Paper ID: CE 0424

Forecasting the annual variability of evaporation using ML-based regression techniques

H. M. Rasel[1], M. A. Saleh[2]

[1]Department of Civil Engineering, RUET, Bangladesh (hmrasel@ce.ruet.ac.bd)
[2]Department of Civil Engineering, RUET, Bangladesh (saleh110308@acc.edu.bd)

Abstract

Evaporation is one of the most significant elements of the water budget. In this study, yearly evaporation over the Ganges-Brahmaputra basin was predicted from 2021 to 2030 using historical data from 2004 to 2020. Random Forest (RF), Gradient Boosting Machine (GBM), and Support Vector Machine (SVM), three machine learning-based regression approaches that were employed in the Python. Rajshahi and Rangpur divisions were considered as the study area. According to the statistical analysis, the RF model showed higher accuracy in cases of correlation (92.19% for Rajshahi division and 66.38% for Rangpur division), model accuracy (85% for Rajshahi division and 44% for Rangpur division), MAE (0.07 mm/day for Rajshahi division and 0.05 mm/day for Rangpur division), MSE (0.01 mm per day for Rajshahi division and 0.006 mm per day for Rangpur division), and agreement of index (0.74 for Rajshahi division and 0.95 for Rangpur division). When comparing the RMSE (0.006 mm per day in Rajshahi and 0.003 mm/day in Rangpur) and MAPE (6.87 mm per day in Rajshahi and 7.61 mm per day in Rangpur) in both divisions, SVM demonstrated greater accuracy. Based on the comparative study of the models, the RF can be concluded as the best model. In Rangpur division, the rate of evaporation will be increasing and a downward trend was found in Rajshahi division. The RF regression model can be further succeeded to forecast the evaporation in the Ganges-Brahmaputra basin.

Keywords: *Evaporation; Water Budget Components; Northwest region of Bangladesh; Regression; Machine Learning.*

1 Introduction

Evaporation is one of the most essential components of any hydrologic basin's water budget. The variability in evaporation rate depends on the soil characteristics, climatic conditions, precipitation, rainfall, humidity, sunshine hours, air quality and so on (Yoon et al., 2019). Evaporation contributes directly to the water budget estimation. The trend of evaporation rate doesn't follow any kind of linear or non-linear patterns exactly, rather the trendline that can be achieved from any empirical equation certainly possess some error. While forecasting the water budget over any area, the evaporation rate needs to be forecasted precisely. Numerous studies had been carried out to forecast the evaporation rate globally. Quantitative remote sensing is a highly effective and economically feasible technology that can offer radiometric observations of various physical quantities at a regional to global scale, which are pertinent to the evaluation of evaporation.

Limited research was conducted on the changes in pan evaporation and its possible impact on the water balance in arid locations(Shen et al., 2010). aimed to analyze the patterns of pan evaporation in the dry regions of China for the last five decades and describe the alterations in the water balance within these regions. (Gaybullaev et al., 2012) utilized empirical data on water volume, precipitation, runoff, evaporation, and salinity to make estimations of water volume and salinity spanning the years 1960 to 2009. The predictive accuracy of the estimated water volume and salinity was evaluated using efficiency coefficients of 0.975 and 0.974, respectively. (Trambauer et al., 2014) calculated the actual evaporation rates over the African continent using a continental adaption of the PCR-GLOBWB global hydrological model, which uses a water balance methodology. In their study, (Ferreira et al., 2014) utilized a combination of GRACE and TRMM to examine the monthly estimations of sink terms, namely evaporation and runoff, through the implementation of the water balance equation. The study was conducted over a period spanning from 2005 to 2010 over the Volta Basin. The issue of water scarcity, resulting from factors such as evaporation from bodies of water and the utilization of water for domestic and

agricultural purposes, has become a significant concern, particularly during the summer season. (Wang et al., 2019) devised a satellite-based algorithm utilizing three sources to estimate instantaneous ET rate (ET) in situations where the sky is clear or non-rainy and cloudy. ANN is distinct from other GP-based techniques, such as Gene-expression programming (GEP), as well as non-GP methods like MLP, RBNN, generalized RNN, and Stephens-Stewart (SS) models(Guven & Kişi, 2011). (Malik et al., 2017) employed four heuristic methodologies to approximate monthly pan-evaporation at two sites situated in India. In comparison to other models, the study's findings showed that the CANFIS and MLPNN models were more effective at estimating monthly pan evaporation. (Sudheer et al., 2002) used ANN to predict Class A pan evaporation. The results suggested that the ANN technique can be used successfully, but with significant error in predictions.

To the authors' best knowledge, there isn't any published research evaluating the efficacy of machine learning-based regression algorithms for forecasting evaporation in Bangladesh's northwest. This study aims to identify the RF, GBM, and SVM regression models that offer the most precise forecasting models. Predicting evaporation in Bangladesh's northwest is the main goal of this research since it is essential to understanding the entire water balance. Additionally, plans for managing water supplies for irrigation and putting measures in place to lessen the effects of drought may be developed based on the expected quantity of evaporation. In order to train the RF, SVM, and GBM models, annual data from 2004 to 2020 were used. The models were trained using the supplied training data, and their accuracy was then evaluated by contrasting projected values with actual values.

2 Study Area and Data Collection

The present study was conducted over the northwest region of Bangladesh that comprised of Rangpur and Rajshahi divisions. The Rangpur division is located between 25.8483° N, 88.9414° E and Rajshahi division is located between 24.7106° N, 88.9414° E. The location is comprised of Ganges and Brahmaputra Basin illustrated in the figure 1. Moreover, the figure 1 also described the inland rivers in the study area. It is a highly drought-prone region in Bangladesh. In addition, any specific study on preventing evaporation in this area has not been conducted before. If the evaporation rate can be estimated, the probable water availability will also be calculated for irrigation and drinking purposes.

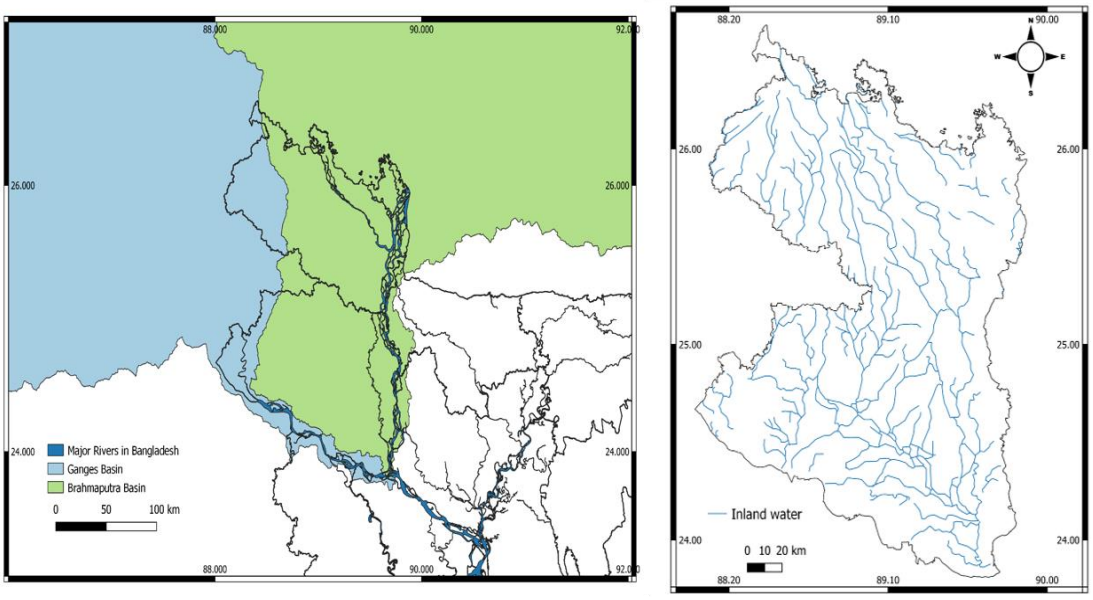

Figure 1. Study area (Northwest region of Bangladesh and the rivers in that region (right))

From the Bangladesh Water Development Board (BWDB) website, evaporation data for the Rajshahi and Rangpur divisions were downloaded. For Rangpur division, the evaporation data was collected from 2003 to 2020 and for the Rajshahi division, the evaporation data was collected from 2004 to 2020. Due to the geographical, climatic, and hydrological varieties between two divisions, the model was developed for two divisions. 30% of the historical evaporation data was chosen as the test data, with the remaining 70% being regarded the training data. The forecast was carried out from 2021 to 2030.

3 Methodology

Using the regression method, three machine learning techniques, namely Random Forest (RF), Gradient Boosting Machine (GBM), and Support Vector Machine (SVM), were employed to forecast evaporation over the study area. All the three models were carried out in Google Collaboratory (https://colab.google/), an open-source python platform for advanced statistical analysis.

3.1 Regression

Statistically, the regression modeling technique is utilized to analyze the correlation between a dependent variable and one or more independent variables. In linear regression, which is the most typical type of regression analysis, the slope and intercept of the equation, which has the shape of a straight line, are estimated. It assists with prediction, improves comprehension and interpretation of data patterns, and sheds light on how variables relate to one another. Mathematically, the simple linear regression can be expressed as the equation (1) below,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1)$$

Where, Y is the predicted value, X is the input variable, βo and β(1) are regression coefficients, and ε(i) is Random error term.

3.2 Random Forest (RF)

To produce predictions for regression assignments, the Random Forest Regression algorithm mixes the concepts of ensemble learning and decision trees. It is a powerful machine learning technique. Nevertheless, the computational cost of this approach may be high, and its interpretability may be inferior to that of more straightforward models such as linear regression. In general, random forest regression is a widely adopted approach for regression tasks owing to its capacity to manage intricate data, mitigate overfitting, and furnish dependable predictions(Dhiman et al., 2020; El-Maghraby et al., 2020).

3.3 Gradient Boosting Machine (GBM)

In predictive modeling and data analysis, the Gradient Boosting Machine (GBM) is a machine learning technique that is frequently employed. The technique is extensively employed in diverse regression applications owing to its capacity to manage intricate associations in the data and generate predictions of superior quality. In order to generate predictions, the input data is sequentially processed through each constituent model within the ensemble, and the resulting predictions are aggregated to yield the ultimate prediction. Mathematically, it can be expressed in equation (2) as follows,

$$\tilde{F}(x) = argmin_{F(x)} L_{y,x}\big(y, F(x)\big) \qquad (2)$$

Where, y is the input variable and $\tilde{F}(x)$ is the predicted value. Nonetheless, the utilization of GBM Regression warrants certain deliberations, including computational complexity and overfitting, thereby necessitating meticulous hyperparameter tuning. In general, GBM Regression is a robust algorithm utilized for regression tasks, renowned for its capacity to manage intricate data, capture non-linear associations, and furnish precise prognostications(Friedman, 2001; Lange & Sippel, 2020).

3.4 Support Vector Machine (SVM)

Support Vector Machines (SVMs) are commonly used for classification and regression tasks since they are supervised machine learning algorithms. The utilization of the margin maximization technique enhances the resilience of Support Vector Machines (SVM) and reduces their susceptibility to outliers. Support Vector Machines (SVM) possess various benefits such as their capability to manage high-dimensional data, proficiency in intricate and non-linear situations, and resilience to overfitting(Cortes & Vapnik, 1995; Raghavendra. N & Deka, 2014; Shrestha & Shukla, 2015).

4 Results and Discussion

Among the three models, the RF model showed best correlation for the study region (92.19% for Rajshahi division and 66.38% for Rangpur division) and the RF model also performed best for the study area while considering the coefficient of determination or R-squared value (85% for Rajshahi division and 44% for Rangpur division). Moreover, RF model showed lowest MAE for the study region (0.07 mm per day for Rajshahi division and 0.05 mm per day for Rangpur division). In addition, the lowest MSE for the study region was shown by the RF model for the evaporation data (0.01 mm per day for Rajshahi division and 0.006 mm per day for Rangpur division). Furthermore, the lowest RMSE for the study region was shown by the SVM model for the evaporation data in Rajshahi with the lowest RMSE (0.006 mm per day) and RF model showed lowest RMSE (0.003 mm per day) for Rangpur division. The lowest MAPE for the study region was shown by the SVM regression model for the evaporation in Rajshahi with the lowest MAPE (6.87 mm per day) and RF regression model showed lowest MAPE value (7.61 mm per day) for Rangpur division. RF model showed best agreement of index for the study

region (0.74 for Rajshahi division and 0.95 for Rangpur division). From the abovementioned analysis, the RF model showed the highest accuracy among the three regression models. In the Rangpur division, the trend of evaporation rate is upward while the Rajshahi division has a downward trend. In the figure 2, the comparison among the three predicted model values and observed values was illustrated. However, the table 1 includes accuracy parameters for the three models used in this study.

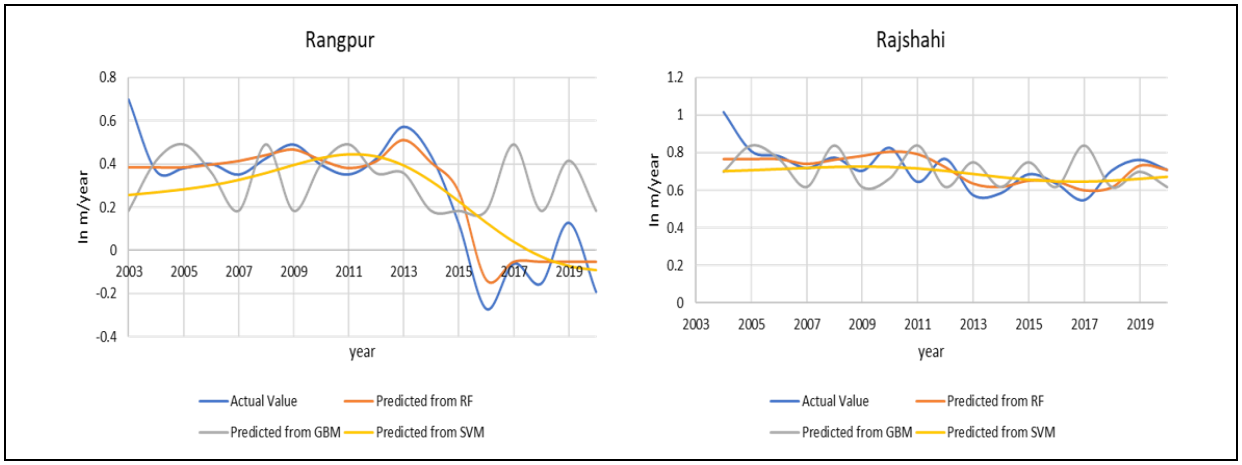

Figure 2. Observed values versus predicted values from RF, GBM, and SVM regression

Table 1: Model Accuracy parameters in the Rangpur and Rajshahi divisions

Regression Methods	Random Forest		Gradient Boosting Machine		Support Vector Machine	
Accuracy Parameters	**Rajshahi**	**Rangpur**	**Rajshahi**	**Rangpur**	**Rajshahi**	**Rangpur**
Correlation	0.92	0.66	0.2	-0.04	0.79	0.46
R^2	0.85	0.44	0.04	0.01	0.62	0.21
MAE	0.07	0.06	0.22	0.12	0.13	0.08
MSE	0.01	0.01	0.08	0.02	0.03	0.01
RMSE	0.01	0	0.04	0.02	0.01	0.01
MAPE	14.02	7.61	-37.4	16.35	6.87	10.49
Agreement of Index, d	0.74	0.95	-0.07	0.25	0.36	0.83

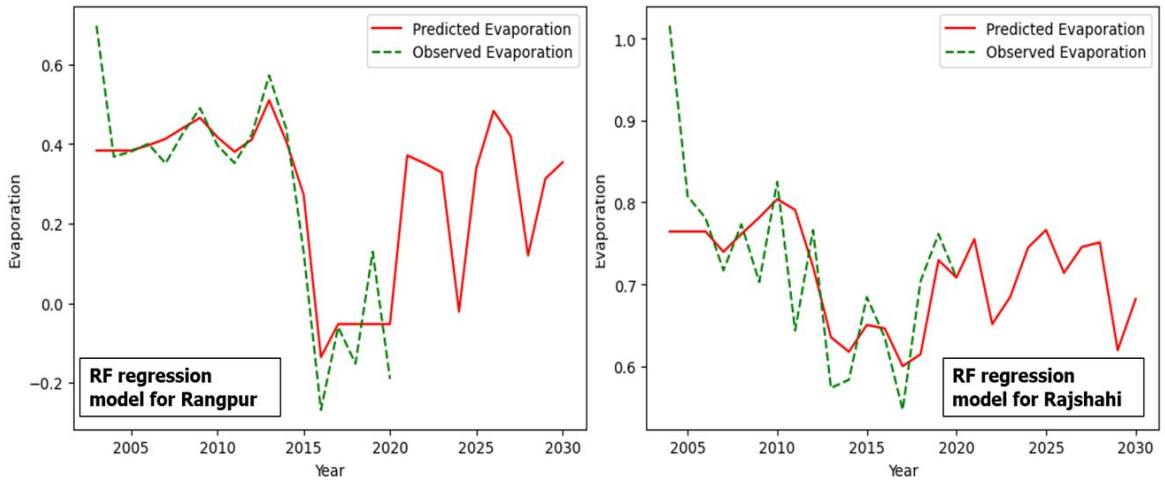

Figure 3. Performance of RF regression models in Rangpur and Rajshahi

In the figure 3 to 5, the performance of RF, GBM and SVM regression models were plotted using observed values and the predicted values with a forecast from 2021 to 2030 where it was detected that the RF and GBM regression captured more accuracy than the SVM regression model.

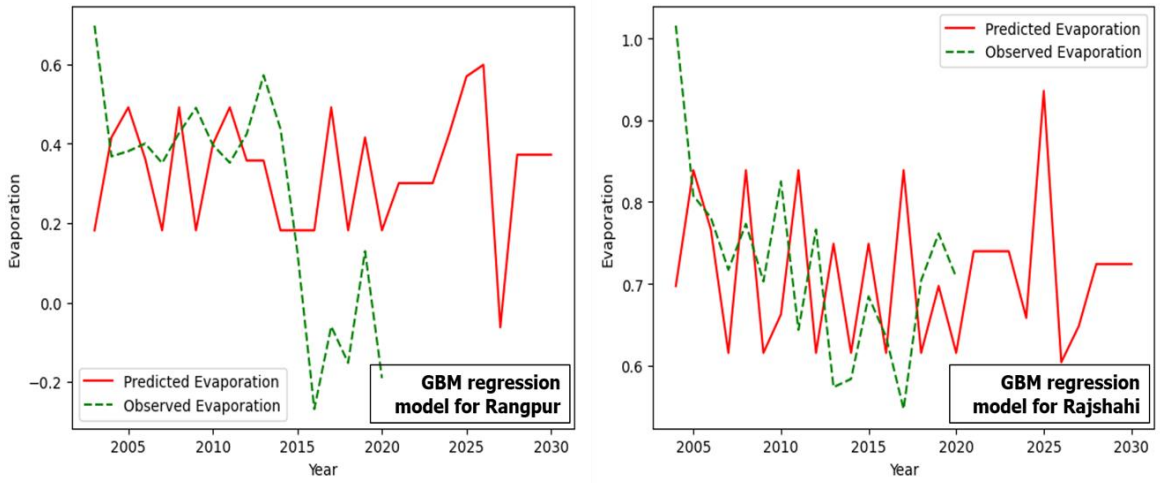

Figure 4. Performance of GBM regression models in Rangpur and Rajshahi

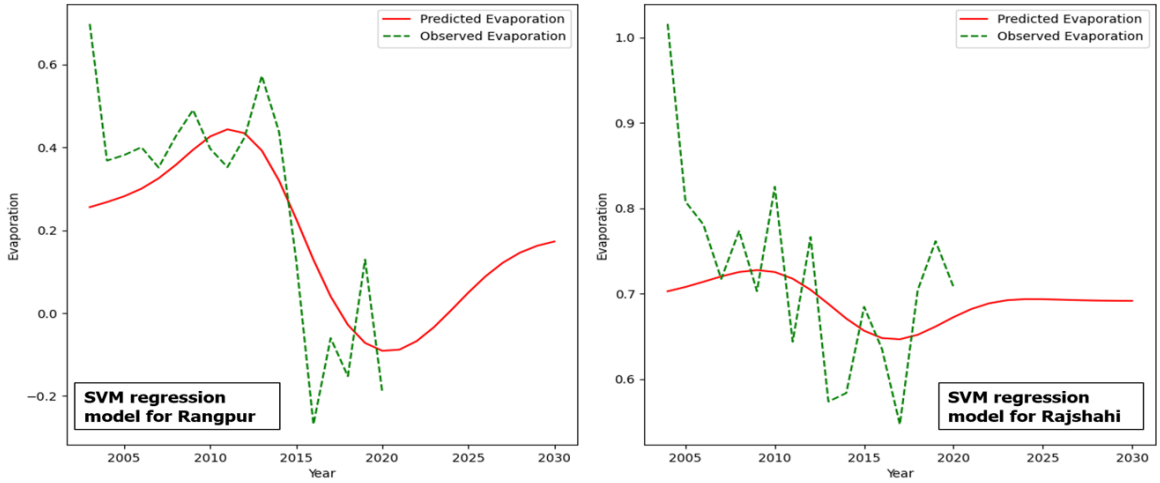

Figure 5. Performance of SVM regression models in Rangpur and Rajshahi

For SVM, outliers can have a significant impact on the decision boundary and therefore the predictions, leading to variations in results. Moreover, large datasets will result in better forecasting. RF and GBM training involve randomness due to the use of bootstrapping and random sampling of features. Thus, overfitting was observed in the forecasting. RF showed relative improvement in capturing the observed values while forecasting.

5 Conclusion

The study found that the random forest (RF) regression exhibited the most consistent performance across the two divisions within the selected study region. The utilization of the Random Forest technique is a highly recommended approach for predicting and forecasting the evaporation within the Ganges-Brahmaputra Basin located in the northwest region of Bangladesh. The Random Forest model produced the lowest values for MAE, MSE, RMSE at 0.06, 0.01, 0.001, and 4.20, correspondingly. The maximum value recorded for the concordance index (d) was 0.95, which is a significant measure of precision.

6 Recommendations

The regression models exhibited limited accuracy in predicting the peaks of the data. Random forest regression can predict complex relationships between characteristics and target variables and handle non-linear data

adequately. Decision trees are immune to outliers, making it robust to them. RF, GBM, and SVM regression are machine learning algorithms that are widely employed for regression tasks. However, they exhibit distinct dissimilarities in their modeling methodology, nonlinear relationship management, feature significance, resistance to outliers, capacity to regulate model complexity, and suitability for high-dimensional data. The selection of an algorithm is contingent upon the problem at hand and the characteristics of the data involved. Accurate forecasting of the maximum and minimum values of hydrologic parameters requires careful consideration, as these parameters do not conform to any discernible pattern over time or across geographic areas. Therefore, this study has the potential to serve as a noteworthy successor to the current research in order to enhance peak and trough prediction and forecasting.

References

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Dhiman, H. S., Deb, D., & Balas, V. E. (2020). Chapter 5 - Decision tree ensemble-based regression models. In H. S. Dhiman, D. Deb, & V. E. B. T.-S. M. L. in W. F. and R. E. P. Balas (Eds.), *Wind Energy Engineering* (pp. 61–73). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-821353-7.00016-8

El-Maghraby, S. S., Alamoud, A., & Al-Harbi, N. (2020). A Comparative Study of Random Forest Regression and Support Vector Regression for Water Budget Forecasting. *Water Resources Management*, *34*(5), 1729–1744. https://doi.org/10.1007/s11269-020-02518-5

Ferreira, V. G., Andam-Akorful, S. A., He, X., & Xiao, R. (2014). Estimating water storage changes and sink terms in Volta Basin from satellite missions. *Water Science and Engineering*, *7*(1), 5–16. https://doi.org/https://doi.org/10.3882/j.issn.1674-2370.2014.01.002

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, *29*, 1189–1232. https://doi.org/10.2307/2699986

Gaybullaev, B., Chen, S.-C., & Kuo, Y.-M. (2012). Large-scale desiccation of the Aral Sea due to over-exploitation after 1960. *Journal of Mountain Science*, *9*. https://doi.org/10.1007/s11629-012-2273-1

Guven, A., & Kişi, Ö. (2011). Daily pan evaporation modeling using linear genetic programming technique. *Irrigation Science*, *29*(2), 135–145. https://doi.org/10.1007/s00271-010-0225-5

J., P., S., A., A., M., A., K., D., H., & R., R. (2009). Daily Pan Evaporation Modeling in a Hot and Dry Climate. *Journal of Hydrologic Engineering*, *14*(8), 803–811. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000056

Lange, H., & Sippel, S. (2020). *Machine Learning Applications in Hydrology* (pp. 233–257). https://doi.org/10.1007/978-3-030-26086-6_10

Malik, A., Kumar, A., & Kisi, O. (2017). Monthly pan-evaporation estimation in Indian central Himalayas using different heuristic approaches and climate based models. *Computers and Electronics in Agriculture*, *143*, 302–313. https://doi.org/https://doi.org/10.1016/j.compag.2017.11.008

Raghavendra. N, S., & Deka, P. C. (2014). Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing*, *19*, 372–386. https://doi.org/https://doi.org/10.1016/j.asoc.2014.02.002

Shen, Y., Liu, C., Liu, M., Zeng, Y., & Tian, C. (2010). Change in pan evaporation over the past 50 years in the arid region of China. In *Hydrological Processes* (Vol. 24, Issue 2, pp. 225–231). https://doi.org/10.1002/hyp.7435

Shrestha, N. K., & Shukla, S. (2015). Support vector machine based modeling of evapotranspiration using hydro-climatic variables in a sub-tropical environment. *Agricultural and Forest Meteorology*, *200*, 172–184. https://doi.org/https://doi.org/10.1016/j.agrformet.2014.09.025

Sudheer, K. P., Gosain, A. K., Mohana Rangan, D., & Saheb, S. M. (2002). Modelling evaporation using an artificial neural network algorithm. In *Hydrological Processes* (Vol. 16, Issue 16, pp. 3189–3202). https://doi.org/10.1002/hyp.1096

Trambauer, P., Dutra, E., Maskey, S., Werner, M., Pappenberger, F., van Beek, L. P. H., & Uhlenbrook, S. (2014). Comparison of different evaporation estimates over the African continent. *Hydrology and Earth System Sciences*, *18*(1), 193–212. https://doi.org/10.5194/hess-18-193-2014

Wang, Y., Li, R., Min, Q., Fu, Y., Wang, Y., Zhong, L., & Fu, Y. (2019). A three-source satellite algorithm for retrieving all-sky evapotranspiration rate using combined optical and microwave vegetation index at twenty AsiaFlux sites. *Remote Sensing of Environment*, *235*, 111463. https://doi.org/https://doi.org/10.1016/j.rse.2019.111463

Yoon, Y., Kumar, S. V., Forman, B. A., Zaitchik, B. F., Kwon, Y., Qian, Y., Rupper, S., Maggioni, V., Houser, P., Kirschbaum, D., Richey, A., Arendt, A., Mocko, D., Jacob, J., Bhanja, S., & Mukherjee, A. (2019). Evaluating the uncertainty of terrestrial water budget components over high mountain Asia. In *Frontiers in Earth Science* (Vol. 7). https://doi.org/10.3389/feart.2019.00120.