Dear Editor and Referees,

Thank you for your time and for your thoughtful and constructive review; they have greatly improved the manuscript. Below, we address the points risen by the three anonymous reviewers and state how we would like to address them in a revised version of the manuscript. Our replies to the reviewers' comments are written in blue and normal font. We believe we have substantially addressed all of the outstanding comments and issues, and we look forward to your second review of the work. A marked-up version of the manuscript can be found right after the point-by-point answers at the end of this document.

On the behalf of all co-authors,

Yours sincerely,

Zhanwei Liu

Reply to Anonymous Referee #1

Thank you very much for your time and for your thoughtful and constructive review. The following are our point-by-point responses to your comments.

1. The manuscript presents multi-step ahead daily inflow forecasting using ERA-Interim reanalysis dataset based on gradient boosting regression trees, which is interesting. It is relevant and within the scope of the journal.

Response: Thank you very much for your positive comments.

2. Full names should be shown for all abbreviations in their first occurrence in texts. For example, ERA in page 1, ECMWF in page 3, etc.

Response: Thank you for your carefulness. We have shown full names for all abbreviations in their first occurrence in the revised manuscript. Please see Section Introduction for more details.

3. For readers to quickly catch your contribution, it would be better to highlight major difficulties and challenges, and your original achievements to overcome them, in a clearer way in abstract and introduction.

Response: Thank you for your suggestion. The major difficulties and challenges are selection of appropriate input variables related to inflow of longer lead times and effective prediction model. This paper proposed a new hybrid inflow forecast framework with ERA-Interim (European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis Interim) data as input, adopting gradient boosting regression trees (GBRT) and the maximal information coefficient (MIC) for multi-step ahead daily inflow forecasting. The proposed inflow forecast framework has three advantages. Firstly, the ERA-Interim dataset provides enough information for the framework to discover inflow for longer lead times. Secondly, MIC can identify effective feature subset from massive features that significantly affects inflow so that the framework can reduce computational burden, distinguish key attributes with unimportant ones and provide a concise understanding of inflow. Lastly, the GBRT is a prediction model in the form of an ensemble of decision trees and has a strong ability to capture nonlinear relationships between input and output in longer lead times more fully. We have made careful modifications in Section Abstract and Introduction of the revised manuscript.

4. It is mentioned in page 1 that ERA-Interim reanalysis data is adopted as input. What are other feasible alternatives? What are the advantages of adopting this particular data over others in this case? How will this affect the results? The authors should provide more details on this.

Response: Thank you for your careful review and suggestion. The ERA-Interim data is the result of assimilating observed data with forecast data, which has less error than observed data and forecast data (Balsamo et al., 2015). ERA-Interim data is produced by a fixed version of numerical weather prediction (NWP) system (Dee et al., 2011). The fixed version ensures there are no spurious trends caused by an evolving NWP system. Therefore, meteorological reanalysis data satisfies the need for long sequences of consistent data and have been used for the prediction of wind speeds (Stopa and Cheung 2014) and solar radiation (Linares-Rodríguez, Ruiz-Arias et al. 2011,

Ghimire, Deo et al. 2019). Meanwhile, ERA-Interim was proved to be one of the best reanalysis data describing atmospheric circulation and elements (Kishore et al., 2011). More details about ERA-Interim data are given in Section Appendix A of the revised manuscript.

5. It is mentioned in page 1 that gradient boosting regression tree is adopted as inflow forecast framework. What are the advantages of adopting this particular soft computing technique over others in this case? How will this affect the results? The authors should provide more details on this.

Response: Thank you for your careful review and suggestion. The gradient boosting regression trees (GBRT) (Friedman 2001, Fienen, Nolan et al. 2018), is a nonparametric machine learning method based on a boosting strategy and the decision trees model. The decision tree robust to outliers is used as a primitive model and boosting algorithm as integration rule is used to improve inflow forecasting accuracy. GBRT was developed and had been used in traffic (Zhan, Zhang et al., 2019) and environmental (Wei, Meng et al., 2019) field and proved to alleviate the problems of being trapped by local minima, over-fitting problems and reduced generalizing performance. More details about GBRT is given in Section 3.2 of the revised manuscript.

6. It is mentioned in page 1 that artificial neural networks, support vector regression and multiple linear regression models are adopted as benchmark for comparison. What are the other feasible alternatives? What are the advantages of adopting these particular models over others in this case? How will this affect the results? More details should be furnished.

Response: Thank you for your careful review and suggestion. The several studies had shown that artificial neural networks (ANN) (Rasouli et al., 2012; Cheng et al., 2015; El-Shafie and Noureldin, 2011; Chau, 2006; Ali Ghorbani et al., 2018) and support vector regression (SVR) (Tongal and Booij, 2018; Luo et al., 2019; Moazenzadeh et al., 2018) are the two powerful models for inflow predicting. They are widely used and very mature algorithms, which are scientific and reasonable compared with them. Please see Section Introduction for more details about compared model.

7. It is mentioned in page 3 that the Xiaowan Hydropower Station is adopted as the case study. What are other feasible alternatives? What are the advantages of adopting this particular case study over others in this case? How will this affect the results? The authors should provide more details on this.

Response: Thank you for your careful review and suggestion. The Xiaowan Hydropower Station in the lower reaches of the Lancang River, which is the longest river with most discard water in Yunnan Province, is chosen as the study site (as shown in Fig. 2 in the revised manuscript). The Xiaowan Hydropower Station is the main controlling hydropower station in the Lancang River and it is very meaningful to adopt the Xiaowan Hydropower Station as the case study. Please see Section 2.1 of the revised manuscript for more details about case study.

8. It is mentioned in page 4 that the maximum information coefficient is adopted to select inputs from 79 potential predictors from reanalysis data. What are the advantages of adopting this particular approach over others in this case? How will this affect the results? The authors should provide more details on this.

Response: Thank you for your careful review and suggestion. The maximal information coefficient (MIC) (Reshef et al., 2011) is a robust measure of the degree of correlation between two variables and has attracted a lot attention from academia (Zhao et al., 2013; Ge et al., 2016; Lyu et al., 2017; Sun et al., 2018), which can identify effective feature subset from massive features that significantly affects inflow so that the framework can reduce computational burden, distinguish key attributes with unimportant ones and provide a concise understanding of inflow. Please see Section 3.1 of the revised manuscript for more details about inputs selection.

9. It is mentioned in page 4 that autocorrelation function is adopted to identify observed inflow and rainfall lags. What are other feasible alternatives? What are the advantages of adopting this particular approach over others in this case? How will this affect the results? The authors should provide more details on this.

Response: Thank you for your careful review and suggestion. The autocorrelation function (ACF) measures the dependency or relationship of observed value with lagged observations of a considered variable. In a long memory time series such as inflow time series, the ACF declines slowly (as shown in Fig. R1). The partial autocorrelation function (PACF) and cross-correlation function (CCF) (as shown in Fig. 6 in the revised manuscript) are two other feasible alternatives. We use the PACF and CCF for modeling, calculation and analysis according to Referee(#3)'s suggestion. We agree to use PACF and CCF replace ACF to determine the model structure. In the revised manuscript, PACF and CCF are adopted to determining the model structures for inflow and rainfall, respectively. 95% confidence interval is used to determine the significant relationships replacing user-defined threshold value. Please see Section 2.2 and Section 4.1 of the revised manuscript for more details.

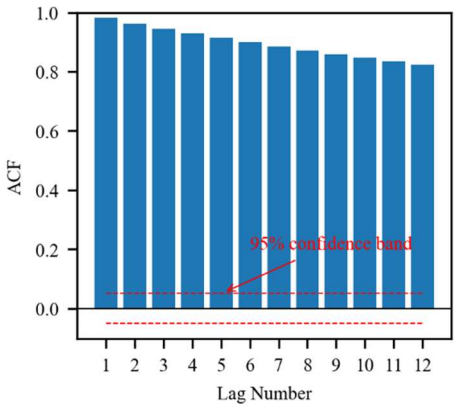


Figure R1. ACF plots of Xiaowan inflow time series.

10. It is mentioned in page 6 that four evaluation criteria are adopted to evaluate the performance of the models. What are the other feasible alternatives? What are the advantages of adopting these particular evaluation criteria over others in this case? How will this affect the results? More details should be furnished.

Response: Thank you for your careful review and suggestion. The root mean squared error (RMSE) and mean absolute error (MAE) are the most commonly used criteria to assess model performance (Luo et al., 2019; Chau, 2005; Chau, 2006). The Pearson correlation coefficient (CORR) is a measure of the strength of the association between observed inflow series and forecasted inflow series. Nash-Sutcliffe efficiency coefficient (NSE) is

replaced by Kling-Gupta efficiency metrics (KGE) and the Index of Agreement (IA) according to results of trials and Referee(#3)'s suggestion. The percent bias in flow duration curve high-segment volume (BHV) is introduced to evaluate the performance of forecasting extreme values for developed model according to results of trials and Referee(#2)'s suggestion. KGE, IA and BHV and are added to compare several model performances in Section 3.3 of the revised manuscript.

11. It is mentioned in page 7 that a grid search algorithm is adopted to optimization model parameters. What are other feasible alternatives? What are the advantages of adopting this particular algorithm over others in this case? How will this affect the results? The authors should provide more details on this.

Response: Thank you for your careful review and suggestion. The grid search algorithm, which is an exhaustive search all candidate parameter combination method, is guided to optimizing model parameters by evaluation of validation set for each lead time (Chicco and Davide, 2017). Grid search is considered as an effective parameter search method, which is widely used (Fienen et al., 2018). Two of other feasible alternatives are randomized search and Bayesian optimization. We have performed some numerical trials to compare grid search, randomized search and Bayesian optimization, and grid search can obtain more reasonable and stable hyperparameter combination. More details about grid search are given in Section 4.2 of the revised paper.

12. It is mentioned in page 9 that grid searching is adopted to tune the hyperparameters of GBRT, GBRT-MIC, ANN-MIC. What are other feasible alternatives? What are the advantages of adopting this particular approach over others in this case? How will this affect the results? The authors should provide more details on this.

Response: Thank you for your careful review and suggestion. Same as question 11, more details about grid search are given in Section 4.2 of the revised paper.

13. It is mentioned in page 9 that Bayesian optimization (Snoek et al., 2012) is adopted to tune the hyperparameters of SVR-MIC. What are other feasible alternatives? What are the advantages of adopting this particular approach over others in this case? How will this affect the results? The authors should provide more details on this.

Response: Thank you for your careful review and suggestion. At present, there are three commonly used methods of hyperparameter selection: grid search, random search and Bayesian optimization (Snoek et al., 2012). We have performed some numerical trials to compare grid search, randomized search and Bayesian optimization, and grid search can obtain more reasonable and stable hyperparameter combination. Bayesian optimization has been replaced by grid search method. Please see Section 3.4 for more details.

14. It is mentioned in page 9 that Python is adopted to perform all computations. What are other feasible alternatives? What are the advantages of adopting this particular software over others in this case? How will this affect the results? The authors should provide more details on this.

Response: Thank you for your careful review and suggestion. Python is an important tool for scientific computing and data analysis, which is powerful, fast and open. Brief introduction about Python is given in Section 4 of the revised manuscript.

15. Some key parameters are not mentioned. The rationale on the choice of the particular set of parameters should be explained with more details. Have the authors experimented with other sets of values? What are the sensitivities of these parameters on the results?

Response: Thank you for your careful review and suggestion. For ANN, A range of 2-20 neurons and four activation functions (as shown in Table 4 in the revised manuscript) are selected by a trail-and-error procedure. The sensitivities of these parameters have been analyzed by trying different parameter combinations (as shown in Fig. 7 in the revised manuscript). Meanwhile, referring to (Fienen et al., 2018; Friedman, 2001; Pedregosa et al., 2011), more wide ranges of the model parameters are used for grid search in Section 4.2 of the revised manuscript.

16. Some assumptions are stated in various sections. Justifications should be provided on these assumptions. Evaluation on how they will affect the results should be made.

Response: Thank you for your careful review and suggestion. The comparison of different models is based on the basic assumption that parameters are optimal. In the revised manuscript, grid search is employed to tune the hyperparameters of model. A lot of models for each lead time are developed to find as possible as optimal parameters and this assumption of optimal parameters can be satisfied. More details about the assumption of optimal parameters are given in Section 4.2 of the revised paper.

17. The discussion section in the present form is relatively weak and should be strengthened with more details and justifications.

Response: Thank you for your careful review and suggestion. KGE, IA and BHV are added to compare several model performances and more details about the discussion of the obtained results have been added. The discussion of the obtained results is enriched in Section 4 of the revised manuscript.

18. Moreover, the manuscript could be substantially improved by relying and citing more on recent literatures about contemporary real-life case studies of soft computing techniques in hydrological prediction such as the followings: ïAn Yaseen, Z.M., et al., "An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction," Journal of Hydrology 569: 387-408 2019. ïAn Fotovatikhah, F., et al., "Survey of Computational Intelligence as ˘Basis to Big Flood Management: Challenges, research directions and Future Work," Engineering Applications of Computational Fluid Mechanics 12 (1): 411-437 2018. ïAn Mosavi, A., et al., "Flood Prediction Using Machine Learning Models: Literature Review," Water 10 (11): article no. 1536 2018. ïAn Moazenzadeh, R., et al., "Coupling a ˘firefly algorithm with support vector regression to predict evaporation in northern Iran," Engineering Applications of Computational Fluid Mechanics 12 (1): 584-597 2018. ïAn Ghorbani, M.A., et al., "Forecasting pan evaporation with an integrated Artificial Neural Network Quantum-behaved Particle Swarm

Optimization model: a case study in Talesh, Northern Iran," Engineering Applications of Computational Fluid Mechanics 12 (1): 724-737 2018. ïAn Chau, K.W., et al., "Use of Meta-Heuristic Techniques in Rainfall-Runoff Modelling" Water 9(3): article no. 186, 6p 2017.

Response: Thanks. We have carefully looked up the mentioned literature, which has been cited in the paper, and we have also extensively looked up other literatures from HESS, JH and other relative journals, added some necessary literatures. Please see Section References.

19. Some inconsistencies and minor errors that needed attention are: ïAn Replace "... was supply to depict…" with "…was supplied to depict ..." in line 86 of page 3 ïAn Replace "... into train set, validation set, and test set..." with "...into training set, validation set, and testing set..." in line 206-207 of page 7 ïAn Replace "... test set..." with "... testing set..." in line 209 of page 7 ïAn Replace "... more accuracy inflow forecasting..." with "... more accurate inflow forecasting ..." in line 283 of p10 ïAn Replace "... arisen in in areas..." with "... arisen in areas..." in line 288 of page 10 ïAn Replace "... for train, validation and test set ..." with "... for training, validation and testing set ..." in line 294 of page 10 ïAn Replace "... According to compare the forecasted results of …" with "... According to the comparison of forecasted results of ... " in line 330 of page 11

Response: Thank you for your careful review. We agree with all of the minor changes above, and we will go through carefully the manuscript to check and correct any errors. Those typos have been corrected in the revision.

20. In the conclusion section, the limitations of this study, suggested improvements of this work and future directions should be highlighted.

Response: Thank you for your careful review and suggestion. We have carefully checked the conclusion of the article and added the limitations and future directions of this study in the revised manuscript. Please see page Section 5 for more details.

Reply to Anonymous Referee #2

I went through the manuscript. Generally, the manuscript has been well organized.

Response: We thank you very much for reviewing the manuscript and giving the positive comment. The following are our point-by-point responses to your comments.

1. P1, L6, In abstract the authors stated that "The impacts of climate change and human activities make accurate inflow prediction increasingly difficult, especially for longer lead times". As far as I know, the climate change deals with long term trends, say the climate variation over 20 years. I cannot understand relevance of the abovementioned with climate change impacts and human activities.

Response: Thank you for your careful review. Climate variation affects the streamflow directly. For example, changing precipitation patterns and intensity, together with changing temperatures, will greatly modify the streamflow. Human activities such as land use change, water withdrawal, and hydraulic structures have substantial impacts on streamflow. We fully approve that climate variation and human activities can generate large effect to streamflow of medium and long term. For short term streamflow forecasting, climate variation and human activities also have some effect, so we need to calibration parameter according to meteorological factors. Please see Section Introduction for more details.

2. The authors have to clearly indicate which model was developed for the inflow forecasting. At first they have to demonstrate if they used conceptual models or data driven models. What is the advantage of the developed model?

Response: Thank you for your careful review. The gradient boosting regression trees (GBRT) is used to forecast daily streamflow. The model is a data-driven model. Compared with artificial neural network (ANN), GBRT has two main advantages. Firstly, GBRT can rank features according to their contribution to model scores, which is of great significance for reducing the complexity of the model. Secondly, GBRT is a white box model and can be easily interpreted. The more details and advantages of model developed are given in Section 3.2 of the revised manuscript according to your suggestion.

3. The input selection for multi-day ahead forecasting should be discussed according to available literature. It is essential why the input structure of the longer period is not updated following literature the earlier stage forecasts.

Response: Thank you for your careful review and suggestion. We have carefully reviewed more literatures about the input selection for multi-day ahead forecasting. There are mainly two strategies that you can use for multi-step forecasting for single-output, namely, Static (Direct) multi-step forecast and Recursive multi-step forecast (Gianluca et al., 2013; Taieb et al., 2012). Recursive forecast strategy is biased when the underlying model is nonlinear; it is sensitive to the estimation error instead of actual ones since estimated values are more and more used when we get further in the future (Bontempi et al., 2012). Thus, the Static multi-step forecasting strategy is

employed in this paper. Since the Static strategy does not use any approximated values to compute the forecasts, it is not prone to any accumulation of errors. The model structure of one-step and two-step forecasting of Static strategy is listed below (as shown in Section 3.4) which has different model parameters.

$$prediction(t+1) = model1\big(obs(t-1), obs(t-2), \ldots, obs(t-n)\big)$$

$$prediction(t+2) = model2(obs(t-1), obs(t-2), \ldots, obs(t-n))$$

where $obs(t-1)$ is the observed value at the $t-1$ period and $prediction(t+1)$ is the predicted value of one-step at the t period. More details about multi-step forecasting are shown in Section 3.4 of the revised manuscript.

4. Literature should be updated discussing on more papers addressing multi-step ahead forecasting.

Response: Thank you for your careful review. We have carefully reviewed more literatures from HESS, JH and relative journals about multi-step ahead forecasting and the more references about multi-step ahead forecasting have been updated in Section Reference of the revised manuscript.

5. The authors employed gradient boosting regression trees as an ensemble framework. More explanations required about ensemble members.

Response: Thank you for your careful review and suggestion. The ensemble member is the decision tree model. More details about decision tree model are given in Section 3.3.1 in the revised manuscript.

6. Uncertainty analysis should be carried out to show how much the predictions are confident. As the lead time increases, the metrics reveal errors are increasing drastically. Moreover, high uncertainties are expected to associate with such models. Please discuss this issues accordingly.

Response: Thank you for your careful review. We agree that uncertainty analysis in predictions is significant. As far as we know, medium and long-term forecasting is more uncertainty, for example, monthly or yearly. This paper focuses on improving inflow prediction accuracy of lead times of one to ten days by developing new model and importing ERA-Interim reanalysis data, which aims to providing reference for reducing discard water. As time goes on, model often needs to be rebuilt and parameters of model need to be recalibrated according to the actual flow and meteorological data. The uncertainty analysis of medium and long-term inflow forecasting will be further studied in the next study.

7. Concerning inflow predictions, please indicate efficiency of the proposed model to simulate and predict extreme values which are of great importance.

Response: Thank you for your careful review and suggestion. We introduce the percent bias in flow duration curve high-segment volume (BHV) to evaluate the performance of catching extreme values for developed model and more details have been given in Section 3.3 of the revised manuscript.

Reply to Anonymous Referee #3

We thank you very much for reviewing the manuscript and giving the positive comment. The following are our point-by-point responses to your comments.

1. In this manuscript, the authors compared several data-driven models for multi-step forecasting of inflow. The employed models include gradient boosting regression trees (GBRT), artificial neural networks (ANN), support vector regression (SVR), and multiple linear regression (MLR) models. The models were developed by considering (1) streamflow and rainfall record, and (2) ERA-Interim reanalysis data. Further, the maximum information coefficient and autocorrelation functions were utilized to construct the input structures of the models. The authors concluded that the developed methodology that considers ERA-Interim reanalysis data considerably gives better results in the forecasting of inflows at lead times of 5-10 days. The manuscript is well written and organized. However, there is not a significant novelty in the manuscript except using ERA-Interim dataset. Further, there are severe weaknesses in the developing of the model input structures.

Response: Thank you very much for your time and for your thoughtful and constructive review, and also thank you for giving some positive comments. We have made careful modifications in Abstract and Introduction of the revised manuscript for readers to quickly catch our contribution. This paper focuses on improving prediction accuracy by three significant measures. Firstly, ERA-Interim reanalysis data is introduced to provide enough information for the model to discover inflow for longer lead times. Secondly, gradient boosting regression trees (GBRT) is adopted to implement inflow forecasting and GBRT has been used to achieve multi-step inflow forecasting. Thirdly, most widely used models are developed to compare with GBRT for multi-step inflow forecasting, which demonstrates that developed model improves inflow forecasting accuracy. In order to make it easier for the reader to grasp the innovation of this paper, we have modified the Abstract and Introduction carefully to make the innovation more prominent. More details are given in the revised version.

2. The authors made a significant mistake in using the autocorrelation function (ACF) in determining the model structures. They should have employed cross-correlation and partial autocorrelation functions (or other measures) to establish the relationship between the observed records and inflow. The ACF only measures the dependency or relationship of observed value with lagged observations of a considered variable. In a long-dependent series such as inflow time series, the ACF will decay slowly. Therefore, statistically significant relationships between the observed and lagged values could not be determined. To determine the significant relationships, the authors employed user-defined threshold value. The obtained inflow and rainfall values for the input structures of the models include only three lagged-day values as could be expected. This number could be higher based on the selected threshold. However, this finding does not convey any meaningful relationship between the observed records (i.e. inflow and rainfall) and the inflow values. The PACF should have been used for determining the lagged relationships of inflows since the inflow time series mainly shows the long-memory feature where the correlation decays after a long observation period. Further, all statistically significant lagged variables should have been included in the model structures found in PACF. Using a user-defined threshold value is a serious mistake in this situation.

Response: Thank you for your careful review and nice comments. According to your suggestions, we use the partial autocorrelation function (PACF) and cross correlation function (CCF) for modeling, calculation and analysis, and find that your suggestions are effective. We agree to replace the autocorrelation function (ACF) to determine the model structure and 95% confidence levels obtained by hypothesis test is used to replace user-defined threshold value to determine the significant relationships (as shown in Fig. 6 in the revised manuscript). The all calculation results have been updated accordingly (as shown in Table 3 in the revised manuscript). In Section 4.1 of the revised manuscript, PACF and CCF to determining the model structures for inflow and rainfall, respectively.

3. The authors claimed that the proposed methodology "significantly" improves the accuracy of inflow prediction for longer lead times. However, I do not agree with this comment. Because, as the authors mentioned, there is only about 1% and 5% improvement in two-day and 10-day ahead forecasting. Therefore, the results do not seem convincing about the superiority of ERA-Interim dataset over the common dataset, especially ill-conditioned input structures with conventional observed inflow and rainfall dataset.

Response: Thank you for your careful review. Revised input structures are used to compare with developed model with ERA-Interim dataset. Table 8 of the revised paper shows performance indices of model in the testing set. The results indicate that the developed method generally performs better than other models and improves the accuracy of inflow forecasting in four and ten-day ahead forecasting. It should be noted that the results of "Supplementary response to Referee 3" has some mistakes because of a tight deadline; they have been corrected in the revised version of manuscript. More discussion about results of models are given in Section 4.4 of the revised manuscript.

4. The authors found that three-day lagged values of inflow and rainfall have less impact on 10-day ahead forecasting of inflow in Section 4.5. This is a clue that more lagged values of input variables should have been included in the models' structure.

Response: Thank you for your careful review and suggestion. According to your suggestion, PACF and CCF are used to determining the model structures for inflow and rainfall, respectively. The results indicate that one-day and four-day lagged values of inflow and one to six-day lagged values of rainfall are included in the model's structure in the revised manuscript. Please see Section 4.1 of the revised manuscript for more details.

5. The employed performance indices, specifically the coefficient of determination, seems insufficient to compare several model performances. More distinctive performance indices such as degree of agreement and Kling-Gupta efficiency metrics should have been used.

Response: Thanks. The Pearson correlation coefficient (CORR) is a measure of the strength of the association between observed inflow series and forecasted inflow series. The root mean squared error (RMSE) and mean absolute error (MAE) are the most commonly used criteria to assess model performance (Luo et al., 2019; Chau, 2005; Chau, 2006). According to Referee (#2)'s and your suggestions, Nash-Sutcliffe efficiency coefficient (NSE) is removed, Kling-Gupta efficiency metrics (KGE), the percent bias in flow duration curve high-segment volume

(BHV) and the Index of Agreement (IA) are introduced as supplements. Please see Section 3.3 of the revised manuscript for more details.

6. It is not clear how the multi-step forecasting scheme (i.e., recursive or static) was employed? Please give more details about this issue.

Response: Thank you for your careful review and suggestion. Recursive forecast strategy is biased when the underlying model is nonlinear and is sensitive to the estimation error, since estimated values, instead of actual ones, are more and more used when we get further in the future (Bontempi et al., 2012). Thus, the Static multi-step forecasting strategy is employed and the models of different lead times have different model parameters. The model structure of one-step and two-step forecasting of Static strategy is listed below which has different model parameters.

$$prediction(t+1) = model1\big(obs(t-1), obs(t-2), \dots, obs(t-n)\big)$$
$$prediction(t+2) = model2(obs(t-1), obs(t-2), \dots, obs(t-n))$$

where $obs(t-1)$ is the observation value at the $t-1$ period and $prediction(t+1)$ is the predicted value of one-step at the t period. More details about multi-step forecasting are shown in Section 3.4 of the revised manuscript.

7. The selected ranges of the model parameters seem highly subjective. Please justify the selected ranges of the model parameters, especially in Section 4.2.

Response: Thank you for your careful review and suggestion. Specifying the selected ranges of the model parameters is the trickiest part of hyperparameter optimization. For gradient boosting regression trees (GBRT), we refer to (Fienen et al., 2018; Friedman, 2001; Pedregosa et al., 2011) to inform our choices of hyperparameter distributions. It can be difficult to figure out the interaction between hyperparameters. We have used wide ranges of the model parameters (as shown in Table 5 and Table 6 in the revised manuscript) and the model parameters have been justified in Section 4.2 of the revised manuscript.

8. The range for the number of hidden neurons (i.e. 2–20) seems too high. Please justify this from a hydrological perspective. Because using a high number of hidden neurons could lead to overfitting that resulted in a poor performance in multi-step forecasting.

Response: Thank you. Specifying the number of hidden nodes is a difficult task (Badrzadeh et al., 2013) and the number of hidden nodes is determined by a trial-and-error procedure in the original paper. In cases where we are not sure about the best number of hidden nodes, we use wide ranges and let the trial-and-error procedure do the reasoning for us. It is found that the optimal number of neurons is 2, 3 or 4. Please see Section 4.2 for more details.

9. The discussion of the obtained results should be improved with more details, especially giving necessary citations to previous studies.

Response: Thank you for your careful review and suggestion. KGE, IA and BHV have been added to compare several model performances and more details about the discussion of the obtained results has been added. In addition, we have carefully looked up the related literature from HESS, JH and other relative journals and some necessary citations to previous studies are discussed. Please see Section 5 and Reference for more details.

10. It is not clear how Fig. 1 was obtained. Please give the necessary information about this figure.

Response: Thank you for your careful review. We cooperate with production unit for a long time and the data of Fig. 1 from production unit has been obtained from public website. We give the source link of the data in the revised manuscript. Please see Section Introduction for more details.

11. Please give more details on the Lines 78–82.

Response: Thanks. More details about ERA-Interim dataset have been introduced in Section Appendix A the revised manuscript.

12. Please give the definitions and meanings of the variables in the ERA-Interim dataset in the Appendix.

Response: Thank you for your suggestion. The definitions and meanings of the variables in the ERA-Interim dataset are given in Section Appendix A of the revised manuscript.

13. Please justify using the feature scaling in Line 108.

Response: Thank you for your suggestion. The "data scaling" has been replaced by "feature scaling". Please see Section 2.2 for more details.

14. What do you mean with "invalid variables" in Line 116?

Response: Thanks. The "invalid variables" in Line 116 mainly demonstrate the weak-correlated variables which has a weak correlation and cannot interpret inflow very well. The "invalid variables" has been modified to "redundant feature information" in the revised manuscript. Please see Section 2.2 for more details.

15. Please prefer "maximal" or "maximum" information criterion throughout the manuscript.

Response: Thank you for your careful review. All "maximum" information criterion in the original manuscript has been modified to "maximal" information criterion in the revised manuscript. Please see Section Abstract for more details.

16. Please check the term MI(D,X,Y) in Eq. (5) since you defined MI*(D,x,y) in Line 130.*

Response: Thank you for your careful review. The term MI*(D,X,Y) has been modified to MI*(D,x,y). Please see Eq. (5) in Section 3.1.

17. The definition of B(n) was given in Line 133; however it is not clear where this parameter is used.

Response: Thank you for your careful review. B(n) is the maximal grid size which is a function of sample size and we usually set $B = n^{0.6}$. Some details about B(n) are added in the revised manuscript. Please see Section 3.1.

18. Please check the terms in Eq. (7). Will they be R1(i,s) or R1(j,s)?

Response: Thank you for your careful review. R1(i,s) and R2(i,s) in Eq. (7) has been modified to R1(j,s) and R2(j,s) in the revised manuscript. Please see Section 3.1.

19. Please check the notations in Line 144; n features with N samples or n samples with N features according to the given definition.

Response: Thanks. To avoid ambiguity, the sentence has been corrected in the revised version of the manuscript. The notation shows N features with n samples. Please see Section 3.1.

20. There is little information about the structure of ERA-Interim dataset. Please give more details about this dataset.

Response: Thank you for your careful review. There are detailed introductions for ERA-Interim dataset in the https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/ era-interim. According to your suggestion, more detailed information about variables of ERA-Interim dataset has been added in the revised manuscript. More details about ERA-Interim dataset are given in Section Appendix A of the revised manuscript.

21. There is not any information about grid searching methodology.

Response: Thank you for your careful review and suggestion. Grid search is considered as an effective parameter search method, which is widely used (Fienen et al., 2018). The grid search methodology is introduced in detail in Section 4.2 of the revised manuscript.

22. Please add "activation function" after "relu" in Line 248.

Response: Thank you for your suggestion. In order to show the results more clearly, this sentence has been replaced with "The results of the trials show *tanh* and *logistic* function are two more robust activation function and ANN with fewer nodes is inclined to obtain lower error" in the revised manuscript. Please see Section 4.2.

23. The comments in Lines 278–280 are vague.

Response: Thank you for your careful review. The comments in Lines 278–280 indicate the relationship between performance indices and lead times in the testing set (2017-2018). We mainly discuss the trend of performance indices as increasing lead time. The comments about the relationship between performance indices and lead times are given more details in Section 4.3 and Section 4.4.

24. The authors did not discuss the reasons why NSE values for lead times of 6-7-8-9-day is worse than the value of lead time of 10-day.

Response: Thank you for your careful review. In original manuscript, it should be noted that NSE values for lead times of 6-7-8-9-day is worse than the value of lead time of 10-day in the validation set. We consider the possible reasons are inadequate parameter optimization and different model structure. MAE is the objective function of parameter optimization instead of NSE. According to the results of trials and your suggestion, we replace NSE with KGE and more discussion are added in Section 4 of the revised manuscript. BHV and KGE have no consistent trend as increasing lead times. Please see Section 4.4 for more details.

25. It is not clear how top k features were selected according to the chosen threshold value. Did the authors employ several threshold values? Please give more details on this issue.

Response: Thank you for your careful review and suggestion. The original manuscript totally employs three threshold values. Two of these thresholds are used to determine the model input structures with inflow and rainfall. Another threshold value is used to determine the model input structures with ERA-Interim dataset.

In the revised manuscript, MIC is employed to select inputs from 26 candidate predictors from reanalysis data; observed inflow and rainfall lags are identified by partial autocorrelation function (PACF) and cross-correlation function (CCF) of the inflow time series. The corresponding 95% confidence interval is used to identify significant correlations. Furthermore, when correlation coefficient slowly declines and cannot fall into confidence interval, a trial-and-error procedure is used to determine the optimum lag, i.e., starting from one-lag and then modifying the external inputs by successively adding one more lagged time series into inputs (Amiri 2015; Shoaib et al., 2015). Consider the subjectivity of user-defined thresholds, the three threshold values are modified by hypothesis test and the trail-and-error procedure to determine input structures of model. Please see Section 2.2 and Section 4.1 for more details.

Multi-step ahead daily inflow forecasting using ERA-Interim reanalysis dataset based on gradient boosting regression trees

Shengli Liao[1], Zhanwei Liu[1], Benxi Liu[1], Chuntian Cheng[1], Xinfeng Jin[1], Zhipeng Zhao[1]
[1]Institute of Hydropower System and Hydroinformatics, Dalian University of Technology, Dalian 116024, China
Correspondence to: Zhanwei Liu (337891617@qq.com)

Abstract. Inflow forecasting plays an essential role in reservoir management and operation. The impacts of climate change and human activities make accurate inflow prediction increasingly difficult, especially for longer lead times. In this study, a new hybrid inflow forecast framework with ERA-Interim reanalysis ~~data~~dataset as input, adopting gradient boosting regression trees (GBRT) and the ~~maximum~~maximal information coefficient (MIC) ~~was~~is developed for multi-step ahead daily inflow forecasting. Firstly, the ERA-Interim reanalysis dataset provides ~~enough~~more information for the framework to discover inflow for longer lead times. Secondly, MIC can identify effective feature subset from massive features that significantly affects inflow so that the framework can ~~avoid over-fitting~~reduce computational burden, distinguish key attributes with unimportant ones and provide a concise understanding of inflow. Lastly, the GBRT is a prediction model in the form of an ensemble of decision trees and has a strong ability to capture nonlinear relationships between input and output ~~in long~~at longer lead times more fully. The Xiaowan hydropower station located in Yunnan Province, China is selected as the study area. ~~Four~~Six evaluation criteria, the mean absolute error (MAE), the root mean ~~square~~squared error (RMSE), the ~~Nash-Sutcliffe efficiency coefficient (NSE) and the~~ Pearson correlation coefficient (CORR), ~~were~~Kling–Gupta efficiency scores (KGE), the percent bias in flow duration curve high-segment volume (BHV) and the Index of Agreement (IA) are used to evaluate the established models using historical daily inflow data (1/1/2017-31/12/2018). Performance of the presented framework ~~was~~is compared to that of artificial neural networks (ANN), support vector regression (SVR) and multiple linear regression (MLR) models. The ~~experimental~~ results indicate that the ~~developed method generally performs better than other models and significantly improves the accuracy of inflow forecasting at lead times of 5-10 days. The~~ reanalysis data ~~also~~ enhances the accuracy of inflow forecasting ~~except~~ for ~~forecasts that are one-day ahead.~~all lead times studied (1-10 days) and the developed method generally performs better than other models, especially for the extreme values and longer lead times (4-10 days).

Keywords Inflow forecasting, Gradient boosting, Regression trees, ~~Maximum~~Maximal information coefficient, ERA-Interim

1 Introduction

Reliable and accurate inflow forecasting 1-10 days in advance is significant for efficient utilization of water resources, reservoir operation and flood control, especially in areas with concentrated rainfall. Rainfall in southern China is usually concentrated for several days at a time due to strong convective weather, ~~such as~~for example, typhoons. Low accuracy inflow predictions can easily cause the failure of power stations to make reasonable power generation plans 7-10 days ahead of disaster events and lead to unnecessary water abandonment and even substantial economic losses. Fig. 1 shows the ~~loss~~losses of electric quantity due to discarded water (LEQDW) in Yunnan and Sichuan Provinces, China from 2011 to 2016~~.~~ (Sohu, 2017; in-en, 2018). The total amount of LEQDW in Yunnan and Sichuan Provinces increased from 1.5 billion kWh to ~~47.9~~45.6 billion kWh from 2011 to 2016, with an average annual growth rate of ~~99.9~~98.0%. In recent years, due to the increased number of hydropower ~~station~~stations and installed hydropower ~~capacity~~capacities, the problem of discarding water caused by inaccurate inflow forecasting is becoming increasingly serious, which has also ~~had~~produced a negative impact on the development of hydropower in China.

The main ~~challenges~~challenge in inflow forecasting caused by climate change and human activities at present are low accuracy, especially for longer lead times (Badrzadeh et al., 2013; El-Shafie et al., 2007). ~~To address the problem~~Meanwhile, due to streamflow variation by reason of climate change and human activities, inflow forecasting model often needs to be rebuilt and the model parameters need to be recalibrated according to the actual inflow and meteorological data within one or two years. To address these problems, a variety of models and approaches have been developed. These approaches can be divided into

three categories: statistical methods (Valipour et al., 2013), physical methods (Duan et al., 1992; Wang et al., 2011; Robertson et al., 2013), and machine learning methods (Chau et al., 2005; Liu et al., 2015; Rajaee et al., 2019; Zhang et al., 2018; Yaseen et al., 2019; Fotovatikhah et al., 2018; Mosavi et al., 2018; Chau, 2017). Each method has its own conditions and scope of application. Statistical methods are usually based on historical inflow records and mainly include ~~use of~~ the autoregressive model, the autoregressive moving average (ARMA) model and the autoregressive integrated moving average (ARIMA) model (Lin et al., 2006)~~, which~~. Statistical methods assume that the inflow series is stationary~~,~~ and the relationship between input and output is simple. However, real inflow series ~~are~~is complex, nonlinear and chaotic ~~disturbances~~ (Dhanya and Kumar, 2011), making it difficult to obtain high-accuracy predictions using statistical models. Physical methods which have clear mechanisms are implemented using theories of inflow generation and confluence~~, which have clear mechanisms~~. These methods can reflect the characteristics of catchment but are very strict with initial conditions and input data (Bennett et al., 2016). Meanwhile, these methods ~~are~~ used for flood forecasting have a shorter lead time and cannot be used to acquire long-term forecasting results due to input uncertainty. Machine learning methods, having a strong ability to handle the nonlinear relationship between input and output and recently shown excellent performance in inflow prediction, are widely used for medium and long-term inflow forecasts. In particular, several studies had shown that artificial neural networks (ANN) (Rasouli et al., 2012; Cheng et al., 2015; El-Shafie and Noureldin, 2011) and support vector regression (SVR) (Tongal and Booij, 2018; Luo et al., 2019; Moazenzadeh et al., 2018) are the two powerful models for inflow predicting. However, these models still have some inherent disadvantages. For example, ANN is prone to being trapped by local minima~~,~~ and both ANN and SVR suffer from over-fitting problems and reduced generalizing performance. ~~Recent~~In recent years, gradient boosting regression trees (GBRT) (Fienen et al., 2018; Friedman, 2001), a nonparametric machine learning method based on a boosting strategy and decision trees, was developed and had been used in traffic (Zhan et al., 2019) and environmental (Wei et al., 2019) field and proved to alleviate these problems mentioned above. Thus, GBRT ~~were~~is selected for daily inflow prediction with a lead ~~time~~times of 1-10 days in this paper. Compared with ANN and SVR, GBRT also has two other advantages. Firstly, GBRT can rank features according to their contribution to model scores, which is of great significance for reducing the complexity of the model. Secondly, GBRT is a white box model and can be easily interpreted. To the best of our knowledge, GBRT has not been used for daily inflow prediction with a lead ~~time~~times of 1-10 days before. For comparison purposes, ~~and~~, SVR and multiple linear regression (MLR) have ~~also~~ been employed to forecast daily inflow~~,~~ and ~~multiple linear regression (MLR) was used~~are considered as ~~a benchmark~~bench mark models in this study.

In addition to forecasting models, a vital reason why many approaches cannot attain higher accuracy for inflow predictions is that inflow is influenced by various factors (Yang et al., 2019), such as rainfall, temperature, humidity, ~~pressure, dew point,~~ etc. Thus, it is very difficult to select appropriate features for inflow forecasting. Current feature selection methods for inflow forecasting mainly include two methodologies. The first method is the model-free method (Bowden et al., 2005; Snieder et al., 2019) which employs a measure of the correlation coefficient ~~method~~criterion (He et al., 2011; Badrzadeh et al., 2013; Siqueira et al., 2018; Pal et al., 2013)~~, the stepwise selection method (Wei, 2016), and the Gamma test method (Chang and Tsai, 2016), etc. These methods have limited ability for capturing nonlinear relationships or tend to need much more computation resource. In order to select effective input factors accurately and quickly, the maximum~~ to characterise the correlation between a potential model input and the output variable. The second method is the model-based method (Snieder et al., 2019) which usually utilizes the model and search strategies to determine optimal input subset. Common search strategies include forward selection, backward elimination et al (May et al., 2011). The correlation coefficient has limited ability for capturing nonlinear relationships and exhaustive search tend to need the higher computation burden. In order to select effective inputs accurately and quickly, the maximal information coefficient (MIC) (Reshef et al., 2011)~~, was~~ is used to select input factors for inflow forecasting. MIC is a robust measure of the degree of correlation between two variables and has attracted a lot attention from academia (Zhao et al., 2013; Ge et al., 2016; Lyu et al., 2017; Sun et al., 2018). In addition,

sufficient potential input factors are the prerequisite for obtaining reliable and accurate prediction results and it is not enough to use only antecedent inflow ~~values~~series as the input of the model. To enhance the accuracy of inflow forecasting and ~~acquiring~~acquire a longer lead time, increasing amounts of meteorological forecasting data ~~are being~~have been used for inflow forecasting (Lima et al., 2017; Fan et al., 2015; Rasouli et al., 2012). However, with extended lead times, the errors of forecast data continuously increase because the variables obtained by numerical weather prediction (NWP) system are also affected by complex factors (Mehr et al., 2019). Moreover, with the continuous improvement of forecasting systems, it is difficult to obtain consistent~~, and~~ long series of forecasting data (Verkade et al., 2013). To mitigate these problems, the reanalysis data generated by ERA-Interim (European Centre for Medium-Range Weather Forecasts (ECMWF ~~(ERA-)~~ Re-Analysis Interim) (Dee et al., 2011), which was proved to be one of the best methods for reanalysis of data describing atmospheric circulation and elements (Kishore et al., 2011), has been used as an input. The reanalysis data ~~are~~which has less error than observed data and forecast data is the result of assimilating observed data with ~~forecast data, which has less error than observed data and~~ forecast data. ERA-Interim shows the results of a global climate reanalysis from 1979 to date, which are produced by a fixed version of ~~a~~ NWP system. The fixed version ensures that there are no spurious trends caused by an evolving NWP system. Therefore, meteorological reanalysis data satisfies the need for long sequences of consistent data and ~~have~~has been used for the prediction of wind speeds (Stopa and Cheung, 2014) and solar radiation (Ghimire et al., 2019; Linares-Rodríguez et al., 2011).

This study aims to provide a reliable inflow forecasting framework with ~~a~~ longer lead ~~time~~times for daily inflow forecasting. The framework adopts the ERA-Interim reanalysis dataset as the input which ensured ample information ~~was supply~~is supplied to depict inflow. MIC ~~was~~is used to select appropriate features ~~so that avoiding~~to avoid over-fitting and waste of computing resources caused by feature redundancy. GBRT, which is robust to outliers and has strong non-linear fitting ability, ~~was~~is used as the prediction model to improve inflow forecasting accuracy of longer lead times.

This paper is organized as follows: Section 2 describes a case study and collected data~~,~~. Section 3 introduces the theory and process of methods used, including MIC and GBRT. Section 4 shows the results and discussion of the data, followed by the conclusions in Section 5.

2 Data

2.1 Study area and collected data

The Xiaowan Hydropower Station in the lower reaches of the Lancang River ~~was~~is chosen as the study site (Fig. 2). The Xiaowan Hydropower Station is the main controlling hydropower station in the Lancang River and it is very meaningful to adopt the Xiaowan Hydropower Station as the case study. The Lancang River is approximately 2000 km long and has a drainage area of 113300 km^2 above the Xiaowan Hydropower Station. The Lancang River which is also known as the Mekong River originates in the Tibetan Plateau and runs through China, Myanmar, Laos, Thailand, Cambodia, and Vietnam ~~and is also known as the Mekong River.~~. The major source of water flowing into the Lancang River in China comes from melting snow on the Tibetan plateau (Commission Mekong River, 2005).

We collected ~~EAR~~ERA-Interim reanalysis dataset, observed daily inflow and rainfall data for Xiaowan for 8 years (January 2011 to December 2018). Fig. 3 depicts the daily inflow series. The data from January 2011 to December 2014 (1461 days, approximately 50% of the whole dataset), from January 2015 to December 2016 (731 days, approximately 25% of the whole dataset) and from January 2017 to December 2018 (730 days, approximately 25% of the whole dataset) ~~were~~are used ~~for~~as training, validation and testing set, respectively. The reanalysis dataset can be downloaded from *https://apps.ecmwf.int/datasets/data/ interim-full-daily/levtype=sfc/* and is provided every ~~six~~12 hours on a spatial grid size of 0.~~75~~25° × 0.~~75°. According to~~25°. Based on the ~~physical meaning~~expert knowledge and on the basis of ~~the variables~~available literature, the near-surface ~~79~~26 variables (Table A1) from the reanalysis data are considered as potential selected predictors for inflow forecasting, which include the total precipitation (tp), the 2 meter temperature (t2m), the total column water (tcw), etc~~.~~. More details about ERA-Interim dataset are presented in the Appendix A.

2.2 ~~Data~~Feature scaling and feature selection

Feature scaling is necessary for machine learning methods and all features ~~were~~are scaled to the range between 0 and 1 before taking part in the calculation, as follows:

$$x_{scale} = \frac{x_{original} - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x_{scale} and x_{original} indicate the scaled and original data, respectively~~ and.~~ x_{max} and~~ ~~x_{min} represent the maximum and minimum of inflow series, respectively.

Reasonable selection of input variables can ~~accelerate~~reduce the ~~calculation speed~~computational burden and improve the prediction accuracy of the model by removing redundant feature information and reducing the dimensions of the features. If too many features are selected, model will become very complex, which will cause trouble when adjusting parameters, resulting in over-fitting and difficult convergence. Moreover, natural patterns in the data will be blurred by noise (Zhao et al., 2013). On the other hand, if ~~too few~~irrelevant features are chosen, there will ~~be not enough information for inflow forecasting. After eliminating invalid variables,~~add noise into the model and also hinder the learning process. MIC ~~was~~is employed to select inputs from ~~79 potential~~candidate predictors from reanalysis data~~ and observed~~. The lagged inflow and rainfall ~~lags~~series are identified by partial autocorrelation function (~~ACF~~PACF) and cross-correlation function (CCF). The corresponding 95% confidence interval is used to identify significant correlations. Furthermore, when correlation coefficient slowly declines and cannot fall into confidence interval, a trial and error procedure is used to determine the optimum lag, i.e., starting from one-lag and then modifying the external inputs by successively adding one more lagged time series into inputs (Amiri 2015; Shoaib et al., 2015).

~~2~~3 Methodology

3.1 Feature selection via maximal information coefficient

The calculation of MIC is based on concepts of the mutual information (MI) (Kinney and Atwal, 2014). For a random variable X, such as observed inflow, the entropy of X is defined as

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{2}$$

where $p(x)$ is the probability density function of X = x. Furthermore, for another random variable Y, such as observed rainfall, the conditional entropy of X given Y may be evaluated from the following expression

$$H(X \mid Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x \mid y) \tag{3}$$

where $H(X \mid Y)$ is the uncertainty of X given knowledge, $p(x, y)$ and $p(x \mid y)$ are the joint probability density and the conditional probability of X = x and Y = y, respectively. The reduction of the original uncertainty of X, due to the knowledge of Y, is called the MI (Amorocho and Espildora, 1973; Chapman, 1986), defined by

$$MI(X,Y) = H(X) - H(X \mid Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x) p(y)} \tag{4}$$

The calculation of MIC is divided into three steps. Consider given a dataset D, including variable X and Y with a sample size n. Firstly, drawing scatter plots of X and Y and drawing grids for partitioning which is called an x-by-y grid. Let D|G denote the distribution of D divided by one of x-by-y grids as G. $MI^*(D, x, y) = \max MI(D \mid G)$, where $MI(D \mid G)$ is the mutual information of D|G. Secondly, characteristic matrix is defined as

$$M(D)_{x,y} = \frac{MI^*(D, X, Y)}{\log(\min(x, y))} \tag{5}$$

$$M(D)_{x,y} = \frac{MI^*(D,x,y)}{\log(\min(x,y))}$$

Lastly, MIC is introduced as the maximum value of characteristic matrix, that is, $MIC(D) = \max_{xy<B(n)} M(D)_{x,y}$, where B($n$) is the upper bound of the grid size ~~need to be considered~~which is a function of sample size, defined B = $n^{0.6}$.

We perform feature selection from ERA-Interim reanalysis dataset in two steps via MIC. First, compute MIC ~~value~~ of each reanalysis ~~factors~~variables and observed inflow. Then, sort features based on MIC in a descending order and ~~select~~determine the optimum inputs by using a trail-and-error procedure, i.e. starting from the top ~~k features according to the set threshold~~one feature and then modifying the external inputs by successively adding one more feature into model inputs. The selected k features from reanalysis data are used as part of input to the model.

3.2 Gradient boosting regression trees

~~Gradient boosting regression trees~~GBRT is an ensemble model which mainly includes two algorithms: decision tree algorithm and the boosting algorithm. The decision tree robust to outliers is used as a primitive model and boosting algorithm as integration rule is used to improve inflow forecasting accuracy.

3.3.1 The decision tree

The decision tree in this paper refers to decision tree learning used in computer science, which is one of the predictive ~~modeling~~modelling approaches used in machine learning. A decision tree consists of branch nodes (the tree structure) and leaf nodes (the tree output).

Supposing a training dataset is given in a feature space with ~~n~~N features and each feature with ~~N~~n samples, $\{(X_1, y_1), (X_2, y_2), \ldots, (X_N, y_N)\}$ ($X_i = (x_1, x_2, \ldots, x_n)$, $i = 1, 2, \ldots, N$). In the input space where the training set is located, each region is recursively divided into two subregions and the output value of each subregion is used to construct a binary decision tree. The top-down cyclic branch learning of the decision tree adopts a greedy algorithm where each branch node only cares about its own objective function. By traversing all features and all segmentation points of each feature, the best feature j and segmentation points s can be found by minimizing ~~square~~squared loss:

$$\min_{j,s}\left[\min_{c_1}\sum_{X_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{X_i\in R_2(j,s)}(y_i-c_2)^2\right] \tag{6}$$

where

$$\begin{aligned} R_1(i,s) &= \left\{X_i \mid x^{(j)} \leqslant s, i=1,2,\cdots,N\right\} \\ R_2(i,s) &= \left\{X_i \mid x^{(j)} > s, i=1,2,\cdots,N\right\} \\ R_1(j,s) &= \left\{X_i \mid x_i^{(j)} \leqslant s, i=1,2,\cdots,N\right\} \\ R_2(j,s) &= \left\{X_i \mid x_i^{(j)} > s, i=1,2,\cdots,N\right\} \end{aligned} \tag{7}$$

$$c_m = \frac{1}{N_m}\sum_{X_i\in R_m(j,s)} y_i \quad (m=1,2) \tag{8}$$

y_i is the observed value and $R_1(j, s)$ and $R_2(j, s)$ are the results of partitioning. c_1 and c_2 are output values of $R_1(j, s)$ and $R_2(j, s)$, respectively. Fig. 4 shows an example of a decision tree model with a max depth and number of leaf nodes of 3 and 5, respectively. If the threshold of loss is set as the stopping condition of the decision tree, it will easily lead to over-fitting problems. Hence, we set the following parameters to alleviate the over-fitting problem of the decision tree model: the maximum depth of the tree, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node and the number of leaf nodes. These parameters are also the ones used for optimization when using the decision tree.

3.3.2 The boosting algorithm

The idea of gradient boosting originated in the observation by Breiman (Breiman, 1997) and can be interpreted as an optimization algorithm based on a suitable cost function. Explicit regression gradient boosting algorithms ~~were~~are subsequently developed (Friedman, 2001; Mason et al., 2000). The boosting algorithm used is described here. Supposing a training dataset with ~~N~~*n* sample $\{(X_1,y_1),(X_2,y_2),\cdots,(X_N,y_N X_n,y_n)\}$, a ~~square~~squared loss function is used to train the decision tree:

~~$L(y,f(X))=\sum_{i=1}^{N}(y-f(X_i))^2$~~

$$L(y,f(X))=\sum_{i=1}^{n}(y-f(X_i))^2 \tag{9}$$

The core of the GBRT algorithm is the iterative process of training the decision with a residual method. The iterative training process of GBRT with *M* decision trees is as follows:

1) Initialization $f_0(x)=\arg\min_c\sum_{i=1}^{N}L(y_i,c)$ $f_0(x)=\arg\min_c\sum_{i=1}^{n}L(y_i,c)$.

2) For m-th (m=1, 2, ..., M) decision trees:

a) Operating i-th (i=1, 2, ..., ~~N~~*n*) sample points. Using the negative gradient of the loss function to replace the residual in the current model $r_{mi}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$.

b) Fitting a regression tree with $\{(x_i,r_{mi})\}$. The *i*-th regression tree with R_{mt} (t = 1, 2, ..., T) as its corresponding leaf node region is obtained, where *t* is the number of leaf nodes of regression.

c) For each leaf region t = 1, 2, ..., T, and the best fitting value is calculated by $c_{mt}=\arg\min_c\sum_{x_i\in R_{mt}}L(y_i,f_{m-1}(x_i)+c)$.

d) The fitting results are updated by adding the obtained fitting values to the previous ones using $f_{mt}(x_i)=f_{m-1}(x_i)+\sum_{t=1}^{T}c_{mt}I\quad(x_i\in R_{mt})$.

3) Finally, a strong learning method is obtained $\hat{f}(x_i)=f_M(x_i)=\sum_{m=1}^{M}\sum_{t=1}^{T}c_{mt}I\quad(x_i\in R_{mt})$.

According to the above introduction to GBRT, the parameters of the GBRT can be divided into two categories: boosting parameters and ~~decision~~ tree parameters. The boosting parameters include the learning rate and the number of weak learners (*learning_rate* and n_estimators). The learning rate setting is used for reducing the gradient step. The learning rate influences the overall time of training, and the smaller the value is, the more iterations are required for training. There are four tree parameters: max_leaf_nodes, min_samples_leaf, min_samples_split and max_depth. Hence, GBRT has six parameters control model complexity (Fienen et al., 2018), ~~five of~~ which we adjusted for tuning~~, except for learning rate determined~~ by using a trial ~~in advance~~-and-error procedure.

3.3 Evaluation criteria of the models

It is critical to carefully define the meaning of performance and to evaluate the performance on the basis of the forecasting and fitted values of the model compared with historical data. The root mean ~~square~~squared error (RMSE) and mean absolute error (MAE) are the most commonly used criteria to assess model performance and are calculated using Eq. (10) and Eq. (11), respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Q}_i - Q_i)^2} \quad (10)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Q}_i - Q_i\right| \quad (11)$$

where $\hat{Q}_i$ and Q_i are the inflow estimation and observed value at time *i*, respectively and *n* is the number of samples. The RMSE is more sensitive to extremes in sample sets and thus it ~~was~~is used to evaluate the model's ability to simulate flood peaks.

~~The Nash-Sutcliffe efficiency coefficient (NSE) (Nash and Sutcliffe, 1970) is commonly for evaluating the performance of hydrological models and it is one of the best performance metrics for reflecting the overall fit of a hydrograph. The NSE is calculated using Eq. (12)~~

The Pearson correlation coefficient (CORR) is a measure of the strength of the association between observed inflow series and forecasted inflow series; it is calculated according to Eq. (12).

$$\text{~~NSE~~} = 1 - \frac{\sum_{i=1}^{n}(\hat{Q}_i - Q_i)^2}{\sum_{i=1}^{n}(Q_i - \bar{Q})^2}$$

$$CORR = \frac{\sum_{i=1}^{n}(Q_i - \bar{Q})(\hat{Q}_i - \bar{\hat{Q}})}{\sqrt{\sum_{i=1}^{n}(Q_i - \bar{Q})^2}\sqrt{\sum_{i=1}^{n}(\hat{Q}_i - \bar{\hat{Q}})^2}} \quad (12)$$

where ~~$\bar{Q}$~~ $\bar{\hat{Q}}$ is the mean of the ~~observed values. An NSE = 1 means the model is perfect and an NSE < 0 reflects that the model forecasts are a worse~~ estimation ~~than the mean value of observed values.~~

~~The Pearson correlation coefficient (CORR) is a good measurement of the average error which is calculated according to Eq. (13).~~

~~$$CORR = \frac{\sum_{i=1}^{n}(Q_i - \bar{Q})(\hat{Q}_i - \bar{\hat{Q}})}{\sqrt{\sum_{i=1}^{n}(Q_i - \bar{Q})^2}\sqrt{\sum_{i=1}^{n}(\hat{Q}_i - \bar{\hat{Q}})^2}} \quad (13)$$~~

~~where $\bar{\hat{Q}}$ is the mean of the estimation values~~series. The range of the CORR is between 0 and 1 and values close to 1 demonstrate a perfect estimation result.

Kling–Gupta efficiency scores (KGE) (Knoben et al., 2019) is also a widely used evaluation index. It can be provided as following Eq. (13) and (14).

$$KGE = 1 - \sqrt{(CORR - 1)^2 + (\frac{\hat{\sigma}}{\sigma} - 1)^2 + (\frac{\bar{\hat{Q}}}{\bar{Q}} - 1)^2} \quad (13)$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Q}_i - \bar{\hat{Q}})^2}, \quad \sigma = \quad (14)$$

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i-\bar{Q})^2}$$

where σ is the standard deviation of the observed values, $\hat{\sigma}$ is the standard deviation of the inflow estimation, μ is the mean of the observed series and $\hat{\mu}$ is the mean of the inflow estimation series.

The percent bias in flow duration curve high-segment volume (BHV) (Yilmaz et al., 2008; Vogel and Fennessey, 1994) is presented to estimate prediction performance of extreme value for model. It can be provided as following Eq. (15).

$$BHV=\frac{\sum_{h=1}^{H}(\hat{Q}_h-Q_h)}{\sum_{h=1}^{H}Q_h}\times 100 \quad (15)$$

where h = 1, 2, . . .H are the inflow indices for inflows with exceedance probabilities lower than 0.02. In this paper, the inflow threshold of exceedance probabilities equalling 0.02 is 1722 m^3/s.

The Index of Agreement (IA) (Willmott, 1981) plays a significant role in evaluating the degree of the agreement between observed series and inflow estimation series. Similar to CORR, its range is between 0 (no agreement at all) and 1 (perfect fit). It is given by:

$$IA=1-\frac{\sum_{i=1}^{n}(\hat{Q}_i-Q_i)^2}{\sum_{i=1}^{n}(|\hat{Q}_i-\bar{Q}|+|Q_i-\bar{Q}|)^2} \quad (16)$$

3.4 Overview of framework

Fig.5 illustrates the overall structure of framework presented. This structure consists of two major models: GBRT and GBRT-MIC.

In GBRT, we measure the relevance of ~~-~~different lags observed inflow and rainfall with observed inflow at the time of forecast via partial autocorrelation function (~~ACF~~PACF) and cross-correlation function (CCF) (Badrzadeh et al., 2013) and select appropriate lags as predictors of model by ~~the set threshold.~~hypothesis test and trial-and-error procedures. Then, data ~~preprocessing~~pre-processing and feature scaling ~~were~~are carried out for selected predictors. Next, dividing the dataset into ~~train~~training set, validation set, and ~~test~~testing set according to the length of each data set specified in advance (in Section 2.2). A grid search algorithm ~~was~~, which is an exhaustive search all candidate parameter combination method, is guided to optimization model parameters by evaluation of validation set for each lead time (Chicco and Davide, 2017). Lastly, prediction results are evaluated based on ~~test~~testing set. Compared with GBRT, GBRT-MIC adds reanalysis data which ~~were~~are selected via MIC (in Section 3.1) as the input of the model. Moreover, GBRT-MIC also calculates the importance of features according to the prediction results and ranks the features~~. The model structures of GBRT and GBRT-MIC are as follows:~~ (Louppe, 2014).

It is difficult to perform multi-step forecasting by the reason of accumulation of errors, reduced accuracy, and increased uncertainty. The current state of multi-step ahead forecasting is reviewed, there are mainly two strategies that you can use for multi-step forecasting for single-output, namely, Static (Direct) multi-step forecast and Recursive multi-step forecast (Bontempi et al., 2013; Taieb et al., 2012). Recursive forecast strategy is biased when the underlying model is nonlinear and is sensitive to the estimation error, since estimated values, instead of actual ones, are more and more used when we get further in the future (Bontempi et al., 2012). Thus, the Static multi-step forecasting strategy is employed in this paper. Since the Static strategy does not use any approximated values to compute the forecasts, it is not prone to any accumulation of errors. The model structures of GBRT and GBRT-MIC are as follows:

$$\hat{Q}_{t+T}^{I} = f(\vec{\theta}_t^{I}; Q_t, Q_{t-1}, ..., Q_{t+1-p}, R_t, R_{t-1}, ..., R_{t+1-q}) \quad (T = 1, 2, ..., 10) \tag{14 17}$$

$$\hat{Q}_{t+T}^{II} = f(\vec{\theta}_t^{II}; Q_t, Q_{t-1}, ..., Q_{t+1-p}, R_t, R_{t-1}, ..., R_{t+1-q}, E_t^1, E_t^2, ..., E_t^k) \quad (T = 1, 2, ..., 10) \tag{15 18}$$

where $\hat{Q}_{t+T}^{I}$ and $\hat{Q}_{t+T}^{II}$ are the forecasted value of GBRT and GBRT-MIC at the lead time T of current time t, respectively. $\vec{\theta}_t^{I}$ and $\vec{\theta}_t^{II}$ are parameters of GBRT and GBRT-MIC at the lead time T of current time t, respectively. p and q are lags of observed inflow and rainfall determined via ~~ACF~~PACF and CCF, respectively. E_t is the features from reanalysis data at the current time t and k is the number of features from reanalysis data determined via MIC.

4. Experimental results and discussion

In order to compare with GBRT-MIC, the ANN-MIC, SVR-MIC and MLR-MIC, obtained by replacing GBRT in the framework with ANN, SVR and MLR, respectively, ~~were~~are also employed for inflow forecasting with lead times of 1-10 days. As mentioned ~~earlier,~~previously, six indices, i.e. the MAE, RMSE, ~~MAE, NSE~~CORR, KGE, BHV and ~~CORR of each model were~~IA, are calculated to ~~compare with~~evaluate the performance of models based on the ~~test~~testing set. We also explored the feature importance based on the GBRT-MIC model~~.~~ (Louppe, 2014). All computations of this paper are performed on a ThinkPad P1 workstation containing an Intel Core i7-9850H CPU with 2.60 GHz and 16.0 GB of RAM, using the version 3.7.10 of Python (Python Software Foundation, 2020), which is powerful, fast and open, and scikit-learn package (Pedregosa et al., 2011).

4.1 Feature selection

Fig. ~~Table 1 shows the ACF of a lag of 1-10 days of inflow and rainfall. The results show the ACF decreases as the lag order increases. Considering the problems of over-fitting and computational time caused by large amounts of input data, thresholds of inflow and rainfall are set as 0.931, 0.226, respectively and threshold of MIC is set as 0.625. And thus a lag of one day of 10 predictors from the reanalysis data and a lag of 1-3 days of inflow and rainfall were selected as the model inputs, that is k=10, p=3 and q=3. Simultaneously, according to the causes of inflow, the total precipitation was also selected as a predictor. Finally, a total of 17~~ 6 shows the PACF, CCF and the corresponding 95% confidence interval from lag 1 to lag 12. The PACF shows significant autocorrelation at lag one and lag four, respectively (Fig. 6(a)), and thus, inflow series one and four-day lag are selected as the inputs of model. CCF between inflow and rainfall gradually decreases as increasing the time lag (Fig. 6(b)) and cannot fall into 95% confidence interval. Therefore, a trial-and-error procedure is used to determine optimal selection of lagged rainfall series. 13 input structures are tried (Table 1) and the trial results are shown in Fig. A1. The results indicate that 7th input structure obtains best performance. Accordingly, rainfall series from one to six-day lag are selected as the inputs of model. As mentioned previously, based to MIC between inflow and the reanalysis variable (Table A1), a trial-and-error procedure is used to determine optimal input subset. 26 input structures are tried (Table 2) and the trial results are shown in Fig. A2. The results show that 8th input structure obtains best performance and thus the No.1 to 8 predictors in Table A1 are selected as the model input.

Finally, a total of 16 variables including ~~6~~8 observed variables and ~~11~~8 reanalysis variables ~~were~~are selected as the model inputs (Table ~~2).~~3). As shown in Table 3, No. ~~7~~9 to ~~17~~18 are reanalysis variables and the range of MIC of the reanalysis variables selected is 0.~~625~~643 to 0.~~853 except for the total precipitation. No. 7 to 10~~847. Furthermore, No. ~~15~~9 and No. 13 to 16 are variables related to ~~solar radiation or~~ temperature ~~that are used to represent the melting of snow. Albedo (No. 7) is a measure of the reflectivity of the Earth's surface. Typically, the highest Albedo is found on snow and ice, followed by land and the lowest in the ocean. Albedo has the highest correlation (negative correlation) with inflow in terms of MIC.~~ Soil temperature level 3 (No. ~~8~~9) is the temperature of the soil in layer 3 (28-100 cm, the surface is at 0 cm). The temperature of the

snow layer (No. ~~15~~13) gives the temperature of the snow layer from the ground to the snow-air interface. ~~Surface thermal radiation downwards (~~No. ~~16) is~~ 10 to 12 are variables related to the ~~amount of thermal (also known as longwave or terrestrial) radiation emitted by~~ water content of the atmosphere. 2 meter dewpoint temperature (No. 10) is a measure of the humidity of the air. Combined with temperature and pressure, it can be used to calculate the relative humidity. The total column water vapor (No. ~~clouds that reaches the Earth's surface.~~11) is only the total amount of water vapor, which is a fraction of the total column water. Total column water (No. ~~11~~12) is the sum of water vapor, liquid water, cloud ice, rain and snow in a column extending from the surface of the Earth to the top of the atmosphere. Volumetric soil water layer 1 (No. 19) is the volume of water in soil layer 1.~~The total column water vapor (No. 12) is only the total amount of water vapor, which is a fraction of the total column water. Runoff (No. 13) is a measure of the availability of water in the soil, including at the surface (surface runoff) and under the ground (subsurface runoff). Volumetric soil water layer 3 (No. 14) is the volume of water in soil layer 3. Total precipitation (No. 17) is the accumulated liquid and frozen water, including rain and snow, that falls to the Earth's surface.~~ In summary, all the selected predictors are interpretable and have a good physical connection with inflow.

4.2 Hyperparameter optimization

~~For machine learning methods, hyperparameters are parameters that are set before starting the learning process, rather than parameters obtained through training. In order to improve the learning performance, it is imperative to tune the hyperparameters of models. Grid searching was employed to tune the hyperparameters of GBRT, GBRT-MIC, ANN-MIC and Bayesian optimization (Snoek et al., 2012) was employed to tune the hyperparameters of SVR-MIC. All computations of this paper are performed in the version 3.7.10 of Python, using the scikit-learn package (Pedregosa et al., 2011).~~

~~Besides the training algorithms~~For machine learning methods, hyperparameters are a king of parameters that are set before training and cannot be directly learned from the regular training process. In order to improve the performance of models, it is imperative to tune the hyperparameters of models. Grid search is employed to tune the hyperparameters of GBRT, GBRT-MIC, ANN-MIC and SVR-MIC.

Reviewing to the basis of available literature (Badrzadeh et al., 2013; Rasouli et al., 2012), an optimizer in the family of quasi-Newton methods, namely *L-BFGS* is used as the training algorithm of ANN and the number of hidden layers ~~of the ANN, there are also~~is fixed to 3. Another two parameters ~~controlling~~, namely activation function and the ~~model structure which~~number of nodes of the hidden layer need to be adjusted. A range of 2-20 neurons and four commonly used activation functions (Table ~~3~~4) are selected by grid ~~searching~~search. To alleviate the influence of random initialization of weights, 50 ANN-MIC models ~~were~~are trained for each parameter combination ~~of~~. Optimal activation function and the number of nodes of the hidden ~~neurons. We obtained activation functions and the number of hidden neurons~~layer are determined by selecting the ~~optimal median of the~~minimal MAE of the validation set for each lead time. The results of the trials show ~~identity is most~~*tanh* and *logistic* function are two more robust activation function (Fig. ~~6~~7) and ~~less number of~~ANN with fewer nodes is inclined to obtain lower error ~~for except for relu.~~. The optimal ~~combinations of number of hidden layer and activation function~~ parameter combination for each lead time ~~are~~is listed in Table ~~4~~5. It ~~is found~~can be seen that the optimal number of ~~neurons~~nodes is 2, 3 or ~~3~~4 and the optimal activation function is either ~~logistic~~*tanh* or ~~tanh~~*logistic* function.

For SVR, ~~there~~according to Lin et al. (2006) and Dibike et al. (2001), radial basis function (RBF) outperforms other kernel functions for runoff modelling and thus RBF is used as the kernel function in this study. There are three parameters ~~that~~ need to be ~~optimized, the regularization parameter C, the loss function tolerance threshold ϵ and the parameter of RBF γ. The search space of ϵ, C and γ are [0, 2], [e^{-5}, e^{5}], [e^{-13}, e^{-1}], respectively. The~~adjusted. Firstly, an appropriate tuning ~~parameters for each lead time are listed in Table 4 and the~~ range of ~~C, ϵ and γ are 0.51 to 144.48, 0.0004 to 0.0085 and 0.0012 to 0.2150, respectively.~~

~~According to~~ parameter is determined by a trial ~~and~~ error~~, the learning rate of GBRT and GBRT-MIC are all decided to be 0.01. For~~ procedure. And then, to reach at an optimal choice of these parameters, the MAE is used to optimize the parameters

by grid search. Optimal tuning parameters *max_depth*, *max_leaf_nodes*, *min_samples_leaf* values between 3 and 9 with an increment of 2 were evaluated, for *min_samples_split* values between 10 and 50 with an increment of 10 were evaluated, and *n_estimators* between 200 and 1200 with an increment of 50 were evaluated. We developed 13440 models for each lead time and Table 5 lists the parameters selected for GBRT and GBRT-MICof SVR are shown in Table 5.

As mentioned earlier, for GBRT, there are six parameters need to be adjusted. In order to obtain an optimal parameter combination as soon as possible, we optimize all parameters in two steps. Firstly, *n_estimators* and *learning_rate* are fixed to 100 and 0.1, respectively. The *max_leaf_nodes*, *min_samples_leaf*, *max_depth* and *min_samples_split*, four tuning parameters generate 40000 models at each lead time. Secondly, after the tree parameters are determined, *learning_rate* is modified to 0.01 and *n_estimators* is determined by grid search. To accommodate the computational burden, all models are distributed among about 12 central processing units (CPUs) and total wall time for the runs is about 7 hours for GBRT_MIC and GBRT. Table 6 lists optimal tuning parameters of GBRT and GBRT-MIC.

4.3 Inputs comparison

Fig. 78 illustrates the relationship between performance indices and lead times of GBRT and GBRT-MIC inon the testtesting set (2017/01/01-2018)./12/31) at lead times of 1-10 days. It is obvious that the use of reanalysis data selected by MIC as model inputs makes a great improvement on the GBRT forecasting. For at both short and long lead times. In particular, for the longer lead times prediction of GBRT-MIC is significantly outperform GBRT. For Fig. 8Fig. 7(a), the MAE of GBRT-MIC decreases from 171175 to 167172, a decrease of 2.41.74% for two-day- ahead forecasting and decreases from 270273 to 241237, a decrease increasing to 10.713.18% for ten-day- ahead forecasting compared with GBRT. For Fig. 78(b), the RMSE of GBRT-MIC achieves 41.4% and 910.6% reduction for two and ten-day- ahead forecasting, respectively, compared with GBRT. For Figs. 7Fig. 8(c), 8(d) and 7(d8(f), the NSECORR, KGE and CORRIA of GBRT-MIC increase by 0.2%, 2.2%, 1.1%, 1.10% for two-day- ahead forecasting and 5.2%, 2.3.4%, 7.8% and 2.2% for ten-day- ahead forecasting, respectively. Thus,Fig. 8(e) compares the above analysis BHV of GBRT and GBRT-MIC which indicates the improvements increase as the lead time increases and reanalysis data significantly improves the accuracy of inflow prediction for longer lead times in terms of all four different evaluation measures in the test setcan enhance forecasting of extreme values.

Fig. 89(a) shows the tenfive-day- ahead forecasted inflow of GBRT-MIC and GBRT versus the observed inflow in the testtesting set. The slopeslopes of fitting curve of GBRT-MIC and GBRT are 0.7889 and 0.7281, respectively, which also demonstrates that GBRT-MIC can obtain more accuracyaccurate inflow forecasting than GBRT. Fig. 89(b) illustrates the distribution of the forecast errors of GBRT and GBRT-MIC. The results show the prediction error of two models approximate to normal distribution. It demonstrates that the prediction error contains less information that is not extracted by the model and more errors of forecasted inflow concentrate at 0 around by GBRT-MIC than GBRT. Fig. 89(c) provides the ten-day-ahead forecasted resultsinflow time series (from the testing set) of GBRT-MIC and GBRT in the test set.at lead time of five-day. It can be seen that forecast error of peak value of GBRT-MIC significantly decreases thanprovides great performance compare to GBRT., especially for the extreme values. This reveals that the problem of inaccurate peak flow extreme value prediction arisen in in areas with concentrated rainfall for the GBRT model could be mitigated by incorporating the reanalysis data identified by MIC as model inputs.

4.4 Model comparison

GBRT-MIC, SVR-MIC, ANN-MIC with obtained optimal model parameters wereare employed for inflow forecasting of one to ten-day- ahead. Summarized results for train, validationtraining and testtesting set are presented in Table 6, Table 7 and Table 8, respectively. Notably, the To avoid local minima problems, 50 ANN-MIC models wereare trained 50 times for each lead time, and the median of the performance indices are listed in these tables.predictions of the 50 models gives the final prediction. It is clear from Table 67 that the GBRT-MIC are more efficient in the trainttraining set than other models at lead times of 1-10 days, andwhich demonstrates that GBRT-MIC has a powerful fitting ability. Meanwhile, all machine learning models obtain better simulationforecasted results than MLR-MIC. However, GBRT-MIC did not always perform best for the

~~test set at lead times of 1-10 days. From Table 8, we notice that SVR-MIC outperforms the other models for a lead time of one day. Furthermore, ANN-MIC performs best in terms of RMSE and NSE at lead times of 2-4 days, and~~ which cannot capture nonlinear relationship. It should be noted that ANN-MIC has best performance for extreme values in terms of BHV in the training set.

As shown in Table 8, GBRT-MIC performs best for ~~longer~~the testing set at lead times ~~(5~~of 4-10 days~~).~~ in terms of six indices. At a lead time of ten days, the ~~NSE~~KGE of GBRT-MIC even reached 0.~~8084.~~8317. At the lead times of 1-3 days, three machine learning models obtain approximate performance but outperform MLP-MIC. The machine learning models can acquire enough information to perform forecasting at the short lead time (1-3 days).

The ~~relationship between~~ performance indices ~~and lead times~~ of these four models in the ~~test~~testing set (2017-2018) ~~is~~ at the lead times of 1-10 days are presented in Fig. ~~9~~10. The results indicate the performance of these four models decreases (higher MAE, RMSE and ~~RMSE~~BHV, and lower CORR, KGE and ~~NSE~~IA) as the lead time increases ~~and~~. As mentioned earlier, the four models perform equally well for one- to ~~four~~three-day- ahead forecasting, whereas significant differences among their performances are found as the ~~forecasting~~ lead time exceeds ~~four~~three days. It clearly indicates that the GBRT produce much higher CORR, KGE and ~~NSE~~IA, and lower MAE, RMSE and ~~RMSE~~BHV than the other three models for ~~five~~four to ten-day- ahead forecasting except that the ~~SVR~~ANN-MIC perform ~~best in terms of MAE~~ nearly to GBRT-MIC for ten-day- ahead forecasting. ~~As mentioned earlier (Section 4.3), the prediction error of these four models approximate to normal distribution and~~It should be noted that SVR performs worst according to ~~(Chai and R., 2014), the RMSE is more appropriate for representing model performance than MAE when the error distribution is expected to be Gaussian. Thus, GBRT-MIC preforms best for five~~BHV and KGE, which demonstrates that SVR cannot capture extreme values. On the contrary, GBRT-MIC significantly outperform other models in terms of BHV at lead times of 1-10 days, which indicates that GBRT-MIC is able to ~~ten-day-ahead forecasting~~obtain extreme values among all models developed in this paper.

4.5 Feature importance

A benefit of using gradient boosting is that after the boosted trees are constructed, relative importance scores for each feature can be acquired to estimate the contribution of each feature to inflow forecasting. Fig. ~~10~~11 shows the feature importance based on GBRT-MIC for lead times of one and ten days. The ~~lagged~~one-day lag observed ~~values (e.g.,~~time series (~~Q_{t-1} and Q_{t-2}) are~~ Q_{t-1}) is more important for shorter lead times (Fig. ~~10~~11(a)), which demonstrates that the historical observed values are essential ~~for~~to inflow forecasting ~~of~~at shorter lead times.

The features (e.g., $stl3_{t-10}$ and $d2m_{t-10}$) from the reanalysis data have a high relative importance ~~for~~at longer lead times (Fig. ~~10~~11(b)). Based on the analysis of the concepts of $stl3_{t-10}$ and ~~$d2m_{t-10}$~~ tcw_{t-10} (Section 4.1), we infer that the temperature near the ground effects the inflow by affecting the melting of snow which is consistent with the fact that the Lancang River is a snow-melt river. The ten-day lag observed time series (Q_{t-10}) is also very important which indicate the long memory of inflow series (Salas 1993). Meanwhile, it is found that the reanalysis data provides important information for inflow forecasting at longer lead times.

5. Conclusion

In this study, GBRT-MIC ~~are~~is employed to make inflow forecasts for lead times of 1-10 days and ANN-MIC, SVR-MIC and MLR-MIC are developed to compare with GBRT-MIC. The reanalysis data selected by MIC, the antecedent inflow and the rainfall records selected by ~~ACF~~PACF and CCF are used as predictors to drive the models. These models are compared using ~~four~~six evaluation criteria, the MAE, RMSE, ~~MAE, NSE and~~ CORR, KGE, BHV and IA. It is shown that GBRT-MIC, ANN-MIC and SVR-MIC outperform MLR-MIC at lead times of 1-10 days, and GBRT-MIC performs best at lead times of ~~5~~4-10 days, especially for forecasting of extreme values.

According to ~~compare~~comparison the forecasted results of GBRT and GBRT-MIC, we ~~concluded~~conclude that GBRT-MIC ~~could~~can be used for more accurate and reliable inflow forecasting ~~for 1-10 day~~at lead times~~,~~ of 1-10 days and ~~the use of~~ reanalysis data selected by MIC ~~as model inputs~~ makes a great improvement on the GBRT forecasting, especially for ~~5~~lead times of 4-10 ~~day lead times.~~days. In addition, the feature importance achieved by GBRT-MIC demonstrates that soil temperature, the total amount of water vapour in a column and dewpoint temperature near the ground ~~significantly affects~~contribute to increase the prediction accuracy of inflow ~~for~~at longer lead times.

In summary, the developed framework that integrates GBRT and reanalysis data~~,~~ selected MIC and ~~GBRT~~ can well perform inflow forecasting ~~for~~at lead times of 1-10 days. The results of this study are of significance to assist power stations in making power generation plans 7-10 days in advance in order to reduce LEQDW and flood disasters.

Another direction of improving the results could be considering heuristic methods (for example, Grey Wolf algorithm) to optimize model parameters, which could search for more wide range of hyper parameters and get optimization parameters more quickly.

Acknowledgements
This research ~~was~~is supported by National Natural Science Foundation of China (No. 51979023, No. U1765103) and the Liaoning province Natural Science Foundation of China (No. 20180550354). We are grateful for reanalysis data provided by European Centre for Medium-Range Weather Forecasts.

Code/Data availability
Request for materials should be addressed to Zhanwei Liu.

Appendix A

The ERA-Interim is a reanalysis product of the global atmospheric forecasts at ECMWF which is produced through data assimilation system, called as the Integrated Forecast System (IFS). The system includes a 4-dimensional variational analysis (4D-Var) with a 12-hour analysis window. The spatial resolution of the data set is approximately 80 km (0.72°) on 60 levels in the vertical from the surface up to 0.1 hPa. (Berrisford et al., 2011). This reanalysis meteorological products of from 0.125° to 2.5° are generated by interpolation. This reanalysis meteorological products from the ERA-Interim such as rainfall, maximum and minimum temperatures, and wind speed at 0.25° (latitude) × 0.25° (longitude) spatial and 12-hour temporal resolutions for the study period 2011-2018 are downloaded from the ECMWF webpage.

Table A1. Description and notations of the ECMWF Reanalysis Fields.

Number	Variable	MIC	Description	Units
1	stl3	0.847	Soil temperature level 3	K
2	d2m	0.781	2 metre dewpoint temperature	K
3	tcwv	0.699	Total column water vapour	kg m^{-2}
4	tcw	0.699	Total column water	kg m^{-2}
5	stl2	0.689	Soil temperature level 2	K
6	mn2t	0.684	Minimum temperature at 2 metres since previous post-processing	K
7	tsn	0.664	Temperature of snow layer	K
8	stl4	0.643	Soil temperature level 4	K
9	stl1	0.631	Soil temperature level 1	K
10	ro	0.619	Runoff	m
11	swvl1	0.614	Volumetric soil water layer 1	m^3 m^{-3}
12	swvl2	0.610	Volumetric soil water layer 2	m^3 m^{-3}
13	swvl3	0.610	Volumetric soil water layer 3	m^3 m^{-3}
14	t2m	0.571	2 metre temperature	K
15	swvl4	0.550	Volumetric soil water layer 4	m^3 m^{-3}
16	mx2t	0.539	Maximum temperature at 2 metres since previous post-processing	K
17	sf	0.470	Snowfall	m of water equivalent
18	cp	0.426	Convective precipitation	m
19	tp	0.416	Total precipitation	m
20	rsn	0.408	Snow density	kg m^{-3}
21	lsp	0.358	Large-scale precipitation	m
22	sd	0.337	Snow depth	m of water equivalent
23	smlt	0.252	Snowmelt	m of water equivalent
24	istl1	0.112	Ice temperature layer 1	K
25	istl3	0.109	Ice temperature layer 3	K
26	istl2	0.100	Ice temperature layer 2	K

13 input structures from observed data are tried and 50 trials are performed for each input structure. The results (Fig. A1) show 7th input structure is the optimal input subset for GBRT.

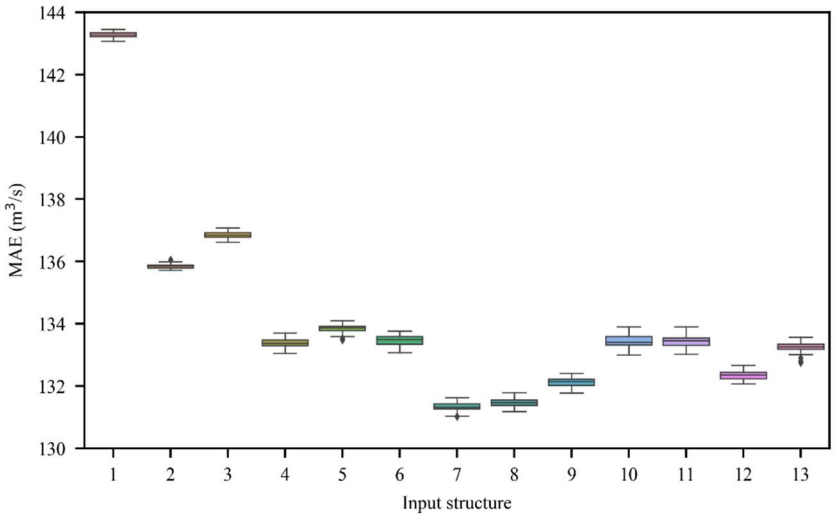


Figure A1. Trial results of 13 input structures from observed data.

26 input structures from reanalysis data are tried and 50 trials are performed for each input structure. The results (Fig. A2) show 8th input structure is the optimal input subset for GBRT-MIC.

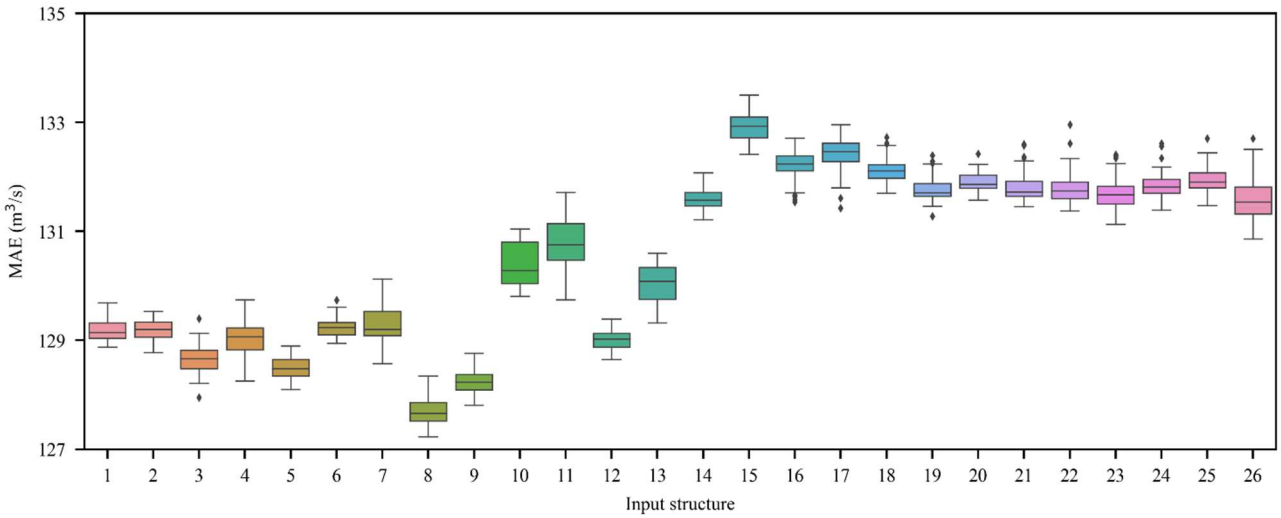


Figure A2. Trial results of 13 input structures from reanalysis data.

Author contributions

S.L. carried out the study design, the analysis and interpretation of data, and drafted the manuscript. Z.L. and B.L. participated in the study design, data collection, analysis of data, and preparation of the manuscript. C.C. and Z.Z. carried out the experimental work and the data collection and interpretation. X.J. participated in the design and coordination of experimental work, and acquisition of data. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no conflict of interest.

References

Ali Ghorbani M, Kazempour R, Chau K W, et al.: Forecasting pan evaporation with an integrated artificial neural network quantum-behaved particle swarm optimization model: a case study in Talesh, Northern Iran. Eng. Appl. Comp. Fluid Mech., 12(1), 724-737, https://doi.org/10.1080/19942060.2018.1517052, 2018.

Amiri, E.: Forecasting daily river flows using nonlinear time series models. J. Hydro., 527, 1054–1072, https://doi.org/10.1016/j.jhydrol.2015.05.048, 2015.

Amorocho, J., and Espildora, B.: Entropy in the assessment of uncertainty in hydrologic systems and models, Water Resour. Res., 9, 1511-1522, https://doi.org/10.1029/WR009i006p01511, 1973.

Badrzadeh, H., Sarukkalige, R., and Jayawardena, A.: Impact of multi-resolution analysis of artificial intelligence models inputs on multi-step ahead river flow forecasting, J. Hydro., 507, 75-85, https://doi.org/10.1016/j.jhydrol.2013.10.017, 2013.

Bennett, J. C., Wang, Q. J., Li, M., Robertson, D. E., and Schepen, A.: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model, Water Resour. Res., 52, 8238–8259, https://doi.org/10.1002/2016WR019193, 2016.

Berrisford P, Kallberg P, Kobayashi S, et al.: Atmospheric conservation properties in ERA-Interim. Quarterly Journal of the Royal Meteorological Society, 137(659), 1381-1399, https://doi.org/10.1002/qj.864, 2011.

Bontempi G, Taieb S B, Le Borgne Y A.: Machine learning strategies for time series forecasting, European business intelligence summer school, edited by: Marcel P. and Zimányi E., Springer, Berlin, Heidelberg, 62-77, https://doi.org/10.1007/978-3-319-61164-8, 2012.

Bowden, G.J., Dandy, G.C., Maier, H.R.: Input determination for neural network models in water resources applications. Part 1 - background and methodology. J. Hydro. 301, 75–92, https://doi.org/10.1016/j.jhydrol.2004.06.021, 2005.

Breiman, L.: Arcing the edge, Technical Report, Statistics Department, University of California,~~Technical Report~~ 486, 1997.

~~Chai, T., and R., D. R.: Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, Geosci. Model Dev., 7, 1247-1250, 10.5194/gmd-7-1247-2014, 2014.~~

Chang, F. J., and Tsai, M.-J.: A nonlinear spatio-temporal lumping of radar rainfall for modelling multi-step-ahead inflow forecasts by data-driven techniques, J. Hydro., 535, 256-269, https://doi.org/10.1016/j.jhydrol.2016.01.056, 2016.

Chapman, T. G.: Entropy as a Measure of Hydrologic Data Uncertainty and Model Performance, J. Hydro., 85, 111-126, https://doi.org/10.1016/0022-1694(86)90079-X, 1986.

Chau, K., Wu, C., and Li, Y.: Comparison of several flood forecasting models in Yangtze River, J. Hydro. Eng., 10, 485-491, https://doi.org/10.1061/(ASCE)1084-0699(2005)10:6(485), 2005.

Chau K.: Use of meta-heuristic techniques in rainfall-runoff modelling, Water, 9(3), 186, https://doi.org/10.3390/w9030186, 2017.

Cheng, C. T., Zhong-Kai, F., Wen-Jing, N., and Sheng-Li, L.: Heuristic Methods for Reservoir Monthly Inflow Forecasting: A Case Study of Xinfengjiang Reservoir in Pearl River, China, Water-Sui, 7, 4477-4495, https://doi.org/10.3390/w7084477, 2015.

Chicco, and Davide: Ten quick tips for machine learning in computational biology, Biodata Min., 10, 35, https://doi.org/10.1186/s13040-017-0155-3, 2017.

Commission Mekong River: Overview of the Hydrology of the Mekong Basin, Mekong River Commission, Vientiane, 82, 73, 2005.

Dee, D. P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., and Bauer, d. P.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q. J. Roy. Meteor. Soc., 137, 553-597, https://doi.org/10.1002/qj.828, 2011.

Dhanya, C. T., and Kumar, D. N.: Predictive uncertainty of chaotic daily streamflow using ensemble wavelet networks approach, Water Resour. Res., 47, W06507, https://doi.org/10.1029/2010WR010173, 2011.

Dibike, Y.B., Velickov, S., Solomatine, D., Abbott, M.B.: Model induction with support vector machines: introduction and applications, J. Comput. Civil. Eng., 15 (3), 208–216, https://doi.org/10.1061/(ASCE)0887-3801(2001)15:3(208), 2001.

Duan, Q., Sorooshian, S., and Gupta, V.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resour. Res., 28, 1015-1031, https://doi.org/10.1029/91WR02985, 1992.

El-Shafie, A., Taha, M. R., and Noureldin, A.: A neuro-fuzzy model for inflow forecasting of the Nile river at Aswan high dam, Water Resour. Manag., 21, 533-556, https://doi.org/10.1007/s11269-006-9027-1, 2007.

El-Shafie, A., and Noureldin, A.: Generalized versus non-generalized neural network model for multi-lead inflow forecasting at Aswan High Dam, Hydrol. Earth Syst. Sc., 15, 841–858, https://doi.org/10.5194/hess-15-841-2011, 2011.

Fan, F. M., Schwanenberg, D., Collischonn, W., and Weerts, A.: Verification of inflow into hydropower reservoirs using ensemble forecasts of the TIGGE database for large scale basins in Brazil, J. Hydro., 4, 196-227, https://doi.org/10.1016/j.ejrh.2015.05.012, 2015.

Fienen, M. N., Nolan, B. T., Kauffman, L. J., and Feinstein, D. T.: Metamodeling for groundwater age forecasting in the Lake Michigan Basin, Water Resour. Res., 54, 4750-4766, https://doi.org/10.1029/2017WR022387, 2018.

Fotovatikhah F, Herrera M, Shamshirband S, et al.: Survey of computational intelligence as basis to big flood management: Challenges, research directions and future work, Eng. Appl. Comp. Fluid Mech., 12(1), 411-437, https://doi.org/10.1080/19942060.2018.1448896, 2018.

Friedman, J. H.: Greedy function approximation: a gradient boosting machine, Ann. Stat., 29, 1189-1232, https://doi.org/10.2307/2699986, 2001.

Ge, R., Zhou, M., Luo, Y., Meng, Q., Mai, G., Ma, D., Wang, G., and Zhou, F.: McTwo: a two-step feature selection algorithm based on maximal information coefficient, Bmc Bioinformatics, 17, 142, https://doi.org/10.1186/s12859-016-0990-0, 2016.

Ghimire, S., Deo, R. C., Downs, N. J., and Raj, N.: Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cites of queensland Australia, J. Clean. Prod., 216, 288-310, https://doi.org/10.1016/j.jclepro.2019.01.158, 2019.

in-en.com : https://www.in-en.com/article/html/energy-2266458.shtml , last access: 9 February 2020, 2018.

Kinney, J. B., and Atwal, G. S.: Equitability, mutual information, and the maximal information coefficient, P. Natl. Acad. Sci. USA, 111, 3354-3359, https://doi.org/10.1073/pnas.1309933111, 2014.

Kishore, P., Ratnam, M. V., Namboothiri, S. P., Velicogna, I., Basha, G., Jiang, J. H., Igarashi, K., Rao, S. V. B., and Sivakumar, V.: Global (50°S–50°N) distribution of water vapor observed by COSMIC GPS RO: Comparison with GPS radiosonde, NCEP, ERA-Interim, and JRA-25 reanalysis data sets, J. Atmos. Sol.-Terr. Phy., 73, 1849-1860, https://doi.org/10.1016/j.jastp.2011.04.017, 2011.

Knoben W J M, Freer J E, Woods R A.: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrol. Earth Syst. Sc., 23(10), 4323-4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

Lima, A. R., Hsieh, W. W., and Cannon, A. J.: Variable complexity online sequential extreme learning machine, with applications to streamflow prediction, J. Hydro., 555, 983-994, https://doi.org/10.1016/j.jhydrol.2017.10.037, 2017.

Lin, J. Y., Cheng, C.-T., and Chau, K.-W.: Using support vector machines for long-term discharge prediction, Hydrology SCI J, 51, 599-612, https://doi.org/10.1623/hysj.51.4.599, 2006.

Linares-Rodríguez, A., Ruiz-Arias, J. A., Pozo-Vázquez, D., and Tovar-Pescador, J.: Generation of synthetic daily global solar radiation data based on ERA-Interim reanalysis and artificial neural networks, Energy, 36, 5356-5365, https://doi.org/10.1016/j.energy.2011.06.044, 2011.

Liu, Z., Zhou, P., and Zhang, Y.: A Probabilistic Wavelet–Support Vector Regression Model for Streamflow Forecasting with Rainfall and Climate Information Input*, J. Hydrometeorol, 16, 2209-2229, https://doi.org/10.1175/JHM-D-14-0210.1, 2015.

Louppe G.: Understanding random forests: From theory to practice. arXiv preprint arXiv:1407.7502, 2014.

Luo, X., Yuan, X., Zhu, S., Xu, Z., Meng, L., and Peng, J.: A hybrid support vector regression framework for streamflow forecast, J. Hydro., 568, 184-193, https://doi.org/10.1016/j.jhydrol.2018.10.064, 2019.

Lyu, H., Wan, M., Han, J., Liu, R., and Cheng, W.: A filter feature selection method based on the Maximal Information Coefficient and Gram-Schmidt Orthogonalization for biomedical data mining, Comput. Biol. Med., 89, 264-274, https://doi.org/10.1016/j.compbiomed.2017.08.021, 2017.

Mason, L., Baxter, J., Bartlett, P. L., and Frean, M. R.: Boosting algorithms as gradient descent, Adv. Neur. in., 2000, 512-518, https://doi.org/10.1.1.126.8716, 2000.

May, R., Dandy, G., Maier, H.: Review of Input Variable Selection Methods for Artificial Neural Networks, InTech, Artificial Neural Networks - Methodological Advances and Biomedical Applications, 10, 16004, https://doi.org/10.5772/16004, 2011.

Mehr, A. D., Jabarnejad, M., and Nourani, V.: Pareto-optimal MPSA-MGGP: A new gene-annealing model for monthly rainfall forecasting, J. Hydro., 571, 406-415, https://doi.org/10.1016/j.jhydrol.2019.02.003, 2019.

Nash, J. E., and Suteliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, J. Hydro., 10, 282-290, 10.1016/0022-1694(70)90255-6, 1970.

Mosavi A, Ozturk P, Chau K: Flood prediction using machine learning models: Literature review, Water, 10(11), 1536, https://doi.org/10.3390/w10111536, 2018.

Moazenzadeh R, Mohammadi B, Shamshirband S, et al.: Coupling a firefly algorithm with support vector regression to predict evaporation in northern Iran, Eng. Appl. Comp. Fluid Mech., 12(1), 584-597, https://doi.org/10.1080/19942060.2018.1482476, 2018.

Pal, I., Lall, U., Robertson, A. W., Cane, M. A., and Bansal, R.: Predictability of Western Himalayan river flow: melt seasonal inflow into Bhakra Reservoir in northern India, Hydrol. Earth Syst. Sc., 17, 2131-2146, https://doi.org/10.5194/hess-17-2131-2013, 2013.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V.: Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12, 2825-2830, 2011.

Python Software Foundation. Python Language Reference, version 3.7. Available at http://www.python.org, 2020.

Rajaee, T., Ebrahimi, H., and Nourani, V.: A review of the artificial intelligence methods in groundwater level modeling, J. Hydro., 572, 336-351, https://doi.org/10.1016/j.jhydrol.2018.12.037, 2019.

Rasouli, K., Hsieh, W. W., and Cannon, A. J.: Daily streamflow forecasting by machine learning methods with weather and climate inputs, J. Hydro., 414, 284-293, https://doi.org/10.1016/j.jhydrol.2011.10.039, 2012.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C.: Detecting novel associations in large data sets, science, 334, 1518-1524, https://doi.org/10.1126/science.1205438, 2011.

Robertson, D., Pokhrel, P., and Wang, Q.: Improving statistical forecasts of seasonal streamflows using hydrological model output, Hydrol. Earth Syst. Sc., 17, 579-593, https://doi.org/10.5194/hess-17-579-2013, 2013.

Salas J D.: Analysis and modelling of hydrological time series. Handbook of hydrology, 19, 1993.

Shoaib, M., Shamseldin, A.Y., Melville, B.W., Khan, M.M.: Runoff forecasting using hybrid Wavelet Gene Expression Programming (WGEP) approach. J. Hydro. 527, 326–344, https://doi.org/10.1016/j.jhydrol.2015.04.072, 2015.

Siqueira, H., Boccato, L., Luna, I., Attux, R., and Lyra, C.: Performance analysis of unorganized machines in streamflow forecasting of Brazilian plants, Appl. Soft. Comput., 68, 494-506, https://doi.org/10.1016/j.asoc.2018.04.007, 2018.

Snoek, J., Larochelle, H., and Adams, R. P.: Practical bayesian optimization of machine learning algorithms, Adv. Neur. in., 2012, 2951-2959, 2012.

Snieder E, Shakir R, Khan U T.: A comprehensive comparison of four input variable selection methods for artificial neural network flow forecasting models. J. Hydro., 124299, https://doi.org/10.1016/j.jhydrol.2019.124299, 2019.

Sohu: https://www.sohu.com/a/209379703_357198, last access: 9 February 2020, 2017.

Stopa, J. E., and Cheung, K. F.: Intercomparison of wind and wave data from the ECMWF Reanalysis Interim and the NCEP Climate Forecast System Reanalysis, Ocean Model., 75, 65-83, https://doi.org/10.1016/j.ocemod.2013.12.006, 2014.

Sun, G., Li, J., Dai, J., Song, Z., and Lang, F.: Feature selection for IoT based on maximal information coefficient, Future Gener. Comp. Sy., 89, 606-616, https://doi.org/10.1016/j.future.2018.05.060, 2018.

Taieb S B, Bontempi G, Atiya A F, et al.: A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. Expert systems with applications, 39(8), 7067-7083, https://doi.org/10.1016/j.eswa.2012.01.039, 2012.

Tongal, H., and Booij, M. J.: Simulation and forecasting of streamflows using machine learning models coupled with base flow separation, J. Hydro., 564, 266-282, https://doi.org/10.1016/j.jhydrol.2018.07.004, 2018.

Valipour, M., Banihabib, M. E., and Behbahani, S. M. R.: Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir, J. Hydro., 476, 433-441, https://doi.org/10.1016/j.jhydrol.2012.11.017, 2013.

Verkade, J., Brown, J., Reggiani, P., and Weerts, A.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, J. Hydro., 501, 73-91, https://doi.org/10.1016/j.jhydrol.2013.07.039, 2013.

Vogel, R. M., and N. M. Fennessey: Flow duration curves: I. A new interpretation and confidence intervals, J. Water Resour. Plan. Manage., 120(4), 485–504, https://doi.org/10.1061/(ASCE)0733-9496(1994)120:4(485), 1994.

Wang, E., Zhang, Y., Luo, J., Chiew, F. H., and Wang, Q.: Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and historical weather data, Water Resour. Res., 47, W05516, https://doi.org/10.1029/2010WR009922, 2011.

Wei, C.-C.: Comparing single-and two-segment statistical models with a conceptual rainfall-runoff model for river streamflow prediction during typhoons, Environ. Modell. Softw., 85, 112-128, https://doi.org/10.1016/j.envsoft.2016.08.013, 2016.

Wei, Z., Meng, Y., Zhang, W., Peng, J., and Meng, L.: Downscaling SMAP soil moisture estimation with gradient boosting decision tree regression over the Tibetan Plateau, Remote Sens. Environ., 225, 30-44, https://doi.org/10.1016/j.rse.2019.02.022, 2019.

Willmott C J.: On the validation of models. Physical geography, 2(2), 184-194, https://doi.org/10.1080/02723646.1981.10642213, 1981.

Yang, Q., Zhang, H., Wang, G., Luo, S., Chen, D., Peng, W., and Shao, J.: Dynamic runoff simulation in a changing environment: A data stream approach, Environ. Modell. Softw., 112, 157-165, https://doi.org/10.1016/j.envsoft.2018.11.007, 2019.

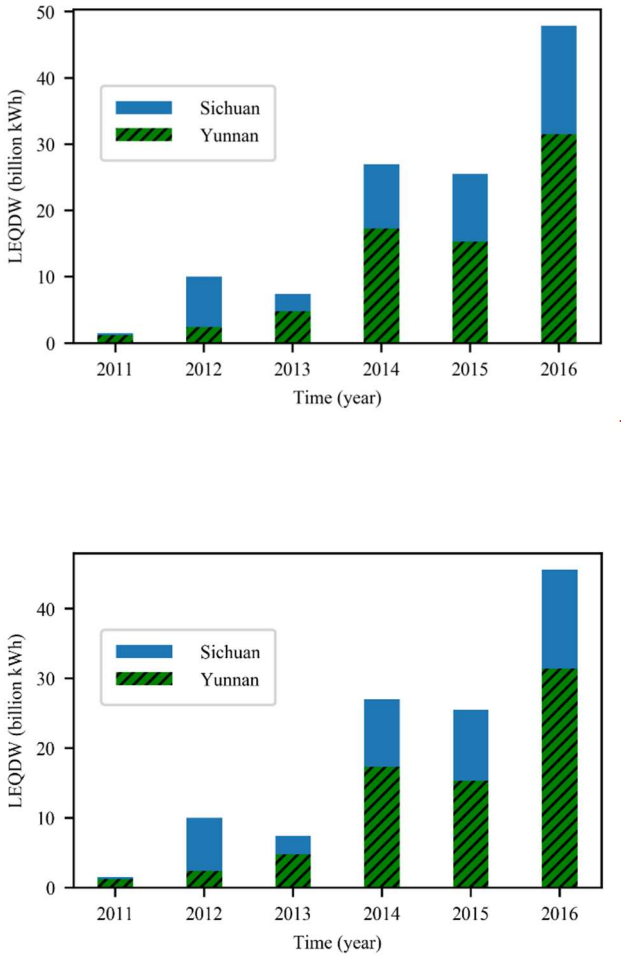
Yaseen Z M, Sulaiman S O, Deo R C, et al.: An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction, J. Hydro., 569: 387-408, 10.1016/j.jhydrol.2018.11.069, 2019.

Yilmaz K K, Gupta H V, Wagener T: A process‐based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. Water Resour. Res., 44(9), https://doi.org/10.1029/2007WR006716, 2008.

Zhan, X., Zhang, S., Szeto, W. Y., and Chen, X.: Multi-step-ahead traffic speed forecasting using multi-output gradient boosting regression tree, J. Intell. Transport. S., 1547-2442, https://doi.org/10.1080/15472450.2019.1582950, 2019.

Zhang, D., Lin, J., Peng, Q., Wang, D., Yang, T., Sorooshian, S., Liu, X., and Zhuang, J.: Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm, J. Hydro., 565, 720-736, https://doi.org/10.1016/j.jhydrol.2018.08.050, 2018.

Zhao, X., Deng, W., and Shi, Y.: Feature selection with attributes clustering by maximal information coefficient, Procedia Comput. Sci., 17, 70-79, https://doi.org/10.1016/j.procs.2013.05.011, 2013.



1085 **Figure 1~~: Loss.~~** Losses of electric quantity due to discarded water (LEQDW~~).~~) in the Sichuan and Yunnan province.

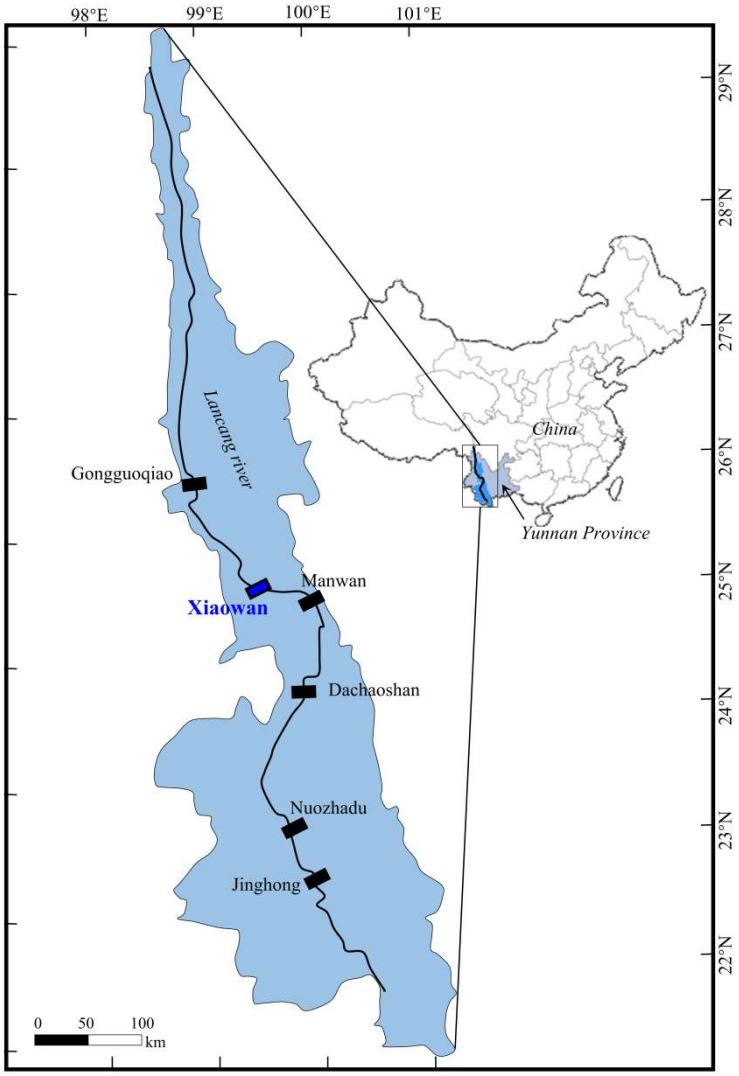
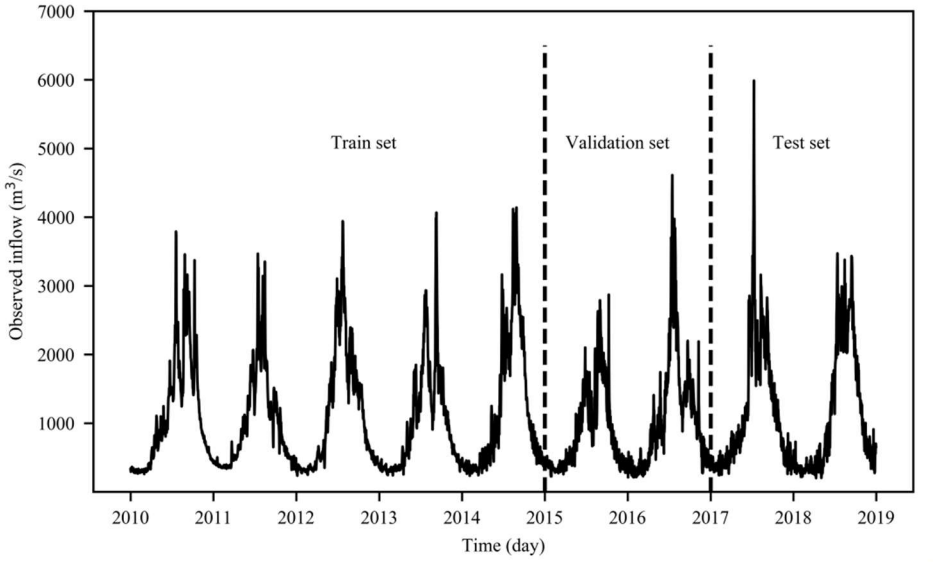


Figure 2: Location of the Xiaowan hydropower station.



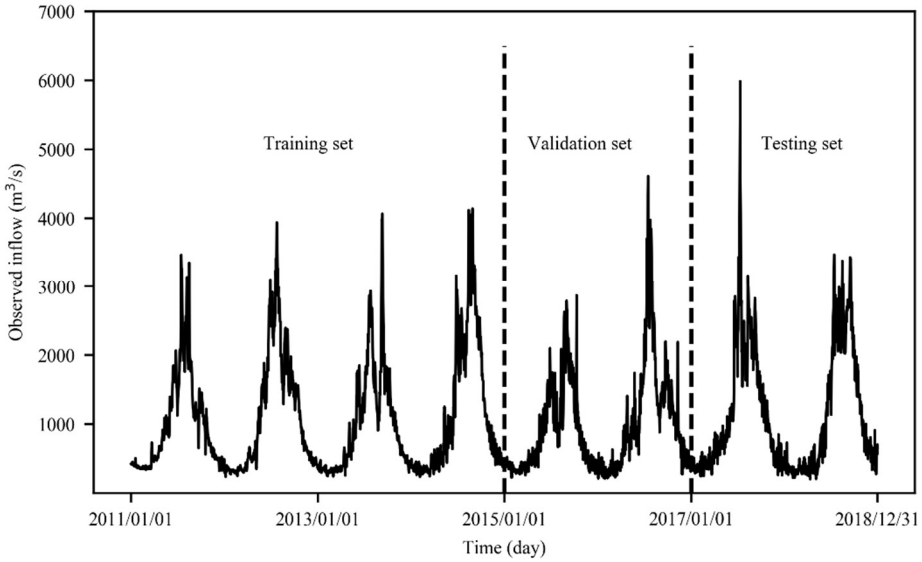


Figure 3: Daily inflow ~~data~~series of the Xiaowan hydropower station.

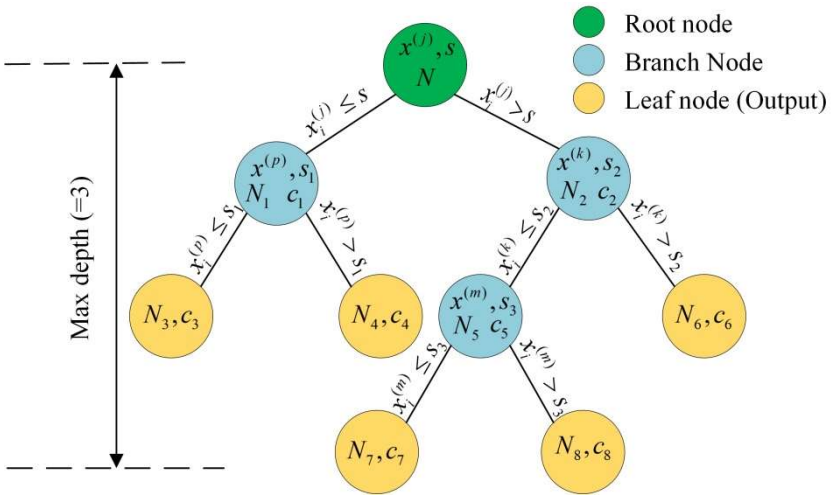
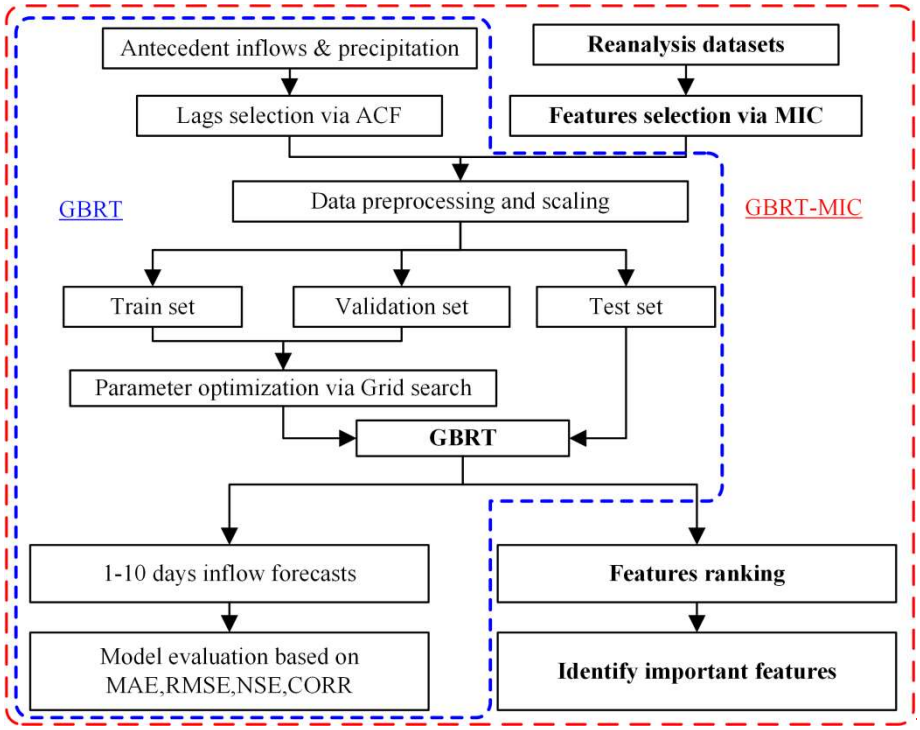


Figure 4: The structure of decision tree model.



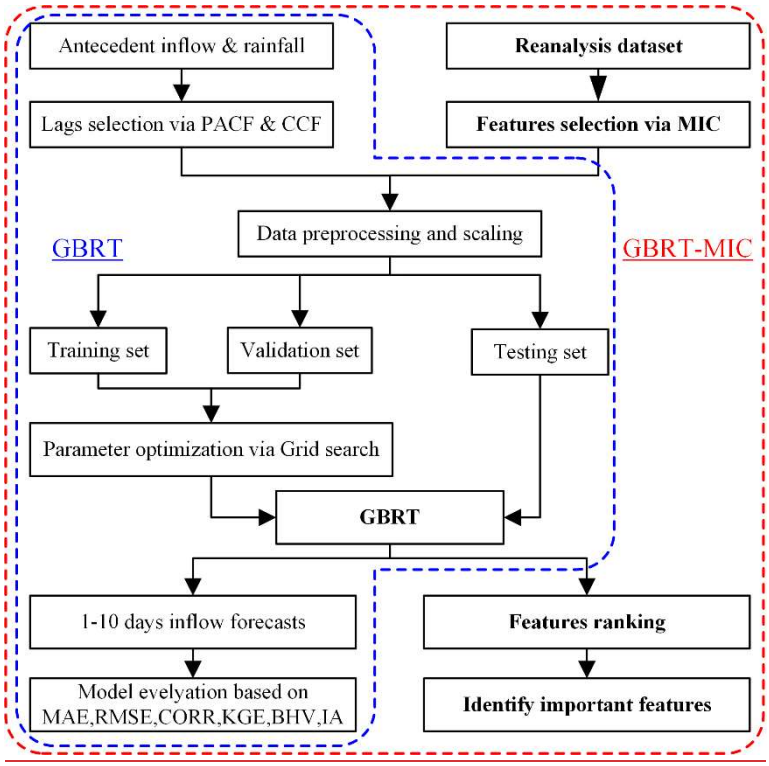


Figure 5: Overview of the framework.

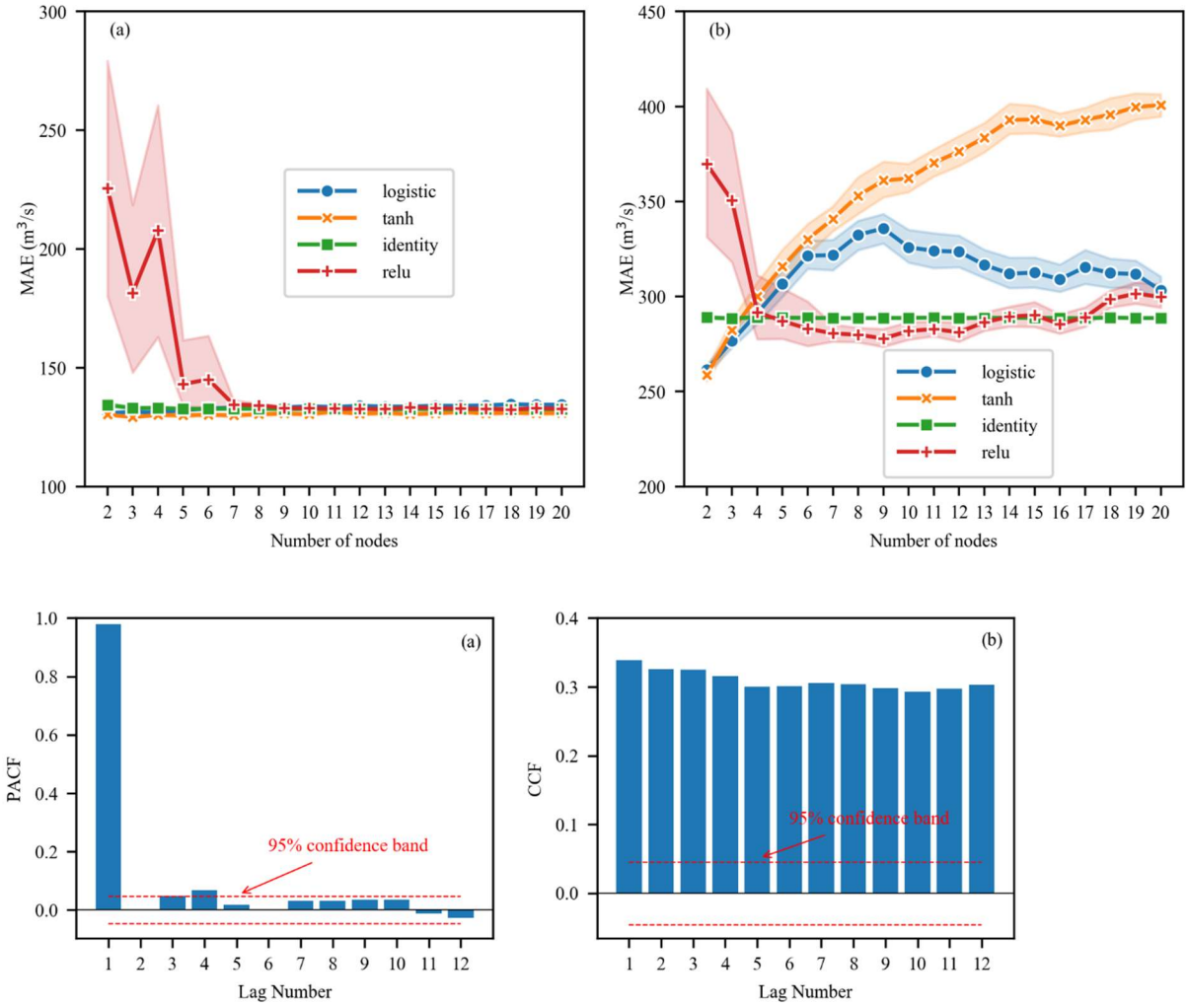


Figure 6~~:~~. PACF plot of Xiaowan daily inflow and CCF of Xiaowan rainfall and inflow. (a) PACF (b) CCF.

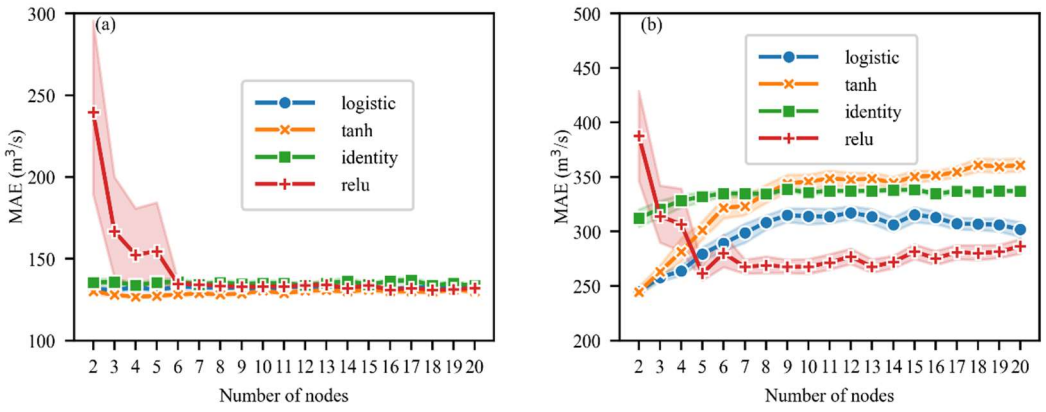
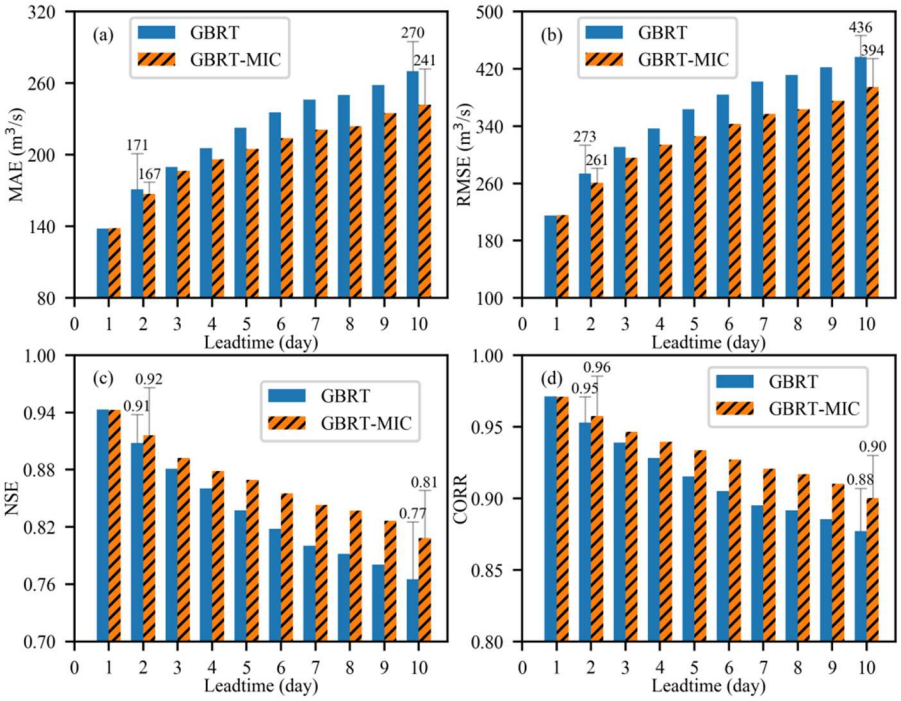


Figure 7. Sensitivity of the ~~number of nodes and~~ activation function and the number of nodes in the hidden layer on the MAE of ANN-MIC, the shadow part is 95% confidence interval obtained by bootstrap of 50 trials. (a) One-day- ahead (b) Ten-day- ahead.


(a)
GBRT
GBRT-MIC
MAE (m3/s)
Leadtime (day)
171
167
270
241
(b)
RMSE (m3/s)
273
261
436
394
(c)
NSE
0.92
0.91
0.81
0.77
(d)
CORR
0.96
0.95
0.90
0.88

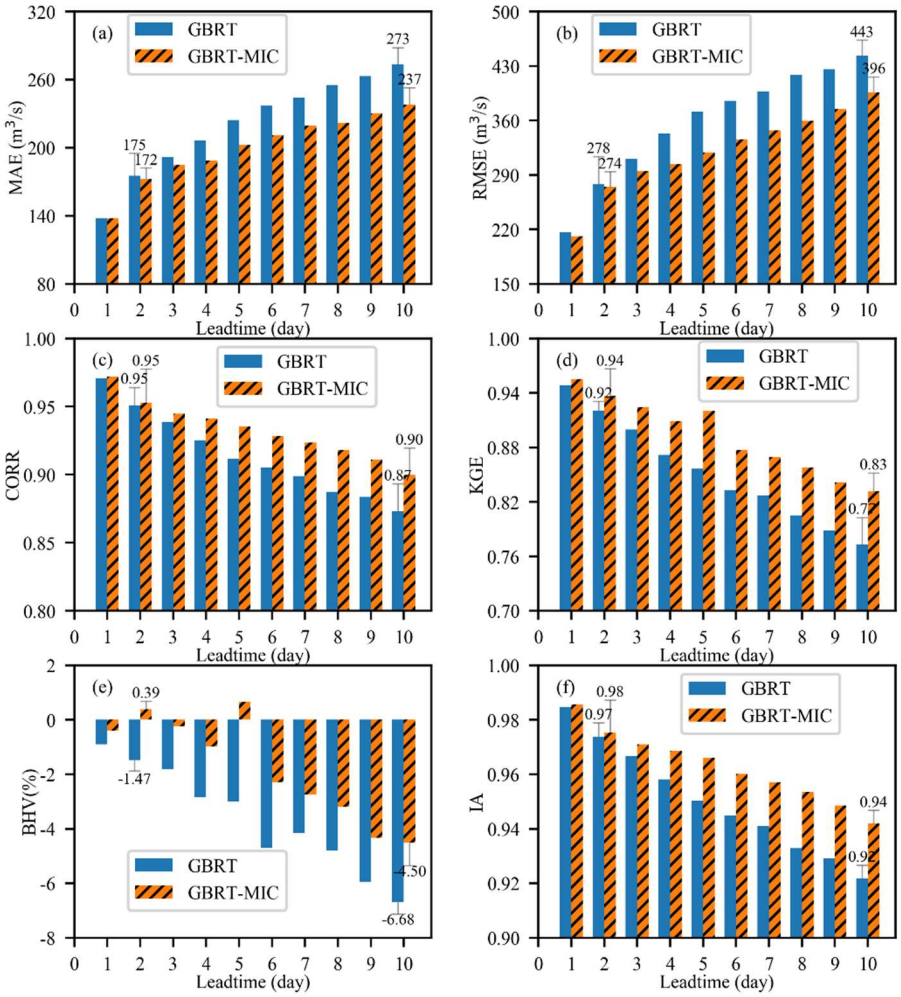
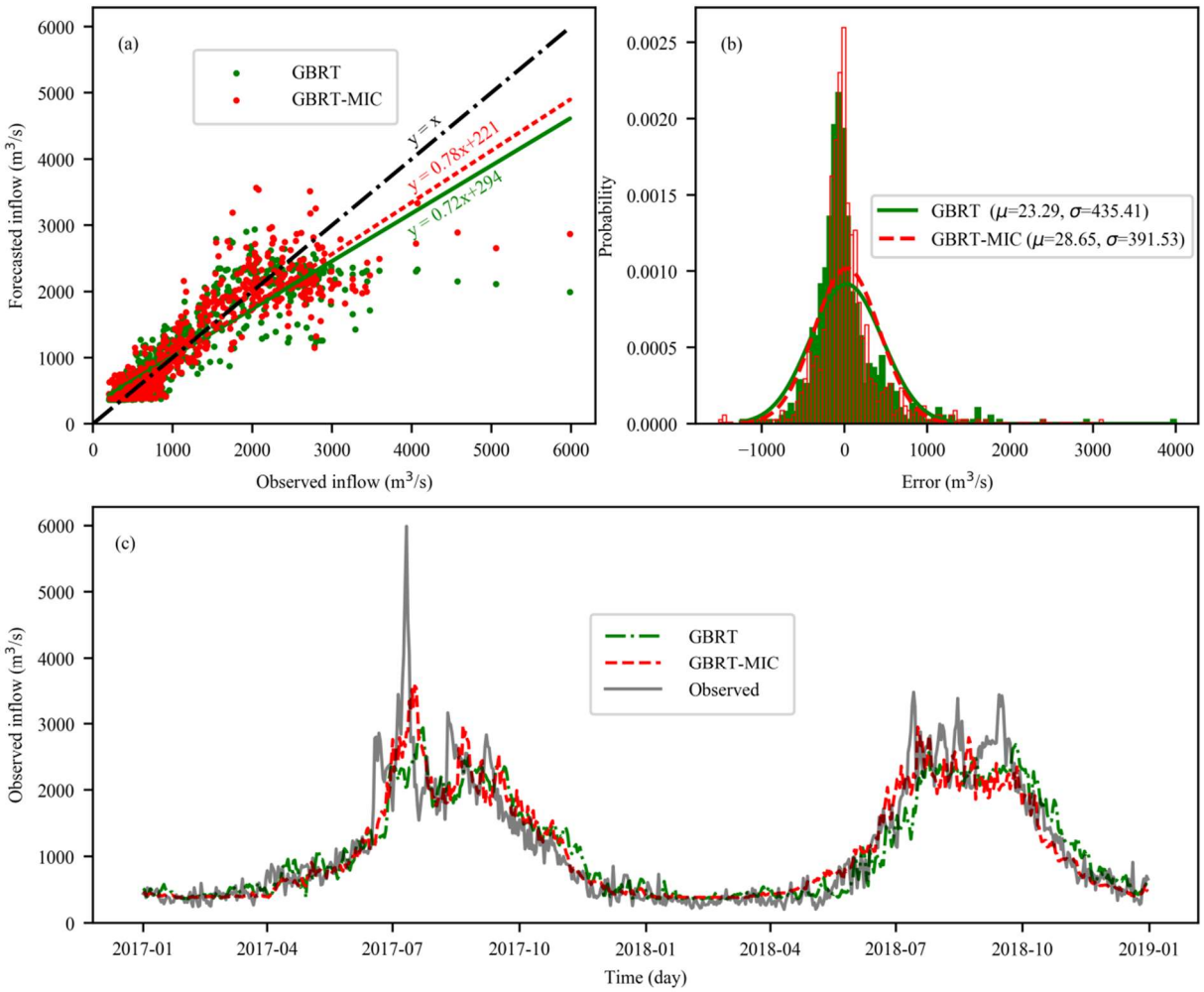


Figure ~~7:~~8. Performance of GBRT and GBRT-MIC for the ~~test~~testing set (2017-2018) in term of ~~four~~six indices. (a) MAE (b) RMSE (c) ~~NSE~~CORR (d) ~~CORR.~~


(a)
GBRT
GBRT-MIC
y = x
y = 0.78x+221
y = 0.72x+294
Forecasted inflow (m3/s)
Observed inflow (m3/s)
(b)
GBRT (μ=23.29, σ=435.41)
GBRT-MIC (μ=28.65, σ=391.53)
Probability
Error (m3/s)
(c)
GBRT
GBRT-MIC
Observed
Observed inflow (m3/s)
Time (day)

KGE (e) BHV (f) IA.

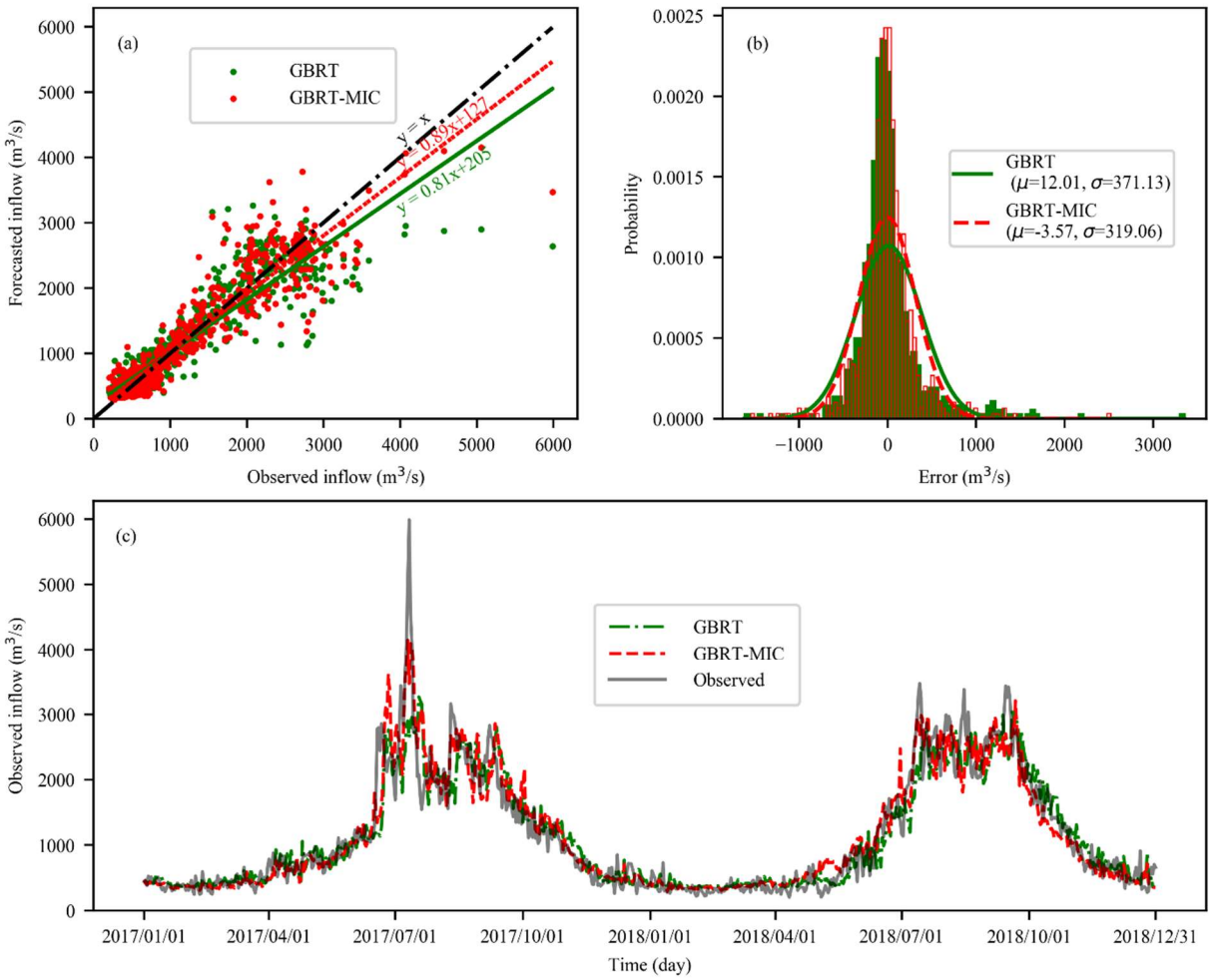
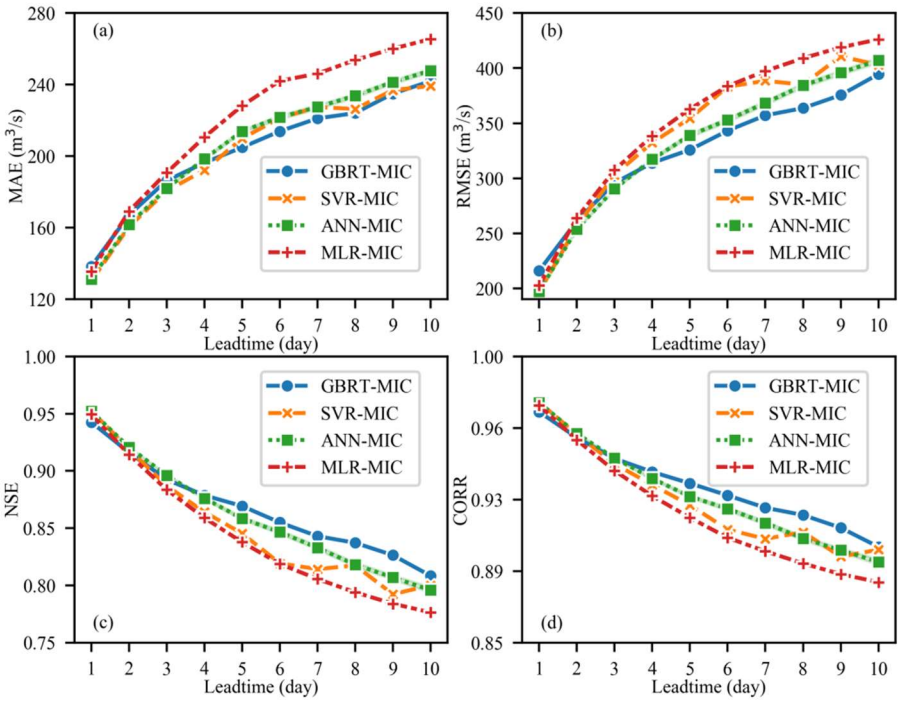


Figure ~~8: Ten~~9. Five-day- ahead inflow forecasts of the GBRT and GBRT-MIC for ~~test~~the testing set (2017-2018, 730 days~~)~~). (a) Observed versus forecasted inflow. (b) The histogram of predicting error of ~~test~~testing set (c) Comparison of the observed and forecasted inflow.



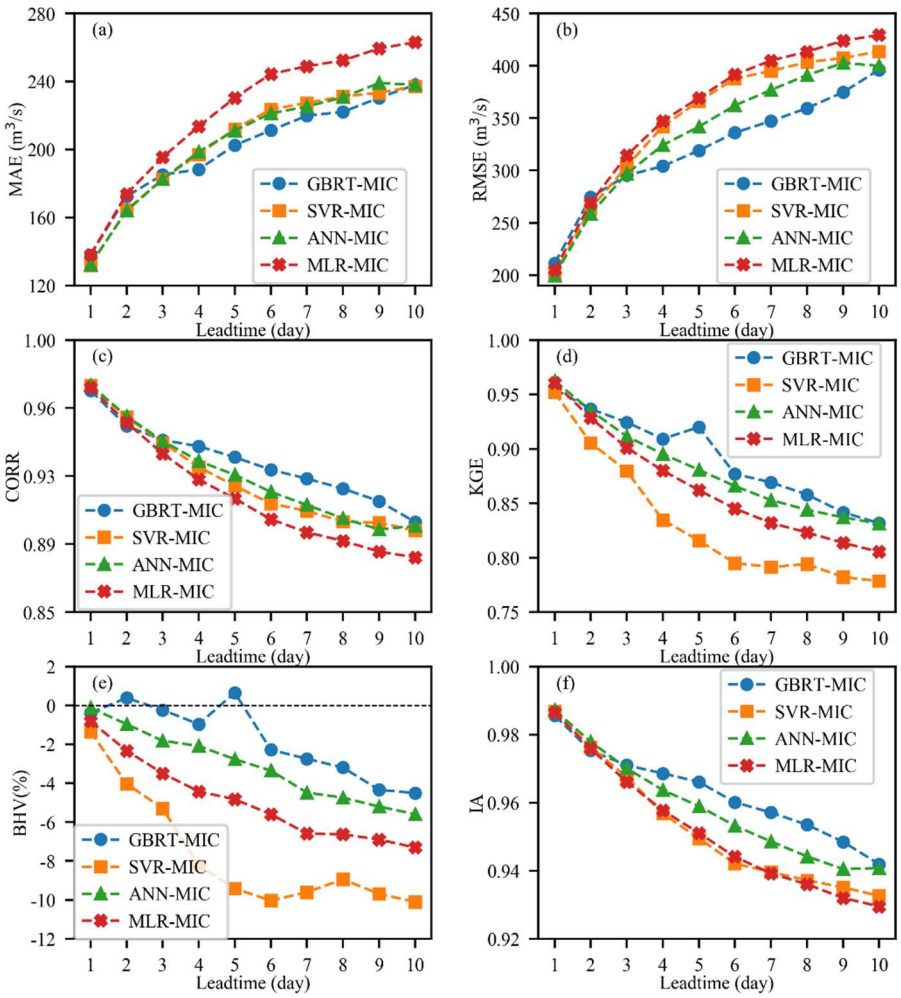
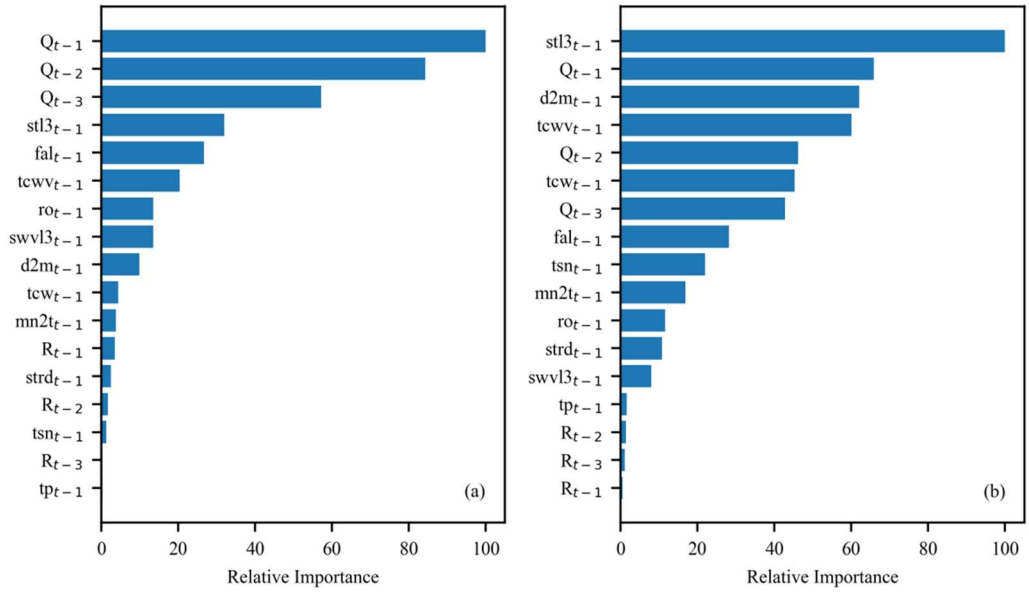


Figure ~~9:~~10. Performance of GBRT-MIC, SVR-MIC, ANN-MIC and MLR-MIC for the ~~test~~testing set (2017-2018) in term of ~~four~~six indices. ~~**Bootstrap 95% confidence intervals of ANN-MIC for 50 trials.**~~ (a) MAE (b) RMSE (c) ~~**NSE**~~CORR (d) ~~**CORR**~~KGE (e) BHV (f) IA.



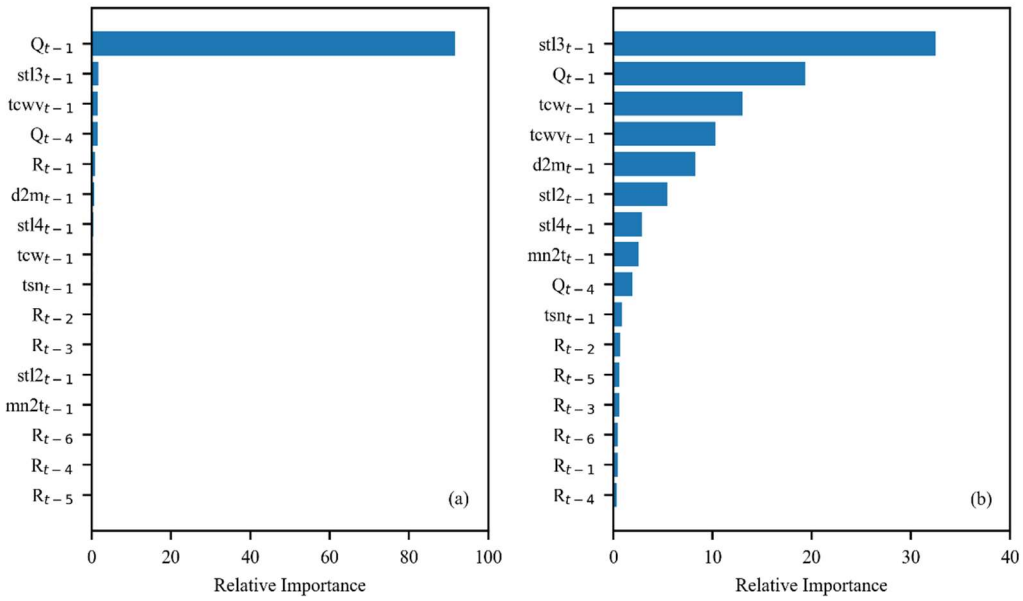


Figure ~~10:~~11. Feature importance ~~for~~obtained by GBRT-MIC. (a) One-day- ahead (b) Ten-day- ahead.

Table 1 ~~ACF of Xiaowan daily inflow~~. The candidate inputs via PACF and ~~rainfall (2011-2014).~~CCF.

Number	Input								
~~ACF~~ ~~t~~ ~~$t-$~~	~~$t-2$~~ Q_{t-1}, Q_{t-4}	~~$t-3$~~	~~$t-4$~~	~~$t-5$~~	~~$t-6$~~	~~$t-7$~~	~~$t-8$~~	~~$t-9$~~	
1									
~~Q_t~~ 2	~~**0.974**~~ $Q_{t-1}, Q_{t-4}, R_{t-1}$	~~0.950~~	~~0.931~~	~~0.914~~	~~0.899~~	~~0.883~~	~~0.870~~	~~0.858~~	~~0.847~~	~~0.838~~
~~R_t~~ 3	~~**0.504**~~ $Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}$	~~0.273~~	~~0.226~~	~~0.233~~	~~0.195~~	~~0.174~~	~~0.153~~	~~0.170~~	~~0.171~~	~~0.174~~
4	$Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}$									
5	$Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}$									
6	$Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}$									
7	$Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, R_{t-6}$									
8	$Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, R_{t-6}, R_{t-7}$									
9	$Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, R_{t-6}, R_{t-7}, R_{t-8}$									
10	$Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, R_{t-6}, R_{t-7}, R_{t-8}, R_{t-9}$									
11	$Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, R_{t-6}, R_{t-7}, R_{t-8}, R_{t-9}, R_{t-10}$									
12	$Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, R_{t-6}, R_{t-7}, R_{t-8}, R_{t-9}, R_{t-10}, R_{t-11}$									
13	$Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, R_{t-6}, R_{t-7}, R_{t-8}, R_{t-9}, R_{t-10}, R_{t-11}, R_{t-12}$									

Table 2. The candidate inputs from reanalysis data via MIC.

Number	Input
1	$obs, stl3_{t-1}$
2	$obs, stl3_{t-1}, d2m_{t-1}$
3	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}$
4	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}$
5	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}$
6	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}$
7	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}$
8	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}$
9	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}$
10	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}$
11	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}$
12	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}$
13	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}$
14	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}$
15	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}, swvl4_{t-1}$
16	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}, swvl4_{t-1}, mx2t_{t-1}$
17	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}, swvl4_{t-1}, mx2t_{t-1}, sf_{t-1}$
18	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}, swvl4_{t-1}, mx2t_{t-1}, sf_{t-1}, cp_{t-1}$
19	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}, swvl4_{t-1}, mx2t_{t-1}, sf_{t-1}, cp_{t-1}, tp_{t-1}$
20	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}, swvl4_{t-1}, mx2t_{t-1}, sf_{t-1}, cp_{t-1}, tp_{t-1}, rsn_{t-1}$
21	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}, swvl4_{t-1}, mx2t_{t-1}, sf_{t-1}, cp_{t-1}, tp_{t-1}, rsn_{t-1}, lsp_{t-1}$
22	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}, swvl4_{t-1}, mx2t_{t-1}, sf_{t-1}, cp_{t-1}, tp_{t-1}, rsn_{t-1}, lsp_{t-1}, sd_{t-1}$
23	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}, swvl4_{t-1}, mx2t_{t-1}, sf_{t-1}, cp_{t-1}, tp_{t-1}, rsn_{t-1}, lsp_{t-1}, sd_{t-1}, smlt_{t-1}$
24	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}, swvl4_{t-1}, mx2t_{t-1}, sf_{t-1}, cp_{t-1}, tp_{t-1}, rsn_{t-1}, lsp_{t-1}, sd_{t-1}, smlt_{t-1}, istl1_{t-1}$
25	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}, swvl4_{t-1}, mx2t_{t-1}, sf_{t-1}, cp_{t-1}, tp_{t-1}, rsn_{t-1}, lsp_{t-1}, sd_{t-1}, smlt_{t-1}, istl1_{t-1}, istl3_{t-1}$
26	$obs, stl3_{t-1}, d2m_{t-1}, tcwv_{t-1}, tcw_{t-1}, stl2_{t-1}, mn2t_{t-1}, tsn_{t-1}, stl4_{t-1}, stl1_{t-1}, ro_{t-1}, swvl1_{t-1}, swvl2_{t-1}, swvl3_{t-1}, t2m_{t-1}, swvl4_{t-1}, mx2t_{t-1}, sf_{t-1}, cp_{t-1}, tp_{t-1}, rsn_{t-1}, lsp_{t-1}, sd_{t-1}, smlt_{t-1}, istl1_{t-1}, istl3_{t-1}, istl2_{t-1}$

Note: obs represents the selected observed optimal input set, $obs = \{Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, R_{t-6}\}$

Table 3. List of inputs of GBRT-MIC. There are of two types, observed and reanalysis variables. The reanalysis variables are available ~~four~~two time a day at 00:00 UTC~~, 06:00 UTC,~~ and 12:00 UTC ~~and 18:00 UTC.~~. The cumulative variable (e.g., Total column water) is the sum of ~~four~~two periods and the instantaneous variable (e.g. 2 meter dewpoint temperature) is the mean of ~~four~~two periods.

~~No.~~Number	~~Variable~~Description	Index	Unit	MIC	Type
1	Inflow at day t – 1	Q_{t-1} ~~Q_t~~	~~$m^3 \cdot s^{-1}$~~ $m^3 s^{-1}$	-	Obs~~.~~.
2	Inflow at day t ~~– 1~~ – 2	Q_{t-4} Q_{t-1}	~~$m^3 \cdot s^{-1}$~~ $m^3 s^{-1}$	-	Obs.
3	~~Inflow~~Rainfall at day t – ~~2~~1	~~Q_{t-2}~~ R_{t-1}	~~$m^3 \cdot s^{-1}$~~mm	-	Obs.
4	Rainfall at day t – 2	~~R_t~~ R_{t-2}	mm	-	Obs.
5	Rainfall at day t – ~~1~~3	~~R_{t-1}~~ R_{t-3}	mm	-	Obs.
6	Rainfall at day t – ~~2~~4	~~R_{t-2}~~ R_{t-4}	mm	-	Obs.
7	~~Forecast albedo~~Rainfall at day t – 5	~~fal_t~~ R_{t-5}	~~-~~mm	~~0.853-~~	~~ERA-I~~Obs.
8	~~Soil temperature level 3~~Rainfall at day t – 6	~~$stl3_t$~~	~~K~~mm	~~0.814-~~	~~ERA-I~~Obs.
9	~~2 meter dewpoint~~Soil temperature level 3	~~$d2m_t$~~	K	0.~~721~~847	ERA-I
10	~~Minimum~~2 meter dewpoint temperature ~~at 2 meters~~	~~$mn2t_t$~~	K	0.~~660~~781	ERA-I
11	Total column water vapour	~~tcw_t~~	~~$kg \cdot m^{-2}$~~ kg m^{-2}	0.~~660~~699	ERA-I
12	Total column water ~~vapour~~	~~$tcwv_t$~~	~~$kg \cdot m^{-2}$~~ kg m^{-2}	0.~~659~~699	ERA-I
13	~~Runoff~~Soil temperature level 2	~~ro_t~~	~~m~~K	0.~~653~~689	ERA-I
14	~~Volumetric soil water layer 3~~Minimum temperature at 2 meters	~~$swvl3_t$~~	~~$m^3 \cdot m^{-3}$~~ K	0.~~637~~684	ERA-I
15	Temperature of snow layer	~~tsn_t~~	K	0.~~631~~664	ERA-I
16	~~Surface thermal radiation downwards~~Soil temperature level 4	~~$strd_t$~~	~~$J \cdot m^{-2}$~~ K	0.~~625~~643	ERA-I
~~17~~	~~Total precipitation~~	~~tp_t~~	~~m~~	~~0.391~~	~~ERA-I~~

Table ~~3~~

4. Four commonly used activation functions for ANN-MIC.

Name	Functional expression
~~logistic~~Logistic	$f(x) = \frac{1}{1+e^{-x}}$
~~tanh~~Tanh	$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
~~identity~~Identity	$f(x) = x$
~~relu~~Relu	$f(x) = max(0, x)$

Table 4

5. Tuning parameters of ANN-MIC and SVR-MIC.

Model	Tuning parameter	Tuning range	1	2	3	4	5	6	7	8	9	10
ANN-MIC	Structure		19-4-1	19-2-1	19-3-1	19-2-1	19-2-1	19-2-1	19-2-1	19-2-1	19-2-1	19-2-1
ANN-MIC	Structure		Tanh	17-3-1 tanh	17-2-1 logistic	17-2-2 logistic	17-2-3 logistic	17-2-4 logistic	17-2-5 logistic	17-2-6 logistic	17-2-7 logistic	17-2-8 logistic	17-2-9 tanh	tanh
SVR-MIC			8.9693	7.6987	6.2105	53.1206	1.0000	9.8590	1.0000	44.2134	1.0000	81.4581	11.4211	120.2356	1.0000	0.2587	1.0000	144.4810	6.2105	0.5105	1.0000	6.2105
	epsilon	(**0.0030**, 0.1, 60)	0.0007	0.0028	0.0066	0.0006	0.0079	0.0015	0.0017	0.0004	0.0012	0.0027	0.0085	0.0043
	gamma	(**0.0265**, 0.1, 0323)	0.0201	0.0037	0.0120	0.0051	0.0062	0.0012	0.0021	0.0088	0.0315	0.0067	0.1815

Note: The bold parts, (min, max, step) represent $[\underline{\ min+\frac{max-min}{step-1}\times 0\ }, \underline{\ min+\frac{max-min}{step-1}\times 1\ }, \ldots, \underline{\ min+\frac{max-min}{step-1}\times(step-1)\ }]$.

Table 5

6. Tuning parameters of GBRT and GBRT-MIC.

Tuning parameter	Tuning range	Optimal parameters (the lead times of 1-10 days)	
		GBRT	GBRT-MIC
max_leaf_nodes	[3,5,7,92, 4, 6, …, 40]	7,7,3,5,9,3,7,9,7,38, 4, 4, 4, 4, 2, 4, 2, 2, 2	3,3,3,7,3, 9, 13, 7,7,9, 15, 4, 5,9 4, 4, 17
min_samples_leaf	[3,5,7,91, 6, 11, …, 46]	9,5,9,5,5,5,7,3,3,36, 31, 1, 1, 1, 31, 6, 1, 6, 1	9,9,9,7,5,5,5,5,5,7 2, 7, 2, 4, 2, 1, 10, 10, 8, 1
max_depth	[1, 2, 3,5,7,9 …, 10]	3, 2, 2, 2, 3,9,9,9, 1, 3,3,5,9,5 1, 1, 1	4, 6, 8, 5, 9, 9, 2, 2, 7,9,9,5,7,5,7,7 2
min_samples_split	[10,20,30,2, 4, 6, …, 40,50]	30,10,40,50,10,20,50,50,50,4018, 2, 16, 16, 24, 2, 16, 2, 2, 2	50,10,20,10,10,20,30,50,10,5018, 15, 12, 13, 8, 3, 19, 3, 19, 8
n_estimators	[100, 200, 250, …, 1200300, …, 4000]	600,600,450,450,550,500,500,550,600,6001100, 900, 1200, 700, 700, 1200, 600, 1100, 900, 900	1150,550,350,350,450,350,350,400,500,5003800, 2700, 1300, 900, 1000, 700, 1400, 2000, 1300, 1200

Table ~~6~~ 7. Performance indices of the ~~train~~training set.

Indice	Model	1	2	3	4	5	6	7	8	9	10
MAE (m³/s)	GBRT-MIC	**~~70~~56**	**~~98~~63**	**~~126~~78**	**~~141~~122**	**~~150~~89**	**~~167~~163**	**~~173~~161**	**~~172~~155**	**~~170~~161**	**~~169~~172**
	SVR-MIC	~~86~~98	~~116~~126	~~138~~144	~~151~~162	~~161~~173	~~177~~183	~~176~~188	~~182~~194	~~192~~197	~~192~~203
	ANN-MIC	~~90~~99	~~119~~129	~~139~~148	~~153~~162	~~163~~172	~~174~~184	~~181~~192	~~187~~196	~~194~~203	~~198~~205
	MLR-MIC	~~92~~103	~~124~~136	~~146~~159	~~162~~175	~~173~~187	~~185~~198	~~195~~207	~~205~~215	~~213~~221	~~219~~228
RMSE (m³/s)	GBRT-MIC	**~~103~~77**	**~~152~~87**	**~~197~~107**	**~~224~~185**	**~~248~~124**	**~~272~~257**	**~~285~~255**	**~~288~~245**	**~~286~~254**	**~~283~~278**
	SVR-MIC	~~139~~153	~~195~~212	~~240~~247	~~267~~280	~~285~~300	~~314~~319	~~318~~329	~~329~~337	~~345~~344	~~344~~353
	ANN-MIC	~~138~~151	~~190~~206	~~226~~240	~~248~~264	~~270~~284	~~287~~304	~~307~~318	~~319~~328	~~331~~334	~~337~~339
	MLR-MIC	~~142~~157	~~195~~214	~~233~~250	~~259~~275	~~281~~295	~~305~~315	~~322~~330	~~337~~342	~~348~~352	~~359~~361
CORR~~N SE~~	GBRT-MIC	**0.~~984~~9952**	**0.~~965~~9940**	**0.~~942~~6908**	**0.~~925~~7724**	**0.~~909~~0877**	**0.~~890~~8464**	**0.~~880~~4468**	**0.~~877~~2951**	**0.~~879~~1947**	**0.~~881~~6936**
	SVR-MIC	0.~~971~~9811	0.~~943~~6411	0.~~914~~9511	0.~~894~~7380	0.~~880~~2866	0.~~854~~0186	0.~~850~~8126	0.~~840~~2073	0.~~824~~5039	0.~~824~~8977
	ANN-MIC	0.~~971~~8185	0.~~946~~6533	0.~~924~~8528	0.~~909~~1944	0.~~892~~4331	0.~~878~~5232	0.~~861~~0156	0.~~849~~5101	0.~~838~~5066	0.~~832~~4036
	MLR-MIC	0.9801	0.9628	0.9485	0.9376	0.9278	0.9172	0.9090	0.9019	0.8959	0.8900
KGE	~~MLR~~GBRT-MIC	**0.~~970~~9884**	**0.~~943~~8987**	**0.~~919~~6738**	**0.~~900~~9439**	**0.~~883~~3642**	**0.~~862~~9009**	**0.~~847~~2069**	**0.~~832~~4099**	**0.~~820~~6900**	**0.~~810~~0887**
	SVR-MIC	0.9618	0.9207	0.8982	0.8613	0.8445	0.8266	0.8223	0.8247	0.8149	0.8103
	ANN-MIC	0.9735	0.9508	0.9325	0.9177	0.9048	0.8907	0.8800	0.8724	0.8668	0.8611
	MLR-MIC	0.9718	0.9473	0.9272	0.9117	0.8979	0.8829	0.8713	0.8613	0.8528	0.8444
~~CORR~~BHV (%)	GBRT-MIC	-0.~~992~~3025	-0.~~983~~2682	-0.~~972~~8986	**~~0.964~~**-1.3422	**~~0.954~~**-1.4019	**~~0.946~~**-1.5485	**~~0.941~~**-1.7486	**~~0.939~~**-1.7692	**~~0.939~~**-2.6647	**~~0.941~~**-3.0375
	SVR-MIC	~~0.985~~-1.3488	~~0.972~~-3.3959	~~0.959~~-4.0686	~~0.950~~-6.9058	~~0.942~~-7.5421	~~0.929~~-8.2216	~~0.925~~-6.9950	~~0.920~~-6.1996	~~0.913~~-6.2406	~~0.913~~-5.6687
	ANN-MIC	**-0.~~985~~1814**	**-0.~~973~~1286**	**-0.~~962~~0710**	**-0.~~953~~7723**	**-0.~~945~~2249**	**-0.~~937~~6815**	**-0.~~928~~6878**	**-0.~~922~~5821**	**-0.~~916~~6487**	**-0.~~913~~1239**
	MLR-MIC	-0.~~985~~1468	~~0.971~~-1.0527	~~0.959~~-1.5863	~~0.949~~-1.9709	~~0.940~~-1.9477	~~0.929~~-2.1634	~~0.921~~-2.0182	~~0.913~~-1.8074	~~0.907~~-2.0454	~~0.901~~-1.7473
IA	GBRT-MIC	**0.9976**	**0.9969**	**0.9952**	**0.9854**	**0.9935**	**0.9706**	**0.9712**	**0.9734**	**0.9712**	**0.9650**
	SVR-MIC	0.9902	0.9804	0.9727	0.9636	0.9574	0.9506	0.9472	0.9449	0.9421	0.9386
	ANN-MIC	0.9906	0.9820	0.9752	0.9695	0.9643	0.9586	0.9541	0.9509	0.9487	0.9468
	MLR-MIC	0.9898	0.9807	0.9729	0.9668	0.9613	0.9551	0.9502	0.9460	0.9423	0.9387

~~None~~*Note :* The bold numbers represent the values of performance criterion for the best fitted models.

Table ~~7~~ 8. Performance indices of the ~~validation~~testing set.

Indice	Model	1	2	3	4	5	6	7	8	9	10
MAE (m^3/s)	GBRT-MIC	**96**137	**127**172	**153**185	**170188**	**177202**	**190211**	**194219**	**194222**	**196230**	**189**237
	SVR-MIC	120**131**	149164	169182	183197	193212	201223	202227	200231	206233	207**237**
	ANN-MIC	122132	154**163**	174**182**	188198	199211	206221	214225	215230	218239	216238
	MLR-MIC	127138	161173	185195	202213	212230	220244	228248	233252	236259	240263
RMSE (m^3/s)	GBRT-MIC	**132**211	**188**274	**230295**	**254304**	**266319**	**286336**	**294347**	**296359**	**293374**	**276396**
	SVR-MIC	178200	240263	274303	294342	308366	326387	325395	326403	329407	330413
	ANN-MIC	177**199**	240**258**	271296	290324	308341	317362	332376	337391	337402	333399
	MLR-MIC	184205	247268	281314	304347	319369	332391	342404	346413	349423	353429
~~NSE~~COR R	GBRT-MIC	0.**9669**9722	0.**9324**9526	**0.8991944 9**	**0.8766941 4**	**0.8647935 4**	**0.8440928 5**	**0.8348923 6**	**0.8325918 1**	**0.8364911 2**	**0.8544899 7**
	SVR-MIC	0.9395975 1	0.8898957 5	0.8570943 4	0.8352930 0	0.8181919 6	0.7969909 9	0.7982905 8	0.7965899 9	0.7926899 3	0.7922895 0
	ANN-MIC	**0.9398975 2**	**0.8900958 0**	0.8594944 4	0.8398933 3	0.8185925 7	0.8075916 3	0.7893909 1	0.7830901 7	0.7824895 6	0.7876897 5
	MLR-MIC	0.9349973 8	0.8832954 5	0.8486937 4	0.8236923 1	0.8052912 6	0.7889901 2	0.7766894 0	0.7708889 3	0.7666883 4	0.7620880 2
KGE	GBRT-MIC	0.9550	**0.9367**	**0.9244**	**0.9092**	**0.9200**	**0.8769**	**0.8693**	**0.8580**	**0.8417**	**0.8317**
	SVR-MIC	0.9520	0.9055	0.8797	0.8347	0.8158	0.7950	0.7915	0.7941	0.7822	0.7786
	ANN-MIC	**0.9625**	0.9352	0.9115	0.8953	0.8808	0.8658	0.8530	0.8440	0.8371	0.8313
	MLR-MIC	0.9605	0.9284	0.9011	0.8800	0.8620	0.8452	0.8319	0.8232	0.8137	0.8054
BHV (%)	GBRT-MIC	-0.3826	**0.3880**	**-0.2319**	**-0.9629**	**0.6566**	**-2.2766**	**-2.7422**	**-3.1924**	**-4.3363**	**-4.5040**
	SVR-MIC	-1.3382	-4.0253	-5.3037	-8.2410	-9.4167	-10.0357	-9.6049	-8.9452	-9.6886	-10.1058
	ANN-MIC	**-0.1228**	-0.9608	-1.8150	-2.0839	-2.7642	-3.3509	-4.4831	-4.7424	-5.1999	-5.5886
	MLR-MIC	-0.8093	-2.3244	-3.4945	-4.4210	-4.8268	-5.5955	-6.5914	-6.6302	-6.8944	-7.3080
IA~~CORR~~	GBRT-MIC	0.**9835**9856	0.**9661**9753	**0.9493971 0**	**0.9376968 6**	**0.9311966 1**	**0.9203960 1**	**0.9154957 1**	**0.9140953 5**	**0.9163948 5**	**0.9260941 9**
	SVR-MIC	0.9693986 9	0.9434976 3	0.9257967 6	0.9139956 8	0.9046949 5	0.8927942 1	0.8939939 6	0.8928937 2	0.8903935 1	0.8906932 6
	ANN-MIC	**0.9697987 2**	**0.9444977 9**	0.9289970 1	0.9185963 7	0.9081959 0	0.9015953 2	0.8925948 6	0.8895944 2	0.8890940 5	0.8912940 8
	MLR-MIC	0.9673986 5	0.9412975 9	0.9238966 1	0.9112957 7	0.9018951 1	0.8935944 1	0.8871939 2	0.8839936 0	0.8818932 0	0.8798929 5

~~*None* : The bold numbers represent the values of performance criterion for the best fitted models.~~

Table 8

Performance indices of the test set.

Indice	Model	1	2	3	4	5	6	7	8	9	10
MAE	GBRT-MIC	138	167	186	196	**205**	**214**	**221**	**224**	**235**	242
	SVR-MIC	**130**	**160.5**	**180.6**	**192**	210	221	227	226	237	**239**
	ANN-MIC	131	161.4	181.4	199	214	221	228	235	242	247
	MLR-MIC	135	169	191	211	228	242	246	254	260	266
RMSE	GBRT-MIC	216	261	296	**314**	**326**	**343**	**357**	**364**	**376**	**395**
	SVR-MIC	**196.7**	256	303	332	354	383	388	385	411	403
	ANN-MIC	197.1	**253**	**290**	318	339	354	368	386	398	406
	MLR-MIC	203	264	307	338	363	383.7	397	409	419	426
NSE	GBRT-MIC	0.9426	0.9161	0.8924	**0.8787**	**0.8693**	**0.8550**	**0.8430**	**0.8372**	**0.8264**	**0.8084**
	SVR-MIC	**0.9524**	0.9190	0.8870	0.8641	0.8454	0.8193	0.8142	0.8175	0.7923	0.8001
	ANN-MIC	0.9522	**0.9211**	**0.8966**	0.8756	0.8582	0.8457	0.8337	0.8165	0.8052	0.7967
	MLR-MIC	0.9494	0.9143	0.8836	0.8590	0.8378	0.8188	0.8056	0.7940	0.7840	0.7764
CORR	GBRT-MIC	0.9710	0.9575	0.9465	**0.9395**	**0.9335**	**0.9272**	**0.9207**	**0.9169**	**0.9102**	**0.9002**
	SVR-MIC	**0.9760**	0.9591	0.9437	0.9327	0.9221	0.9093	0.9044	0.9079	0.8951	0.8987
	ANN-MIC	0.9758	**0.9598**	**0.9470**	0.9358	0.9265	0.9197	0.9132	0.9037	0.8974	0.8926
	MLR-MIC	0.9744	0.9562	0.9401	0.9269	0.9155	0.9050	0.8979	0.8915	0.8859	0.8817

NoneNote : The bold numbers represent the values of performance criterion for the best fitted models.