Dam Inflow Time Series Regression Models Minimising Loss of Hydropower Opportunities

Yasuno Takato[1]

[1] Yachiyo Engineering Co., Ltd., Asakusabashi 5-20-8, Taito-ku, Tokyo, Japan
tk-yasuno@yachiyo-eng.co.jp

Abstract. Recently, anomalies in dam inflows have occurred in Japan and around the world. Owing to the sudden and extreme characteristics of rainfall events, it is very difficult to predict dam inflows and to operate dam outflows. Hence, dam operators prefer faster and more accurate methods to support decision-making to optimise hydroelectric operation. This paper proposes a method to predict dam inflows for the target three-hour forecast using time series data with a unit interval of 15 minutes. This method can predict the time of rise to the peak, maximum level at the peak, and hydrograph shapes to estimate the volume. This paper presents several experiments applied to the 20,627 time stamped recorded dam inflow time series data. The input data contains hundreds of time series data features, such as data from rain gauges in the dam regions and upstream river level sensors, and their time windows from the previous three hours. I provide additional practices to minimise the loss of dam hydroelectric opportunities more efficiently.

Keywords: Dam Inflow Time Series, Upstream River Sensor, Time Series Regression, Hydropower Opportunities Loss.

1 Objective

1.1 Related papers

In recent years, extreme flooding has caused massive damages worldwide [1,2]. In Japan, sudden rainfalls and extreme river floods have occurred in regions such as Joso City at the Ibaraki prefecture [3]. Owing to such extreme and sudden rainfall events, it is difficult to predict the dam inflow and to operate the dam outflow. There are several approaches for comparing water resources time series prediction modelling with statistical time series models and modern machine learning methods. For example, artificial neural network (ANN) models are applied to monthly reservoir inflow time series [4,5]. Furthermore, the data are 500 monthly streamflows covering a period of 40 years, including climate and land cover data where rainfall-runoff modelling simulates the streamflow of watersheds. Here, several models are compared with the Gaussian linear regression, Gaussian generalised additive models (GAMs), multivariate additive regression splines (MARSs), ANN models, random forest (RF), and regression tree mod-

els [6]. They suggest that GAMs and RF can effectively capture some non-linear relationships. Regarding daily lake water level forecasting, the data were collected at five water gauges at the lake for 50 years, and several models were compared, such as the RF, support vector regression, ANNs, and the linear model. The results suggested that the RF can obtain a more reliable and accurate lake water level based on daily forecasting than others [7]. The variable selection approach can improve the daily reservoir discharge forecasting efficiency by machine learning methods. The data collected were 2,854 daily records over a period of eight years. After training and testing five methods, such as the ANN, the instance-based classifier, k-nearest-neighbour classifier, RF, and the random tree, key variables that influence the reservoir water level were selected and better models were built. These experimental results indicate that the RF forecasting model, when applied to variable selection with full variables, has better performance than other models [8]. Furthermore, regarding monthly reservoir inflow forecasting, hybrid processes are proposed where they are combined with a linear model and a non-linear model; for example, the former is the seasonal autoregressive integrated moving average (SARIMA) and the latter is the ANN and the genetic programming (GEP). The data were collected at the hydrometric stations and calculated as 360 monthly inflow series for a period of 30 years. After modelling and predicting, a comparison of the two hybrid models indicated the superiority of SARIMA-GEP over the SARIMA-ANN [9].

In these studies, the perspective was decision support for water resource management and planning on monthly and daily terms. Monthly dam inflows prediction is important for adopting a reservoir water control plan to meet agricultural and hydropower demands. However, short term daily dam inflows prediction could be crucial for real-time reservoir operations to prevent extreme floods and ensure a consistent water supply by maintaining a safe reservoir level.

1.2 Dam Inflow Prediction for Hydropower Operation

When the dam region for hydropower is narrow, it takes a short time for the main river stream to flow from upstream to the dam inflow. Extreme rainfall events occur suddenly at real-time streams whose data range is hourly or every minute. Recent high-level dam inflow events are different from normal inflow patterns and are rarely observed. Owing to these extreme and sudden dam inflow events, it is difficult to predict dam inflows and operate dam outflows for hydropower generation. To support decisions concerning outflow operations in response to sudden and extreme high-level inflow events, there is a need for more accurate dam inflow time series modelling and for hourly interval predictions.

This paper proposes a method to predict dam inflows for a target 3-hour forecast using time series data with a unit interval of 15 minutes. It will enable decision-makers to predict the time of rise to the peak, maximum level at the peak, and hydrograph shapes to estimate the volume. Furthermore, it will facilitate knowledge of clustering dam inflows and minimise the loss of hydropower opportunities. This paper discusses several experiments applied to the 20,000 time stamped recorded dam inflow time series data. The inputs are the hundreds of time series data features, such as rain gauges in the dam region and upstream river level sensors, and their time windows. Additional

practices are provided to minimise the loss of dam hydropower opportunities more efficiently.

2 Modelling

2.1 Minimising the Loss of Hydropower Opportunities

Using MetaCost Minimise Hydropower and Flood Damage. Based on the inflow classification of the annotated boundary between high and low water levels, the task is to build a classification model using decision tree, the support vector classifier, and so forth. The classification model could calculate a confusion matrix that describes the overall accuracy, precision, and recall for each inflow classification.

Table 1 lists two types of dam inflow prediction errors. On the upper right side of the confusion matrix, the misclassification is important for minimising the risk of flood damage owing more to the under-prediction than actual inflows [10,11]. On the bottom left side of the confusion matrix, the misclassification is critical to minimise the loss of hydropower opportunities influenced by the excess outflow operations whose forecast is due more to the over-prediction than the actual inflows [12]. Incorporating these costs, the MetaCost algorithm proposed a method for creating cost-sensitive classifiers [13,14].

Table 1. Two types of dam inflow prediction errors

		True, actual value	
		Low water	High water
Prediction	Low water	True Low water	**Under-prediction Error (Flood Damage Risk)**
	High water	**Over-prediction Error (Hydropower Opportunity Loss)**	True High water

Loss Index to Compute Over-prediction Error for Hydropower. This paper provides the indexes to compute the loss of hydropower opportunities and the risk of flood damages. First, the over-prediction error at time t is indicated as:

$$ope_t = \hat{y}_t - y_t, \text{ if positive then } \hat{y}_t > y_t. \quad (1)$$

Here, $\hat{y}_t$ is the dam inflow prediction at time t, and y_t is the actual inflow value. The sum of over-prediction errors among a term $t \in \{1, \dots, \mathrm{T}\}$ is formulated as follows:

$$sope = \sum_{t=1}^{T} \max(\hat{y}_t - y_t, 0). \quad (2)$$

This positive value can be computed as the predicted case when there are only excess inflows over the actual level. If over-predictions frequently occur, i.e. the value of $sope_t$ is large, then the operator may carry out an excessive outflow more than an actual required level. These errors would correspond to the loss of hydropower opportunities.

Risk Index to compute Under-prediction error for Flood Damage. Next, the under-prediction error at time t is indicated as:

$$upe_t = \hat{y}_t - y_t, \text{ if negative then } \hat{y}_t < y_t. \quad (3)$$

This situation appears to be an optimistic forecast at time t, where $\hat{y}_t$, the dam inflow prediction, is less than y_t , the actual inflow value. The sum of under-prediction errors among a term $t \in \{1, \dots, \mathrm{T}\}$ is formulated as follows:

$$supe = \sum_{t=1}^{T} \min(\hat{y}_t - y_t, 0) \quad (4)$$

This negative value can be computed as the predicted case when there are only lower inflows under the actual level. If under-predictions frequently occur, i.e. the absolute value of $supe_t$ is due largely to these negative values, then the mitigation policy may be insufficient for the risk of an over-flow scenario. These errors would correspond to the risk of flood damages.

2.2 Inflow Prediction Model

Prepare Windowing Features and Inflow Moving Average Filter. To predict 3-hour forecasts, the models are applied to manipulate windowing [15] from the past three hours for the rain gauge features and river level sensor features. Incorporating the trends in dam inflow movements, the models are applied to invert the dam inflow moving average filer [16,17] for vibration reduction.

Generalised Linear Model and Additive GLM Regression. This paper discusses several applications via the linear model regression approach, such as the generalised linear model (GLM) [18-20]. The GLM with Gaussian family is useful for predicting a dam inflow time series. Furthermore, they extend the additive regression model based on the Gaussian GLM [21,22].

Regression Tree and Gradient Boosting Machine. This paper discusses some applications via non-linear models, such as the regression tree [23,24] and the gradient boosting machine (GBM) [25-28]. They enable optimising parameters like the number of trees and the maximum depth of a tree using grid searches.

3 Applied results

3.1 Dataset of Target Inflow and Features

Figure 1 shows the case study of a dam and a river region where three river height sensors are located. The purpose of the dam is hydropower generation and the dam type is PG which means concrete gravity. The height is 32 m and the reservoir capacity is 1.5 million m^3. The dam river region has an area scale of 112 km^2. It has a yearly electrical capacity of 279 MWh. The total arrival time takes 90 minutes from sensor-1 to the dam. In addition, the middle arrival takes time 60 minutes from sensor-2 to the dam. Third, the short arrival time takes time 20 minutes from sensor-3 to the dam.

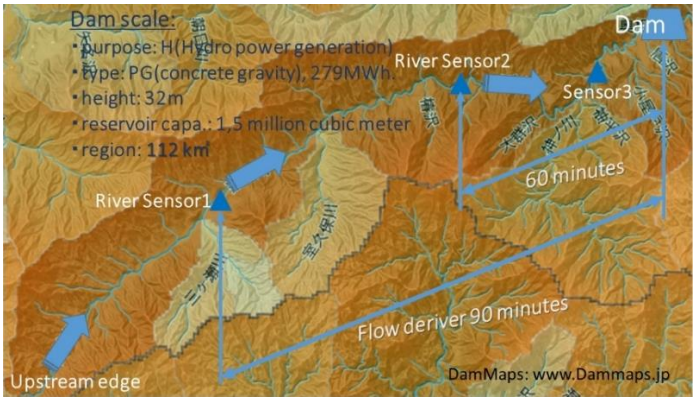

Fig. 1. Case of dam and river region with three sensors (http://www.dammaps.jp/?d=0699)

Table 2. Data name, role, and variable profile

Data name	Data role and feature profile
Dam-inflow time series	Target variable. The volume of dam-inflow time series from the upstream region. The interval is 15 minutes.
Rain gauge	Rain features. The quantity of rainfall measured at nine points over the upstream region.
River height sensor	Main stream river features. The level of the height sensed at the three points within the main river.

Table 2 lists the target dam inflow variable and several features, such as rain gauges and river level sensors. Here it is targeted to predict the forward time horizon to set at 12, i.e. the 3-hour forward forecast. Regarding the features, each of these window sizes is set at 12, i.e. they are the historical data of the past three hours. The unit time intervals are always 15 minutes.

3.2 Training data and Test set

The case study area is in the Kanto region in Japan. The data are from 10 years of data collected from 2006 to 2015, containing the 55 flood events with high water flows. The

number of target inflow value data is 20,627 in unit intervals of 15 minutes. Figure 2 shows an image of the training data and the test set split with linear sampling. The training set contains the 52 flood events from 2006 to 2014. In contrast, the test set contains the three flood events that recently occurred in 2015.

Training dataset	Test set
Flood events # 52	Flood events # 3 (53th-55th)
2006 **2014**	**2015**

Fig. 2. Training data and test set split with linear sampling

3.3 Prediction Results

The case study carried out several numerical experiments, where the computing environment is 64-bit under 8 GB RAM. This study was implemented using RapidMiner Studio version 8.1.

Figure 3 shows the accuracy comparison among the four prediction models based on the root mean square error (RMSE) and the correlation between each prediction and actual inflow. Table 3 lists the 5-fold cross validation results, each split under linear sampling. Both results suggested that the GLM extended additive regression model is more preferable than the other models, such as original GLM, the regression tree, and the gradient boosting machine.

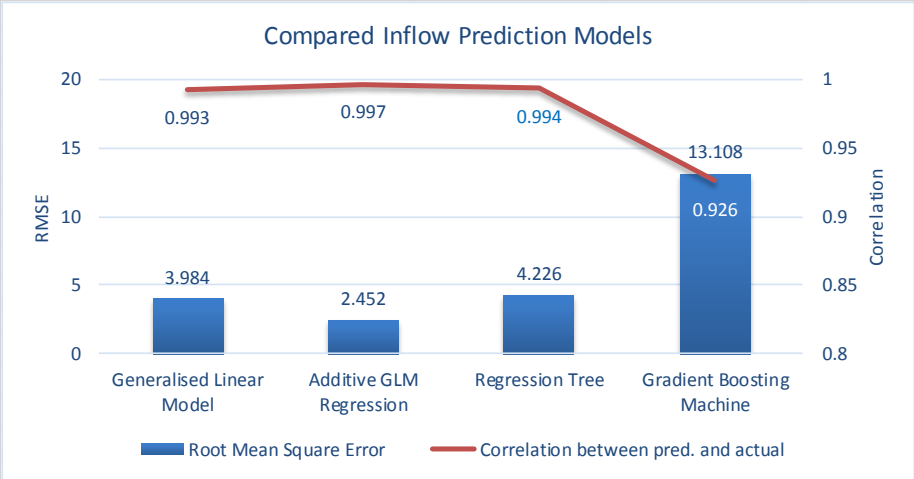

Fig. 3. Compared accuracy among prediction models based on root mean square error and correlation between prediction and actual inflow

The RMSE of the additive GLM regression is 2.452+/-0.594 with both the lowest mean error and the smallest deviation. Furthermore, the correlation between the actual

and predicted value is the highest value at 0.997+/-0.001. The execution time of 20 seconds is faster and practical. Hence, additive GLM regression is both an accurate and practical fast inflow prediction method for the applications compared in the experiments in this study.

Table 3. Cross validation results among prediction models

Prediction Model	Root Mean Square Error	Correlation between pred. and actual	TOC (Time of computation)
Generalised Linear Model (gaussian family, identity link)	3.984+/-0.638	0.993+/-0.004	2s
Additive GLM Regression (iteration=10, shrinkage=1.0)	**2.452+/-0.594**	**0.997+/-0.001**	**20s**
Regression Tree (number of depth=15)	4.226+/-1.894	0.994+/-0.001	18s
Gradient Boosting Machine (max depth=10, ntree=2,000)	13.108+/-4.128	0.926+/-0.044	8m54

Figures 4 through 6 show the plots of actual and prediction values using classification models, such as the original GLM under the Gaussian distribution, additive GLM regression, regression tree, and gradient boosting machine.

The GLM selected the best family compared with others such as the Poisson, gamma family, and so forth. The Tweedie family has the same accuracy as the Gaussian; however, the author selected the Gaussian distribution as the most usable and acceptable. Although GLM approximates good output like the complete hydrograph shape, we should be concerned about the weakness where the several jumped up points do not match the peak level far from the actual inflow, exactly at Figure 4.

The additive GLM regression model could almost approximate the same shape as the actual inflow series. We also focus on the strong points where the maximum inflow predictions are always below the actual peak inflow without over-prediction. This model only required 20 seconds for 5-fold cross validation. This result maintains that the additive GLM regression model is practical for dam inflow time series prediction tasks without over-prediction. In addition, such an advantage would correspond to minimising the loss of hydropower opportunities. Furthermore, there is little under-prediction owing to the smoothness of the additive function and the flexible formulation [21].

The gradient boosting machine (GBM) has parameters where the number of trees is 2000 and maximum depth is 10. This model only set two parameters to maximise the accuracy using grid search (e.g. the number of trees in an interval from five to 20 and maximum depth from 100 to 2000). This result means that the GBM could approximate almost the same shape of the actual inflow time series, although there are more biases than the additive GLM regression model because it has many fluctuations at the high inflow level close to the peak. The regression tree has little over-prediction; however,

there are some under-predictions that are less than the actual inflows, as shown in Figure 5. These under-predictions would correspond to the risk of flood damages.

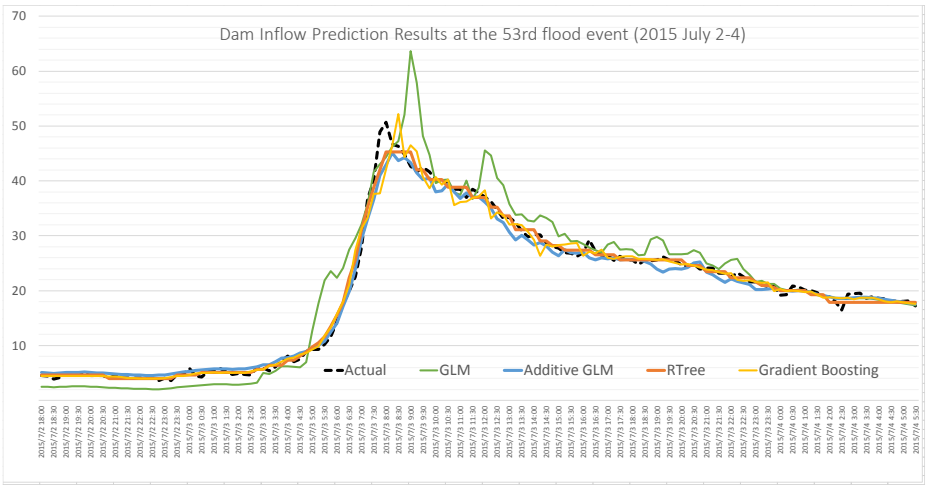

Fig. 4. The 53rd event of actual and prediction value using dam inflow time series regression

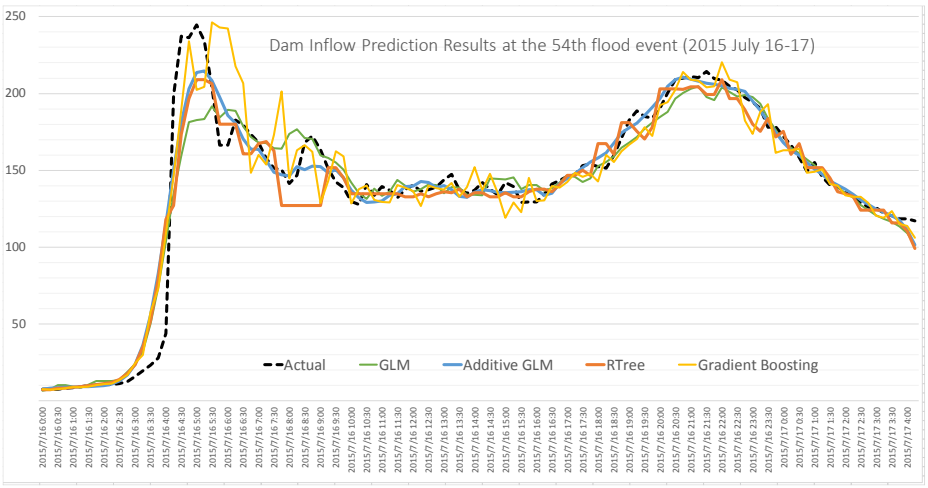

Fig. 5. The 54th event of actual and prediction value using dam inflow time series regression

Table 4 lists the calculated results of the prediction error indexes regarding over-prediction and under-prediction errors. In the left side three columns are the sum of over-prediction error, where the additive GLM regression model had very low scores. Therefore, this model had the advantage of the smallest loss of hydropower opportunities. On the right-side column is the sum of under-prediction error, where the additive GLM regression model is preferable at the lower scores based on the absolute values. Hence, this model had the advantage of a minimal risk of flood damages.

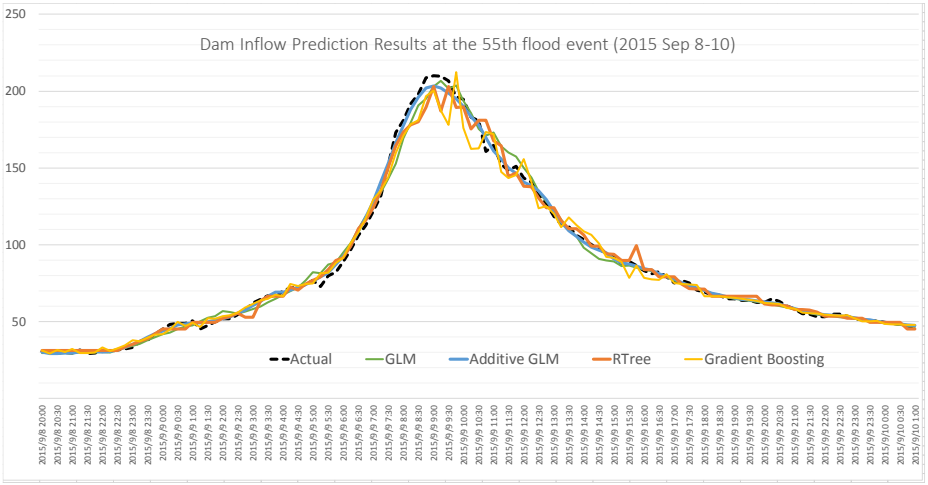

Fig. 6. The 55th event of actual and prediction value using dam inflow time series regression

Table 4. Computed results of loss indexes regarding over/under-prediction

	sum of over predictuion error (sope)			sum of under predictuion error (supe)		
test data of flood event	53th	54th	55th	53th	54th	55th
Generalised Linear Model (gaussian family, identity link)	250.8	490.7	172.3	-110.6	-609.8	-194.1
Additive GLM Regression (iteration=10, shrinkage=1.0)	**42.3**	400.5	**114.9**	-92.3	**-428.7**	**-104.3**
Regression Tree (number of depth=15)	45.3	**385.1**	169.7	**-53.0**	-758.8	-206.9
Gradient Boosting Machine (max depth=10, ntree=2,000)	60.0	704.2	177.5	-73.9	-607.6	-286.7

4 Concluding Remarks

4.1 Prototyping Dam Inflow Data Mining Process for Prediction

This paper proposed a method to predict dam inflows for the target of a 3-hour forecast using a time series data with a unit interval of 15 minutes. This study implemented numerical experiments applied to the 20,000 time stamped recorded dam inflow time series data. They contained the input of hundreds of time series data features, such as nine rain gauges in the dam region and three upstream river level sensors. These experiment results suggested that the GLM extended additive regression model is more preferable than other models such as the original GLM, regression tree and gradient boosting machine. The RMSE of the additive GLM regression was 2.452+/-0.594. Furthermore, the correlation between the actual and prediction value is the highest value at

0.997+/-0.001. The execution time of 20 seconds is fast and practical. Hence, the additive GLM regression is both an accurate and fast inflow prediction method in these experiments. The advantage is the smoothness of the additive function; therefore, there were few over-predictions and rare under-predictions. These points would correspond to minimise the loss of hydropower opportunities and the risk of flood damages.

4.2 Lesson learned for Efficient Hydropower Operations and Further Variations

If we use the proposed additive GLM Regression model method, we could approximate a maximum level of inflow prediction close to the actual peak level. It remains the more accurate and the faster possibility formulated by multiple predictors with the dam inflow and river height sensors via the vector generalised additive models [29]. This paper provided limited experience focusing on the Kanto area in Japan. Further variations will be obtained in cases where dam inflow time series mining and predictive machine learning are used to operate more efficient and safe hydropower generation. This case study contained only three main river sensors; however, there are opportunities to measure sensitive locations such as the edge of the upstream river which are broadly spanned at each sub rivers owing to reinforcement learning for hydropower generation performances. Further, we need research concerning the threshold for high and low inflow levels to predictively detect the boundary between normal inflows and anomaly inflows. For example, the one-class support vector machine [30] enhances the outlier inflows and computes the outlier scores to classify anomaly inflows and other flows. Furthermore, long-term 30-year data mining needs to be continued so that we can develop another accurate and fast dam inflow prediction application.

Acknowledgements. I wish to thank the DaMEMO committee and referees for their time and valuable comments. I also wish to thank Yachiyo Engineering Co., Ltd. for various support based on big data and AI project outcomes since 2012. I appreciate Mizuno Takashi for giving development opportunities and Amakata Masazumi for helping me the river and dam engineering domain knowledge and experiences.

References

1. UN News Centre: EM-DAT International Disaster Database-www.emdat.be & Reliefweb-reliefweb.int/disaster (2016).
2. Munich Re, Geo Risks Research: NatCatSERVICE Homepage, https://www.iii.org/fact-statistic/facts-statistics-global-catastrophes, last accessed 2017/12/31.
3. 2016 September 10, Japan floods: City of Joso hit by 'unprecedented' rain. BBC Homepage, http://www.bbc.com/news/world-asia-34205879, last accessed 2017/12/31.
4. Othman, F. et al.: Reservoir Inflow Forecasting using Artificial Neural Network. International Journal of the Physical Sciences 6(3), 434–440 (2011).
5. Attygalle D. et al.: Ensemble Forecast for Monthly Reservoir Inflow; A Dynamic Neural Network Approach. doi:10.5176/2251-1938_ors16.22 (2016).

6. Shortridge, J.E. et al.: Machine Learning Methods for Empirical Streamflow Simulation: A Comparison of Model Accuracy, Interpretability and Uncertainty in Seasonal Watersheds. Hydrology and Earth System Sciences 20, 2611–2628 (2016).
7. Li, B. et al.: Comparison of Random Forests and Other Statistical Methods for the Prediction of Lake Water Level: Case Study of Poyang Lake, China. Hydrology Research 47.S1 (2016).
8. Yang, J.-H. et al.: A Time-series Water Level Forecasting Model Based on Imputation and Variable Selection Method. Computational Intelligence and Neuroscience, 1–11 (2017).
9. Moeeni, H. et al.: Monthly Reservoir Inflow Forecasting using a New Hybrid SARIMA Genetic Programming Approach. Journal of Earth System Science 126:18), Indian Academy of Science (2017).
10. Hofmann, M., Klinkenberg, R.: RAPID MINER: Data Mining Use Cases and Business Analytics Applications. Data Mining and knowledge Discovery Series, CRC Press (2014).
11. Kotu, V., Deshpande, B.: Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner, Morgan Kaufmann, Elsevier USA (2015).
12. International Hydropower Association (iha): Cost-benefit & Economic Performance. Hydropower Sustainability Assessment Protocol, last accessed 2018/2/16.
13. Mattmann, M. et al.: Hydropower Externalities: A Meta-analysis. Energy Economics 57, 66–77 (2016).
14. International Renewable Energy Agency (IRENA), Renewable Energy Technologies: Cost Analysis Series. Hydropower vol. 1: Power Sector (2012).
15. Domingos, P.: MetaCost: A General Framework for Making Classifiers Cost-sensitive. CONFERENCE 1999, ACM KDD, pp. 165–174 (1999). doi=10.1.1.15.7095
16. Ling, C.X. et al.: Cost-Sensitive Learning and the Class Imbalance Problem. In : Sammut, C. (eds.) Encyclopedia of Machine Learning, Springer (2008).
17. Pavlyshenko, B.M.: Linear, Machine Learning and Probabilistic Approaches for Time Series Analysis. IEEE International Conference on Data Stream Mining & Processing (2016).
18. Wedderburn, R.W.M.: Quasi-likelihood Functions, Generalized linear Models and the Gauss-Newton Method. Biometrika *61(3)*, 439–447 (1974).
19. McCullagh, P., Nelder, J.: Generalized Linear Models. Monographs on Statistics and Applied Probability, Chapman & Hall (1983).
20. Hardin, J.W., Hilbe, J.M.: Generalized Linear Models and Extensions. 3rd ed., Stata (2012).
21. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning; Data Mining, Inference and Prediction. 2nd ed., Springer (2009).
22. Hastie, T., Tibshirani, R.: Generalize Additive Models. Statistical Science 1, 297–318 (1986).
23. Brieman, L., Friedman, J. et al.: Classification and Regression Trees. Wadsworth (1984).
24. Ripley, B.D.: Pattern Recognition and Neural Networks. Cambridge Univ. Press (1996).
25. Friedman, J., Hastie, T., Tibshirani, R.: Additive Logistic Regression: A Statistical View of Boosting (with discussion). Annals of Statistics 28, 307–337 (2000).
26. Friedman, J.: Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics 29(5), 1189–1232 (2001).
27. Scholkopf, B., Freund, Y.: Boosting: Foundations and Algorithms, MIT Press (2012).
28. Efron, B., Hastie, T.: Computer Age Statistical Inference: Algorithms, Evidence and Data Science. Cambridge Univ. Press (2016).
29. Yee, T.W., Wild, C.J.: Vector Generalized Additive Models. Journal of the Royal Statistical Society, Series B 58(3), 481–493 (1996).
30. Amer, M., Goldstein, M. et al.: Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection, 11th ODD'2013, Chicago (2013).

(Submitted 23 March 2018, Proceedings 21 Dec 2018)