https://doi.org/10.36375/prepare_u.iei.a183

Comparison of Data Driven Modelling Techniques For Rainfall Runoff Modelling of Kosi River

Nikhil Jadhav, Pankaj Kumar and Abhinav Kumar Singh
Department of Soil and Water Conservation Engineering
(Govind Ballabh Pant University Agriculture &Technology, Pantnagar)
(Pantnagar, Uttarakhand 263145, India)
{E-mail:nikhiljadhavmoto@gmail.com}
{E-mail:*pankaj591@gmail.com*}
{E-mail:avi4913@gmail.com}

https://doi.org/10.36375/prepare_u.iei.a183

Abstract

In the recent past, use of various machine learning techniques in predicting runoff from the catchment has become very popular. In this study, three empirical rainfall runoff models are employed to predict the discharge of the Kosi River for 11 years (2005-2015). The machine learning techniques such as support vector regression (SVR), multivariate adaptive regression splines (MARS) and random forest (RF) are employed for rainfall runoff modelling of Kosi watershed. The performances of all three prediction models have been successfully compared. Daily rainfall-runoff data for the period of 2005 to 2015 was collected for the Kosi River at Ramnagar barrage. It was seen that RF model outperformed over other two models. The gamma test was successfully applied in determination of the best input variables. The performance of the models is evaluated in terms of efficiency measures such as coefficient of determination (R^2), root mean squared error (RMSE) and Nash Sutcliffe Efficiency (NSE). The results revealed that random forest with R^2 value 0.95 in testing phase performed superior than other two models. The performance of MARS model was satisfactory while SVR model resulted very poor values. Therefore, RF model can be considered as most accurate model for prediction of discharge.

Keywords: Rainfall-runoff model; Gamma test; SVR; MARS; RF

Introduction

Rainfall-runoff models develop relationship between rainfall and runoff. The rainfall-runoff relationship is one of the most intricate hydrologic phenomena to comprehend due to its immense spatial and temporal variability of watershed characteristics and precipitation patterns, and the number of variables involved in the modelling of the physical process. For the decision makers, rainfall-runoff modelling produces a means of quantitative prediction. Modelling of runoff benefits to gain a better understanding of hydrologic phenomena as well as how changes affect the hydrological cycle (Xu, 2002). Modelling surface runoff can be difficult, as per as complex calculation is concerned and contain number of interconnected variables. The model add general components such as inputs, governing equations, boundary conditions or parameters, model processes, and outputs (Singh, 1995). The results of surface runoff modelling helps to understand catchment yields and response, estimate water availability, changes over time and forecasting.

https://doi.org/10.36375/prepare_u.iei.a183

Of the different models available, empirical models or black box models develop empirically identified statistical relationships between rainfall and runoff, without endeavouring to characterize and comprehend the physical processes invoked in the transformation. Empirical models are data driven models and are simple to use. These type of models do not consider physical processes prescribed in the system. The precision of model predictions is greatly subject to ability, knowledge of user and user's understanding of the model and watershed characteristics.

Models based on machine learning are capable of providing a useful alternative to deal with the multivariate and complicated nature of the phenomena of rainfall and runoff.

Support Vector Regression (SVM)

Support Vector Regression technique was firstly made known by Vapnik in 1992. Support Vector Regression (SVR) is basically a nonlinear regression method based on Support Vector Machines (SVM). It can be said as a sub part of Support Vector Machines (SVM). Support Vector Regression maps the data lower dimensional data into a higher dimensional feature space by using various kernel functions and then after solves a linear regression problem in the newly developed higher dimensional space. The SVR algorithm works with the goal to create the best line that can segregate n- dimensional space into classes.

Multivariate adaptive regression spline (MARS)

Multivariate adaptive regression spline (MARS) model is a newer non-parametric regression method. Friedman in 1991, firstly established this non-linear regression method. Nonlinear response between systems's input and outputs are recorded by the model by constructing several splines and also creates number of knots between the splines constructed (Friedman 1991). It works in two stages viz. forward stage and backward stage. In forward stage, MARS algorithm take the whole data and then takes the sub sample from the dataset and tries to fit linear regression line on those sample data set. When these lines are getting fitted on those sample data set, the algorithm just try to connect all those linear regression lines fitted by the algorithm. Thereafter, the algorithm joins all those regression line with knot. Each knot marks the end of one region of data and the beginning of another data. In the model, number of knots are

selected randomly. These knots occurs in pairs and are called as basis function. In backward stage, the algorithm removes the basis function which does not contribute to model accuracy or removes the model error.

Random Forest (RF)

Random forest is very popular ensemble machine learning technique. It was systematically developed by Breiman in 2001. The algorithm of Random Forest are very stable, straightforward and flexible. Random forest classifier or regressor is basically a bagging technique. Number of models called as based learners or decision trees are created using some samples of rows and features from the complete dataset. Sampling with replacement method is used while performing sampling. Decision tree have two properties viz. low bias and high variance. Each decision tree gets trained on the particular dataset used thereby becoming able to give accuracy or prediction. In case of classifying problems, majority of votes given by various decision trees are considered as outputs. Whereas, in case of regression problems, mean of the outputs given by various decision trees is considered as output. Mean of the outputs of all decision tree causes conversion of high variance possessed by individual decision tree into low variance in overall decision trees.

This paper illuminates application of SVM, MARS and RF in rainfall runoff modelling. This study is been carried out with an intent to evaluate the performance of machine learning techniques viz. SVM, MARS and RF in modelling the runoff using statistical parameters such as root mean square error (RMSE), coefficient of determination (R^2) and Nash Sutcliffe efficiency (NSE) etc. for the Kosi watershed. Also, this paper compares values of statistical analysis. On the basis of values computed, the study puts forward the best model to use in rainfall runoff modelling among these three empirical models.

Material and methods

Description of the study area and data

The study area is located on the Kosi River, a Himalayan river which originates at Rudradhari in Almora district of Uttarakhand state. It confluences to river Ramganga river near village Chamraul (Uttar Pradesh). This study area lies spatially between 33°21′53″N to 34°27′52″ N latitude and

74° 24′ 09″ E to 75° 35′ 36″ E longitude. The total area is about 3420 sq. km. It covers about all the physiographic parts of the Kashmir Valley. The area is drained by the important tributaries of Jhelum River. The rainfall and runoff data at Ramnagar barrage gauging station from 2005 to 2015 was procured, comprising of 4013 days. The data sets for the years 2005 to 2013 were used for training of models and these models were finally validated for various data sets achieved for 2013 to 2015. The whole dataset were work out in Gamma test software for best input selection. Gamma test creates different models with combination of different inputs. Then it calculates gamma value for each model. The model having least gamma value is selected as best input model. The best inputs illustrated by gamma test were further used for modelling.

Support Vector Regression

The Support Vector Machine (SVM) is a nonlinear generalization algorithm. It was introduced by Vapnik in 1992 and is used for classification and regression problems. The rainfall-runoff phenomenon is itself non-linear in nature, thus creating non-linearly separable points in space. The regression model can be constructed by mapping non-linear mapping function. The nonlinearly separable problem can be converted into linearly separable by mapping the original input data into higher dimensional space. The goal of the SVR algorithm is to construct a function $y = f(x)$ representing the dependence of the output y_i on the input x_i. This function can be expressed in the form as given below,

$$y = \omega^T \Phi(x) + b \quad (1)$$

Where,

ω is a weight vector and b is bias

$\Phi(x)$ is non-linear mapping function of inputs

Multivariate Adaptive Regression Splines (MARS)

Multivariate adaptive regression spline (MARS) records the nonlinear response between the inputs and output of a system by constructing several splines. The model creates number of knots between these splines. Each knot marks the end of one region of data and the beginning of another data. There is no need for any specific assumption about the elemental functional relationship between the inputs and output in a MARS model.

https://doi.org/10.36375/prepare_u.iei.a183

$$GCV = \frac{\cdot b^{d} \cdot P}{2 : \frac{\dot{l}\ ,\ 2\ ,\ \dot{J}\dot{l}\ 3}{\dot{J}\cdot}}$$

(2)

Where,

MSE= Mean squared error,

f = number of basis functions,

p = Basis function penalty and

n = number of observations

Random forest (RF)

Random Forest model is a decision tree model which handle complex relationships of independent and dependent variable without any assumption. The algorithm deals well with over fitting of the data and they can operate in parallel computing mode (Dayal el al. 2021). Considering a training set X = x_1, x_2, . . ., x_n, responses Y = y_1, y_2, . . ., y_n, and B times repeated bagging, a random sample (X_b,Y_b) is selected replacing the training set, which is fitted to a regression tree (f_b), for b = 1, 2, . . ., B. After training, the unseen samples (say, x') can be predicted by averaging all the individual regression trees' predictions on x' as:

$$\hat{f} = \frac{2}{\dot{b}} \sum_{>2} \dot{L}_{\ b}(x')$$

(3)

Results and discussion

The least gamma value was found for the combination of seven inputs as following: R, $R_{(t-1)}$, $R_{(t-2)}$, $Q_{(t-3)}$, $Q_{(t-2)}$, $Q_{(t-1)}$ and Q where R and Q are the rainfall and runoff data calculated from t^{th} day. (t-1), (t-2) and (t-3) represents the lagging done from one, two and three days before t^{th} day, respectively. Minimum gamma value was computed to be 0.0959. These input variables were further used in SVM, MARS and RF model. The complete dataset were divided into two parts. First 80 % data i.e. for the period of 2005 to 2013 were used for training of the models and remaining 20 % dataset i.e. for the period of 2013 to 2015 were used for testing of the models. The maximum discharge for the Kosi River was recorded to be 2180.341 cumec. The values of statistical indices such as minimum, maximum, mean, first quartile, and third quartile of the training, testing, and the complete dataset is given in Table 1.

https://doi.org/10.36375/prepare_u.iei.a183

Table 1. Values of statistical parameters of training, testing and complete daily dataset R, $R_{(t-1)}$, $R_{(t-2)}$, $Q_{(t-3)}$, $Q_{(t-2)}$, $Q_{(t-1)}$ and Q of the study area. X_{min}, X_{max}, X_{mean}, 1st Q, 3rd Q are minimum value, maximum value, mean, first quartile, and third quartile.

Statistical parameters	Input Variables						
	R (mm)	$R_{(t-1)}$ (mm)	$R_{(t-2)}$ (mm)	$Q_{(t-3)}$ (cumec)	$Q_{(t-2)}$ (cumec)	$Q_{(t-1)}$ (cumec)	Q (cumec)
Training (2005-2013)							
X_{min}	0.00	0.00	0.00	0.00	0.00	0.00	0.00
X_{max}	140.00	140.00	140.00	2180.341	2180.341	2180.341	2180.341
X_{mean}	3.012	3.012	3.012	27.864	27.868	27.872	27.876
Median	0.00	0.00	0.00	6.966	6.966	6.980	6.994
1st Q	0.00	0.00	0.00	3.709	3.709	3.709	3.709
3nd Q	0.00	0.00	0.00	21.606	21.606	21.606	21.606
Testing (2013 - 2015)							
X_{min}	0.00	0.00	0.00	2.379	2.379	2.379	2.379
X_{max}	142.400	142.400	142.400	731.198	731.198	731.198	731.198
X_{mean}	2.472	2.472	2.472	27.776	27.759	27.743	27.726
Median	0.00	0.00	0.00	11.383	11.383	11.355	11.298
1st Q	0.00	0.00	0.00	6.173	6.159	6.145	6.131
3nd Q	0.00	0.00	0.00	22.101	22.101	22.101	22.101
Complete Data (2005-2015)							
X_{min}	0.00	0.00	0.00	0.00	0.00	0.00	0.00
X_{max}	142.400	142.400	142.400	2180.341	2180.341	2180.341	2180.341
X_{mean}	2.958	2.958	2.958	27.846	27.846	27.846	27.846
Median	0.00	0.00	0.00	7.929	7.929	7.929	7.929
1st Q	0.00	0.00	0.00	4.247	4.247	4.247	4.247
3nd Q	0.00	0.00	0.00	21.634	21.634	21.634	21.634

In case of SVR model, the value of Root Mean Square Error (RMSE) for training and testing phase was found to be 58.28 and 33.28, respectively.

Also, NSE values were -1.28 and 0.00 for training and testing phase respectively. Coefficient of determination (R^2) of 0.66 was observed for testing phase. Overall, the performance of SVR model was poor as compared to MARS and RF model. The agreement between the observed and predicted discharge is unsatisfactory.

The MARS model resulted better values than conventional SVR model. Root Mean Square Error (RMSE) value was 47.36 for training and 17.96 for testing phase. The NSE and coefficient of determination between observed and simulated discharge was found to be in the range of 0.49 to 0.88 and 0.66 to 0.89, respectively. Thus, big hike observed in R^2 value in testing phase. The performance of MARS model is intermediate between SVR and RF model.

Random forest model came up with superior values among all three models. The model performance during training and testing period is found to be very good. RMSE value for training and testing phase was found to be 28.52 and 12.98, whereas NSE value was found to be 0.81 and 0.92, respectively. Highest coefficient of determination is obtained by RF model is 0.95. The agreement between observed and predicted discharge is very satisfactory.

Comparison of the performance values of RMSE, NSE and R^2 for both the models is shown in table 2. From the comparison, it is clear that Random Forest model outperformed than SVR and MARS model.

Table 2 Comparison of SVM, MARS and RF model using statistical indices

Models	Training			Testing		
	RMSE	NSE	R^2	RMSE	NSE	R^2
SVM	58.28	-1.28	0.57	33.28	0.00	0.66
MARS	47.36	0.49	0.66	17.96	0.88	0.89
RF	28.52	0.81	0.90	12.98	0.92	0.95

Conclusion

In this study, comparison of three empirical rainfall-runoff models have been successfully done. Random forest model outperformed the other two model and thus, it is best suited for the prediction of runoff. Decision trees present in random forest model learns better from the data, thereby creating good correlation between

observed and predicted values. Due to tremendous variation in the data, regression line has limitation to fit well. It causes SVR and MARS model to perform in the range of poor to satisfactory.

References

[1]. Deen Dayal, Praveen K. Gupta and Ashish Pandey., (2021) Streamflow estimation using satellite-retrieved water fluxes and machine learning technique over monsoon-dominated catchments of India. Hydrological Sciences Journal, 66:4, 656-671.

[2]. Maity Rajib, Bhagwat Parag and Bhatnagar Ashish (2010). Potential of support vector regression for prediction of monthly streamflow using endogenous property. Hydrological Processes, 24, 917–923.

[3]. Zhenliang Yin, Qifeng, Xiaohu Wen, Ravinesh C. Deo, Linshan Yang, Jianhua Si, Zhibin He (2018). Design and evaluation of SVR, MARS and M5Tree models for 1, 2 and 3-day lead time forecasting of river flow data in a semiarid mountainous catchment. Stochastic Environmental Research and Risk Assessment.

[4]. Dinh Tuan Vu, Xuan Linh Tran, Minh-Tu Cao, Thien Cuong Tran, Nhat-Duc Hoang (2020). Machine Learning Based Soil Erosion Susceptibility Prediction Using Social Spider Algorithm Optimized Multivariate Adaptive Regression Spline.

[5]. Joshi Jignesh, Patel Vinod M., (2011) Rainfall-Runoff Modeling Using Artificial Neural Network. National Conference on Recent Trends in Engineering & Technology

[6]. J.H. Friedman, C.B. Roosen., (1995) An Introduction to Multivariate Adaptive Regression Splines, Statistical Methods in Medical Research, 4 197-217

[7]. J.H. Friedman, Multivariate Adaptive Regression Splines, (1991) The Annals of Statistics, 19 1-67

[8]. Hsu K Gao X, Sorooshain S., and Gupta H.V. (1995). Artificial Neural Network Modelling of the Rainfall-runoff Process Water Resources, 3, 2517-2530.

[9]. Keith Beven., (2012) Rainfall-Runoff Modelling The Primer. 1-5

[10]. Xiao M, Zhang Q, Singh VP, Chen X., (2017) Probabilistic forecasting of seasonal drought

behaviors in the Huai river basin, China. Theor Appl Climatol 128:667–677. https://doi.org/10.1007/ s00704-016-1733-x

[11]. Seth Richa, Singh Prashant, Manindra Mohan, Singh Rakesh, Gupta Vinod K., Uniyal D.P., Dobhal Rajendra and Gupta Sanjay (2013). Assessment of Water Quality of Kosi River, Almora, Uttarakhand (India) for Drinking and Irrigation Purposes, TACL 3 (4), 287 – 297

[12]. Zhaoli Wang, Chengguang Lai, Xiaohong Chen, Bing Yang, Shiwei Zhao and Xiaoyan Bai. (2015). Flood hazard risk assessment model based on random forest. Journal of Hydrology 527, 1130-1141.