Article title: Analyzing trend and prediction of precipitation in

Germany using non-parametrical
tests and machine

learning algorithms

Authors: Minhaz Rahman Talukder[1]

Affiliations: Environmental and Resource Management, Brandenburg University of Technology, Cottbus, Germany[1]

Orcid ids: 0000-0002-6001-1866[1]

Contact e-mail: talukder.minhaz@outlook.de

License information: This work has been published open access under Creative Commons Attribution License <http://creativecommons.org/licenses/by/4.0/>, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Conditions, terms of use and publishing policy can be found at <https://www.scienceopen.com/>.

Preprint statement: This article is a preprint and has not been peer-reviewed, under consideration and submitted to ScienceOpen Preprints for open peer review.

Links to data: <https://github.com/talukmdm/Trend-analysis-and-prediction-of-precipitation>

DOI: 10.14293/S2199-1006.1.SOR-.PPCPS00.v1

Preprint first posted online: 21 November 2021

Keywords: precipitation, non-parametrical test, machine learning

Analyzing trend and prediction of precipitation in Germany using non-parametrical tests and machine learning algorithms

Md Minhaz Rahman Talukder — Matriculation Number 3814344¹

¹Student, Environmental and Resource Management, Brandenburg University of Technology, Cottbus, 03046, Germany

ABSTRACT

The study explores analyzes the temporal changes in precipitation using the data from 1881 to 2020 across Germany at regional level. Man-Kendall and Hamad-Rao modification test were employed to analyze the precipitation trend ,while Pettit test was used for detecting the change point in the time frame. Machine learning methods like k-nearest neighbor, Support vector machine and Random forest algorithms were applied for prediction. Most of the regions showed increasing trend annually and seasonally in 0.05 significance level while some negative can be seen in summer. Furthermore, Based on Pettit test most of change points detected after 1940 in several regions. In prediction of precipitation, k-NN algorithm showed better performance in terms of mean absolute error rather than Support vector machine and Random forest algorithms.

trend analysis, non-parametrical test, change point analysis, machine learning, k-nn, support vector machine, random forests

Introduction

Precipitation is a vital part of the hydrologic cycle and perhaps the most influential element in defining climate [1; 2]. Agriculture, forestry, the energy industry, and natural heritages can be directly affected by the alteration of its pattern [3]. Precipitation pattern variation resulting from climate change concerns water resource managers, climatologists, and hydrologists [1]. Srivista et al. and Islam et al. explained that the variation of precipitation quantities and frequencies directly affect the pattern of streamflow and its demand, spatiotemporal allocation of runoff, groundwater reserves, and soil moisture [4] [5]. Consequently, due to the extreme precipitation pattern changes, the probability of drought and flood-like hazards increases[1]. The drought event in 2003 across Europe had a significant impact on various sectors, and reduced water availability led to a vast economic loss on the order of 1.5 Billion Euros [6]. Therefore, extensive research on climate change or, more particularly, on precipitation patterns and frequencies is very much needed. Most importantly, a better understanding of the precipitation pattern in the changing environment will assist better decision-making and facilitate the communities' adapting capacity to sustain extreme weather events[1].

Several scholars asserted that the recurrence and intensity of the precipitation increase due to global warming and are also expected to increase further [7] [8] [9]. Various studies of trend analysis of extreme rainfall across Germany provide concrete evidence for significant changes in the frequency and intensity of precipitation [10] [11] [12]. Trend analysis for central-eastern Germany showed increasing precipitation in Winter and a decreasing pattern in summer [11]. Therefore, the research on trend analysis of precipitation could be advantageous since it provides essential information about the probability of future changes. Parametric and Non-parametric tests [13], like a regression test [14] [15], Man-Kandell test[16] [17], Kendall rank correlation test [18], and Spearmen rank correlation test [19], are the most widely used methods for trend analysis by researchers. In the present study, the Man-Kandell test was used to track down the precipitation pattern trend since it can be used on independent time series data and not sensitive to outliers [20]. Hensel et al. and Schonweise et al. also used the Man-Kandell test to detect Central-eastern Germany and Europe's Precipitation trend, respectively [11] [12].

Although trend analysis helps climatologists, meteorologists, and hydrologists understand present and past climate changes, the forecasting model tends to be more useful for planners to administer good planning considering upcoming climate variables [21]. Only two possible approaches exist in the literature for precipitation forecasting[22] [23] . The physical-based approach based on physical laws of precipitation is modeled by studying the precipitation process. When the prediction is generated at

smaller than 2-3 km scales, a computational problem can occur in the physical model [24]. The Physical models are both time and money-consuming, although they provide a reliable result. Due to high spatial variability and noise in the precipitation data, pattern recognition methods have become more popular among researchers for prediction [25]. Furthermore, Previous studies suggest that pattern recognition methodology based on machine learning techniques positively results in precipitation forecasting [26] [27] [28]. The motivation for applying machine learning methods for prediction comes from the increasing volume and availability of data. Nowadays, a vast amount of literature can be found where artificial intelligence like machine learning methods has been used for prediction since it does not need an enormous amount of information and can handle large and complicated datasets [29]. Researchers have been spending much effort developing precipitation forecasts using numerous machine learning techniques that yield good forecasts.

Huang et al. developed a novel approach for precipitation forecast by an improved K-nearest neighbor algorithm, which offers robustness against different choices of the neighborhood size k, particularly in the case of the irregular class distribution of the precipitation dataset [3]. Hong tried to implement a support vector machine (SVM) in rainfall forecasting in Northern Taiwan, and the empirical results revealed that this approach yields well-forecasting performance [26]. Pham et al. used random forest methods to forecast short-term daily streamflow in rainfall and snowmelt-driven watersheds [30]. There is a minimal implementation of the machine learning approach in precipitation forecasting in Germany. In the present study, I tried to implement k-NN, SVM, and RF algorithms to predict the precipitation in April, August, and December every year.

Therefore, based on previous literature and research opportunities, the present study's objectives are 1) variability and trend analysis in the overall annual and seasonal precipitation in German regions. 2) Change point analysis by non parametric Pettitt test 3) predicting precipitation using k-NN, SVM, and RF algorithms

Data source and method description

The geographical location of Germany is in Western and Central Europe, with sixteen states. Although the whole country experiences a moderate climate, it has four seasons like summer (June- August), Fall(September – November), Winter (December – February), and Spring (March-May). The data here used historical monthly regional averages for Germany and its federal states from 1881 to 2020, available on the DWD website[31].The regional averages result from the averaging of the respective areas of the 1km gridded data.

Mann-Kendall trend test [16] based on the ranks and sequence of a time series correlation is nonparametric test that works for all types of distribution. For a given time series $X_i, i=1,2,\dots,n$, the null hypothesis for this test is that the data is independently distributed, and the alternate hypothesis is that a trend exists. The test statistics S is can be calculated by

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn} (M_j - M_i) \quad (1)$$

Where, M_j and M_i are the values of the sequence i, j n is the length of the time series and

$$\text{sgn} (M_j - M_i) = \begin{pmatrix} 1 & \text{if } (M_j - M_i) > 0 \\ 0 & \text{if } (M_j - M_i) = 0 \\ -1 & \text{if } (M_j - M_i) < 0 \end{pmatrix} \quad (2)$$

$$\text{var} (s) = \frac{n(n-1)(2n+5) - \sum_{y=1}^x t_y(t_y-1)(2t_y+5)}{18} \quad (3)$$

The Z_{mk} Value is used to find out the time-series information is demonstrating a significant trend or not. The Z_{mk} value is computed using equation 4

$$Z_{mk} = \begin{pmatrix} \frac{S-1}{\sqrt{\text{var}(S)}} & \text{if } S > 0 \\ 0 & \text{if } S < 0 \\ \frac{S+1}{\sqrt{\text{var}(S)}} & \text{if } S < 0 \end{pmatrix} \quad (4)$$

It is very common and effective to use Pettitt [32] test to figure out a single change point in climate time series data It tests the H_0 : The T variables follow one or more distributions that have the same location parameter (nochange), against the alternative:

a change point exists. The non-parametric statistic is defined as:

$$K_T = \max |U_{t,T}| \quad (5)$$

where

$$U_{t,T} = \sum_{i=1}^t \sum_{j=t+1}^T \text{sgn}(X_i - X_j) \quad (6)$$

The change-point of the series is located at K_T , provided that the statistic is significant [32]. The significance probability of K_T is approximated for $p \leq 0.05$ with

$$p \simeq 2 \exp \left(\frac{-6K_T^2}{T^3 + T^2} \right) \quad (7)$$

Prediction methodologies

The supervised machine learning approach's primary goal is to predict some output variable associated with each input item. In this case, given the precipitation data for the last three months (input), predict the precipitation in the next consecutive month (output). Input and output data are then divided into two sets, i.e., the Training set is used to construct the model, and the test set is to verify the model built. Data in training set excluded from the test set. The ratio for splitting training and testing data used here is 80/20. I used mean absolute error (MAE) to test the model. Figure 1 represents the total workflow for the prediction methodology.

I have used three different algorithms, e.g., k- Nearest Neighbor, Support Vector Machine, and Random Forest algorithms,

Figure 1. Prediction methodologies work flow

which are described below

k-NN (k-nearest neighbor) algorithm is a classifier algorithm that classifies data points based on most similar to it. If training set, $T = \{x_n \in R\}_{n=1}^d$ M classes comprise N training samples in d-dimensional feature space, and the class label of one sample x_n is c_n [3]. If x is a query point, the KNN algorithm works as follows

Find k nearest neighbors from the set T for the unknown data point x , and let $\bar{T} = \{(x_i^{NN}, c_i^{NN})\}_{i=1}^k$ impart the set of k nearest neighbors for x . The distance between x and the neighbor x_i^{NN} calculated by Euclidean distance metric

$$d(x, x_i^{NN}) = \sqrt{(x - x_i^{NN})^T (x - x_i^{NN})} \quad (8)$$

The majority voting of its neighbors predicts the class label of the query point x

$$\hat{c} = \operatorname{argmax} \sum_{(x_i^{NN}, c_i^{NN}) \in T} \delta(c = c_i^{NN}) \quad (9)$$

Where c is a class label and c_i^{NN} indicates the class label for the i -th nearest neighbor from its k nearest neighbors. $\delta(c = c_i^{NN})$, an indicator function, takes a value of one the class c_i^{NN} of the neighbor x_i^{NN} is the same as class c and zero otherwise.

Support Vector Regression is a supervised machine learning algorithm that can be utilized for classification and regression problems. The SVR algorithm's main idea is plotting each data in n -dimensional space (where n is the number of features) [33]. Every value of each feature indicates the value of particular coordinates. It classifies the values by obtaining the best hyperplane that differentiates the two classes in the best possible way. The hypothesis function h can be illustrated as

$$h(x_i) = f(x) = \begin{cases} +1 & \text{if } w \cdot x + b \geq 0 \\ -1 & \text{if } w \cdot x + b < 0 \end{cases} \quad (10)$$

Where w is the direction of the vector x . For example, If vector $x = (x_1, x_2)$

$$w = \left(\frac{x_1}{\|x\|}, \frac{x_2}{\|x\|} \right) \quad (11)$$

and b is the bias term

Random forests develop many individual decision trees during the training process. Each tree makes one prediction, and after that, all predictions from the trees are pooled to make the final prediction [34]. It is also called ensemble techniques since it uses a collection of results to decide. If we assume, there are only two child nodes, the equation for calculating the new node is

$$ni_j = w_j C_j - w_{\text{left}(j)} C_{\text{left}(j)} - w_{\text{right}(j)} C_{\text{right}(j)} \quad (12)$$

Where ni_j is the importance of node j , w_j is the weighted number of samples reaching node j , C_j is the impurity value of node j and $\text{left}(j)$ and $\text{right}(j)$ is the child node from left and right split on node j , respectively.

Mean absolute error is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set [35]. Each prediction error is the

difference between the true value and the predicted value for the instance.

$$mae = \frac{\sum_{i=1}^n \text{abs}(y_i - \lambda(x_i))}{n} \quad (13)$$

where y_i is the true target value for test instance, x_i , $\lambda(x_i)$ is the predicted target value for test instance x_i , and n is the number of test instances.

Variables and data analysis

Conda environment and Python 3.8.5 is used for data analysis and machine learning applications. Jupyter notebook was used for code documentation. pandas [36], numpy [37], scikit-learn [38], pyman Kendall [39], pyhomogeneity [40] packages has been used for data analysis, visualization and prediction processes. The data was organized in one dataframe where it has 19 columns and 2240 entries. The first column contains the years which ranges from 1880 to 2020 and second column represents regions of Germany. From third to fourteenth columns contain total precipitation of every month in every year on different regions. The next four columns depicts the total precipitation in summer, winter, spring, fall and the last column contain total precipitation in a year. After preparing the data, trend analysis has been done by Man-Kendall test and Hamad-Rao modification test. After conduction of change point analysis Pettit test, the data was splitting into two different sets (input and output) for machine learning algorithms. Every input and output datasets further split into two sets training and testing purposes. The training datasets contains 691 entries and testing datasets contains 173 entries. After training the model with training datasets, testing datasets was used for validation purpose. Mean absolute error was calculated to calculate the accuracy of the model.

Results

I calculated the descriptive statistics of annual precipitation and seasonal precipitation from 1880 to 2020 for Germany's sixteen federal states. Results indicate that the states in the western part of Germany, i.e., Baden-Wuerttemberg, Bayern, and Saarland, have observed the highest average rainfall (938.96mm, 909.89mm, 886.08mm, respectively). On the other hand, minimum average rainfall has been recorded in the Sachsen-Anhalt and Brandenburg (559.1mm and 566.07mm, respectively). The highest variation (standard deviation) in precipitation is found in Saarland state, where Mecklenburg-Vorpommern has the minimum standard deviation among all other states. The skewness for all federal states ranges from 0.25 to -0.067. The negative skewness found in Rheinland-Pfalz (-0.003), Saarland (-0.054), Thueringen/Sachsen-Anhalt (-0.067), Thueringen (-0.0423), while the rest of the federal states observed positive skewness.

I calculated the annual and seasonal precipitation trend for sixteen federal states using the non-parametric Mann- Kendall test

Figure 2. Boxplot for annual average precipitation of German's regions over 1881-2020

at a 0.05 significance level. Table 1 exhibits the z value for four seasons and annual precipitation in which z values ranges from -1 to 4.37. Brandenburg/Berlin, Brandenburg and Sachsen has the z values ranges from 0 to -1. Baden-Wuerttemberg, Bayern, Hessen, Mecklenburg-Vorpommern, Niedersachsen, Niedersachsen/Hamburg/Bremen, Nordrhein-Westfalen, Rheinland-Pfalz, Schleswig-Holstein, Saarland has z value greater than two, which suggests a highly increasing trend of precipitation. In Winter (December to February), all states show increasing precipitation patterns with the highest z value of 3.60 by Bayern and the lowest value of 1.36 by Sachsen. Moreover, Niedersachsen, Niedersachsen/Hamburg/Bremen, Schleswig-Holstein, and Saarland have shown an increasing pattern between September to November.

Since the Man-Kandell test [16] has some significant drawbacks like the presence of autocorrelation with the datasets, non-linearity could influence the result. Hamed and Rao [41] proposed an innovative trend method to overcome the mentioned drawbacks. Hamed and Rao [41] suggested that this modified MK test can effectively detect the trend in data with serial correlation. Table 4 shows the z value for the MK test proposed by Hamad and Rao with a 0.05 significance level. Although the original MK test results and Hamad Rao Modified MK test imply almost the same trend among the federal states, there are some slight changes in the z values.

Pettitt test [32] is a non-parametric homogeneity test that checks if two or more sets of data are from the same distribution or not. The homogeneity test is applied in the time series data to detect the breakpoint in the series. Table 2 shows the result of the Pettitt test to detect the change point in sixteen federal states. The alternate hypothesis (h) for that change point exists. Change Point selection is based on the performances (p-value) of the test. The h value in several states (e.g., Bayern,

Hessen, Niedersachsen/Hamburg/Bremen, Rheinland-Pfalz, Schleswig-Holstein, Saarland) came True, which means those states exhibits non-homogenous precipitation trend throughout 1880-2020.

Name	DEC-FEB	MAR-MAY	JUN-AUG	SEP-NOV	ANNUAL
Brandenburg/Berlin	2.3042	-0.2901	-0.4432	-0.0180	0.5315
Brandenburg	2.2970	-0.2937	-0.4414	-0.0522	0.4900
Baden-Wuerttemberg	3.5364	1.8412	-1.5403	1.9349	2.1330
Bayern	3.5977	1.5007	-0.6305	1.7529	2.3798
Hessen	2.5510	1.9385	-0.1747	1.1584	2.2105
Mecklenburg-Vorpommern	2.5564	0.3477	-0.0955	0.9602	2.0285
Niedersachsen	2.9023	0.5621	0.3657	2.0988	2.5366
Niedersachsen/Hamburg/Bremen	2.8933	0.5729	0.3693	2.1150	2.5744
Nordrhein-Westfalen	2.5690	1.2287	0.2324	1.4539	2.2591
Rheinland-Pfalz	2.8699	2.1420	-0.9548	1.2160	2.4015
Schleswig-Holstein	3.1419	0.9512	1.2016	2.5222	4.3723
Saarland	3.1815	2.1871	-0.5585	2.0105	2.9815
Sachsen	1.3620	-1.2503	-1.3728	0.1027	-0.9999
Sachsen-Anhalt	1.7782	0.3711	-0.4522	0.1675	0.2630
Thueringen/Sachsen-Anhalt	1.9997	0.5044	-0.4540	0.3909	0.6972
Thueringen	1.8340	0.6305	-0.4198	0.6792	1.2232

Table 1. z value of MK test for annual and seasonal precipitation for 1880-2020

States	Change Point	p	h
Brandenburg/Berlin	1992	0.96605	FALSE
Brandenburg	1992	0.9712	FALSE
Baden-Wuerttemberg	1934	0.0542	FALSE
Bayern	1964	0.046	TRUE
Hessen	1921	0.0223	TRUE
Mecklenburg-Vorpommern	1921	0.11795	FALSE
Niedersachsen	1949	0.05265	FALSE
Niedersachsen/Hamburg/Bremen	1924	0.04715	TRUE
Nordrhein-Westfalen	1921	0.0699	FALSE
Rheinland-Pfalz	1921	0.01745	TRUE
Schleswig-Holstein	1949	0.00005	TRUE
Saarland	1921	0.00535	TRUE
Sachsen	1941	0.3445	FALSE
Sachsen-Anhalt	1991	0.9417	FALSE
Thueringen/Sachsen-Anhalt	1949	0.8948	FALSE
Thueringen	1964	0.53745	FALSE

Table 2. Pettitt test result for annual precipitation over 1881-2020

Precipitation prediction

The analyses mentioned above exhibit most of the regions in Germany experienced a significant increase in precipitation. Therefore, prediction of precipitation has become essential for water resource management. Several machine learning algorithms are available for precipitation prediction, but I applied k-NN, SVR, and RFR. The precipitation data from the year of 2010 to 2015 has taken into account to train and test the model. In the first step, I divided the data into two sets, input and output. Precipitation in January, February, March, May, June, July, September, October, November worked as input variable, and April, August, and December precipitation data worked as output. After that, input and output data were split into two datasets for training and testing purposes. In the end, the training set was fed into the algorithms for building the model. The number of

training and testing data are 691 and 173 respectively. After that, testing dataset has used for calculating model performance. Figure 3 depicts the model performance where x axis represent test size and y axis is showing the precipitation in mm. The comparisons of three simulations with the observed precipitation data has shown on figure 3. In general, RFR algorithm predictions are more accurate than SVR and k-nn algorithms. Mean absolute percentage error for RFR is 26.527 where algorithms like k-NN and SVR has mean absolute error of 28.416 and 31.012 respectively.

Several parameters can be tuned for changing the model performance in these three models. In k-NN model, the k parameter is changed for sensitivity analysis. k is the number of neighbors used for prediction. Figure 4 in additional figures shows the sensitivity analysis of k-NN model in terms of k. There are some slight changes in values of MAE as the value of k changes. When the k has the value 5, the MAE is 28.416 (figure 3 above). Apart from that, when the k value are 1, 20 and 50, the MAE's are 27.282, 30.661, 30.443 respectively. In SVR model, The Regularization parameter often termed as C parameter in python's sklearn library tells the SVR optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points [42]. Figure 6 shows how SVR model reacted with changing the C parameter. It shows there is no significant change the MAE'S value in different C values. In RFR model, the sensitivity analysis has been done by changing number of estimators (nestimator). Number of estimator is the number of trees to be used in the forest. Since Random Forest is an ensemble method comprising of creating multiple decision trees, this parameter is used to control the number of trees to be used in the process [43]. Figure 5 shows the nestimator effect in the RFR model.

Figure 3. Prediction of precipitation by the model k-NN (up), Support vector regressor (middle), Random forest regressor (below)

Discussion and conclusion

Although there is a scope for selecting the proper dataset for analyzing, the methods for trend and change point analysis gave us a good idea of precipitation pattern in different year on the period of 1881–2020. The higher temporal resolution data could give much more convenient result and need to work with spatial temporal variability also. Man-Kendall test does not consider the serial autocorrelation but Hamad Rao modification test considers all significant lags. The choice of splitting the data for prediction can be improved but it gives a clear idea about the three model used here. Prediction model could be optimized by changing parameters of these three models. Moreover, The initiation of using machine learning algorithms in the field of hydrology needs much more appreciation and investigation to use them as useful tool for forecasting.

References

1. Praveen, B. *et al.* Analyzing trend and forecasting of rainfall changes in india using non-parametrical and machine learning approaches. *Sci. Reports* **10**, DOI: [10.1038/s41598-020-67228-7](https://doi.org/10.1038/s41598-020-67228-7) (2020).
2. NASA. Tropical rainfall measuring mission. <https://www.earthobservatory.nasa.gov/features/TRMM/trmm2.php> (1999).
3. Huang, M., Lin, R., Huang, S. & Xing, T. A novel approach for precipitation forecast via improved k-nearest neighbor algorithm. *Adv. Eng. Informatics* **33**, 89–95, DOI: [10.1016/j.aei.2017.05.003](https://doi.org/10.1016/j.aei.2017.05.003) (2017).
4. Srivastava, P. K., Han, D., Rico-Ramirez, M. A. & Islam, T. Sensitivity and uncertainty analysis of mesoscale model downscaled hydro-meteorological variables for discharge prediction. *Hydrol. Process.* **28**, 4419–4432, DOI: [10.1002/hyp.9946](https://doi.org/10.1002/hyp.9946) (2013).
5. Islam, T., Rico-Ramirez, M. A., Han, D., Srivastava, P. K. & Ishak, A. M. Performance evaluation of the TRMM precipitation estimation using ground-based radars from the GPM validation network. *J. Atmospheric Solar-Terrestrial Phys.* **77**, 194–208, DOI: [10.1016/j.jastp.2012.01.001](https://doi.org/10.1016/j.jastp.2012.01.001) (2012).
6. Zink, M. *et al.* The german drought monitor. *Environ. Res. Lett.* **11**, 074002, DOI: [10.1088/1748-9326/11/7/074002](https://doi.org/10.1088/1748-9326/11/7/074002) (2016).
7. Kharin, V. V., Zwiers, F. W., Zhang, X. & Wehner, M. Changes in temperature and precipitation extremes in the CMIP5 ensemble. *Clim. Chang.* **119**, 345–357, DOI: [10.1007/s10584-013-0705-8](https://doi.org/10.1007/s10584-013-0705-8) (2013).
8. Berg, P., Moseley, C. & Haerter, J. O. Strong increase in convective precipitation in response to higher temperatures. *Nat. Geosci.* **6**, 181–185, DOI: [10.1038/ngeo1731](https://doi.org/10.1038/ngeo1731) (2013).
9. Clouds and aerosols. In on Climate Change, I. P. (ed.) *Climate Change 2013 - The Physical Science Basis*, 571–658, DOI: [10.1017/cbo9781107415324.016](https://doi.org/10.1017/cbo9781107415324.016) (Cambridge University Press).
10. Deumlich, D. & Gericke, A. Frequency trend analysis of heavy rainfall days for germany. *Water* **12**, 1950, DOI: [10.3390/w12071950](https://doi.org/10.3390/w12071950) (2020).
11. Hänsel, S., Petzold, S. & Matschullat, J. Precipitation trend analysis for central eastern germany 1851–2006. In *Bioclimatology and Natural Hazards*, 29–38, DOI: [10.1007/978-1-4020-8876-6_3](https://doi.org/10.1007/978-1-4020-8876-6_3) (Springer Netherlands, 2009).
12. Trömel, S. & Schönwiese, C.-D. Extreme value and trend analysis based on statistical modelling of precipitation time series. In *In Extremis*, 184–200, DOI: [10.1007/978-3-642-14863-7_9](https://doi.org/10.1007/978-3-642-14863-7_9) (Springer Berlin Heidelberg, 2010).
13. Sonali, P. & Kumar, D. N. Review of trend detection methods and their application to detect temperature changes in india. *J. Hydrol.* **476**, 212–227, DOI: [10.1016/j.jhydrol.2012.10.034](https://doi.org/10.1016/j.jhydrol.2012.10.034) (2013).
14. Piao, S. *et al.* The impacts of climate change on water resources and agriculture in china. *Nature* **467**, 43–51, DOI: [10.1038/nature09364](https://doi.org/10.1038/nature09364) (2010).
15. Heinisch, O. Snedecor, g. w.: Statistical methods. 5. aufl. the iowa state university press, ames, iowa 1965. XIII + 534 s., preis \$ 7,50. *Biom. Zeitschrift* **9**, 210–210, DOI: [10.1002/bimj.19670090322](https://doi.org/10.1002/bimj.19670090322) (1967).
16. Mann, H. B. Nonparametric tests against trend. *Econometrica* **13**, 245, DOI: [10.2307/1907187](https://doi.org/10.2307/1907187) (1945).
17. M.T.L.B. Exercises in theoretical statistics. by kendall m. g.. [pp. vii + 179. london: Charles griffin and co. ltd. 1954. price 20s.]. *J. Staple Inn Actuar. Soc.* **12**, 247, DOI: [10.1017/s0020269x00005673](https://doi.org/10.1017/s0020269x00005673) (1954).

18. J., N. L. Rank correlation methods. by maurice g. kendall, m.a. [pp. vii + 160. london: Charles griffin and co. ltd., 42 drury lane, 1948. 18s.]. *J. Inst. Actuar.* **75**, 140–141, DOI: [10.1017/s0020268100013019](https://doi.org/10.1017/s0020268100013019) (1949).
19. Ross, S. M. INTRODUCTORY STATISTICS. In *Introductory Statistics*, ibc1–ibc4, DOI: [10.1016/b978-0-12-374388-6.50002-7](https://doi.org/10.1016/b978-0-12-374388-6.50002-7) (Elsevier, 2010).
20. Mondal, A., Khare, D. & Kundu, S. Spatial and temporal analysis of rainfall and temperature trend of india. *Theor. Appl. Climatol.* **122**, 143–158, DOI: [10.1007/s00704-014-1283-z](https://doi.org/10.1007/s00704-014-1283-z) (2014).
21. Pandey, B. K. & Khare, D. Identification of trend in long term precipitation and reference evapotranspiration over narmada river basin (india). *Glob. Planet. Chang.* **161**, 172–182, DOI: [10.1016/j.gloplacha.2017.12.017](https://doi.org/10.1016/j.gloplacha.2017.12.017) (2018).
22. Luk, K. C., Ball, J. & Sharma, A. An application of artificial neural networks for rainfall forecasting. *Math. Comput. Model.* **33**, 683–693, DOI: [10.1016/s0895-7177\(00\)00272-7](https://doi.org/10.1016/s0895-7177(00)00272-7) (2001).
23. Carrier, C. A., Kalra, A. & Ahmad, S. Long-range precipitation forecasts using paleoclimate reconstructions in the western united states. *J. Mountain Sci.* **13**, 614–632, DOI: [10.1007/s11629-014-3360-2](https://doi.org/10.1007/s11629-014-3360-2) (2016).
24. Foresti, L., Pozdnoukhov, A., Tuia, D. & Kanevski, M. Extreme precipitation modelling using geostatistics and machine learning algorithms. In *Quantitative Geology and Geostatistics*, 41–52, DOI: [10.1007/978-90-481-2322-3_4](https://doi.org/10.1007/978-90-481-2322-3_4) (Springer Netherlands, 2010).
25. Antonić, O., Križan, J., Marki, A. & Bukovec, D. Spatio-temporal interpolation of climatic variables over large region of complex terrain using neural networks. *Ecol. Model.* **138**, 255–263, DOI: [10.1016/s0304-3800\(00\)00406-3](https://doi.org/10.1016/s0304-3800(00)00406-3) (2001).
26. Hong, W.-C. Rainfall forecasting by technological machine learning models. *Appl. Math. Comput.* **200**, 41–57, DOI: [10.1016/j.amc.2007.10.046](https://doi.org/10.1016/j.amc.2007.10.046) (2008).
27. Di, Y. *et al.* Developing machine learning tools for long-lead heavy precipitation prediction with multi-sensor data. In *2015 IEEE 12th International Conference on Networking, Sensing and Control*, DOI: [10.1109/icnsc.2015.7116011](https://doi.org/10.1109/icnsc.2015.7116011) (IEEE, 2015).
28. Beheshti, Z., Firouzi, M., Shamsuddin, S. M., Zibarzani, M. & Yusop, Z. A new rainfall forecasting model using the CAPSO algorithm and an artificial neural network. *Neural Comput. Appl.* **27**, 2551–2565, DOI: [10.1007/s00521-015-2024-7](https://doi.org/10.1007/s00521-015-2024-7) (2015).
29. Mukherjee, P., Singh, C. K. & Mukherjee, S. Delineation of groundwater potential zones in arid region of india—a remote sensing and GIS approach. *Water Resour. Manag.* **26**, 2643–2672, DOI: [10.1007/s11269-012-0038-9](https://doi.org/10.1007/s11269-012-0038-9) (2012).
30. Pham, L. T., Luo, L. & Finley, A. O. Evaluation of random forest for short-term daily streamflow forecast in rainfall and snowmelt driven watersheds. DOI: [10.5194/hess-2020-305](https://doi.org/10.5194/hess-2020-305) (2020).
31. Historical monthly station averages precipitation of different regions. http://opendata.dwd.de/climate_environment/CDC/regional_averages_DE/monthly/precipitation/ (2020).
32. Pettitt, A. N. A non-parametric approach to the change-point problem. *Appl. Stat.* **28**, 126, DOI: [10.2307/2346729](https://doi.org/10.2307/2346729) (1979).
33. Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, 61–74 (MIT Press, 1999).
34. Ho, T. K. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, DOI: [10.1109/icdar.1995.598994](https://doi.org/10.1109/icdar.1995.598994) (IEEE Comput. Soc. Press).
35. Sammut, C. & Webb, G. I. (eds.). *Mean Absolute Error*, 652–652 (Springer US, Boston, MA, 2010).
36. pandas development team, T. pandas-dev/pandas: Pandas, DOI: [10.5281/zenodo.3509134](https://doi.org/10.5281/zenodo.3509134) (2020).
37. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362, DOI: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2) (2020).
38. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
39. Hussain, M. M. & Mahmud, I. pymannkendall: a python package for non parametric mann kendall family of trend tests. *J. Open Source Softw.* **4**, 1556, DOI: [10.21105/joss.01556](https://doi.org/10.21105/joss.01556) (2019).

40. pyhomogeneity 1.1. <https://pypi.org/project/pyhomogeneity/> (2020).
41. Hamed, K. H. & Rao, A. R. A modified mann-kendall trend test for autocorrelated data. *J. Hydrol.* **204**, 182–196, DOI: [10.1016/s0022-1694\(97\)00125-x](https://doi.org/10.1016/s0022-1694(97)00125-x) (1998).
42. Chang, C.-C. & Lin, C.-J. LIBSVM. *ACM Transactions on Intell. Syst. Technol.* **2**, 1–27, DOI: [10.1145/1961189.1961199](https://doi.org/10.1145/1961189.1961199) (2011).
43. Breiman, L. *Mach. Learn.* **45**, 5–32, DOI: [10.1023/a:1010933404324](https://doi.org/10.1023/a:1010933404324) (2001).

Acknowledgements

I would like to thank Prof. Dr. rer. nat. Christoph Hinz for supervising and Brian Hafar ,Dipl.-Hydrol. Sabine Schümberg for helping me to complete this report.

Author contributions statement

All the data analysis has been done by Talukder .The scripts of Jupyter notebook can be accessed by going <https://github.com/talukmdm/Trend-analysis-and-prediction-of-precipitation>

Supplements

Name	description	class type
data	contains monthly,annual and seasonal average precipitation	pandas dataframe
trends	contains z values from original Man-kendall test	pandas dataframe
cp_df	contains change point of annual precipitation	pandas dataframe
X_train	input datasets for training phase of k-nn, SVR and RFR model	numpy nd array
X_test	output datasets for training phase of k-nn, SVR and RFR model	numpy nd array
y_train	validation input datasets for testing k-nn, SVR and RFR model	numpy nd array
y_test	validation output datasets for testing k-nn, SVR and RFR model	numpy nd array
predicted_knn	predicted output datasets by knn model	numpy nd array
pred_svr	predicted output datasets by SVR model	numpy nd array
predicted_rfr	predicted output datasets by RFR model	numpy nd array

Table 3. variables used in the data analysis

Name	DEC-FEB	MAR-MAY	JUN-AUG	SEP-NOV	ANNUAL
Brandenburg/Berlin	1.8390	-0.4228	-0.4432	-0.0180	0.6087
Brandenburg	1.8342	-0.4299	-0.4414	-0.0522	0.5618
Baden-Wuerttemberg	3.2474	3.3002	-1.3545	1.9349	1.8487
Bayern	3.9982	1.6748	-0.6010	1.7529	2.0797
Hessen	2.5309	3.4054	-0.1908	1.1584	2.0735
Mecklenburg-Vorpommern	2.2126	1.5252	-0.0833	0.9147	2.0823
Niedersachsen	2.9023	0.5890	0.8249	1.9955	2.3966
Niedersachsen/Hamburg/Bremen	2.8933	0.5999	0.8416	2.0091	2.4318
Nordrhein-Westfalen	2.6083	1.2287	0.5315	1.4539	2.3770
Rheinland-Pfalz	2.7989	3.8202	-0.8478	1.2160	2.2495
Schleswig-Holstein	4.8705	0.9512	1.6476	2.1978	4.0606
Saarland	4.1163	2.0908	-0.6384	2.0105	2.8085
Sachsen	1.3620	-1.0855	-2.3170	0.0981	-0.9999
Sachsen-Anhalt	2.1528	0.5141	-0.3740	0.2038	0.2764
Thueringen/Sachsen-Anhalt	1.9997	0.5450	-0.3909	0.4847	0.7305
Thueringen	1.9624	0.7560	-0.3599	0.9040	1.2232

Table 4. z values by Modified MK test proposed by Hamad and Rao for 1880-2020 in different regions of Germany

Figure 4. Prediction of precipitation by the k-NN algorithm in different k number. k=1(up),k=20(middle),k=50(down)

Figure 5. Prediction of precipitation by the Random forests algorithm in different number of estimator. up (50),middle(150),down(300)

Figure 6. Prediction of precipitation by the Support vector machine algorithm in different value of C parameter $C=100$ (up), $C=200$ (middle), $C=300$ (down)