2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT) | 978-1-6654-3773-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISAECT53699.2021.9668539

ARIMA Based Forecasting of stream flows of Three Georges Dam for efficient Water Resource Planning and Management

Dola Gupta
DST Women Scientist
Jadavpur University
Kolkata, India
dola.ju@gmail.com

Amlan Chakrabarti
A. K. Choudhury School of IT
University of Calcutta
Kolkata,India
acakcs@caluniv.ac.in

Jyoti Gautam
Department of IT
JSS Academy of Technical Education
Noida, India
jyotijssaten@gmail.com

***Abstract*—Stream-flow forecasting is one of the major aspects in improving the efficiency of water resource planning and management for the water reservoirs in various geographical regions. Different forecasting techniques have been implemented for forecasting the inflow rate in the past with Support Vector Machine (SVM) being very popular and accurate among them. Similarly, the outflow forecasting is essential to estimate the usage and also encompasses different models for forecasting. Muskingum model is found to be prevalent in forecasting outflow but it has the limitation of inefficiency in taking care non-linearity. In this paper, both inflow and outflow rate prediction is done by Auto-Regressive Integral Moving Average (ARIMA) model as prediction of stream flow was never modelled as ARIMA in the literature earlier. The performance evaluation parameter considered is the root mean square error (RMSE) for comparison with the existing SVM models. It is found that in case of inflow rate, RMSE obtained by ARIMA model shows a decrease of 36% in the best case and increase of 4.3% in the worst case when compared with the SVM models. Likewise, the outflow rate RMSE when compared with Muskingum model gives better results taking non-linearity into consideration. The results of this study will help in not only planning efficient water resource management but also prediction of flood frequency in future.**

***Index Terms*—Three Georges Dam, forecasting, ARIMA model, Muskingum model, water resource management**

I. Introduction

The Yangtze River is the third longest river in the world and the longest river in Asia which is regulated by the Three Georges Dam (TGD) since the year 2003. The TGD is a hydroelectric gravity dam situated on the Yangtze River nearby Yichang city in China. The storage of a reservoir has hydrological variability at many time zones, hourly to daily and seasonal to yearly. Managing a steady water supply with all time variability and sudden local demands leads to complex water resource planning. Forecasting of hydrological variations can improve water management to a large extent and eventually helps in preventing floods. Study of TGD have been done extensively in the literature due to its impact on the socio-economic contribution to China. Numerous researches may be found encircling forecasting, comparison and analysis of the various aspects of TGD hydrological variations [1].

Real time forecasting and analysing flood frequency is one of the major areas of research. Earlier studies in flood prediction were done considering occurrence of flood independent and identically distributed in each year. The IPCC report in 2014 clearly stated that the frequency of flood will be time-varying under the impact of climatic changes.

The increase in the population growth has an adverse impact on water resources throughout the world. The scarcity of drinking water needs improved water resource management. The reservoir operations should be optimised as they often change the nutrient distribution and degrade the quality of water [2]. Machine learning algorithms also have a great impact in forecasting the hydrological conditions but SVM models have shown best results [3][7]. The discrepancy between the theoretical modelling and real-time allocation of water resources is optimised by modelling the inflow data by Auto-Regressive Integral Moving Average (ARIMA) time series modelling.

Muskingum model is an important hydrological model used for Outflow forecasting based on numerical solution methods. It encompasses optimization and prediction in each step of simulation [4]. The linear model is easy to implement but does not give accuracy if non-linearity exists [5]. The model gives the storage equation which is mainly dependent on the discharges of inflow and outflow rates of a natural river. It has been observed that the conventional hydrological models often rely on regression analysis and hence cannot capture the non-linearity of the water data [9]. This limitation has been overcome by modelling the data by ARIMA model. The temporal variations of the inflow data could be better modelled by the ARIMA model.

A. Motivation and Contribution

- Real-time inflow rate and outflow rate prediction of TGD is modelled as ARIMA time series model and results are found comparatively better than the existing models.
- Considering the case of inflow rate, RMSE calculated for ARIMA model is found always less than existing SVM or modified SVM models.

- In case of outflow forecasting, the RMSE of Muskingum model was found better in linear region but it could not forecast in region of non-linearity whereas ARIMA could predict the whole region with slight increase in RMSE.

II. Time-Series Modelling

A time-series as defined ,[6] is a sequence of observations taken sequentially in time.The time defined here may be annually, monthly, daily or per hour depending on the problem taken for research. The ability of ARIMA model to identify complex data patterns in time series modelling makes it widely used in forecasting temporal datasets.

A. Auto Regressive Integrated Moving Average (ARIMA)

ARIMA model may be considered the combination of Auto-Regressive (AR) model and Moving Average (MA) model with Integrated component. A time-series may get effected by cumulative effect of a feature belongs to a class of integrated process. Mathematically, ARIMA models may be defined as ARIMA(p,q,d), where p is the order of Auto-Regression, q is the order of Moving Average and d is the order of differencing. The predicted value y_t, may be defined as

$$y_t = \mu + \phi_1 y_{t-1} \cdots \phi_p y_{t-p} - \theta_1 \epsilon_{t-1} \cdots - \theta_q \epsilon_{t-q} \quad (1)$$

where μ is a constant, $\phi_p y_{t-p}$ is lagged values of y and $\theta_q \epsilon_{t-q}$ is lagged errors.

B. Model Output parameters

The time-series analysis done in this paper have taken into account the following output parameters for model evaluation and hence forecasting ARIMA models for both inflow and outflow rates.

1) Akaike Information Criterion (AIC): The AIC is the statistical method for selecting a model and can be calculated by:

$$AIC = -2/N * LL + 2 * k/N \quad (2)$$

Where N is the number of samples in the training data set,k is the number of parameters in the model and LL is the log-likelihood of the model on the training data set. The model with least AIC value is chosen for estimation.

2) BIC: Similarly BIC is also a statistical method for selecting a model but it gives higher values to models having more complexity and gives preference to simpler models and can be calculated as follows:

$$BIC = -2 * LL + log(N) * k \quad (3)$$

Where log() has the base-e called the natural logarithm, LL is the log-likelihood of the model, N is the number of examples in the training data set, and k is the number of parameters in the model.

3) Hanan Quinn Information Criterion (HQIC): It is also a measure of the best model fit for a given data set but does not depend upon log-likelihood and penalises the model with more number of parameters lesser the value of HQIC, better the model is with this criterion.

4) ACF: An ACF measures and plots the average correlation between data points in a time series and previous values of the series measured for different lag lengths.

5) PACF: A PACF is similar to an ACF except that each correlation controls for any correlation between observations of a shorter lag length. Thus, the value for the ACF and the PACF at the first lag are the same because both measure the correlation between data points at time t with data points at time t – 1.

III. System Modelling

The Storage of the reservoirs are very important for determining the flood routing and this can be done by solving the continuity equation. Flood Routing may be defined as the process of estimating the timing and shape of a flood wave. The continuity equation states that the water stored is the difference of the inflow rate and the outflow rate of a reservoir in the time interval. A schematic diagram have been shown in Fig 1. Hence determination of the inflow rate and outflow rate contributes significantly in prediction of floods. The prediction of the inflow rate and the outflow rate will be helpful in forecasting the future storage of water assuming that the sudden demand level is within the set limit.

$$I - O = S * T \quad (4)$$

Where I is the inflow rate, O is the outflow rate, S is the change in water level stored and T is the time interval. The storage equation gives a relation between the discharges of inflow rate and the outflow rate.

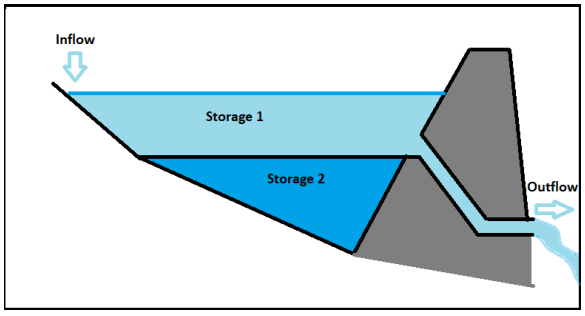

Fig. 1. Schematic diagram of flow rates of a reservoir [7]

A. Data

The data set that have been considered for the time-series analysis have been taken from Yangtze station[8] of TGD. The data set bears the value of inflow rate ,outflow rate , upstream level and downstream level since May 2011 to April 2021 that is a data set of almost 10 years. The correlation plot of the four

variables is shown in Fig2. It may be observed that the inflow rate and the outflow rate data are highly correlated and rest of the parameters hardly have any correlation among each other. The observed data of inflow rate and outflow rate is plotted in Fig 3 and Fig 4. It is observed that the flow rates follow a definite pattern every year with their peak consumption in the mid of every year during the summer season. Prediction of the flow rates will help in prior planning of the water resources efficiently.

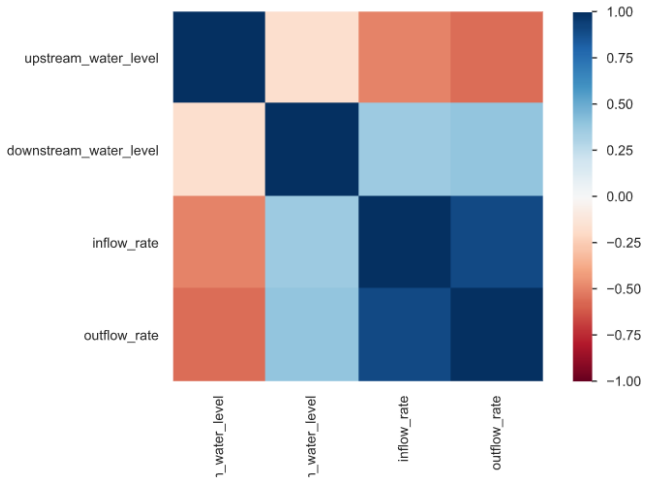

Fig. 2. Correlation plot of the four variables of the data set.

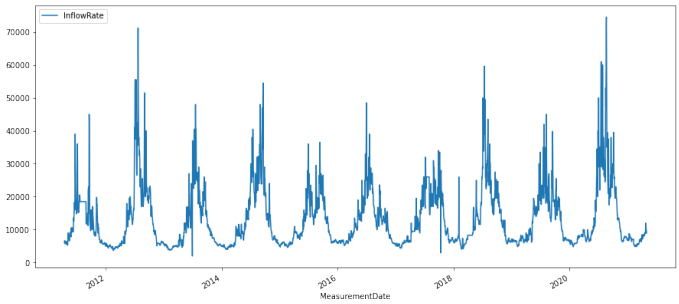

Fig. 3. Plot of inflow rate of TGB from 2011 to 2021.

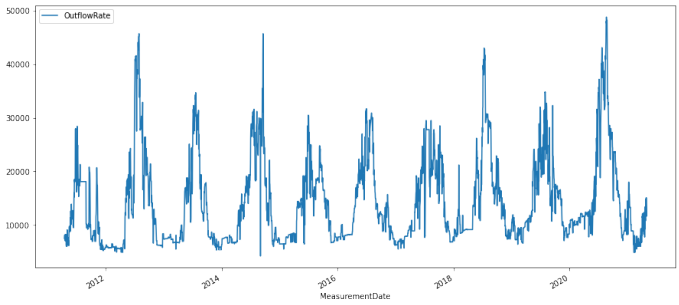

Fig. 4. Plot of outflow rate of TGB from 2011 to 2021.

IV. Simulation and Results

The simulation results of the inflow rate and the outflow rate were done by ARMA and ARIMA models over the 10 years available data. The results of the ARIMA model were found much better than the ARMA model as given in Table I. In [11], ARMA model was used for Inflow prediction where the RMSE error was calculated to 4237 as compared to our RMSE of 2717. Hence only the detailed results of the ARIMA models are given in this paper for both inflow as well as outflow predictions of TGD.

The Augmented Dickey-Fuller (ADF) is done for verifying the stationarity of the data set. As the name suggests the ADF test is the augmented form of Dickey-Fuller test that implies it expands the Dickey-Fuller equation incorporating higher order regressive terms in the model. The p-value obtained in the results must not exceed 0.05 for the data to be stationary.The results of the ADF test of inflow rate data are as follows:

ADF Test Statistic : -9.917636351181333
p-value : 3.049978305758276e-17
Lags Used : 27

The ADF test results of outflow rate data are as follows:

ADF Test Statistic : -4.660744755691467
p-value : 9.96644910429753e-05
Lags Used : 28

It may be observed that the p-value of both the inflow rate and outflow rate data are much less and hence the data is stationary. Corresponding ACF and PCF, as defined in equation 2 and equation 3 are plotted in the Fig 5 and Fig 6 of the inflow and outflow rates respectively. The results were found best with value of lag = 27 for inflow rate and lag = 28 for outflow rate. The auto correlation plot with respect to the corresponding values of lag is given in Fig 7 and Fig 8 for inflow and outflow rate respectively. The inflow rate and outflow rate have been decomposed into the signal components namely trend, seasonal and residue as shown in Fig 9 and Fig 10 for analysis. As mentioned earlier the data is stationary hence no differencing was required for removal of any seasonal component.

The output parametric values of the ARMA and ARIMA model for the inflow rate and the outflow rate have been tabulated in Table I respectively. Four models were considered as ARMA.in, ARIMA.in, ARMA.out and ARIMA.out as two inflow models and two outflow models respectively.The AIC and BIC values of the stream flows of ARIMA model are found better than that of ARMA model. The HQIC value would be better for the model which has least complexity but that would compromise with the accuracy. Summarising all the above discussions done in simulation and result section, the prediction is done by ARIMA model and the predicted values are plotted. It may be observed in Fig 11 and Fig 12,

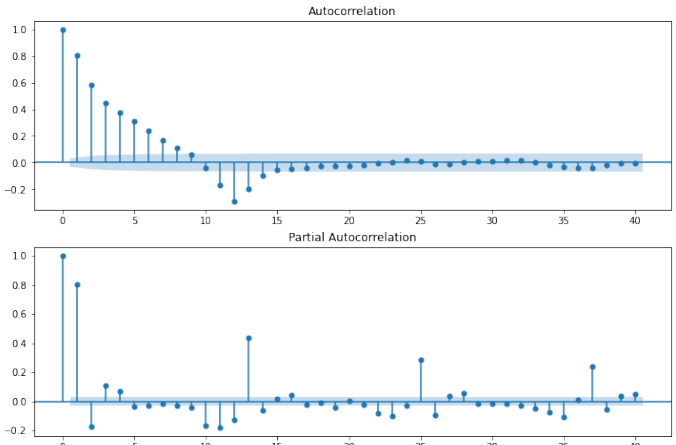

Fig. 5. The ACF and PCF plots of the inflow rate.

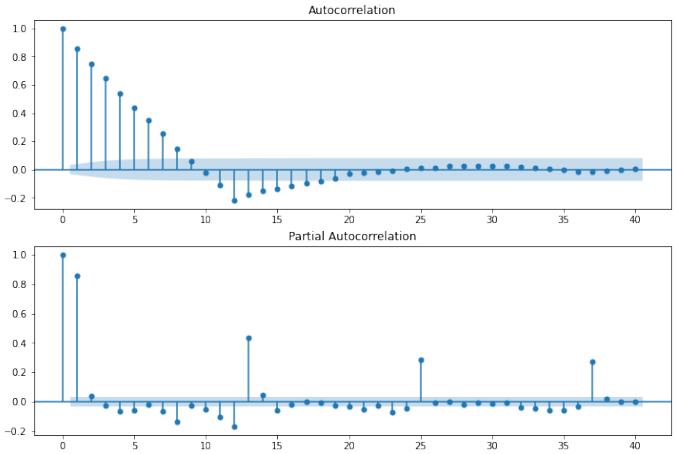

Fig. 6. The ACF and PCF plots of the outflow rate.

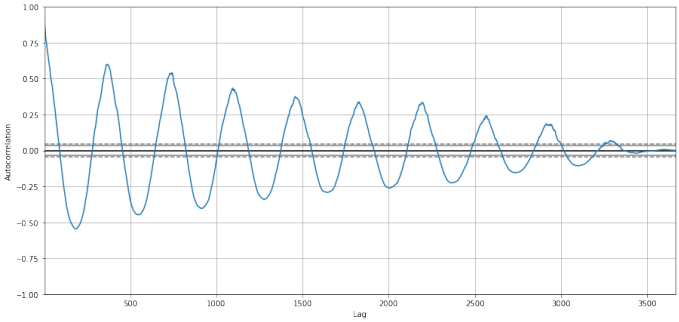

Fig. 7. The auto-correlation plot with respect to lag of the inflow rate

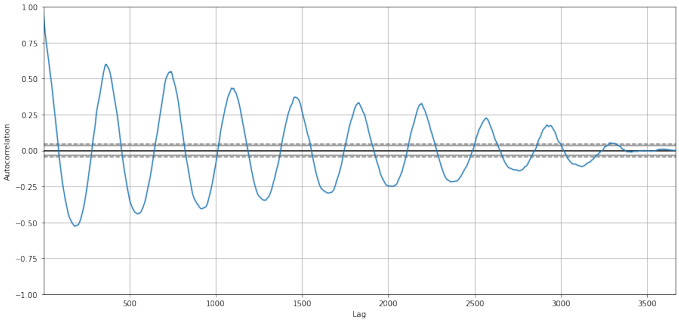

Fig. 8. The auto-correlation plot with respect to lag of the outflow rate

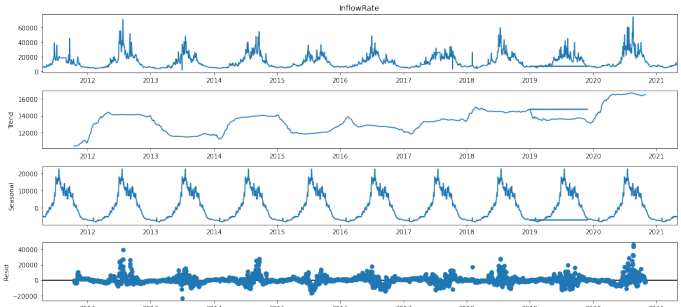

Fig. 9. Seasonal Decomposition of the inflow rate of TGD

the inflow rate and the outflow rate predicted values almost follows the observed value evidently. The predicted plot is shown for the year 2020 for better visualization though the prediction is done for everyday for 10 year available data.

V. Comparison with the Existing Work

The comparison of our work with the existing model are done for inflow rate and outflow rate separately as they have

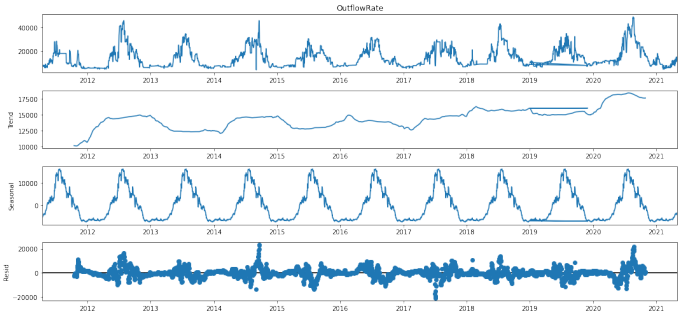

Fig. 10. Seasonal Decomposition of the outflow rate of TGD

TABLE I
OUTPUT COMPARISON OF THE PARAMETRIC VALUES OF ARMA AND ARIMA MODEL OF INFLOW AND OUTFLOW RATES

	MODEL	AIC	BIC	HQIC	LL
ARMA.in	ARMA(5,3)	68402.47	68464.5	68424.57	-34191.23
ARIMA .in	ARIMA(5,1,3)	**68401.43**	**68463.5**	**68423.53**	**-34190.71**
ARMA.out	ARMA(5,3)	65744.35	65806.42	65766.45	-32862.17
ARIMA.out	ARIMA(5,1,3)	**65733.91**	**65795.98**	**65733.91**	**-32856.95**

TABLE II
COMPARISON TABLE OF RMSE VALUES OF EXISTING MODELS WITH ARIMA MODEL FOR INFLOW RATE

	Model	RMSE	%Change
[9]	SSA-SVM	2598	**-4.3**
	SVM	3812	28.7
[10]	SSA-SSR (L=2)	4275	**36.4**
	SSA-SVM (L=12)	2941	7.6
	SSA-GP (L=7)	2805	3.1
Proposed technique	**ARIMA**	**2717**	

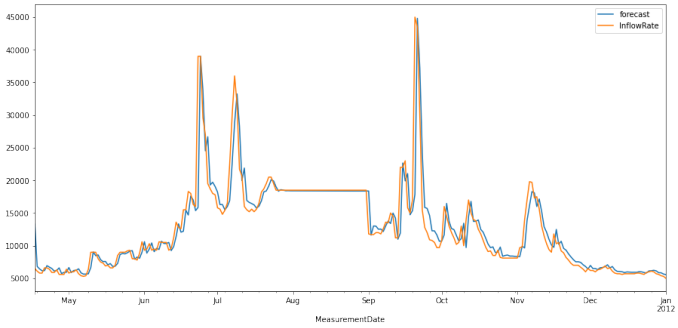

Fig. 11. Plot of Actual and Predicted data for inflow rate.

been compared with different models in different research works. The performance of each of the model is done considering the RMSE of each model.

A. Comparison of the predicted Inflow rate with Existing Models

The work done in [9] predicts the inflow rate with SVM and SSA-SVM where the performance of the SSA-SVM is found to be better. The RMSE error of the SSA-SVM model is calculated to be 2598 that is comparable to the RMSE in our work which is 2717 which is much better than the RMSE of SVM which is 3812.Another study done in [10]the RMSE, taking models SSA-SAR, SSA-SVM and SSA-GP with best results with length of window, L=2,12,7 respectively is 4275, 2941 and 2805 which than RMSE obtained in our study. The RMSE as summarized in Table II, the column "% Change" reflects the variation in RMSE of the state of art techniques [9][10] with our proposed ARIMA model. It may be observed that with respect to ARIMA model the decrease in RMSE in best case is about 36% and an increase of 4.3% in worst case.

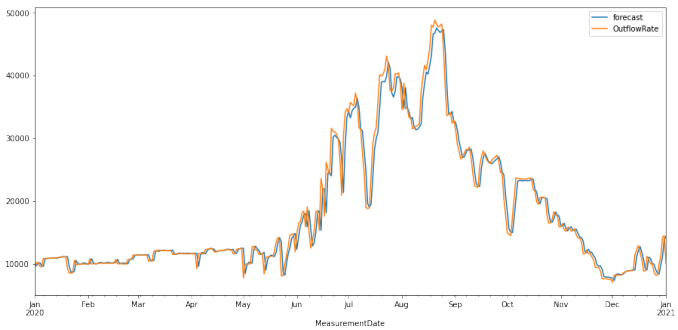

Fig. 12. Plot of Actual and Predicted data for outflow rate.

B. Comparison of the predicted Outflow rate with Existing Models

The linear Muskingum model and ARIMA model were simulated with the same set of outflow data set. The RMSE calculated for Muskingum model was found to be 1499 for the year 2013 when calculated from January to May. If the data were considered for the whole year then the decreasing values from June to December were not captured and an absurd RMSE was obtained as non-linearity cannot be taken care by this model.The uni-variate ARIMA model on the other hand could predict the whole 10 years data with a RMSE error of 1888. The correlation between the inflow and the outflow rate if taken into account with multivariate ARIMA model the RMSE is expected to reduce in ARIMA model.

VI. CONCLUSION AND FUTURE WORK

The purpose of this study is to completely analyse the best model for stream flow prediction which in future leads to determining the best suitable model for flood prediction and water resource management.The predicted values of inflow rate and the outflow rate are done by uni-variate ARIMA model of TGD for 10 years. As discussed earlier there exists a high correlation between the inflow rate and outflow rate which will be taken into account for modelling in future study with time series models.The RMSE calculated in our study in case of inflow was found comparatively lower than the existing models except in one case. In case of outflow the RMSE error when compared to Muskingum linear model was found much better for 10 years data.The RMSE of Muskingum model depends on initial conditions and hence the initial value of the storage highly effects the predicted values. ARIMA model does not need the initial value of storage. Hence we conclude that the time series analysis modelling can significantly predict the real time inflow and outflow rate of a dam.The real time forecasting will be useful for modelling water resource planning as well as flood forecasting.

Analysing the joint correlation between the stream flows taken into account in future work is expected to give better results in future study. Other appropriate time series models will be incorporated and a complete analysis and prediction of the stream flow will be taken into account.Similarly in case of Outflow prediction non-linearity will be taken into consideration for the Muskingum model and the predicted values will be compared with the time series analysis.The effects of annual run-off, evaporation and underground water also would be taken into consideration for modelling the in future study. The study done in the paper will be extended to Dams in India for prevention of flood and effective water resource planning, initially starting with states Bihar and West-Bengal.

REFERENCES

[1] Hu, B., Yang, Z., Wang, H., Sun, X., Bi, N. and Li, G., 2009. Sedimentation in the Three Gorges Dam and the future trend of Changjiang (Yangtze River) sediment flux to the sea. Hydrology and Earth System Sciences, 13(11), pp.2253-2264.

[2] Wu, Y., Wang, X., Zhou, J., Bing, H., Sun, H. and Wang, J., 2016. The fate of phosphorus in sediments after the full operation of the Three Gorges Reservoir, China. Environmental Pollution, 214, pp.282-289.

[3] Su, J., Wang, X., Liang, Y. and Chen, B., 2014. GA-based support vector machine model for the prediction of monthly reservoir storage. Journal of Hydrologic Engineering, 19(7), pp.1430-1437.

[4] Easa, S.M., Barati, R., Shahheydari, H., Nodoshan, E.J. and Barati, T., 2014, November. Discussion: New and improved four-parameter nonlinear Muskingum model. In Proceedings of the Institution of Civil Engineers-Water Management (Vol. 167, No. 10, pp. 612-615). Thomas Telford Ltd.1

[5] Beven, K.J., 2011. Rainfall-runoff modelling: the primer. John Wiley and Sons.

[6] Box, G. E., Jenkins, G. M., Reinsel, G. C., andLjung, G. M. (2015). Time series analysis: forecasting and control. John Wiley and Sons.

[7] https://www.nrcs.usda.gov/

[8] https://www.kaggle.com/konivat/three-gorges-dam-water-data

[9] Yu, Y., Wang, P., Wang, C., Qian, J. and Hou, J., 2017. Combined monthly inflow forecasting and multiobjective ecological reservoir operations model: case study of the three gorges reservoir. Journal of Water Resources Planning and Management, 143(8), p.05017004.

[10] Wang, Y., Guo, S., Chen, H. and Zhou, Y., 2014. Comparative study of monthly inflow prediction methods for the Three Gorges Reservoir. Stochastic Environmental Research and Risk Assessment, 28(3), pp.555-570.

[11] Dawson, C.W., Harpham, C., Wilby, R.L. and Chen, Y., 2002. Evaluation of artificial neural network techniques for flow forecasting in the River Yangtze, China. Hydrology and Earth System Sciences, 6(4), pp.619-626.