Reply to Anonymous Referee #3

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. This paper focuses on improving prediction accuracy by developing new model and importing ERA-Interim reanalysis data and aims to providing reference for reducing discard water. The following are our point-by-point responses to your comments.

Proposed changes to manuscript: N/A.

1. 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) in these days 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 confidence interval obtained by hypothesis test is used to replace user-defined threshold value to determine the significant relationships. The all calculation results will be updated accordingly.

Proposed changes to manuscript: In Section 4.1 of the revised manuscript, PACF and CCF to determining the model structures for inflow and rainfall, respectively. Hypothesis test is used to determine the significant relationships replacing user-defined threshold value. The all calculation results will be updated accordingly in the revised manuscript.

2. 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. There is about 1% and 5% improvement in two-day and ten-day ahead forecasting according to NSE and there is about 2.3% and 10.7% improvement in two-day and ten-day ahead forecasting according to MAE. Considering the inflow forecasting is good for reducing discard water, we think the improvement also is very valuable.

Proposed changes to manuscript: Revised input structures are used to compare with developed model with ERA-Interim dataset. More discussion about results of models will be given in Section 4.4 of the revised manuscript.

3. 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, in these days. Numerical experiment results indicate that more lagged values of input variables will be included in the model's structure.

Proposed changes to manuscript: PACF and CCF will be used to determining the model structures for inflow and rainfall, respectively. More lagged values of input variables will be included in the model's structure in the revised manuscript.

4. 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 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 Pearson correlation coefficient (CORR) is a good measurement of the average error.

Proposed changes to manuscript: Peak flow criterion, degree of agreement and Kling-Gupta efficiency metrics will be added to compare several model performances in Section 3.3 of the revised manuscript.

5. 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 was 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.

Proposed changes to manuscript: More details about multi-step forecasting will be shown in Section 3.4 of the revised manuscript.

6. 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 model, we refer to (Fienen et al., 2018; Friedman, 2001; Pedregosa et al., 2011) to inform our choices of hyperparameter distributions by placing greater probability where we think the best values are. It can be difficult to figure out the interaction between hyperparameters. In cases where we aren't sure about the best values, and let the Bayesian algorithm do the reasoning for us.

Proposed changes to manuscript: We can use wide selected ranges of the model parameters have been justified in Section 4.2 of the revised manuscript.

7. 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 neurons is a difficult task (Badrzadeh et al., 2013) and the number of hidden neurons is determined by trial and error procedure in the original paper. In cases where we aren't sure about the best number of hidden neurons, we can 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 or 3.

Proposed changes to manuscript: The number of hidden neurons will be justified in Section 4.2 of the revised manuscript.

8. 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. Peak flow criterion, degree of agreement and Kling-Gupta efficiency metrics will be added to compare several model performances and more details about the discussion of the obtained results will be discussed.

Proposed changes to manuscript: The discussion of the obtained results will be enriched and some necessary citations to previous studies will be discussed.

9. 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.

Proposed changes to manuscript: We give the source link of the data in the revised manuscript.

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

Response: Thanks.

Proposed changes to manuscript: More details about ERA-Interim dataset have been introduced in the revised manuscript.

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

Response: Thank you for your suggestion.

Proposed changes to manuscript: The definitions and meanings of the variables in the ERA-Interim dataset will be given in Section "Appendix" of the revised manuscript.

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

Response: Thank you for your suggestion.

Proposed changes to manuscript: The "data scaling" has been replaced by "feature scaling".

13. 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.

Proposed changes to manuscript: The "invalid variables" will be modified to "weak-correlated variables" in the revised manuscript.

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

Response: Thank you for your careful review.

Proposed changes to manuscript: All "maximum" information criterion in the original manuscript will be modified to "maximal" information criterion in the revised manuscript.

15. 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.

Proposed changes to manuscript: The term MI*(D,X,Y) has been modified to MI*(D,x,y).

16. 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 $B = n^{0.6}$.

Proposed changes to manuscript: Some details about B(n) will be added in the revised manuscript.

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

Response: Thank you for your careful review.

Proposed changes to manuscript: R1(i,s) and R2(i,s) in Eq. (7) will be modified to R1(j,s) and R2(j,s) in the revised manuscript.

18. 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. The notations in Line 144 shows n features with N samples.

Proposed changes to manuscript: N/A

19. 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 will be added in the revised manuscript.

Proposed changes to manuscript: More details about ERA-Interim dataset will be given in Section "Appendix" of the revised manuscript.

20. 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).

Proposed changes to manuscript: The grid searching methodology will be introduced in detail in Section 4.2 of the revised manuscript.

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

Response: Thank you for your suggestion.

Proposed changes to manuscript: The "activation function" has been added in the revised manuscript.

22. 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 test set (2017-2018). We mainly discuss the trend of performance indices as the lead time increases.

Proposed changes to manuscript: The comments about the relationship between performance indices and lead times will be given more details in Section 4.3 and 4.4.

23. 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. 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 train set and validation set. We consider the possible reasons are parameter optimization and model structure. We have done a lot of numerical experiments in these days and the question will be discussed in Section 4 of the revised manuscript.

Proposed changes to manuscript: More discussion about why NSE values for lead times of 6-7-8-9-day is worse than the value of lead time of 10-day in the train set and validation set will be added in Section 4 of the revised manuscript.

24. 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 were used to determine the model input structures with inflow and rainfall. Another threshold value was used to determine the model input structures with ERA-Interim dataset. Further, we perform input selection in two steps via the maximal information coefficient (MIC). First, compute MIC value of each reanalysis factors and observed inflow. Then, sort features based on MIC in a descending order. And then, select the top k features whose mic are greater than or equal to the set threshold. The selected k features from reanalysis data are used as part of input to the model.

Proposed changes to manuscript: Consider the subjectivity of user-defined thresholds, the three threshold values will be modified by significance test and trail-and-error to determine input structures of model in Section 3.1 of the revised manuscript.