Research Article

Statistical Learning-Based Spatial Downscaling Models for Precipitation Distribution

Yichen Wu, Zhihua Zhang, M. James C. Crabbe, and Lipon Chandra Das

1Climate Modeling Laboratory, School of Mathematics, Shandong University, Jinan 250100, China
2MOE Key Laboratory of Environmental Change and Natural Disaster, Beijing Normal University, Beijing 100875, China
3Wolfson College, Oxford University, Oxford OX2 6UD, UK
4Institute of Biomedical and Environmental Science & Technology, University of Bedfordshire, Luton LU1 3JU, UK
5School of Life Sciences, Shanxi University, Taiyuan 030006, China
6University of Chittagong, Chittagong 4331, Bangladesh

Correspondence should be addressed to Zhihua Zhang; zhangzhihua@sdu.edu.cn

Received 7 April 2022; Revised 12 May 2022; Accepted 20 May 2022; Published 7 June 2022

Copyright © 2022 Yichen Wu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The downscaling technique produces high spatial resolution precipitation distribution in order to analyze impacts of climate change in data-scarce regions or local scales. In this study, based on three statistical learning algorithms, such as support vector machine (SVM), random forest regression (RF), and gradient boosting regressor (GBR), we proposed an efficient downscaling approach to produce high spatial resolution precipitation. In order to demonstrate efficiency and accuracy of our models over traditional multilinear regression (MLR) downscaling models, we did a downscaling analysis for daily observed precipitation data from 34 monitoring sites in Bangladesh. Validation revealed that $R^2$ of GBR could reach 0.98, compared with RF (0.94), SVM (0.88), and multilinear regression (MLR) (0.69) models, so the GBR-based downscaling model had the best performance among all four downscaling models. We suggest that the GBR-based downscaling models should be used to replace traditional MLR downscaling models to produce a more accurate map of high-resolution precipitation for flood disaster management, drought forecasting, and long-term planning of land and water resources.

1. Introduction

Global warming is significantly influencing the environment, hydrology, and ecosystem. Continued warming in the 21st century will significantly impact precipitation and monsoons and lead to the intensification of extreme rainstorm and drought events [1–3]. South Asia is a well-known summer monsoon region. The formation of the South Asian monsoon is mainly caused by the seasonal movement of the pressure belt and wind belt, as well as the influence of thermal differences between land and ocean as well as topographic factors. About 80% of precipitation in South Asia are closely linked with monsoons [4–6]. More than one billion people rely on monsoonal rainfall for agricultural production, hydroelectric generation, and other basic needs [7]. Especially, Bangladesh is located in one of the largest deltas in the world with a dense network of main rivers and their tributaries, resulting Bangladesh being a flood prone country. Due to the reliance on rain-fed agriculture, Bangladesh is extremely sensitive and high vulnerability to climate change. Since Bangladesh has only few and sparse precipitation monitoring stations, it is very important to generate high spatial resolution precipitation data to mitigate climate change impacts. However, only very limited downscaling research in Bangladesh was carried out by now: observed precipitation data in Bangladesh were downscaled by using multilinear regression as the core part of the downscaling algorithm [8, 9]. Since the nonlinear relation between the large and small-scale dynamics in these research studies was ignored, the obtained downscaling accuracy is unstable. Simulated precipitation data from ensemble climate models in Coupled Model...
Intercomparison Project Phase 5 (CMIP5) were down-scaled by using the method of model output statistics [10, 11], but this method can only be applied for simulated climate data.

Generally, for any country with few and sparse precipitation monitoring stations, downscaling is the key technique to generate high spatial resolution precipitation data. Downscaling can be divided into dynamical downscaling and statistical downscaling. Dynamic downscaling mainly depends on physical principles governing the climate system and high-resolution regional climate models, while statistical downscaling is based on statistical relation between local variables and large-scale variables [11]. Compared with dynamic downscaling relying on some local scale models or regional climate models, statistical downscaling uses a multilinear regression model to establish the correlation between local variables and large-scale variables. Since Earth’s climate is a complex, the multidimensional multiscale system with different physical processes acting on different temporal and spatial scales, statistical downscaling cannot reveal complex nonlinear relationships between local variables and large-scale variables [12,13].

Compared with traditional statistical techniques, advanced statistical learning techniques have showed excellent performance on solving problems with complex nonlinear correlations between variables [14]. Statistical learning techniques can map the predictor(s) only rely on the existing relationship between the two rather than the explicit function [15]. Main statistical learning techniques include the following. (a) Support vector machine (SVM) uses a kernel function to map features to a high-dimensional space for classification and regression; the main advantage lies in that SVM can effectively solve small-sample, nonlinear and high-dimensional regression problems. (b) Random forest (RF) is an ensemble learning method based on bagging, which can handle classification and regression problems well. (c) Gradient boosting regressor (GBR) is an ensemble learning model based on boosting, which reduces the loss by fitting the residuals to obtain high prediction accuracy. Compared with other statistical learning techniques (e.g., neural networks), the SVM requires only small amount of samples and RF and GBR can avoid over fitting [13], so in this study, based on SVM, RF, and GBR, we propose a new downscaling approach to produce a finer spatial resolution precipitation map. In order to demonstrate efficiency and accuracy of our models over traditional multilinear regression (MLR) downscaling models, we use a downscaling analysis for daily observed precipitation data from 34 monitoring sites in Bangladesh. Moreover, based on obtained high spatial resolution precipitation distribution, we analyzed patterns and trend of Bangladesh’s precipitation from 1989 to 2018.

2. Downscaling Methods

Based on three statistical learning algorithms, such as support vector machine (SVM), random forest regression (RF), and gradient boosting regressor (GBR), we proposed an efficient downscaling approach to produce high spatial resolution precipitation, especially for any country with few and sparse precipitation monitoring stations.

2.1. Three Known Statistical Learning Algorithms. Support vector machine (SVM) can map the complex data features into a high-dimensional space by using nonlinear mapping algorithms and separate data using optimal linear hyper-plane [16–18]. For given \( n \) training data \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\), the SVM is to find a regression function \( f(x) = \omega \Phi(x) + b \), such that \( f(x) \) has at most \( \varepsilon \) deviation from the actual value \( y_i \), where \( \Phi \) is a kernel function mapping the input data to a high-dimensional space, and the parameters \( \omega \) and \( b \) are the weight term and bias term, respectively. The basic algorithm to search \( f(x) \) is to minimize the regression risk by the following formulas:

\[
R_{reg}(f) = \frac{1}{2} \omega^2 + C \sum_{i=1}^{n} \Gamma \left( f(x_i) - y_i \right),
\]

subject to\( y_i - \omega \Phi(x_i) - b \leq \varepsilon \quad i = 1, 2, \ldots, n \),

where \( \Gamma(\cdot) \) is a cost function, and the parameter \( C \) can balance the prediction error and model complexity to avoid the overfitting of training data.

Random forest (RF) uses the bagging (or bootstrap aggregation) technique and decorrelation technique to combine a series of small-scale decision trees into a single procedure for better regression prediction [19]. RF can overcome the disadvantage of single decision tree in overfitting to training data and can handle data with few missing values. By using one in a randomly chosen subset of \( m \) predictors from a total of \( n \) predictors, a new node in a decision tree of RF can be generated, where the bootstrap resampling technique is used to randomly select \( k \) samples from \( N \) original training samples as its training set, and the remaining \( N-k \) samples (i.e., out-of-bag samples) are used for cross validation. Each decision tree is only trained by \( m \) predictors and \( k \) training samples, and different decision trees are generated by different predictors and training samples which are randomly chosen. In order to reduce the variance of prediction results by decision trees, the optimal prediction by RF is the average of the predictions from all decision trees (i.e., so-called the aggregate procedure). The prediction accuracy and computing efficiency of RF models are mainly affected by the number of decision trees and the number of predictors/training samples in each decision tree [20].

Gradient boosting regressor (GBR) is an ensemble regression tree model which starts from a simple regression tree and adds a new regression tree again and again [21]. The GBR is a weighted sum of regression trees:

\[
F(x) = \sum_{m=1}^{M} \gamma_m h_m(x),
\]

where \( h_m(x) \) is a \( m^{th} \) regression tree for boosting prediction accuracy. The core procedure in GBR is to continuously...
reduce the loss by searching optimal parameters in the new regression tree to fit the negative gradient of the residual error of existing ensemble regression tree model. In detail, the $F(x)$ in GBR can be estimated through an iterative procedure by using the following formula.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x). \quad (4)$$

During each iteration, a new regression tree $h_m(x)$ is constructed to minimize the residual error by using a gradient descent method. The output of GBR can achieve better generalization performance than a single regression tree [22]. The idea behind GBF is very different from RF. The RF is to build all regression trees in parallel and the output of RF is the average of prediction results from all decision trees, while GBR is to build regression trees in a form of sequence and the output of GBR is the sum of prediction results from all regression trees.

2.2. Statistical Learning-Based Downscaling Technique. The widely used statistical downscaling techniques are usually based on traditional multiple linear regression (MLR), which cannot effectively deal with the instability of downscaling time series and the existence of collinearity between downscaling factors and makes the improvement of downscaling performance significantly limited. In this study, based on GBR, RF, and SVM, we propose an efficient downscaling method to produce high spatial resolution precipitation, where daily station-level precipitation data and longitude/latitude/altitude are used as the input of GBR/SVM/RF models. The output is the downscaled precipitation product. Our downscaling models can largely make up for the deficiencies of the MLR downscaling approach.

For the validation of our downscaling method, noticing that available observed precipitation data are small scale, and in order to avoid overfitting and use as much data as possible in model training, we utilized the 5-fold cross validation method [23]. The main model training process was to divide all data into five subsets; each time one subset was used for the test set and the remaining four subsets were used for training set, and finally, the average of five training errors is used as the result. The correlation of determination ($R^2$), mean absolute error (MAE), and root mean square error (RMSE) are used to assess the performance of different downscaling models. To demonstrate accuracy and efficiency of our models with traditional MLR downscaling models, we used a downscaling analysis for daily observed precipitation data from Bangladesh.

3. Study Area and Data

Bangladesh is located on deltas of large rivers flowing from the Himalayas, leading to that its topography is extremely flat (Figure 1). Traditionally, it is divided into seven regions (Figure 2). High humidity, warm temperature, and wide seasonal variability in precipitation are the main climate characteristics of Bangladesh. This climate is mainly caused by geographic location, north-south continental atmospheric pressure gradient, and fluctuation in terrestrial and sea surface temperature [24].
with a dense river network (Figure 1), floods and related disasters take place frequently [25]. Due to an agriculture-based economy, the high spatial resolution precipitation map can play a key role in Bangladesh’s flood control, drought resistance, and water resource management. Since there are few and sparse precipitation monitoring stations in Bangladesh, it is necessary to conduct a downscaling analysis for observed precipitation data in Bangladesh. To achieve this aim, the daily precipitation data in Bangladesh were obtained from 34 monitoring stations (Figure 3) of the Bangladesh Meteorological Department, and the longitude, latitude, and elevation data of Bangladesh were extracted from Google Earth [26]. Based on statistical learning-based downscaling models in Section 2.2, we can produce high spatial resolution precipitation in Bangladesh.

4. Results and Discussion

4.1. Optimal Statistical Learning-Based Spatial Downscaling Models. Based on daily precipitation data during 1989–2018 and longitude/latitude/elevation data in Bangladesh, we used our statistical learning-based downscaling models to produce higher spatial resolution precipitation data. Table 1 provides the validation results of our models and a traditional MLR downscaling model during 5-fold cross-validation processing. Our downscaling models demonstrate good performance over traditional MLR downscaling models. In terms of $R^2$ value, the downscaled data using GBR and RF showed good consistency with the original observation data. In validation analysis, the GBR downscaling model produced the highest $R^2$ (0.98) and the lowest RMSE (9.63) and MAE (7.24). Figure 4 shows the correlation between the downscaled products and the observed precipitation. IQ he GBR downscaling model yielded the highest performance followed by RF, and the SVM downscaling model ranked the last.

In terms of spatial distribution, our downscaling models were better than the traditional MLR model (Figure 5). The spatial distribution maps of downscaled precipitation produced by GBR and RF are in high agreement with observations. The downscaled precipitation produced by SVM revealed only coarse spatial distribution characteristics: the precipitation gradually increased from western to central regions.

In summary, by using our downscaling model (GBR, RF, and SVM), to simulate the relationship between terrain variables and observed precipitation data in Bangladesh, it is clear that the GBR downscaling model performed best, compared with the RF model, the SVM model, and the traditional MLR model.

4.2. Spatial Variation Analysis of Downscaled Precipitation over Bangladesh. In order to analyze the seasonal variation of precipitation in Bangladesh, we used our GBR downscaling model to produce mean seasonal precipitation distribution during 1989–2018 (Figure 5). Bangladesh has significantly high precipitation during the monsoon season and low precipitation during the remaining three seasons (Figure 6). In the winter season, the precipitation is significantly lower and is close to uniform spatial distribution; in the premonsoon season, the highest precipitation occurs in the middle region; in the monsoon season, higher precipitation occurs in the southwestern and southeastern regions; in the postmonsoon season, the precipitation distribution is particularly uneven and has high spatial variability. Relative dry conditions will occur in the northwestern and central regions.

Using downscaled precipitation by our GBR downscaling model, we demonstrated a difference between the seven regions of Bangladesh (Figure 7). The eastern region showed the highest fluctuation, followed by the southeastern region. The F-statistic value exceeds the critical point in analysis of variance (ANOVA) showing that these regional differences are statistically significant.

Based on the Mann–Kendall trend test and Sen’s slope test (Table 2), eastern, southwestern, southern, and southeastern regions showed upward trends during 1989–2018, but these trends were not significant. The remaining three regions showed downward trends, where only one region showed a statistical significance. Among all seven regions, the northern region showed the highest downward trend with $-13.38$ mm/year, while southeastern region shows the highest upward trend with $4.24$ mm/year.

![Figure 3: Location of 34 monitoring stations.](image-url)
Table 1: Validation results of downscaled precipitation data.

| Regression models | $R^2$ | RMSE  | MAE  |
|-------------------|-------|-------|------|
| MLR               | 0.69  | 440.5 | 353.8|
| SVM               | 0.88  | 275.5 | 132.4|
| RF                | 0.94  | 197.7 | 134.24|
| GBR               | 0.98  | 9.63  | 7.24 |

Figure 4: The correlation between downscaled precipitation by (a) GBR, (b) RF, (c) SVM and observed precipitation, respectively.
Figure 5: Continued.
Figure 5: Continued.
mean annual precipitation (mm)

1084 - 1388
1388 - 1614
1614 - 1781
1781 - 1905
1905 - 2073
2073 - 2298
2298 - 2603
2603 - 3014
3014 - 3569

Figure 5: Spatial distribution of mean annual precipitation by (a) observation, (b) GBR, (c) RF, (d) SVM, and (e) MLR in Bangladesh from 1989 to 2018.
Figure 6: Continued.
Figure 6: Mean seasonal precipitation distribution of (a) winter, (b) premonsoon, (c) monsoon, and (d) postmonsoon in Bangladesh during 1989–2018.

Figure 7: Regional difference in precipitation in Bangladesh.
5. Conclusions

For an agriculture-based country like Bangladesh, water resource contributes the most to agricultural planning. Precipitation plays a more important role on agricultural development than other climatic and environmental variables. It can influence flood disaster management, drought resistance, long-term planning of land and water resources, and different kinds of infrastructure. Therefore, to produce high spatial resolution precipitation data is crucial in analyzing climate change impacts, especially for countries with few and sparse precipitation monitoring stations. Downscaling is an effective technique to solve this issue. The widely used statistical downscaling techniques are usually based on traditional MLR, which cannot effectively deal with the instability of downscaling time series and the existence of collinearity between downscaling factors, and make the improvement of downscaling performance significantly limited. In this study, based on GBR, RF, and SVM, we propose an efficient downscaling approach to produce high spatial resolution precipitation from daily station-level precipitation data and longitude/latitude/altitude data. In order to demonstrate the efficiency and accuracy of our models with traditional MLR downscaling models, we did a downscaling analysis for daily observed precipitation data from 34 monitoring sites in Bangladesh. Our downscaling models have clear advantages over traditional multilinear regression (MLR) downscaling models. The GBR-based downscaling model had the best performance in all four downscaling models. Therefore, we suggest that the GBR-based downscaling models should be used to replace traditional MLR downscaling models to produce a more accurate map of high-resolution precipitation for mitigating impacts of climate disasters, especially South Asian countries with few and sparse precipitation monitoring stations.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

Yichen Wu and Zhihua Zhang are the co-first authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the European Commission’s Horizon 2020 Framework Program (861584) and Taishan distinguished professorship fund.

Table 2: Trend analysis in difference regions of Bangladesh.

| Region          | Northern | Northwestern | Central | Eastern | Southwestern | Southern | Southeastern |
|-----------------|----------|--------------|---------|---------|--------------|----------|--------------|
| $Z^*_T$         | −1.78*   | −1.05        | −1.39   | 0.16    | 0.21         | 0.29     | 0.44         |
| Q (mm/year)     | −13.38   | −9.99        | −11.77  | 1.66    | 2.36         | 3.86     | 4.24         |

90% significance level*.

References

[1] IPCC, *Climate Change 2021: The Physical Science Basis*, IPCC, Geneva, Switzerland, 2021.
[2] M. Latifur, N. Janet, S. A. Mansor et al., “Remote sensing an integrated method for identifying present status and risk of drought in Bangladesh,” *Remote Sensing*, vol. 2020, no. 12, p. 2686, 2020.
[3] S. Shahid and H. Behrawan, “Drought risk assessment in the western part of Bangladesh,” *Natural Hazards*, vol. 46, no. 3, pp. 391–413, 2008.
[4] V. H. Jamshadali, M. J. K. Reji, H. Varikoden, and R. Vishnu, “Spatial variability of south Asian summer monsoon extreme rainfall events and their association with global climate indices,” *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 221, Article ID 105708, 2021.
[5] S. S. C. Shenoi, “Differences in heat budgets of the near-surface Arabian sea and bay of bengal: implications for the summer monsoon,” *Journal of Geophysical Research*, vol. 107, no. C6, pp. 3052–3061, 2002.
[6] A. G. Turner and H. Annamalai, “Climate change and the South Asian summer monsoon,” *Nature Climate Change*, vol. 2, no. 8, pp. 587–595, 2012.
[7] M. R. Rahman and H. Lateh, “Climate change in Bangladesh: a spatio-temporal analysis and simulation of recent temperature and rainfall data using GIS and time series analysis model,” *Theoretical and Applied Climatology*, vol. 128, no. 1-2, pp. 27–41, 2017.
[8] M. Alamgir, S. Hadi Pour, M. Mohsenipour, M. Mehedi Hasan, and T. Ismail, “Predictors and their domain for statistical downscaling of climate in Bangladesh,” *Jurnal Teknologi*, vol. 78, no. 6-12, 2016.
[9] S. A. Shahriar, M. A. M. Siddique, and S. M. A. Rahman, “Climate change projection using statistical downscaling model over Chittagong Division, Bangladesh,” *Meteorology and Atmospheric Physics*, vol. 133, no. 4, pp. 1409–1427, 2021.
[10] J. Fang, J. Du, W. Xu, P. Shi, M. Li, and X. Ming, “Spatial downscaling of TRMM precipitation data based on the orographical effect and meteorological conditions in a mountainous area,” *Advances in Water Resources*, vol. 61, pp. 42–50, 2013.
[11] S. H. Pour, S. Shahid, E.-S. Chung, and X. J. Wang, “Model output statistics downscaling using support vector machine for the projection of spatial and temporal changes in rainfall of Bangladesh,” *Atmospheric Research*, vol. 213, pp. 149–162, 2018.
[12] Z. Zhang and J. Li, (Monograph) *Big Data Mining for Climate Change*, Elsevier, Amsterdam, Netherlands, 2020.
[13] H. Zhang, P. Wu, A. Yin, X. Yang, M. Zhang, and C. Gao, “Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: a comparison of multiple linear regressions and the random forest model,” *The Science of the Total Environment*, vol. 592, pp. 704–713, 2017.
[14] W. Jing, Y. Yang, X. Yue, and X. Zhao, “A comparison of different regression algorithms for downscaling monthly satellite-based precipitation over north China,” *Remote Sensing*, vol. 8, no. 10, pp. 835–917, 2016.
[15] Y. Mei, V. Maggioni, P. R. Houser, Y. Xue, and T. Rouf, “A nonparametric statistical technique for spatial downscaling of precipitation over high mountain Asia,” Water Resources Research, vol. 56, no. 11, Article ID e2020WR027472, 2020.

[16] V. Vapnik, S. E. Golowich, and A. Smola, “Support vector method for function approximation, regression estimation, and signal processing,” Advances in Neural Information Processing Systems, vol. 9, pp. 281–287, 2008.

[17] A. G. Abo-Khalil and D. C. Lee, “MPPT control of wind generation systems based on estimated wind speed using SVR,” IEEE Transactions On Industrial Electronics, vol. 55, no. 3, pp. 1489-1490, 2008.

[18] P. Tsangaratos, I. Ilia, and I. Matiatos, “Spatial analysis of extreme rainfall values based on support vector Machines optimized by genetic algorithms,” in Spatial Modeling in GIS and R for Earth and Environmental Sciences, H. R. Pourghasemi and C. Gokceoglu, Eds., Elsevier, Amsterdam, Netherlands, pp. 1–19, 2019.

[19] C. d. S. Chagas, W. de Carvalho Junior, S. B. Bhering, and B. Calderano Filho, “Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions,” Catena, vol. 139, pp. 232–240, 2016.

[20] P. T. Guo, M. F. Li, W. Luo, Q. F. Tang, Z. W. Liu, and Z. M. Lin, “Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach,” Geoderma, vol. 237-238, pp. 49–59, 2015.

[21] N. Bagalkot, A. Keprate, and R. Orderløkken, “Combining computational fluid dynamics and gradient boosting regressor for predicting force distribution on horizontal axis wind turbine,” Vibrations, vol. 4, no. 1, pp. 248–262, 2021.

[22] A. Elnashar, H. Zeng, B. Wu et al., “Downscaling TRMM monthly precipitation using google Earth engine and google cloud computing,” Remote Sensing, vol. 12, no. 23, p. 3860, 2020.

[23] S. Yadav and S. Shukla, ”Analysis of K-fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality classification,” in Proceedings of the IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, India, 2016.

[24] A. Neelim and T. Islam, Climate Change in Bangladesh: a closer look to temperature and rainfall data, 2010.

[25] M. A. A. Hoque, B. Pradhan, and N. Ahmed, “Assessing drought vulnerability using geospatial techniques in north-western part of Bangladesh,” The Science of the Total Environment, vol. 705, Article ID 135957, 2020.

[26] http://earth.google.com/.