methods

5.2.1 Methodology

Two ML algorithms, namely, SVR and RF, are used in this study to predict heatwave days (HWDs) for the study area. HWDs are defined as the daily maximum temperature exceeding the 95th percentile of the maximum temperature for at least five consecutive days, using the maximum temperature of the base year as a reference (Khan et al. 2019). The study focused on the months of April, May, and June (AMJ) and analyzed the HWDs for the period between 1990 and 2019 in the study area. In order to predict heatwaves, this study uses five climatic variables, namely, geopotential height, U-wind, V-wind, air temperature, and relative humidity, for four pressure levels, that is, 1000, 925, 850, and 700 hPa, for the time period of 1990–2019, thus concluding with 20 input variables as predictors. The redundancy due to these many numbers of predictors is avoided by the use of the principal component analysis (PCA) that helps to discard the effect of multidimensionality. Figure 5.1 depicts the steps of the overall methodology used in this study.

5.2.1.1 Support Vector Regression (SVR)

The nonlinearity is addressed by the SVR by mapping them into a higher-dimensional space using kernel functions such as polynomial, radial, sigmoid, and linear (Manali Pal et al. 2020). The polynomial and radial kernels are commonly used in SVR-based prediction models. The mathematical representation of SVR

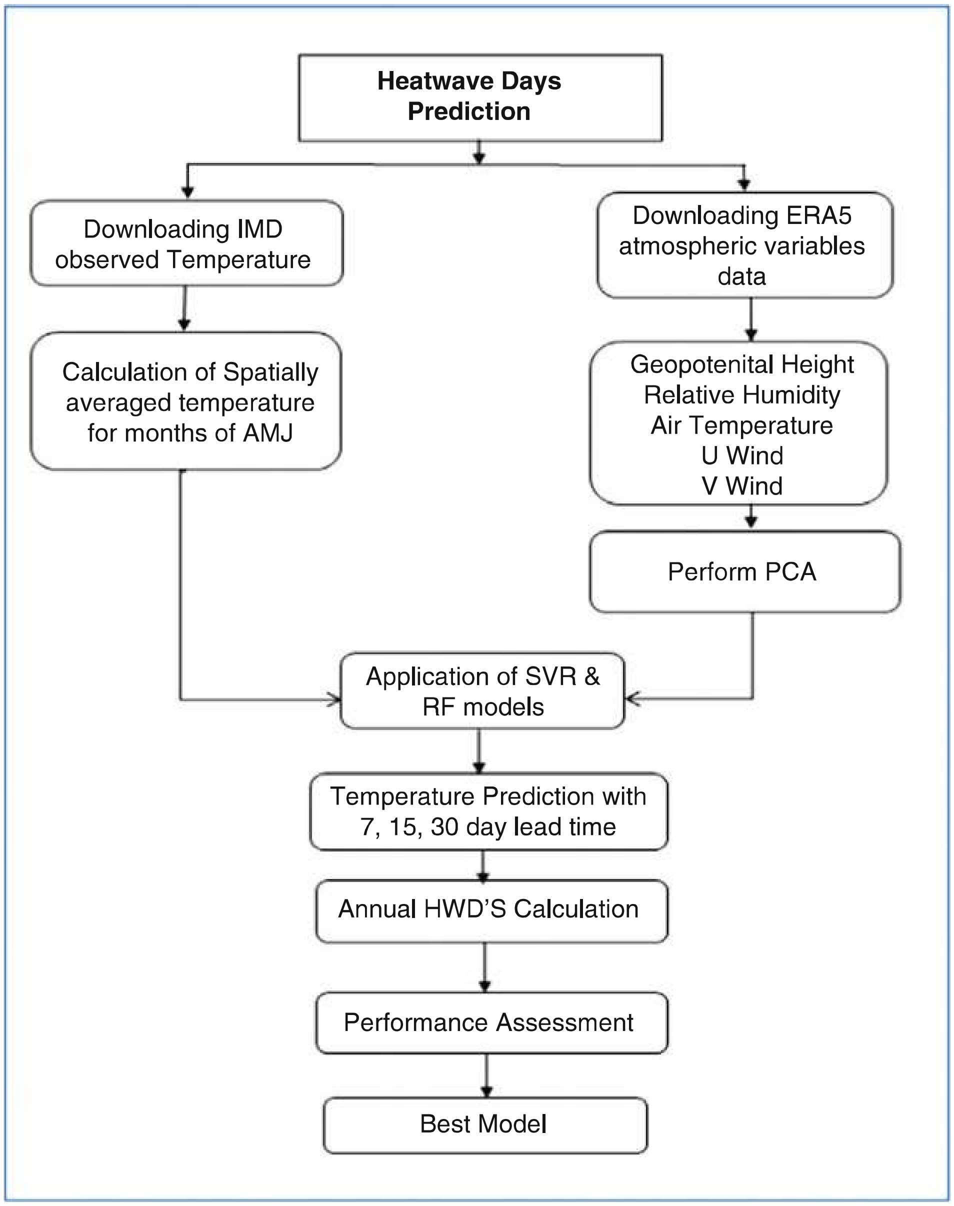
Heatwave Days Prediction
Downloading IMD observed Temperature
Downloading ERA5 atmospheric variables data
Calculation of Spatially averaged temperature for months of AMJ
Geopotenital Height
Relative Humidity
Air Temperature
U Wind
V Wind
Perform PCA
Application of SVR & RF models
Temperature Prediction with 7, 15, 30 day lead time
Annual HWD'S Calculation
Performance Assessment
Best Model

Fig. 5.1 The flowchart of overall methodology

involves finding a regression function, ($f(x) = (w, x) + b$), that describes the observed output y with an error tolerance of $\in$. Here, $[(x_1, y_1), (x_2, y_2). \ldots (x_i, y_i). \ldots (x_l, y_l)]$ represent a training dataset with x_i and y_i as input and output vectors, respectively, and l is the number of data pairs. To achieve this, the original input domain is mapped to a higher-dimensionality space where the function underlying the data is assumed to be linear. The transformed SVR problem in this space is solved by optimizing the following equation (Wang et al. 2007):

$$\text{Minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\varepsilon_i + \varepsilon_i^*\right)$$

$$\text{Subject to}\begin{cases} Y_i - \sum_{j=1}^{K}\sum_{i=1}^{L} w_j x_{ji} - b \leq \in + \varepsilon_i, \\ \sum_{j=1}^{K}\sum_{i=1}^{L} w_j x_{ji} - y_i \leq \in + \varepsilon_i^*, \\ \varepsilon_i, \varepsilon_i^* \geq 0, \end{cases}$$

C is the capacity parameter cost, which is a positive constant that determines the degree of penalized loss when a training error occurs to tune the trade-off between model complexity and tolerance to empirical errors; ε_i, ε_i^* are the slack variables, which measure the distance (in the target space) of the training samples lying outside the insensitive tube from the tube. The equation below represents the functional dependency:

$$f(x) = \sum_{j-1}^{K} w_j x_j + b$$

where K is the number of support vectors. The optimization problem is solved using the dual formulation subject to constraints in the loss function and introducing the Lagrange multipliers, α_{i} and $\alpha_{\mathrm{i}}{}^*$. By solving the optimization problem, the final prediction function is

$$f'(x) = \sum_{i \in N}\left(\alpha_{\mathrm{i}} - \alpha_{\mathrm{i}}^*\right)k(x_{\mathrm{i}}, x) + b$$

where $k(\ldots)$ is kernel function which computes nonlinear dependence between the two input variables x_i and x where x_i are the "support vectors" and b is the bias. In the present study, the radial basis function (RBF) kernel is used in the prediction of HWDs, and it can be mathematically represented with kernel width $-$ γ, as,

$$k(x_i, x) = \exp\left(-\gamma\|x - x_i\|^2\right), \gamma > 0$$

5.2.1.2 Random Forest (RF)

RF is a collection of learning methods that generates several decision trees collectively used to execute a classification or regression. It is a decision tree-based ML algorithm that consists of many individual decision trees that operates as an ensemble. In RF, many uncorrelated trees will outperform any of the individual trees; the low correlation being important as a better result can be achieved. There are two known ensemble methodologies, boosting and bagging (Asadollah et al. 2021). Bagging and boosting are two techniques that are often used in conjunction with RF to improve its performance.

Here is a summary of the steps taken in RF to construct a regression model: (1) bootstrapping samples from training data and random selection of m ($< p$) variables at each split, n number trees $T(\theta_t)$ are constructed, where θ_t denotes the parameter. It controls the growth of the t^{th} tree (2). The prediction is produced from the n^{th} single tree's average output as

$$f(x) = \widehat{E}(Y|X = x) = \sum_{i=1}^{n} \omega_i(x) Y_i \tag{5.1}$$

where

$$\omega_i(x) = \frac{\sum_{t=1}^{n\text{tree}} \omega_i(k, \theta_t)}{n_{\text{tree}}}, \omega_i(x, \theta) = \frac{1(x_i \in R_l(k, \theta))}{\left(j : x_{j \in R_1}(x, \theta)\right)} \tag{5.2}$$

The random selection of m ($<p$) predictors at each split is a key concept of RF, which provides an enhancement over bagging (Khan et al. 2019).

5.2.1.3 The Performance Metrics

The accuracies of the predictions of the models are assessed by the root mean square error (RMSE), mean square error (MSE), and Pearson's coefficient of correlation (R), between the observed and the predicted values. The mathematical expressions of the metrics are provided as follows.

The RMSE is computed by the following equation:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \widehat{x_i})^2}$$

where N is the number of observations, x_i is actual value, and $\widehat{x_i}$ is predicted value.

MSE is the average squared difference between the predicted values and the actual values in a dataset. MSE is calculated by taking the difference between the predicted value and the actual value for each data point, squaring the difference, and then taking the mean of these squared differences. A smaller MSE indicates a better fit of the model to the data, and it can be computed by the following equation:

$$\text{MSE} = \sum \frac{(y_i - \ldots y)^2}{n}$$

where y_i = predicted value and $\ldots y$ = actual value.

The Pearson's correlation coefficient (R) is a measure of the strength and direction of a linear relationship between two variables. It is commonly used to assess the degree to which one variable is related to another and can be used to identify trends or patterns in data. The values of coefficient of correlation vary between -1 and 1, where $-1/1$ indicates a perfect negative/positive correlation, and a value of 0 indicates no correlation. Coefficient of correlation (R) is given by below equation:

$$R = \frac{\sum_{i=1}^{N} (y_i - \overline{y}) \left(\widehat{y_i} - \widehat{\overline{y})}\right.}{\sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} (y_i - \overline{y})^2 \left(\widehat{y_i} - \overline{\widehat{y}}\right)^2}}$$

where y_i is the actual value, $\widehat{y}$ is the predicted value of y, and $\overline{y}$ and $\overline{\widehat{y}}$ are the mean values of the actual and predicted values, respectively.

5.2.2 Study Area and Data Source

5.2.2.1 Telangana

Telangana, a state located in the southern part of India, is known for its hot and dry climate and state stretches between 15°46′N to 19°47′N latitude and 77°16′E to 81° 43′E longitude. Telangana is one of the regions in India that is highly susceptible to heatwaves due to its geographical location, which makes it prone to the scorching heat of the sun. The state has been witnessing an increase in the frequency and intensity of heatwaves in recent years, which is a cause for concern. According to the India Meteorological Department (IMD), in 2019, the state witnessed over 200 heatwave days, which is significantly higher than the previous years (Fig. 5.2).

5.2.2.2 Data Collection

The study utilizes the Indian Meteorological Department's (IMD) data, that is, maximum daily temperature during summer (April, May, June) from 1990 to 2019. The data has a spatial resolution of $1^{\mathrm{o}} \times 1^{\mathrm{o}}$. To predict the temperature, the study employs a daily time series of five meteorological variables – geopotential height, U-wind, V-wind, air temperature, and relative humidity. These variables were extracted from the European Centre for Medium-Range Weather Forecasts Reanalysis Version 5 (ERA5) for four pressure levels – 1000, 925, 850, and

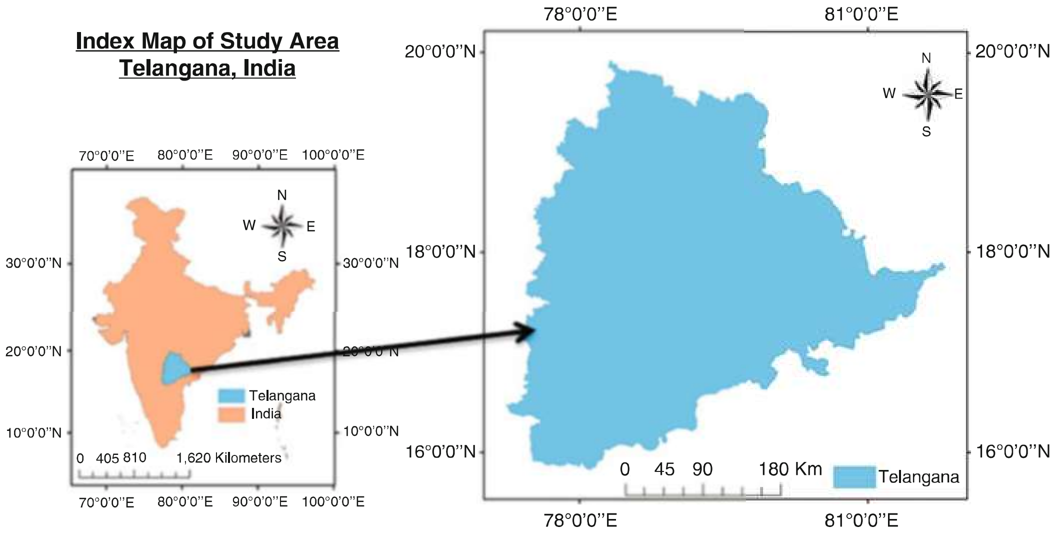

Fig. 5.2 The study area

Table 5.1 Atmospheric variables used for the selection of heatwave predictors

Atmospheric variables	Symbol	Pressure level (hPa)
Air temperature	Air	1000, 925, 850, and 700
Geopotential height	hgt	1000, 925, 850, and 700
Relative humidity	rhum	1000, 925, 850, and 700
U-wind	uwnd	1000, 925, 850, and 700
V-wind	vwnd	1000, 925, 850, and 700

700 hPa – for the same time period of 1990–2019. ERA5 provides a comprehensive record of the global atmosphere, land surface, and ocean waves since 1950 (Table 5.1).

5.3 Results

5.3.1 SVR Model Development

The SVR model developed by using the 21 years of data, that is, from a period of 1990 to 2010 (training period), and remaining 9 years of data, that is, from 2011 to 2019, is used for testing the models. The model is fitted without loss of the generality by selecting the optimal kernel function, that is, the Gaussian kernel with optimal parameter. It is tuned generally with three parameters, that is, box constraint, kernel scale, and epsilon. The epsilon values represent the error tolerance, and the box constraints are positive numeric values that aid to prevent overfitting. Kernel scale represents the width of kernel. Estimated box constraint, kernel scale, and epsilon values for the developed model are 9.81, 3.36, and 0.0043, respectively, for 30-day lead time; 11.23, 1.71, and 0.21 for 15-day lead time; and 9.08, 3.43, and 0.93 for 7-day lead time.

5.3.2 RF Model Development

The RF model is constructed using 21 years of data spanning from 1990 to 2010 as the training period. The model was then tested using the remaining nine years of data, covering the period from 2011 to 2019. Random forest is generally defined by number of decision trees that model can generate, maximum number of splits that can be made in each decision tree, and number of variables that are randomly sampled for each split in the decision tree. To get the optimum combination of above three parameters, the RF has optimized hyperparameters which randomly selects all three parameters and establish an objective function. The parameters with minimum error in objective function are declared as optimum parameters, and model will be developed using those parameters. The parameters that are optimized in model are as follows: The number of variables to sample is the number of predictors to select at random for each split, which is taken as 2 for regression. Maximum number of splits is maximum number of decision splits (or branch nodes) per tree, and default value is considered as number of observations – 1. It is useful for controlling the complexity of the trees in the model. The number of learning cycles is the number of times that each tree in the model will be trained on the data which is taken as 100 for fitting the model.

5.3.3 Prediction of Maximum Temperature and Annual HWDs

This study aims to evaluate the predictive capabilities of SVR and RF models for predicting heatwaves in Telangana for the months of April, May, and June with different lead times of 7, 15, and 30 days. Spatially averaged temperature time series were generated using observed temperature data from the IMD, which are then predicted using SVR and RF models with the abovementioned lead times. Figures displaying time series plots of both observed (IMD) and predicted temperatures with SVR and RF during the training and testing periods are presented below (Figs. 5.3, 5.4, 5.5, 5.6, 5.7, and 5.8):

According to observed IMD data, HWDs were observed in the years of 1995, 1998, 2003, 2005, 2010, 2012, and 2015. The SVR and RF models, on the other hand, can only capture HWDs in the years 1995, 1998, 2003, and 2015. The year 2015 has the maximum numbers of HWDs (13 days) for the state calculated from IMD observed temperature, during the months of April, May, and June. A number of spatially averaged HWDs predicted by SVR and RF for the same year and time

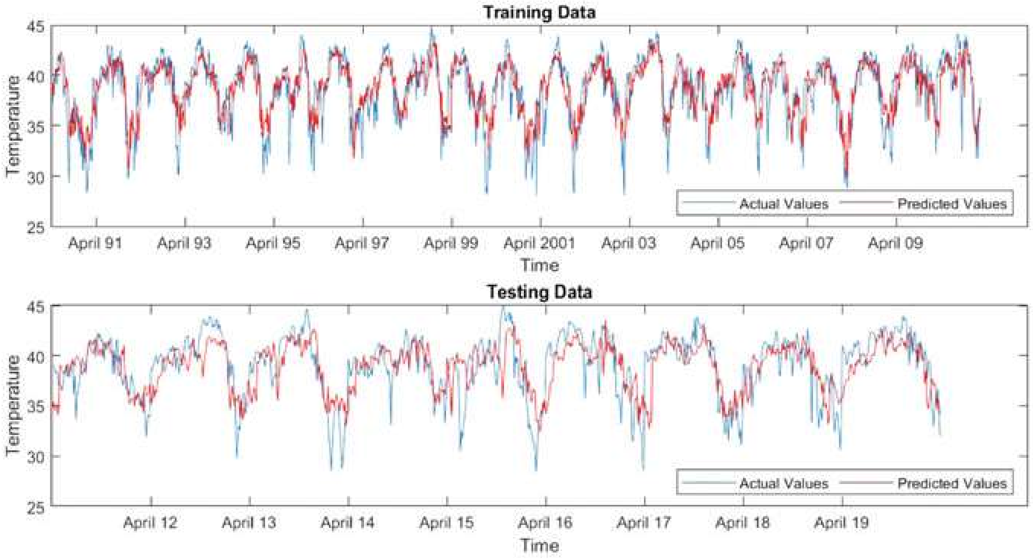

Fig. 5.3 Time series plot of observed and predicted temperature by SVR with seven-day lead time

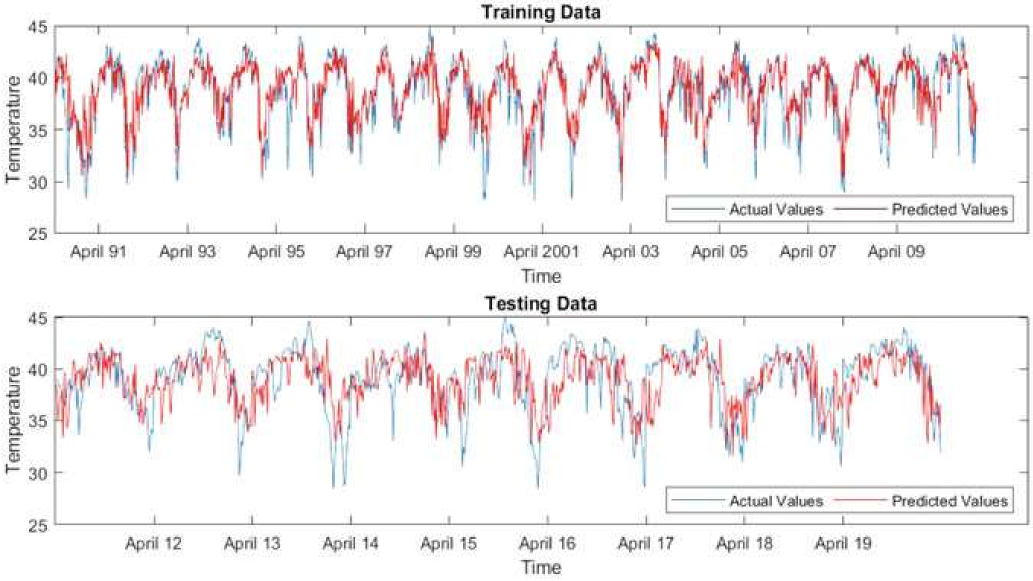

Fig. 5.4 Time series plot of observed and predicted temperature by SVR with 15-day lead time

period (AMJ) are 6 and 8 for seven-day lead time, respectively. The comparison between the numbers of observed and predicted HWDs from both the models has been presented in Fig. 5.9.

To evaluate the performance of SVR and RF models in predicting temperature, three performance metrics – correlation coefficient (R), root mean square error (RMSE), and mean square error (MSE) – are used. During the training periods, the ranges of R for all three lead times are found to be 0.64–0.78 and 0.88–0.89 for SVR and RF, respectively. During model testing periods, the same range is from 0.5–0.67 for SVR and 0.49–0.66 for RF. The ranges of RMSE for all three lead times

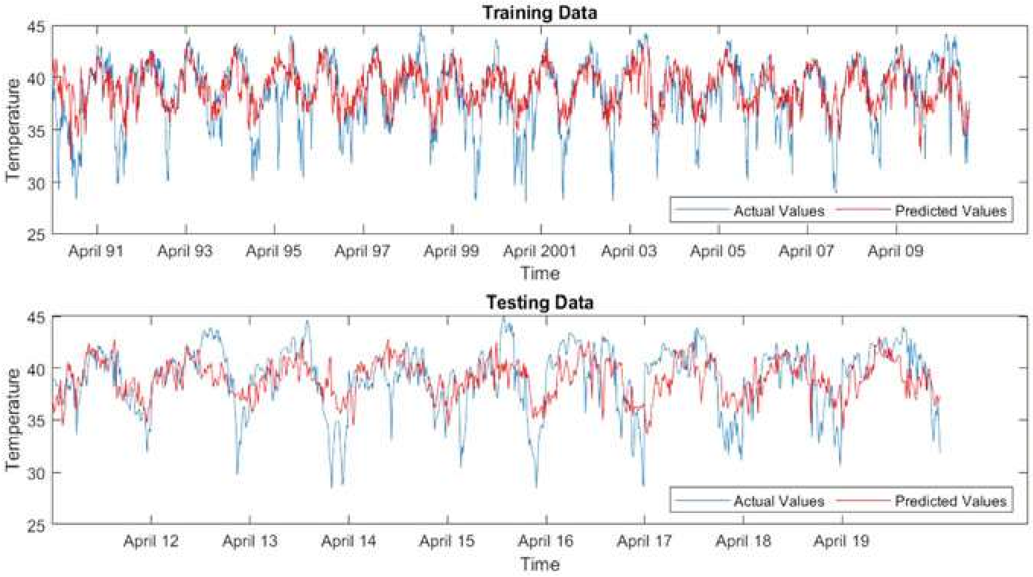

Fig. 5.5 Time series plot of observed and predicted temperature by SVR with 30-day lead time

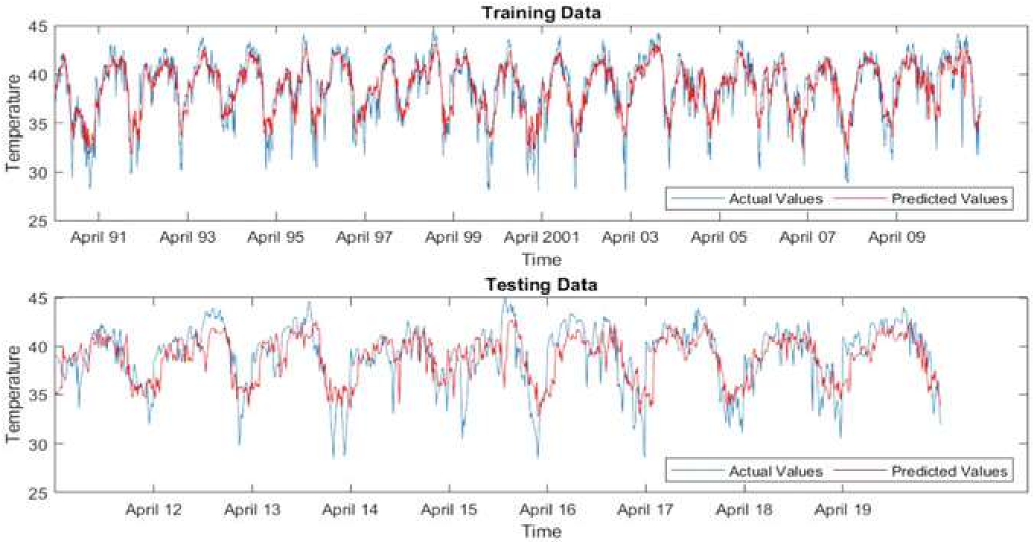

Fig. 5.6 Time series plot of observed and predicted temperature by RF with seven-day lead time

during training periods are 2.01–2.53 and 1.55–1.70 for SVR and RF, respectively. During model testing periods, the same are 2.35–2.70 for SVR and 2.37–2.78 for RF. It is observed that the RMSE values changed in the range of 2.7 °C and 2.78 °C for SVR and RF, respectively, for lead times 7–30 days. The average values of performance metrics for both models are presented in Tables 5.2 and 5.3. It is observed that the model performance decreases with the increase in lead time for both SVR and RF. While comparing between the two models, it is observed that the

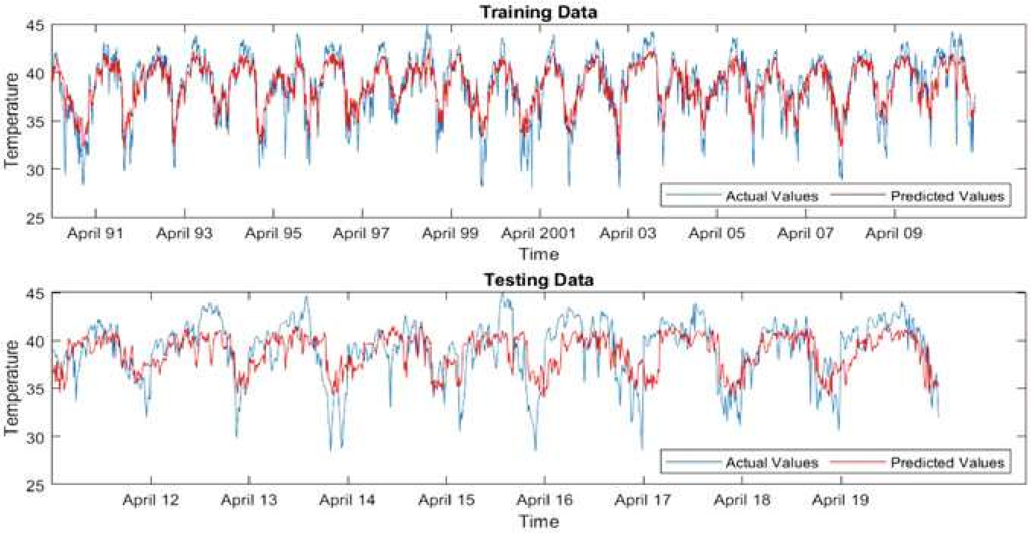

Fig. 5.7 Time series plot of observed and predicted temperature by RF with 15-day lead time

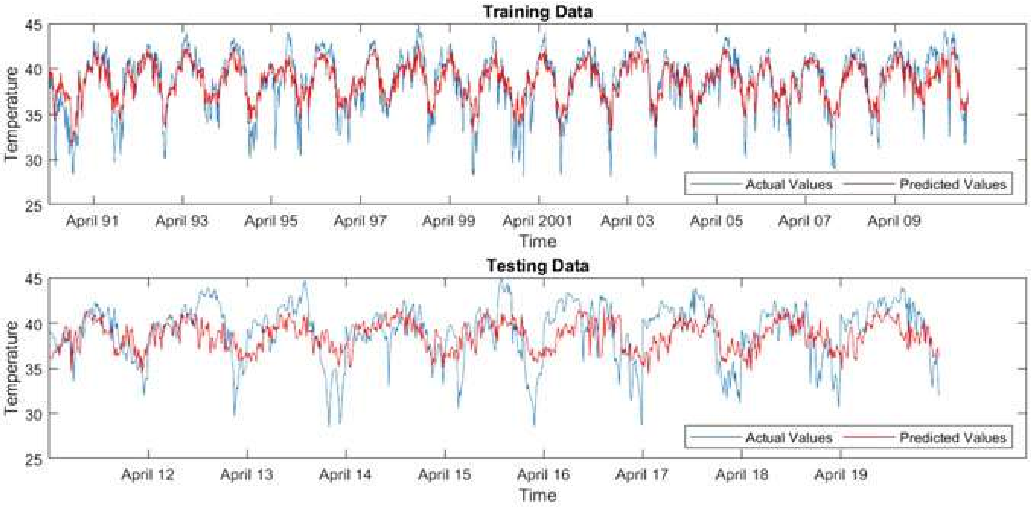

Fig. 5.8 Time series plot of observed and predicted temperature by RF with 30-day lead time

RF tends to overfit as the lead time increases whereas although the performance of SVR decreases, it does not show model overfitting.

Figure 5.10 shows the spatial distribution of observed and predicted temperature from SVR and RF, respectively, for seven-day lead times for the date June 4, 1995, that had experienced the heatwave (according to the definition mentioned above). It shows that both the models are not able to capture the highest observed temperature, that is, 41 °C.

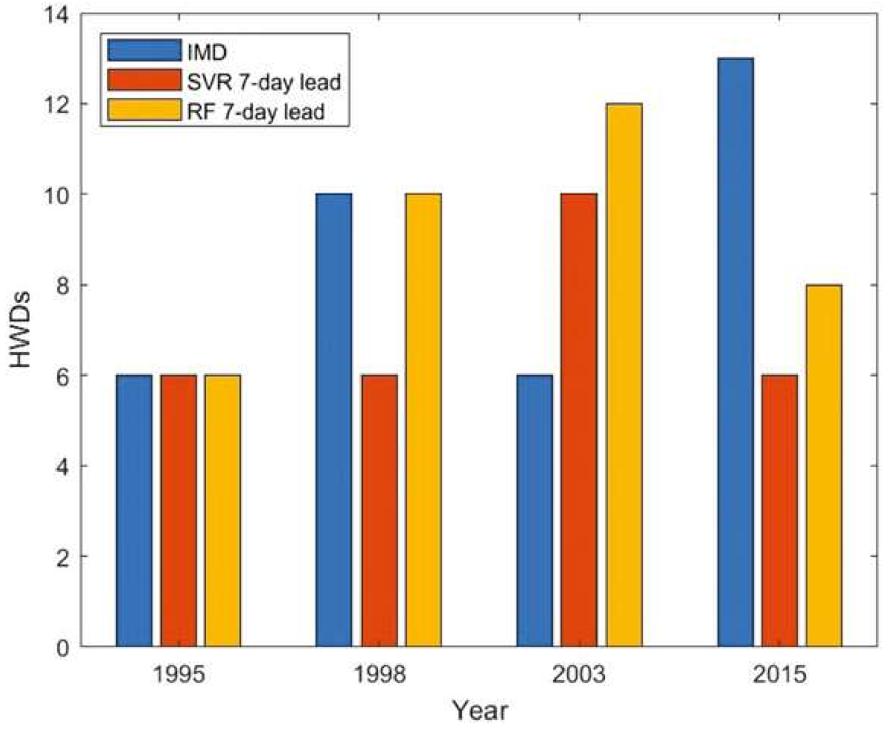

Fig. 5.9 The comparison between the numbers of observed and predicted HWDs from SVR and RF

Table 5.2 Performance metrics of SVR model for the lead times of 7, 15, and 30 days

	CC train	CC test	RMSE train	RMSE test	MSE train	MSE test
7	0.78	0.67	2.02	2.35	4.11	5.50
15	0.78	0.53	2.02	2.73	4.09	7.48
30	0.57	0.50	2.68	2.74	7.21	7.51

Table 5.3 Performance metrics of RF model for the lead times of 7, 15, and 30 days

Lead	CC train	CC test	RMSE train	RMSE test	MSE train	MSE test
7	0.88	0.66	1.55	2.37	2.43	5.62
15	0.89	0.58	1.59	2.58	2.54	6.69
30	0.89	0.49	1.71	2.78	2.91	7.76

5.4 Discussion

In this study, SVR and RF models are employed to predict the maximum temperature and forecast the number of HWDs in Telangana during summer (April to June). These two machine learning techniques have proven to be highly effective in solving a range of intricate problems. The outcomes have demonstrated the ML models'

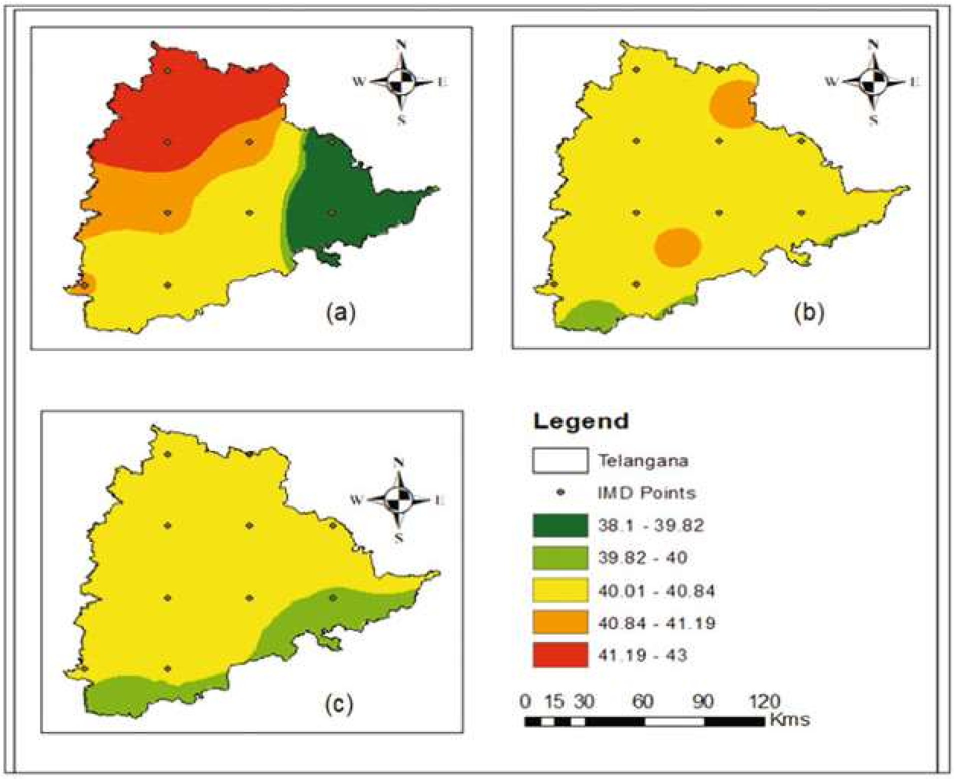

Fig. 5.10 The spatial distribution of (**a**) IMD observed, (**b**) SVR, and (**c**) RF predicted temperature for seven-day lead time on June 4, 1995

capability to predict the maximum temperature and fluctuation of HWDs from year to year. Both the models are able to predict maximum temperature and HWDs accurately up to a lead time of seven days. However, beyond that, the models' performance decreases, that is, it is not able to accurately predict the temperature or HWDs. The evaluation comparing both the models is done by using three performance metrics – correlation coefficient (R), root mean square error (RMSE), and mean square error (MSE) – and these performance metrics demonstrated that SVR possesses a high degree of accuracy in predicting temperature and HWDs compared to RF.

5.5 Conclusions

In this study, the heatwave prediction capabilities of SVR and RF are investigated for Telangana, spanning from 1990 to 2019. The study aims to assess the models' performances for lead times of 7, 15, and 30 days during April, May, and June. The observed temperature data from the IMD are utilized to calculate a time series of the average cumulative annual heatwave days. The same was predicted for SVR and RF models with lead times of 7, 15, and 30 days. The SVR model is able to capture the five heatwave events out of seven whereas the RF model is able to capture four heatwave events. The results of the study suggest that SVR and RF machine learning algorithms can be used to predict heatwave days for Telangana by employing meteorological variables such as air temperature, relative humidity, geopotential height, u-wind, and v-wind as predictors. However, performance metrics indicate that the models' performance decreases with increasing lead times. The study found that both SVR and RF models demonstrated acceptable performances for up to seven days of lead time, with RMSE values of 2.36 °C and 2.37 °C for SVR and RF, respectively, during testing periods. In comparing the performance of the two models, SVR is found to be more efficient for this specific study.

5.6 Future Scope

The limitations of the current study can lead to future scope in terms of enhancing the model. For example, the model development has not considered the temporal variation of spatial correlation between the predictors and predictand along with the influences of teleconnections and that can be incorporated in future studies. These gaps in the study indicate potential areas for future research to address the abovementioned limitations. Additional research can be undertaken to compare the predictive capabilities of SVR and RF with other machine learning models when it comes to forecasting heatwaves. The efficacies of SVR, RF, and other statistical models can be contrasted with that of dynamical models to highlight their respective strengths and weaknesses.

References

Afroz M, Chen G, Anandhi A (2022) Drought- and heatwave associated compound extremes: A review of hotspots, variables, parameters, drivers, impacts, and analysis frameworks. Front Earth Sci 10:1–25. https://doi.org/10.3389/feart.2022.914437

Ahsan Khan N, Sajadul Alam Saimon M, Naqib Jimmy A, Fatema Lubna K, Abul Kalam Mallik M, Sajadul Alam M, Islam T, Ahmad I, Fatema K, Analyst G, Science E, Author A (2020) Study on heat wave and its thermodynamic features over bangladesh using numerical weather prediction model (NWPM). Int J Sci Bus. https://doi.org/10.5281/zenodo.3839997

Amna S, Samreen N, Khalid B, Shamim A (2013) Numerical climate modeling and verification of selected areas for heat waves of Pakistan using ensemble prediction system. J Phys Conf Ser 439(1). https://doi.org/10.1088/1742-6596/439/1/012041

Asadollah S, Khan N, Sharafati AS, Chung S, Wang ES, Xiao J (2021) Prediction of heat waves using meteorological variables in diverse regions of Iran with advanced machine learning models. Stoch Env Res Risk A 7(36):1959–1974

Basha G, Kishore P, Ratnam MV, Jayaraman A, Kouchak AA, Ouarda TBMJ, Velicogna I (2017) Historical and projected surface temperature over India during the 20th and 21st century. Sci Rep 7(1). https://doi.org/10.1038/s41598-017-02130-3. Nature Publishing Group

Christidis N, Jones GS, Stott PA (2015) Dramatically increasing chance of extremely hot summers since the 2003 European heatwave. Nat Clim Chang 5:46–50

Coumou D, Rahmstorf S (2012) A decade of weather extremes. Nat Clim Chang 2(7):491–496

Das J, Nanduri UV (2018) Assessment and evaluation of potential climate change impact on monsoon flows using machine learning technique over Wainganga River basin, India. Hydrol Sci J 63:1020–1046. https://doi.org/10.1080/02626667.2018.1469757

Das J, Umamahesh NV (2022) Heat wave magnitude over India under changing climate: Projections from CMIP5 and CMIP6 experiments. Int J Climatol 42:331–351. https://doi.org/10.1002/joc.7246

Das PK, Podder U, Das R, Kamalakannan C, Rao GS, Bandyopadhyay S, Raj U (2020) Quantification of heat wave occurrences over the Indian region using long-term (1979–2017) daily gridded (0.5° × 0.5°) temperature data—a combined heat wave index approach. Theor Appl Climatol 142(1–2):497–511. https://doi.org/10.1007/s00704-020-03329-7

Das J, Manikanta V, Umamahesh NV (2022) Population exposure to compound extreme events in India under different emission and population scenarios. Sci Total Environ 806. https://doi.org/10.1016/j.scitotenv.2021.150424

De US, Mukhopadhyay RK (1998) Severe heat wave over Indian subcontinent in 1998 in a perspective of global climate. Curr Sci 75:1308–1311

De Perez EC, Van Aalst M, Bischiniotis K, Mason S, Nissan H, Pappenberger F, Stephens E, Zsoter E, Van Den Hurk B (2018) Global predictability of temperature extremes. Environ Res Lett 13(5). https://doi.org/10.1088/1748-9326/aab94a

Dodla VB, Satyanarayana GC, Desamsetti S (2017) Analysis and prediction of a catastrophic Indian coastal heat wave of 2015. Nat Hazards 87(1):395–414. https://doi.org/10.1007/s11069-017-2769-7

Dole R, Hoerling M, Perlwitz J, Eischeid J, Pegion P, Zhang T, Quan XW, Xu T, Murray D (2011) Was there a basis for anticipating the 2010 Russian heat wave? Geophys Res Lett 38(6). https://doi.org/10.1029/2010GL046582

Guha-Sapir, D. & Below, R. Hoyois Ph. EM-DAT: The international Disaster database-www.emdat.be – Université Catholique de Louvain – Brussels – Belgium. (Accessed 17th February 2016)

Imran Khan M, Maity R (2022) Hyrbid deep learning approach for multi-step-ahead prediction for daily maximum temperature and heatwaves. Theor Appl Climatol 149:945–963. https://doi.org/10/1007/s00704-022-04103-7

Jacques-Dumas V, Ragone F, Borgnat P, Abry P, Bouchet F (2022) Deep learning-based extreme heatwave forecast. Front Climate 4:1–20. https://doi.org/10.3389/fclim.2022.789641

Jenamani RK (2012) Analysis of Ocean-Atmospheric features associated with extreme temperature variations over east coast of India- A special emphasis to Orissa heat waves of 1998 and 2005. Mausam 63:401–422

Khan N, Shahid S, Juneng L, Ahmed K, Ismail T, Nawaz N (2019) Prediction of heat waves in Pakistan using quantile regression forests. Atmos Res 221:1–11

Khan N, Shahid S, Ismail TB, Behlil F (2021) Prediction of heat waves over Pakistan using support vector machine algorithm in the context of climate change. Stoch Env Res Risk A 35(7): 1335–1353. https://doi.org/10.1007/s00477-020-01963-1

Mandal R, Joseph S, Sahai AK, Phani R, Dey A, Chattopadhyay R, Pattanaik DR (2019) Real time extended range prediction of heat waves over India. Sci Rep 9(1). https://doi.org/10.1038/s41598-019-45430-6

McMichael AJ, Lindgren E (2011) Climate change: present and future risks to health, and necessary responses. J Intern Med 270(5):401–413

Meehl GA, Tebaldi C, Walton G, Easterling D, McDaniel L (2009) Relative increase of record high maximum temperatures compared to record low minimum temperatures in the US. Geophys Res Lett. 36(23). https://doi.org/10.1029/2009GL040736

Mishra V, Ganguly AR, Nijssen B, Lettenmaier DP (2015) Changes in observed climate extremes in global urban areas. Environ Res Lett 10(2):024005. https://doi.org/10.1088/1748-9326/10/2/024005

Naveena N, Satyanarayana GC, Raju AD, Umakanth N, Srinivas D, Rao KS, Suman M (2021) Prediction of heatwave 2013 over Andhra Pradesh and Telangana, India using WRF model. Asian J Atmos Environ 15(3):1–12

Pai DS, Nair A, Ramanathan AN (2013) Long term climatology and trends of heat waves over India during the recent 50 years (1961-2010). MAUSAM. 64(4). https://doi.org/10.54302/mausam.v64i4.742

Manali Pal, Rajib Maity, Ratnam, J.V.,Masami Nonaka, & Swadin,K.B.(2020). Long-lead prediction of ENSO Modoki index using machine learning algorithms. Sci Rep, DOI: https://doi.org/10.1038/s41598-019-57183-3

Pandey M, Md Azamathulla H (2021) Discussion of "Gene-Expression Programming, Evolutionary Polynomial Regression, and Model Tree to Evaluate Local Scour Depth at Culvert Outlets" by Mohammad Najafzadeh and Ali Reza Kargar. J Pipeline Syst Eng Pract 12:07021001. https://doi.org/10.1061/(asce)ps.1949-1204.0000532

Pandey M, Zakwan M, Sharma PK, Ahmad Z (2020) Multiple linear regression and genetic algorithm approaches to predict temporal scour depth near circular pier in non-cohesive sediment. ISH J Hydraul Eng 26:96–103. https://doi.org/10.1080/09715010.2018.1457455

Pandey M, Jamei M, Ahmadianfar I et al (2022) Assessment of scouring around spur dike in cohesive sediment mixtures: A comparative study on three rigorous machine learning models. J Hydrol 606:127330. https://doi.org/10.1016/j.jhydrol.2021.127330

Rohini P, Rajeevan M, Srivastava AK (2016) On the variability and increasing trends of heat waves over India. Sci Rep 6:26153. https://doi.org/10.1038/srep26153

Sharma A, Goyal MK (2017) A comparison of three soft computing techniques, Bayesian regression, support vector regression, and wavelet regression, for monthly rainfall forecast. J Intell Syst 26:641–655. https://doi.org/10.1515/jisys-2016-0065

Singh UK, Jamei M, Karbasi M et al (2022) Application of a modern multi-level ensemble approach for the estimation of critical shear stress in cohesive sediment mixture. J Hydrol 607:127549. https://doi.org/10.1016/j.jhydrol.2022.127549

Wang W, Men C, Weizhen L (2007) Online prediction model based on support vector machine. Neurocomputing 71:550–558

Zhang Y, Yu C, Bao J, Li X (2017) Impact of temperature on mortality in Hubei, China: a multi-county time series analysis. Sci Rep 7:45093. https://doi.org/10.1038/srep4509