Official Journal of the World Water Council
Water Policy

Water Policy Vol 25 No 12, 1113 doi: 10.2166/wp.2023.078

Determination of annual rainfall in north-east India using deterministic, geospatial, and machine learning techniques

Shivam Agarwal [a,*], Disha Mukherjee[b] and Nilotpal Debbarma[b]

[a] Civil Engineering Department, NIT Silchar, Silchar 788010, India
[b] Civil Engineering Department, NIT Agartala, Barjala, Jirania, Agartala 799046, India
*Corresponding author. E-mail: shivamgupta.agarwal@gmail.com

SA, 0000-0001-7221-1171

ABSTRACT

Analysis of extreme annual rainfall in the six north-east Indian states of Assam, Meghalaya, Nagaland, Manipur, Mizoram, and Tripura using the deterministic interpolation technique of inverse distance weighting (IDW), the geospatial interpolation technique of Ordinary Kriging (OK) and the machine learning prediction technique of generalised additive model (GAM). GAM is used only for prediction and hence the results are then subsequently interpolated by OK to create the rainfall maps. The datasets considered for this study are a training dataset of 171 points which consisted of satellite rainfall and a testing dataset with ground rain gauge data of 33 points which was used for validation of the former. A combined dataset of training + testing was also interpolated and mapped to compare for visual accuracy of each technique. It was seen that OK was a superior and a much more realistic interpolation technique than IDW, since it took the altitude of each site into consideration along with latitude and longitude, unlike IDW, which only interpolated over the *x*–*y* plane and didn't rely on altitude. When the predictions of the training dataset through GAM were mapped using OK, it showed almost parallel contours, which is undesirable for natural phenomenon like rain.

Key words: Generalised additive models (GAMs), Inverse distance weighting (IDW), Kriging, North-east India, Rainfall

HIGHLIGHTS

- Analyse extreme annual rainfall in the six north-east Indian states of Assam, Meghalaya, Nagaland, Manipur, Mizoram and Tripura by using.
- Using the deterministic interpolation technique of Inverse distance weighting method (IDW).
- The geospatial interpolation technique of Ordinary Kriging (OK).
- The machine learning prediction technique of generalised additive model (GAM).5.GAM is used for prediction.

ABBREVIATIONS

COD	coefficients of determination
NSE	Nash–Sutcliffe Efficiency
GAM	generalised additive model
DEM	digital elevation modelling
RMSE	root mean square error
MAE	mean absolute error
OK	Ordinary Kringing

1. INTRODUCTION

Many operations in agriculture, climate forecasting, watershed management, water resource management, irrigation schemes, and water provisioning require accurate precipitation determination. We can use mapping to forecast rainfall patterns for required return periods, especially in places where rain gauges are not available (Singh *et al*., 2022). Using data from nearby, known sources, spatial interpolation can be used to infer incomplete or unavailable data.

North-east (NE) India's precipitation is unique in that it frequently exceeds 1,000 mm, the highest recorded annual rainfall in history. The average annual rainfall in Mawsynram, the wettest place on Earth is currently 11,802.4 mm, which is about 10 times as much rain as the average annual rainfall in India. NE India, on the other hand, exhibits a wide range in localised rainfall due to its high-altitude differences within a relatively compact spatial area. Owing to the Tropic of Cancer running through the territories of Mizoram and Tripura, NE India has a significant tropical impact, especially in the plains. NE India receives the highest annual rainfall in the world, which seldom falls below 1,000 mm. Mawsynram in Meghalaya, which receives an average annual rainfall of about 11,800 mm, is the wettest place on Earth. However, while some parts of NE India, namely parts of Assam and Meghalaya, suffer from devastating floods every year, which disrupts huge masses of lives and livelihood, the other parts of these states along with parts of Tripura, Mizoram, Manipur, and Nagaland face an overall scarcity (Agarwal *et al*., 2021a, 2021b). The terrain of NE India is highly undulating within very short distances, making its climatic and rainfall patterns variable over small stretches and hence becomes hard to ascertain (Agarwal *et al*., 2022). The precipitation of Arunachal Pradesh includes snowfall and is quite different from that of the other six states, due to which it has been excluded from this study.

Among previous studies, Adhikary *et al*., (2017) compared two deterministic interpolation methods (inverse distance weighting (IDW) and radial basis function) for enhanced spatial interpolation of monthly rainfall (Ditthakit *et al*., 2021), as well as three geostatistical interpolation methods (Ordinary Kriging [OK], ordinary cokriging [OCK], and kriging with an external drift [KED]) and observed that in a geostatistical framework, using elevation as an auxiliary variable in addition to rainfall data can greatly improve rainfall estimation over a catchment (Da Silva Monteiro *et al*., 2022). Borges *et al*. (2016) used annual and seasonal precipitation to test different spatial interpolation methods, including IDW, Spline tension (Spline), OK, CoKriging (CoOK), Detrended Universal Kriging (DUK), Multiple linear regression (MReg), and Residual interpolation (MRegIDW and MRegOK). Geographic considerations were included in multivariate models (i.e., altitude, longitude and latitude) and according to the Nash–Sutcliffe Efficiency (NSE), IDW, OK, and multivariate regression with interpolation of residuals by IDW (MRegIDW) and OK (MRegOK) had the lowest errors and the highest correlation. Lemus-Canovas *et al*. (2019), in their area of research, combined the synoptic scale with the local scale and used the Pyrenees regional scale to interpolate mean daily precipitation (MDP) based on a synoptic scale classification of weather types using the general linear model (GLM), generalised additive model (GAM), and Regression Kriging (RK) methods. The GAM and RK approaches offered the best fit for the models. Underwood (2009) in her research showed how GAMs can enhance the flexibility of models to capture seasonal patterns and long-term trends in the occurrence and amount of daily rainfall. The models' smoothed estimations gave excellent graphical depictions of changing rainfall patterns, even for non-linear relationships in the data in his study region, over the last 40 years (Markuna *et al*., 2023).

The present study hence focuses on the application and comparison of three different interpolation and prediction techniques: (1) the deterministic technique of IDW method; (2) the geospatial technique of OK; and (3) the machine learning technique of GAMs, for the evaluation, and assessment of annual extreme rainfall over the aforementioned states of NE India (Figure 1).

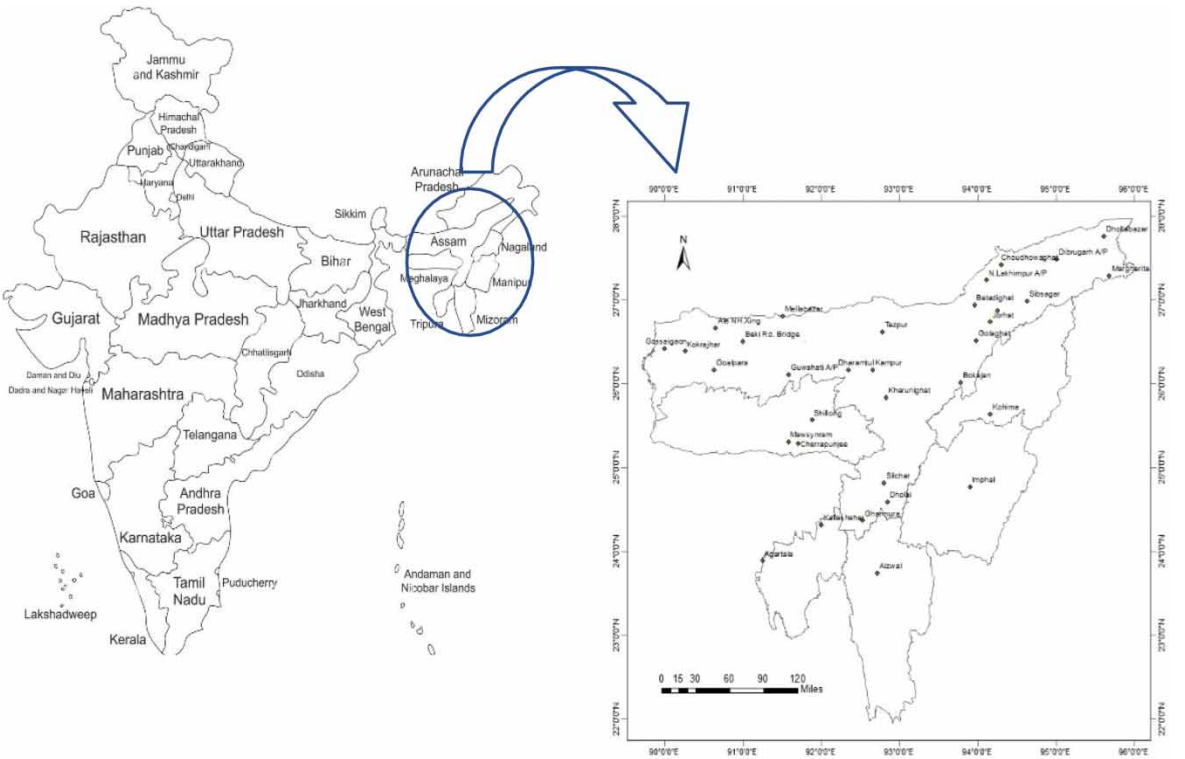

Fig. 1 | Study region showing testing points.

2. STUDY AREA AND DATA COLLECTION

A total of 246 grid points, taken at intersections of latitudes and longitudes over the six states of NE India were considered for the years 1980–2020 (Pai *et al.*, 2014). Latitudes and longitudes have been taken at increase of 0.25° in both horizontal and vertical directions with minimum and maximum latitudes at 22° N and 29° N, respectively, and minimum and maximum longitudes at 90° E and 96° E, respectively, over the six states of Assam, Meghalaya, Nagaland, Manipur, Mizoram, and Tripura. The rainfall values were acquired from the IMD Pune (Mann, 1945) website as daily gridded satellite-based rainfall data files at 0.25×0.25 grid junctions.

Assam has a hot, humid monsoonal climate with 1,800 mm average annual rainfall. Meghalaya receives an average annual rainfall of over 10,000 mm, which includes rainfall at Mawsynram, the wettest place in the world. However, both these states are struck by disastrous floods every year, often in the form of flash floods due to failure of hydraulic structures. Tripura and Mizoram have the Tropic of Cancer passing through them. These states, along with Manipur and Nagaland, have climates varying from tropical monsoonal to subtropical temperate, depending on the altitude and geography. Rainfall in Tripura varies from 1,922 to 2,855 mm every year, spread over April to September, while Mizoram has an average annual rainfall of 2,540 mm during months of May to September. Annual rainfall in Nagaland ranges from 1,800 to 2,500 mm and mostly occurs during the southwest monsoon months of May to September. The annual rainfall in Manipur averages between 1,250 and 2,700 mm. November, December, January, and February are the dry months, while rain occurs over the other 8 months.

To determine the randomness and trend for each grid point, all 246 points were evaluated. The finalised grid points were supposed to display pure randomness and no trend. After processing, 171 of the 246 points were left, all of which were random and lacked any kind of trend. To verify the procedures previously undergone by the training dataset of 171 points, a set of testing data, totalling 33 points, distributed throughout the six states of NE India, was taken into consideration. In contrast to the satellite rainfall data of the training dataset, these 33 points were ground rainfall data from rain gauge stations.

3. METHODOLOGY

From 1980 to 2020, a total of 41 datasets were taken into account. The maximum daily rainfall at each grid point in a year was then calculated using the datasets. With an average of 97.16 mm, the maximum rainfall was recorded at 26.25° N, 90.25° E in 1984 as 795.65 mm and the minimum rainfall was recorded at 22.75° N, 92.75° E in 2009 as 97.16 mm.

The outcomes were analysed to look for trends and randomness at each grid point. After determining the randomness and trends, the grid points with only randomness present and no trends were chosen for future investigations.

The Ljung-Box (Lemus-Canovas *et al.*, 2019) test was employed to ascertain whether randomness existed. The Ljung-Box test's mathematical formula is as follows:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k}$$

Given a time series Y of length n, the test statistic is defined as where r_k^2 is the estimated autocorrelation of the series at lag k, and m is the number of lags being tested.

k and m are the number of lags being tested. $H0$ = The information is disseminated independently (i.e., any observed correlations in the data result from randomness of the sampling process, such that the correlations in the population from which the sample is taken are 0), Ha = The data are serially correlated rather than being distributed randomly. If the null hypothesis (that the model has a significant lack of fit) is rejected, the Box-Ljung test is used, i.e., if $Q > \chi^2_{1-\alpha,h}$ where $\chi^2_{1-\alpha,h}$ is the χ^2 distribution table value with h degrees of freedom and significance level α, the chi square table determines whether two variables are independent of each other or not (Ljung & Box, 1978).

The Mann (Link & Koch, 1970)–Kendall (Box & Pierce, 1970) Pattern Test was used to detect whether or not the time series data had a trend (Kendall 1975). The Mann–Kendall (MK) test determines whether the null hypothesis ($H0$) should be rejected and the alternative hypothesis (H_a) should be accepted. $H0$ = There is no consistent pattern, H_a = There is a monotonic trend. The MK test begins with the assumption that is true, and that data must be convincing beyond a reasonable doubt before being dismissed or accepted. The MK Trend Test compares the indications of data points from earlier and later periods. The assumption is that if a trend exists, the sign values will tend to increase or decrease on a regular basis. Every value in the time series is compared to every value before it, yielding $n(n–1)/2$ pairs of data, where 'n' is the number of observations in the collection.

The presence of randomness and trend was determined by the p-value or probability value. If p-value is less than 0.05, the null hypothesis is rejected, i.e., probability of occurrence is less than 5% and hence very low (Figure 2). So, for randomness to be present, p-value for randomness should be ≥ 0.05. To infer that there is a significant trend, $p < 0.05$ for the trend test is required, which means that only 5% of the time could such a pattern have arisen by chance. Therefore, for no trend to be present, its p-value should be ≥ 0.05. In statistical terms, by randomness, we mean that the values of rainfall should not be showing any noticeable patterns or regularities in

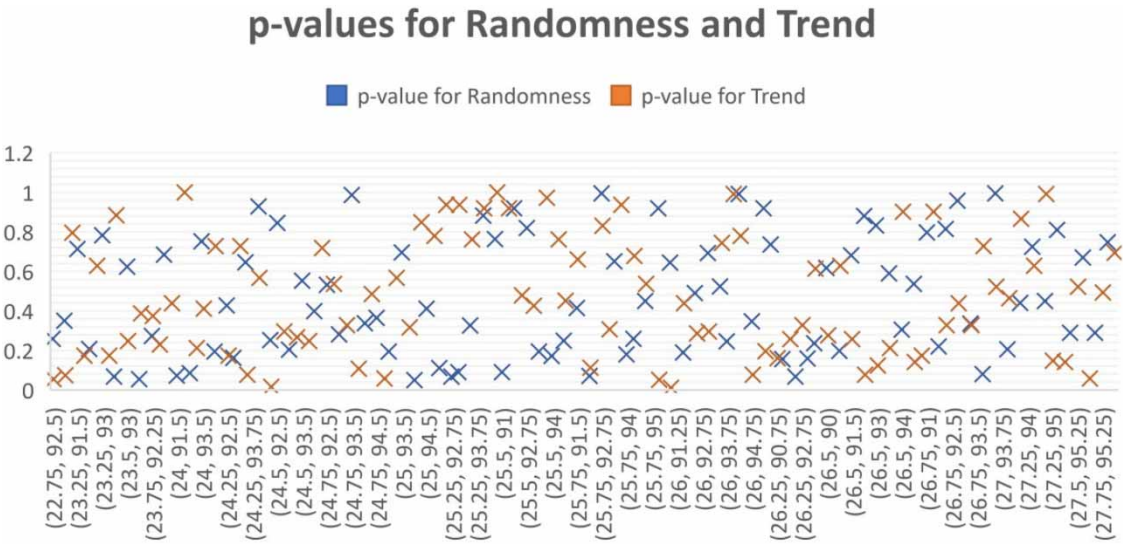

Fig. 2 | *p*-values for randomness and trend for the final 171 points.

their occurrences, for this may hinder the accuracy in our predictions and our forecasts will tend to be biased or influenced. If the rainfall values do show a certain recognisable sequence, we can say there's a trend present in their occurrence and rule them out of our calculations. It is important that the dataset finally taken has no trend present and only possesses randomness. This is because data showing trend will influence our prediction with their interdependence or correlation with time, resulting in our predictions being prejudiced and hampering their independent existence and accuracy. A total of 171 such points remained after this process.

The 171 remaining points were then put on a Digital Elevation Modelling (DEM) image of NE India to further determine the respective geographical elevations at these points.

The average of the annual maximum rainfall for 41 years from 1981 to 2020 was then found out for each of the 171 grid points.

Deterministic and geostatistical interpolation systems are the two basic types of interpolation systems. Deterministic interpolation methods generate interpolations from measured points based on the degree of similarity (e.g., inverse distance weighted) or the smoothing degree (e.g., radial basis functions). The statistical features of the measured sites, such as terrain and geology, are used in geostatistical interpolation techniques (e.g., kriging). Geostatistical procedures account for the spatial layout of the sample points surrounding the prediction location and assess the spatial autocorrelation among measured points.

3.1. IDW method

IDW is a deterministic multivariate interpolation method that uses a known distributed set of points. The required value for a particular point is calculated as a weighted average of the known values available at its surrounding specified points. Thus, the IDW approach computes an average value for unsampled sites using values from neighbouring weighted sites. The weights are proportional to the vicinity of the specified points to the unsampled site and can be specified by the IDW power coefficient. The larger the power coefficient, the stronger the weight of neighbouring points that estimates the value at an unsampled position.

The IDW method was applied to interpolate the rainfall data at the 171 grid points over the six states of NE India and a map was produced through the process. This interpolation was then used to predict the rainfall for the testing dataset of 33 points (yellow dots in the training data map) (Figure 1). The prediction results were then cross-validated with their actual values, and the errors and efficiency of the process calculated. The original testing dataset of 33 points was also interpolated using this method to create a map. Another dataset was created by joining the two datasets, i.e., a combination of satellite and gauge data was created and again interpolated to produce a map. Contour maps of all the interpolations were also created. The equation of the IDW method is as follows:

$$P_j = \frac{\sum_{i=1}^{n} P_i / d_i^j}{\sum_{i=1}^{n} 1 / d_i^j}$$

where P_j indicates the precipitation at the unknown point, P_i indicates the precipitation at known ith station, d_i indicates the distance between the known and unknown points, i.e., P_i and P_j, n indicates the number of known stations.

3.2. Ordinary Kriging

OK generates interpolation values by assuming a known but constant mean value and allowing for local impacts due to nearby values (Wackernagel, 1995). This method assumes that the trend is stable only within a small area. This theory is noteworthy since it assures that regional differences in the field are taken into account. Similar to that of the IDW method, the training dataset of rainfall at the 171 grid points was processed through OK in ArcGIS and a map of its interpolated predictions produced. The prediction through this process was then used to make rainfall predictions at the 33 testing points and the results compared with their original values. OK was also used to interpolate the testing dataset separately and was also executed for the training and testing combined dataset of 171 and $33 = 204$ points.

$Z_j = \sum_{i=2}^{n} \lambda_i Z_i$ is the equation for OK. where Z_j indicates the precipitation at the unknown point, Zi indicates the precipitation at known ith station, n indicates the number of known stations, λi indicates the unknown weight for the measured value at the ith location. For OK, $\lambda = 1$.

OK was chosen in this study over Simple Kriging since its predictions are more accurate than those of Simple Kriging and it does not require knowledge of mean stationarity throughout the entire area. Our dataset is free of trend, i.e., shows no trend locally. Hence Universal Kriging was dismissed. For Cokriging to be executed, each variable should have clear strong and well-defined corelations with each other for both co-variations and cross-variations. Hence, Cokriging was also ruled out. Also, the Gaussian model of semi-variogram has been chosen to be used in this study since it follows the Normal Probability Distribution function. The Normal Probability Distribution function is advantageous for its continuous distribution and symmetric bell curve with the observations clustered around the central peak. It accurately expresses the value distribution for several natural phenomena.

OK was applied to the training dataset (171) points and testing (33) points were predicted (yellow points) from the interpolation of the former set. Similarly, the testing points and a combined dataset of training and testing points (204) was also separately interpolated. Contour maps were also produced for these along with the interpolation maps.

3.3. Generalised additive models

Trevor Hastie and Robert Tibshirani (Da Silva Monteiro *et al.*, 2022) created GAMs in 1990 to combine the properties of generalised linear models with additive models (Hastie & Tibshirani, 1990). GAMs ignore the requirement that the relationship be a simple weighted sum, assuming instead that the outcome can be

represented by a sum of arbitrary functions of each characteristic (Matheron, 1973). GAMs hence allow the modelling of non-linear data while keeping it simple.

A spline is the name for this flexible function. Splines are multi-dimensional functions that let us represent non-linear interactions for each feature. A GAM is formed by the sum of multiple splines. We can use a non-linear combination of variables, represented by s for spline or 'smooth function,' instead of making the assumption that our aim can be determined using a linear combination of variables. The equation for GAM, where 's' is a smooth function. The general equation for GAM is represented by, $Z = s_0x_0$ and s_1x_1 and s_2x_2 and ⋯ ⋯ ⋯ ⋯ and s_nx_n, where s_0 is the value of Z when all of the independent variables (x_1 through x_n) are equal to zero, i.e., $x_0 = 0$, and s_1 through sn are the estimated regression coefficients, where Z is the predicted or expected value of the dependent variable, x_1 through x_n are n distinct independent or predictor variables. Each regression coefficient indicates the change in Z when the independent variable is changed by one unit.

A total of 40 trial models were tried out for determination of optimum predictions and results. Attributes were also checked for results without spline functions in some of the trials. An optimum number of splines denoted by k were determined as 15 for elevation and 10 for latitude and longitude each. The same model was then also applied on the testing data to check for its errors and efficiency. The GAM iterations were executed by default as Gaussian distributions. Chosen Model for GAM: average of maximum annual rainfall was predicted as a dependent variable, depending on a combination of independent attributes of s (Latitude, $k = 10$) and s (Longitude, $k = 10$) and s (Elevation, $k = 15$), where s is the spline function. The models were first tried out on the training dataset and after the final model was chosen, it was then verified over the testing dataset for validation. The predictions through GAM for both training and testing datasets were then processed by OK for further interpolation and creation of respective predicted maps. Also, through the training map, predictions at the testing data points (33 points) were also made. The same GAM model was also used on the training and testing dataset of 204 points and the predictions run through OK to create a map.

3.4. Statistical criteria

The rainfall values at the testing data points were predicted from interpolation of the values of the training dataset by IDW and OK. For GAM, prediction was made twice, firstly directly from the chosen model for validation, and secondly from interpolation by OK of the GAM prediction of the training dataset. These predictions were checked for mean absolute error (MAE), root mean square error (RMSE), coefficient of determination, R^2 (COD) and NSE. These are given by the following equations.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i) \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \tag{2}$$

$$R^2 = \frac{\sum \hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \tag{3}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{m}(\hat{y}_i - \bar{y})^2}{\sum\left(\sum_{i=1}^{m}(\widehat{y}_i - \bar{y})^2\right)} \tag{4}$$

n indicates the number of data points, y_i indicates the actual output value at ith grid point, $\hat{y}_i$ indicates the predicted output value at ith grid point and $\bar{y}$ indicates the mean of observed values, y_i.

The lower the MAE and RMSE, the higher the accuracy of prediction. A COD result of 0% would mean that the calculation completely fails to accurately describe the data, while a result of 100% denotes a flawless match and is therefore a very credible model for future forecasts. NSE > 75% denotes extremely good performance, 65% < NSE < 75% denotes good performance, 50% < NSE < 65% denotes satisfactory performance, and NSE < 50% denotes unsatisfactory performance.

4. RESULTS AND DISCUSSIONS

For IDW, the training maps as expected are much more detailed than the testing maps since the testing data points are very few as compared to the training points. However, the maps for the combined data of training and testing are the most detailed among all the maps of Figure 3. Comparing the training data map in Figure 3 by IDW and the training data map by OK in Figure 4, we can see that both the maps are similar in their interpolation, but interpolations by OK are more intricate in the sense that some of the locations showing only a few contours for IDW are divided into a greater number of contours for OK for the same dataset, showing a more detailed prediction.

The training maps in Figure 5 show almost parallel contours of values, being highest in the west and gradually decreasing as we move towards the east. However, although parallel, the curved contours are seen to have definitely addressed the issue of linearity. A factor which might have influenced the interpolated map to have parallel contours might be the usage of satellite data instead of more accurate ground rainfall data, since satellite data is also a form of interpolated dataset and also the fact that the equation of GAM is actually a modification of the equation of linearity. The testing maps are of a more realistic nature than that. However, due to availability of fewer known points (33), the interpolated map is still much more inaccurate. Training and testing data produces a much more natural map than that of the training dataset, which foregoes the almost parallel contouring. This is because it is a mix of satellite (training) and rain gauge (testing) data.

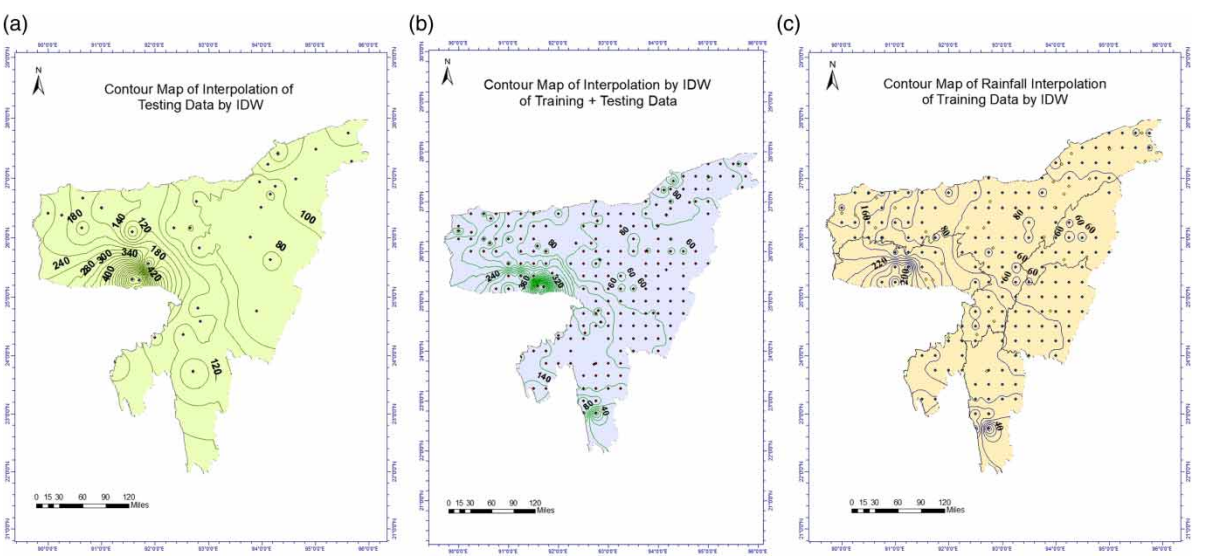

Fig. 3 | Interpolated rainfall maps by IDW.

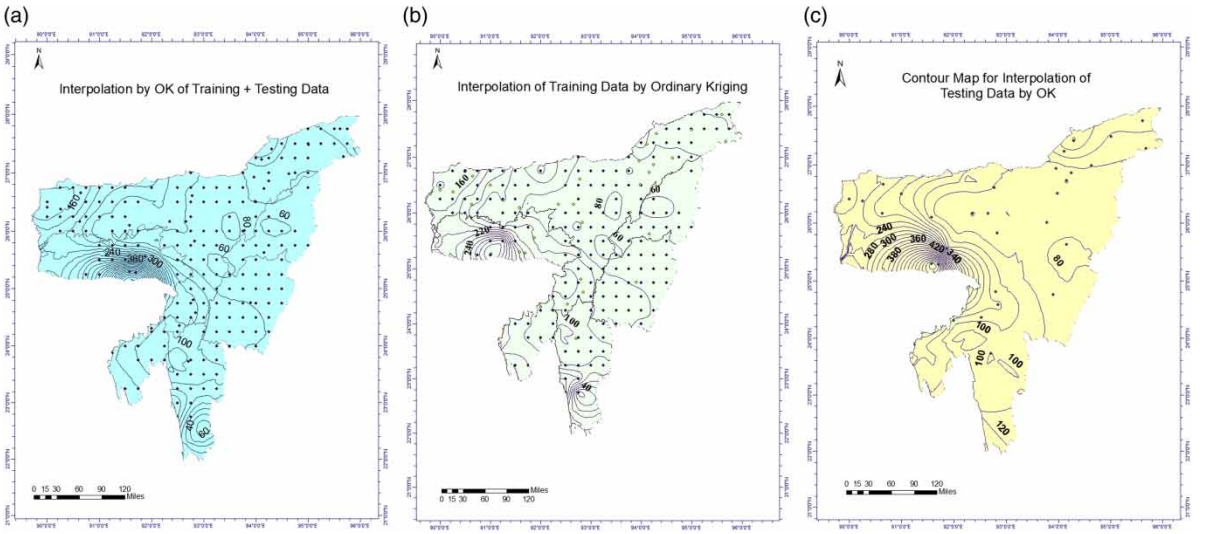

Fig. 4 | Interpolated rainfall maps by OK.

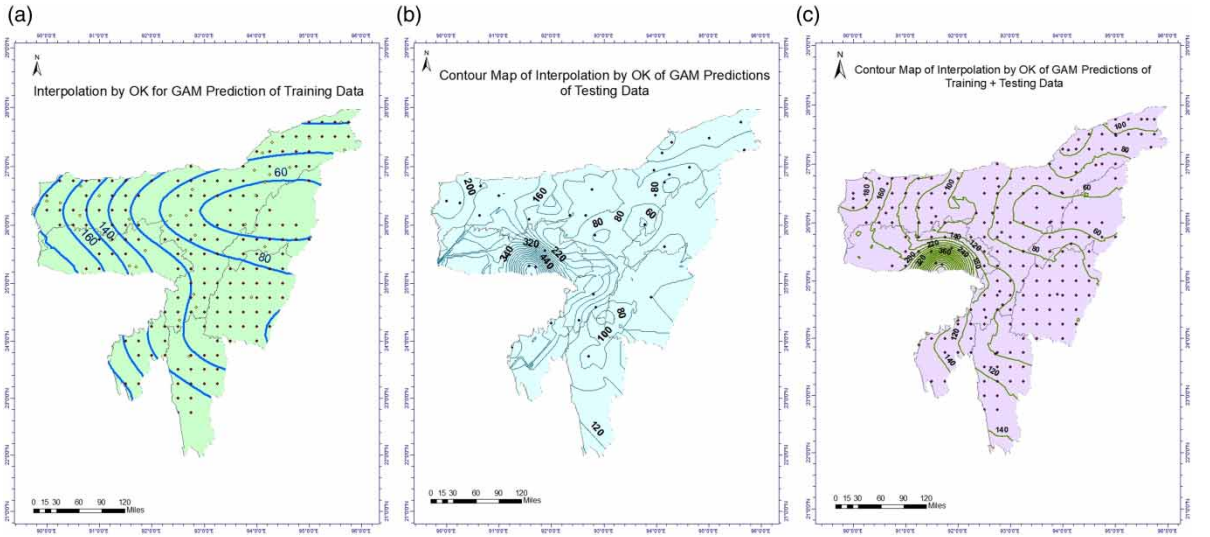

Fig. 5 | Interpolated rainfall maps by OK for GAM predictions.

Table 1 shows that for predictions, GAM has the least MAE and also RMSE in a comparatively low range. OK shows exceptionally low RMSE value but a significant MAE value. MAE and RMSE for IDW and GAM predictions and OK interpolations for the testing dataset are in the same ranges and also significantly high. COD and NSE for GAM are predictably above 90% and hence proved to be very efficient. Since RMSE of OK is quite low, its NSE is correspondingly high. However, its COD is only 30% and below acceptable range. IDW and GAM and OK also have low CODs and extremely low NSEs, thus indicating these techniques to be highly inefficient.

Table 1 | Statistical criteria of various predictions.

Dataset	Method	MAE (mm)	RMSE (mm)	COD, R^2	NSE (%)
Testing (33 points)	IDW	52.47	107.89	31.10	4.69
Testing (33 points)	OK	51.63	10.16	30	99.15
Testing (33 points)	GAM	17.70	23.71	90.40	95.40
Testing (33 points)	GAM and OK	52.98	110.35	31.40	0.29

5. CONCLUSIONS

(i) The maps for data interpolation by OK are substantially more detailed and accurate than those for IDW or GAM and OK.

(ii) Interpolated OK maps for the GAM exhibits curved contours which are almost parallel to each other. The use of splines in the linear model allows GAMs to get around linearity constraints. GAMs automatically learn a non-linear relation between each dependent variables and the outcome variable, and then add these effects, as well as the intercept, linearly. This style of interpolation is clearly inaccurate, because natural phenomena such as rainfall do not occur in such precise and orderly variations.

(iii) Comparing the MAE and RMSE of the testing data prediction set of IDW, OK, GAM and GAM and OK, it is evident that GAM has the least error overall. IDW and GAM and OK have high MAE and RMSE. RMSE is the lowest for OK, while MAE is still on the higher side. Also, MAE $<$ RMSE for all the processes.

(iv) OK shows exceptional good NSE of above 99%, followed by GAM with NSE above 95%. Although OK has high NSE, its COD is much lower and hence cannot be considered very effective. GAM on the other hand also has a high COD of 90.4% for the testing data predictions. The other methods fail to show such considerable good performance as GAM. Even GAM and OK has 31.4% COD and only 0.29% NSE. IDW has moderate COD and NSE.

(v) Hence, GAM is determined to be the best method of rainfall prediction in our study area, while OK remains the best method for interpolated mapping, among the methods used in this study.

6. FUTURE SCOPE OF THE STUDY

Fine-tuning of GAM: Future research could explore the optimisation and fine-tuning of GAMs to address the issue of curved and almost parallel contours observed in the interpolated maps. This could involve experimenting with different model configurations or regularisation techniques to improve the accuracy of GAM-based predictions.

Hybrid approaches: Researchers might investigate hybrid approaches that combine the strengths of both OK and GAM. For instance, integrating the flexibility of GAM for prediction with the spatial accuracy of OK for interpolation could lead to improved rainfall mapping techniques.

Advanced machine learning models: Consideration of more advanced machine learning models beyond GAM could be explored. Researchers might explore deep learning techniques or ensemble methods to further enhance the accuracy of rainfall predictions.

Spatial covariates: Expanding the list of covariates used in the analysis could lead to improved predictions. Incorporating additional spatial covariates, such as land use, land cover, or vegetation indices, could provide more comprehensive input data for the models.

Temporal analysis: Extending the study to include temporal analysis of rainfall patterns could provide valuable insights into seasonal variations and long-term trends. Time series analysis and forecasting models could be applied to enhance predictive capabilities.

Climate change considerations: Considering the potential impact of climate change on rainfall patterns in the study area could be a critical future research direction. Investigating how climate change might affect the accuracy of the chosen methods is essential for more robust predictions.

Data integration: Integrating data from various sources, including remote sensing data, weather radar, and climate models, could enhance the accuracy and spatial resolution of rainfall predictions.

Visualisation techniques: Research could explore advanced visualisation techniques to represent rainfall patterns more intuitively and accurately. This could aid in conveying complex information to stakeholders effectively.

Validation studies: Conducting additional validation studies in different regions or under varying climatic conditions could validate the findings and determine the generalizability of the chosen methods.

Real-Time applications: Investigating the feasibility of real-time rainfall prediction and mapping using the identified methods could have practical applications in disaster management, agriculture, and water resource management.

Overall, there are several promising future directions of this research, including method refinement, model optimisation, and broader applications in the context of climate and environmental studies.

FUNDING

Funding data are not available.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories: https://www.imdpune.gov.in/cmpg/Griddata/Rainfall_25_Bin.html.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Adhikary, S. K., Muttil, N. & Yilmaz, A. G. (2017). Cokriging for enhanced spatial interpolation of rainfall in two Australian catchments. *Hydrological Processes 31*(12), 2143–2161.

Agarwal, S., Roy, P. J., Choudhury, P. S. & Debbarma, N. (2021a). River flow forecasting by comparative analysis of multiple input and multiple output models form using ANN. *H2Open Journal 4*(1), 413–428. https://doi.org/10.2166/h2oj.2021.122.

Agarwal, S., Roy, P. J., Choudhury, P. & Debbarma, N. (2021b). Flood forecasting and flood flow modeling in a river system using ANN. *Water Practice and Technology 16*(4), 1194–1205. https://doi.org/10.2166/wpt.2021.068.

Agarwal, S., Roy, P., Choudhury, P. & Debbarma, N. (2022). Comparative study on stream flow prediction using the GMNN and wavelet-based GMNN. *Journal of Water and Climate Change 13*(9), 3323–3337. https://doi.org/10.2166/wcc.2022.226.

Borges, P. D. A., Franke, J., da Anunciação, Y. M. T., Weiss, H. & Bernhofer, C. (2016). Comparison of spatial interpolation methods for the estimation of precipitation distribution in Distrito Federal, Brazil. *Theoretical and Applied Climatology 123*(1), 335–348.

Box, G. E. P. & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association 65*(332), 1509–1526.

da Silva Monteiro, L., de Oliveira-Júnior, J. F., Ghaffar, B., Tariq, A., Qin, S., Mumtaz, F., Correia Filho, W. L. F., Shah, M., da Rosa Ferraz Jardim, A. M., da Silva, M. V., de Barros Santiago, D., Barros, H. G., Mendes, D., Abreu, M. C., de Souza, A., Pimentel, L. C. G., da Silva, J. L. B., Aslam, M. & Kuriqi, A. (2022). Rainfall in the urban area and its impact on climatology and population growth. *Atmosphere 13*(10), 1610. https://doi.org/10.3390/atmos13101610.

Ditthakit, P., Pinthong, S., Salaeh, N., Binnui, F., Khwanchum, L., Kuriqi, A., Khedher, K. M. & Pham, Q. B. (2021). Performance evaluation of a Two-Parameters monthly rainfall-runoff model in the Southern Basin of Thailand. *Water 13*(9), 1226. https://doi.org/10.3390/w13091226.

Hastie, T. J. & Tibshirani, R. J. (1990). Generalized additive models. In *Monographs on Statistics and Applied Probability* (Bunea, F., ed.). Routledge, New York, Vol. 43, p. 335.

Kendall, M. (1975). *Rank Correlation Methods*. 4th edn. Griffin, San Francisco, CA, London, Vol. 8, p. 875.

Lemus-Canovas, M., Lopez-Bustins, J. A., Trapero, L. & Martin-Vide, J. (2019). Combining circulation weather types and daily precipitation modelling to derive climatic precipitation regions in the Pyrenees. *Atmospheric Research 220*, 181–193.

Link, R. F. & Koch, G. S. (1970). *Experimental Designs and Trend-Surface Analsysis, Geostatistics, A Colloquium*. Plenum Press, New York.

Ljung, G. M. & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika 65*(2), 297–303.

Mann, H. B. (1945). Nonparametric tests against trend. *Journal of Economics 13*(3), 245–259.

Markuna, S., Kumar, P., Ali, R., Vishwakarma, D., Kushwaha, K., Kumar, R., Singh, V., Chaudhary, S. & Kuriqi, A. (2023). Application of innovative machine learning techniques for long-term rainfall prediction. *Pure and Applied Geophysics 180*, 335–363. doi:10.1007/s00024-022-03189-4.

Matheron, G. (1973). The intrinsic random functions, and their applications. *Advances in Applied Probability 5*, 439–468.

Pai, D. S., Rajeevan, M., Sreejith, O. P., Mukhopadhyay, B. & Satbha, N. S. (2014). Development of a new high spatial resolution (0.25× 0.25) long period (1901–2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region. *Indian Journal Meteorology, Hydrology and Geophysics 65*(1), 1–18.

Singh, S. K., Vishwakarma, D. K., Kashyap, P. S., Al-Ansari, N., Kumar, A., Kumar, P., Kumar, R. & Elbeltagi, A. (2022). Soil erosion control from trash residues at varying land slopes under simulated rainfall conditions. doi:10.20944/preprints202204.0302.v1. PPR:PPR488937.

Underwood, F. M. (2009). Describing long-term trends in precipitation using generalized additive models. *Journal of Hydrology (Amsterdam, Neth.) 364*(3–4), 285–297.

Wackernagel, H. (1995). Ordinary kriging. In: *Multivariate Geostatistics* (Wackernagel, H., ed.). Springer, Berlin, Heidelberg.

First received 27 March 2023; accepted in revised form 25 October 2023. Available online 11 November 2023