# Statistical Learning-based Spatial Downscaling Models for Precipitation Distribution**

Yichen Wu[a], Zhihua Zhang[a,b*], M. James C. Crabbe[c], Lipon Chandra Das[a,d]

a) Climate Modeling Laboratory, Shandong University, Jinan, 250100, China.
b) MOE Key Laboratory of Environmental Change and Natural Disaster, Beijing Normal University, 100875, China
c) Wolfson College, Oxford University, Oxford OX2 6UD, UK
d) University of Chittagong, Chittagong, 4331, Bangladesh

*Corresponding Author: Distinguished Professor Zhihua Zhang, Email: zhangzhihua@sdu.edu.cn

## Abstract

The downscaling technique produces high spatial-resolution precipitation distribution in order to analyze impacts of climate change in data-scarce regions. In this study, based on three statistical learning algorithms: Support Vector Machines (SVM), Random Forest Regression (RF) and Gradient Boosting Regressor (GBR), we propose a novel downscaling approach to produce high spatial-resolution precipitation. In order to demonstrate efficiency and accuracy of our models over traditional multi-linear regression (MLR) downscaling models, we used a downscaling analysis for daily observed precipitation data from 34 monitoring sites in Bangladesh. Validation analysis revealed that the $R^2$ of GBR could reach 0.98, compared with the RF (0.94), SVM (0.88) and multi-linear regression (MLR) (0.69) models. We found the GBR model had the best performance in downscaling analysis.

**Keywords:** Downscaling Analysis; Statistical Learning; Gradient Boosting Regressor; Support Vector Machine, Random Forest.

## 1. Introduction

Global warming is significantly influencing the environment, hydrology and ecosystem. Continued warming in the 21st century will significantly impact precipitation, monsoons and lead to the intensification of extreme rainstorm and drought events (IPCC, AR6). 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

** *The second author was supported by Horizon2020 Program 861584 of European Commission and Taishan Distinguished Professorship Fund.*

between land and ocean as well as topographic factors. About four-fifths of precipitation in South Asia are closely linked with monsoons. The volatile summer monsoon rainfall at different timescales affects human lives across South Asia, e.g. more than one billion people rely on monsoonal rainfall for agricultural production, hydroelectric generation and other basic needs.

For south Asian countries with few and sparse precipitation monitoring stations, it is very important to generate high spatial-resolution precipitation data for analyzing climate change mechanisms and impacts. Dynamical downscaling and statistical downscaling are two techniques which can achieve this aim. 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. Compared with traditional statistical methods (e.g., multi-linear regressions), statistical learning has shown excellent performance on problems with complex nonlinear correlations between variables (Jing et al. 2016) since they can use unknown correlation information hidden in data (Mei et al. 2020). Main statistical learning techniques include:

- Support Vector Machines (SVM) use a kernel function to map features to 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.
- Random Forest (RF) is based on bagging (or bootstrap aggregation) for decision trees. RF can handle classification and regression issues well.
- Gradient Boosting Regressor (GBR) is an ensemble of regression trees through boosting. GBR is to iteratively add a new regression tree to reduce the loss and obtains high prediction accuracy.

In this study, based on SVM, RF and GBR, we propose a new downscaling approach to produce finer spatial-resolution precipitation. In order to demonstrate efficiency and accuracy of our models over traditional multi-linear 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 trends of Bangladesh's precipitation from 1989-2018.

## 2. 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 hyperplane (Vapnik et al. 1997). For given $n$ training data $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, the SVM is to find a regression function

$$f(x) = \langle \omega, \Phi(x) \rangle + b$$

such that $f(x_i)$ 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:

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

$$\text{subject to} \quad |y_i - \langle\omega, \Phi(x_i)\rangle - b| \leq \varepsilon \qquad \mathrm{i} = 1,2,\ldots,\mathrm{n}$$

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 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 (Chagas et al. 2016). 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 the 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 (Guo et al. 2015).

**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. The GBR is a weighted sum of regression trees (Bagalkot et al. 2021):

$$F(x) = \sum_{m=1}^{M}\gamma_m h_m(x)$$

where $h_m(x)$ is a *m*th regression tree for boosting predication 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 the GBR can be estimated through an iterative procedure:

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

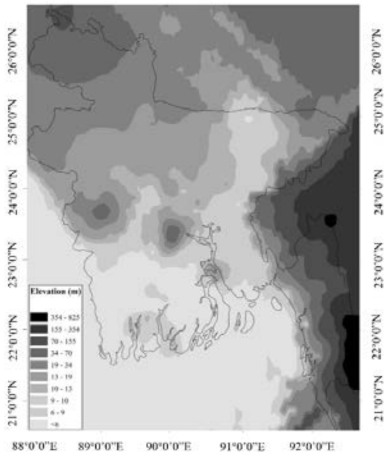
During each iteration, a new regression tree $h_m(x)$ is constructed to minimize the residual error by using a gradient descent method. The output of GBR can achieve better generalization performance than a single regression tree (Elnashar et al. 2020). The idea behind GBF is very different from RF: The RF is to build all regression tree 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.

## 3 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 (Yadav and Shukla 2016). 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 in Bangladesh. The daily precipitation data in Bangladesh was obtained from 34 monitoring sites (Figure 2) of the Bangladesh Meteorological Department. The longitude, latitude and elevation data of Bangladesh were extracted from Google Earth. Due to few and sparse precipitation monitoring stations in Bangladesh, it is necessary to operate a downscaling analysis for observed precipitation data in Bangladesh. Bangladesh has a climate with significantly high precipitation in the monsoon season. Floods and related disasters take place frequently since most of Bangladesh lies in the flat and low delta plain with a dense river network (Figure 1) (Hoque et al. 2020). With an agriculture-based economy, the obtained high spatial-resolution precipitation map can play a key role in Bangladesh’s flood control, drought resistance, and water resource management.



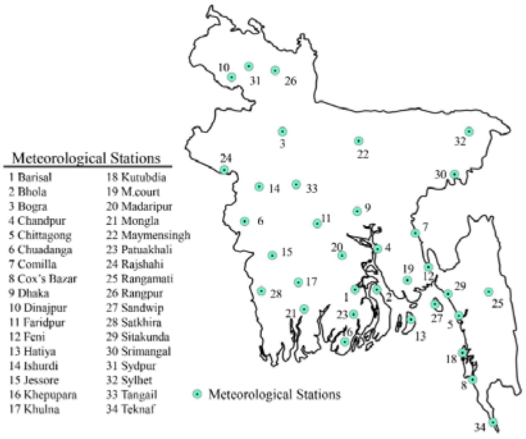


**Figure 1.** *Elevation map of Bangladesh*

**Figure 2.** *Location of 34 Monitoring Sites*

Based on daily precipitation data during 1989-2018 and longitude/latitude/elevation data in Bangladesh, we used our downscaling models to produce higher spatial-resolution precipitation data. Table 1 demonstrates 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 3 presents the correlation between the downscaled products and the observed precipitation. The GBR downscaling method yielded the highest performance followed by the RF, the SVM downscaling model ranked the last.

**Table 1.** *Validation results of our Models and MLR*

| **Regression Models** | $\boldsymbol{R^2}$ | **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 |

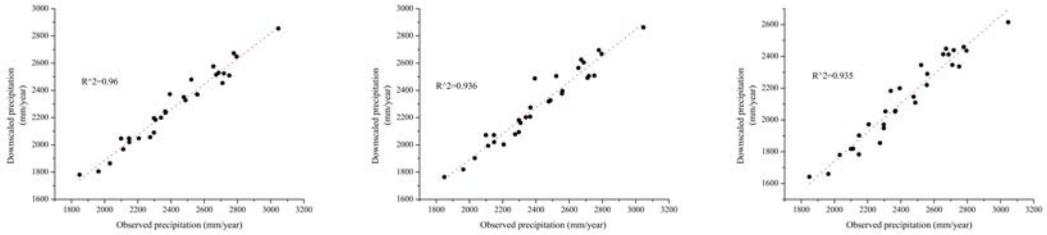


**Figure 3.** *The correlation between downscaled precipitation by GBR/RF/SVM and observed precipitation, respectively*.

In terms of spatial distribution, our downscaling models were better than the traditional MLR model (Figure 4). The spatial distribution maps of downscaled precipitation produced by GBR and RF are in high agreement with observations. The downscaling precipitation produced by SVM revealed only some spatial distribution characteristics of precipitation in western and central regions of Bangladesh.

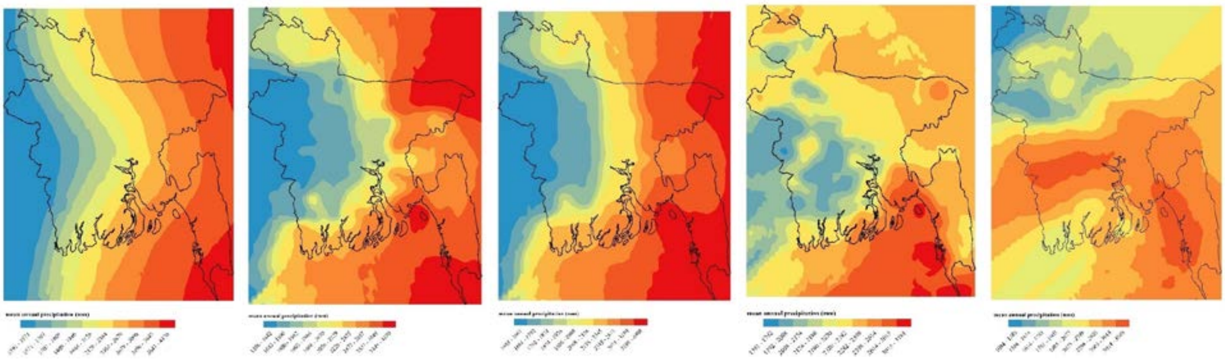


**Figure 4.** *Spatial pattern of mean annual precipitation in Bangladesh from 1989 to 2018. From left to right: (a) observation, (b) GBR, (c) RF, (d) SVM, (e) MLR.*

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 method performed best, compared with the RF model, the SVM model, and the traditional MLR model.

## 4. Spatial Variation Analysis of Downscaled Precipitation in Bangladesh

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 (Islam and Neelim 2010). Bangladesh has four seasons: winter season, pre-monsoon season, monsoon season and post-monsoon season.

*4.1. Seasonal Variations*

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. In the winter season, the precipitation is significantly lower and is close to uniform spatial distribution; in the pre-monsoon season, the highest precipitation occurs in the middle region; in the monsoon season, higher precipitation occurs in the southwestern and southeastern regions; in the post-monsoon season, the precipitation distribution is particularly uneven and has high spatial variability. Drying conditions will occur in the northwestern and central regions.

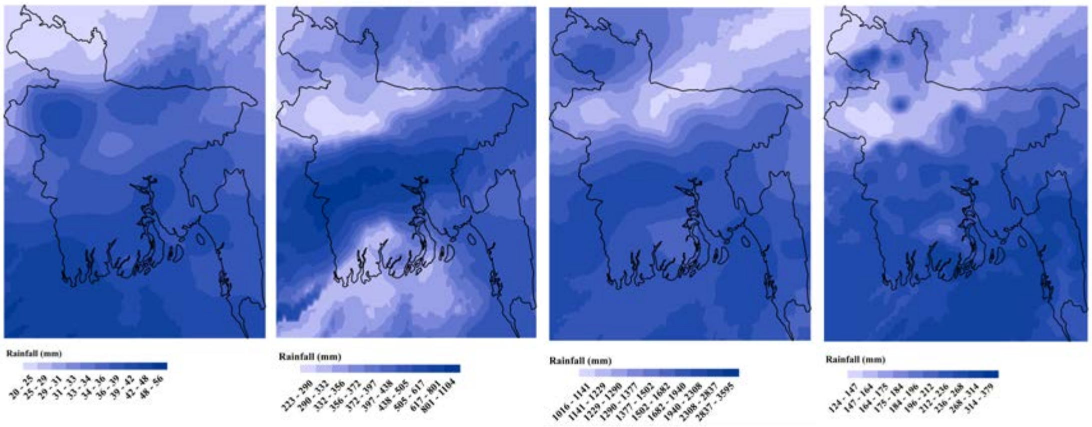
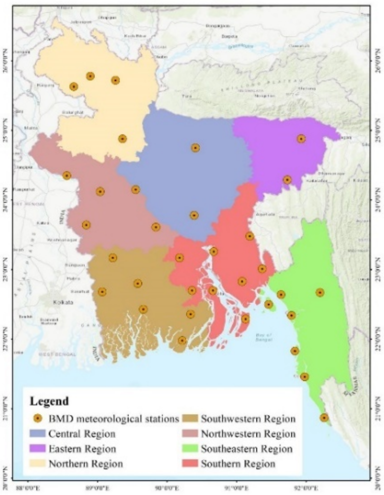


**Figure 5.** *Mean seasonal precipitation distribution in Bangladesh during 1989-2018. From left to right: winter, pre-monsoon, monsoon and post-monsoon.*

*4.2. Regional difference*

Bangladesh consists of seven regions (Figure 6). 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 exceeded the critical point in analysis of variance (ANOVA) (Table 2) showing that these regional differences were statistically significant.



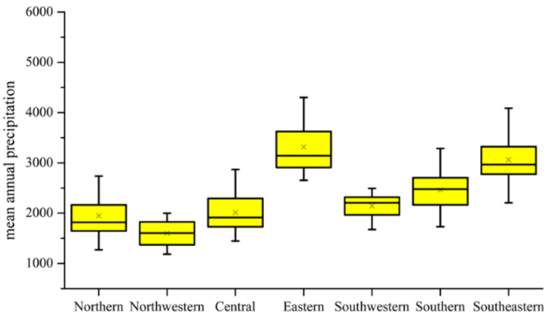


**Figure 6.** *Seven regions of Bangladesh*

**Figure 7.** Regional difference in precipitation in Bangladesh

**Table 2.** *Significance of Regional difference by ANOVA*

| | Sum of Squares | df | Mean Square | F | P-value | F crit |
|---|---|---|---|---|---|---|
| **Between Regions** | 69888694 | 6 | 11648116 | 79.03 | 2.06e-50 | 2.14 |
| **Residuals** | 29921117 | 203 | 147395 | | | |

Based on the Mann–Kendall trend test and Sen's slope test (Table 3), 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 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.

**Table 3**. *Trend analysis in difference regions of Bangladesh*

| **Region** | Northern | Northwestern | Central | Eastern | Southwestern | Southern | Southeastern |
|---|---|---|---|---|---|---|---|
| $Z_S$ | -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 *

## 5. Conclusions

For an agriculture-based countries 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 construction 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 makes 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 used a downscaling analysis for daily observed precipitation data from 34 monitoring sites in Bangladesh. Our downscaling models have clear advantages over traditional multi-linear regression (MLR) downscaling models. The GBR model had the best performance in all downscaling analyses.

## References

Abo-Khalil AG, Lee DC (2008) MPPT Control of Wind Generation Systems Based on Estimated Wind Speed Using SVRR. IEEE Transactions on Industrial Electronics, 55(3), 1489–1490

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., Notarnicola, C. (2015). Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Remote Sensing,7(12), 16,398–16,421. https://doi.org/10.3390/RS71215841

Bagalkot N, Keprate A, Orderløkken R (2021) Combining Computational Fluid Dynamics and Gradient Boosting Regressor for Predicting Force Distribution on Horizontal Axis Wind Turbine. Vibration, 4(1), 248–262

Chagas, Junior, Bhering, SB, Filho BC (2016) Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena 139, 232–240

Denissen AK (2012) Climate change & its impacts on Bangladesh, the National Commission for International Cooperation and Sustainable Development Foundation (NCDO) 63, 1092.

Elnashar A, Zeng H, Wu B, Zhang N, Tian F, Zhang M, … Sun Z (2020) Downscaling TRMM Monthly Precipitation Using Google Earth Engine and Google Cloud Computing. Remote Sensing, 12(23), 3860.

Fang J, Du J, Xu W, Shi P, Li M, Ming X (2013) Spatial downscaling of TRMM precipitation data based on the orographical effect and meteorological conditions in a mountainous area. Advances in Water Resources, 61, 42–50

Guo PT, MF Luo, Tang, & QF, et al (2015) Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. GEODERMA, 2015,237(-), 49-59

Habiba U, Shaw R, Hassan AWR (2013) Drought risk and reduction approaches in Bangladesh.

Disaster Risk Reduction Approaches in Bangladesh, pp:131–164

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2015). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma, 265, 6 2–77.

Hoque MA-A, Pradhan, B., Ahmed N (2020) Assessing drought vulnerability using geospatial techniques in northwestern part of Bangladesh. Sci Total Environ 705:135957

IPCC, Sixth Assessment Report (AR6), 2022

Islam MN, Tamanna S, Amstel A, Noman, M, Ali M, Aparajita DM, … Ashiquzzaman M (2021) Climate Change Impact and Comprehensive Disaster Management Approach in Bangladesh: A Review. Springer Climate, 1–39

Islam T, Neelim A (2010) Climate change in Bangladesh: a closer look into temperature and rainfall data. University Press, Dhaka

Jing W, Yang Y, Yue X, Zhao X (2016) A Comparison of Different Regression Algorithms for Downscaling Monthly Satellite-Based Precipitation over North China. Remote Sensing, 8(10), 835

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22

Martin MP, Wattenbach M, Smith P, Meersmans J, Jolivet C, Boulonne L, Arrouays D (2011) Spatial distribution of soil organic carbon stocks in France. Biogeosciences 8, 1053–1065

Mei, Y., Maggioni, V., Houser, P. R., Xue, Y., & Rouf, T. (2020). A Nonparametric Statistical Technique for Spatial Downscaling of Precipitation over High Mountain Asia. Water Resources Research, 56(11)

Naghibi, S. A., & Pourghasemi, H. R. (2015). A comparative assessment between three machine learning models and their performance comparison by bivariate and multivariate statistical methods in groundwater potential mapping. Water Resources Management, 29(14), 5217–5236.

Pour SH, Shahid S, Chung ES, Wang XJ (2018) Model output statistics downscaling using support vector machine for the projection of spatial and temporal changes in rainfall of Bangladesh. Atmospheric Research, 213, 149–162

Rahman MA, Yunsheng L, Sultana N (2017) Analysis and prediction of rainfall trends over Bangladesh using Mann–Kendall, Spearman’s rho tests and ARIMA model. Meteorology and Atmospheric Physics, 129(4), 409–424

Rodell M, Beaudoing HK, L’Ecuyer TS, Olson WS, Famiglietti JS, Houser PR, … Chambers D (2015) The Observed State of the Water Cycle in the Early Twenty-First Century. Journal of Climate, 28(21), 8289–8318

Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. Isprs

Journal of Photogrammetry and Remote Sensing, 67(67), 93–104

Sarker M (2015) Environmental change and its impact on migration in Bangladesh. University of Sheffield

Sarker M, Nichol J, Siti M, Baharin A, Shamsuddin S, Chung E-S, Reid J, Siswanto E (2020) An integrated method for identifying present status and risk of drought in Bangladesh. Remote Sens 12:17. 2686

Shahid S (2010) Rainfall variability and the trends of wet and dry periods in Bangladesh. International Journal of Climatology, 30(15), 2299–2313

Shahid S, Behrawan H (2008) Drought risk assessment in the western part of Bangladesh. Natural Hazards46(3): 391–413

Shenoi S, Shankar D, Shetye SR (2002) Differences in heat budgets of the near-surface Arabian Sea and Bay of Bengal: Implications for the summer monsoon. Journal of Geophysical Research, 107

Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. Hydrology and Earth System Sciences, 20(7), 2611–2628.

Smola, AJ., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), 199–222

Tsangaratos, P., Ilia, I., Matiatos, I., (2019) Spatial Analysis of Extreme Rainfall Values Based on Support Vector Machines Optimized by Genetic Algorithms: The Case of Alfeios Basin, Greece. Spatial Modeling in GIS and R for Earth and Environmental Sciences, 1-19

Turner AG, Annamalai H (2012) Climate change and the south asian summer monsoon. Nature Climate Change, 2(8), 587-595

Vapnik V, Golowich S, Smola A (1997) Support vector method for function approximation, regression estimation and signal processing. Advances in Neural Information Processing Systems, Vol. 9, MIT Press, Cambridge, MA

Yadav S and Shukla S (2016) Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. 2016 IEEE 6th International Conference on Advanced Computing (IACC). IEEE

Zhang H, Wu P, Yin A, Yang X, Zhang M, Gao C (2017) 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. Science of The Total Environment, 592, 704–713