Reply to Anonymous Referee #1

Thank you very much for your time and for your thoughtful and constructive review. The following are our supplementary reply for a lot of research has been done to your comments.

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 select effective input factors accurately and quickly. According to your suggestion, we adjusted the selection procedure of reanalysis variables. We perform feature selection in two steps via MIC. First, compute MIC value of each reanalysis factors and observed inflow. Then, sort features based on MIC in a descending order and determine the optimum inputs using trail-and-error method, i.e. starting from the top one feature and then modifying the external input feature by successively adding one more feature into model input (Moosavi et al., 2013; Shoaib et al., 2015). Finally, the top 14 reanalysis variables are selected as the input (Table 1). More details will be given in the revised version.

Table 1: Selected reanalysis variables

No.	Variable	Index	Unit	MIC
1	Forecast albedo	fal_t	-	0.865
2	Soil temperature level 3	$stl3_t$	K	0.846
3	2 meter dewpoint temperature	$d2m_t$	K	0.781
4	Total column water vapour	$tcwv_t$	$kg \cdot m^{-2}$	0.699
5	Total column water	tcw_t	$kg \cdot m^{-2}$	0.699
6	Soil temperature level 2	$stl2_t$	K	0.689
7	Minimum temperature at 2 meters	$mn2t_t$	K	0.683
8	Surface thermal radiation downwards	$strd_t$	$J \cdot m^{-2}$	0.669
9	Temperature of snow layer	tsn_t	K	0.664
10	Soil temperature level 4	$stl4_t$	K	0.642

11	Soil temperature level 1	$stl1_t$	K	0.631
12	Surface net thermal radiation, clear sky	$strc_t$	$J \cdot m^{-2}$	0.620
13	Runoff	ro_t	m	0.619
14	Volumetric soil water layer 1	$swvl1_t$	$m^3 \cdot m^{-3}$	0.614

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. We use the partial autocorrelation function (PACF) and cross-correlation function (CCF) in these days for modeling, calculation and analysis in these days according to Referee (#3)'s suggestion. Figure 1 shows the PACF, CCF and the corresponding 95% confidence bands from lag 1 to lag 10.

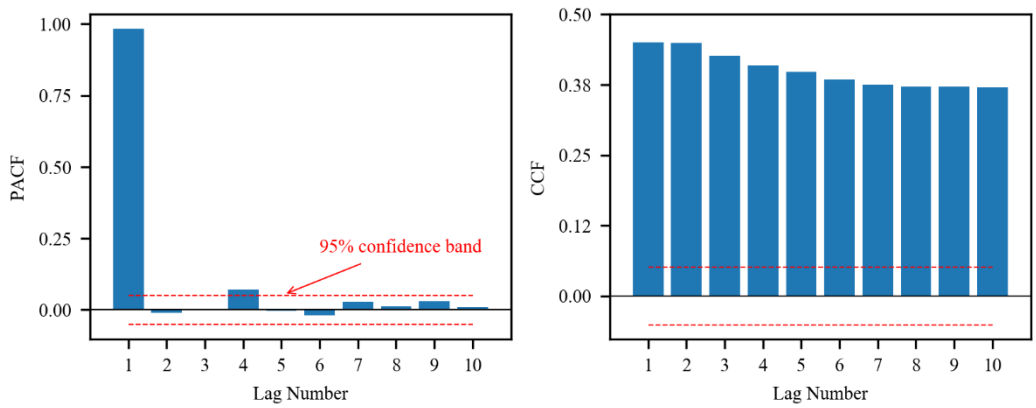

Figure 1: PACF of Xiaowan daily inflow and CCF of rainfall (2011-2014).

The PACF show significant autocorrelation at lag one and lag four, respectively. Therefore, one-day and four-day lag can be selected as input of the model. According to CCF of Xiaowan daily inflow and rainfall, ten-day lag all are significant. are selected as the input. And thus, trail-and-error method is used to determine the optimum inputs. The following five inputs are used as the model input successively.

1. Q_{t-1}, Q_{t-4}

2. $Q_{t-1}, Q_{t-4}, R_{t-1}$

3. $Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}$

4.$Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}$

5. $Q_{t-1}, Q_{t-4}, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}$

Finally, the fourth input is selected as the model input. More details will be given in the revised version.

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 square 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 good measurement of the average error. According to Referee (#2) and (#3)'s suggestions, 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. Kling–Gupta efficiency scores (KGE) (Knoben et al., 2019) is also a widely used evaluation index. It can be provided as following Eq. (1) and (2).

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

$$\mathrm{CORR} = \frac{\sum_{i=1}^{n}(Q_i - \bar{Q})\left(\widehat{Q_i} - \bar{\widehat{Q}}\right)}{\sqrt{\sum_{i=1}^{n}(Q_i - \bar{Q})^2}\sqrt{\sum_{i=1}^{n}\left(\widehat{Q_i} - \bar{\widehat{Q}}\right)^2}}, \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Q_i} - \bar{\widehat{Q}}\right)^2}, \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i - \bar{Q})^2} \quad (2)$$

where $\widehat{Q_i}$ and Q_i are the inflow estimation and observed value at time *i*, respectively and *n* is the number of samples. $\bar{\hat{Q}}$ is the mean of the estimation values. σ is the standard deviation of the observed values, $\hat{\sigma}$ is the standard deviation of the inflow estimation.

The percent bias in flow duration curve high-segment volume (BHV) (Yilmaz et al., 2008; Vogel and Fennessey, 1994) is used to evaluate performance of peak inflow forecasting. It can be provided as following Eq. (3).

$$BHV = \frac{\sum_{h=1}^{H}(\hat{Q}_h - Q_h)}{\sum_{h=1}^{H} Q_h} \times 100 \quad (3)$$

where h = 1, 2,. . .H are the flow indices for flows with exceedance probabilities lower than 0.02.

The Index of Agreement (IA) (Willmott, 1981) plays a significant role in evaluating the degree of the agreement between observed values and inflow estimation. It is given by Eq. (4).

$$IA = 1 - \frac{\sum_{i=1}^{n}(\widehat{Q_i} - Q_i)^2}{\sum_{i=1}^{n}(|\hat{Q}_i - \bar{Q}| + |Q_i - \bar{Q}|)^2} \quad (4)$$

More details will be given in the revised version.

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. The grid search is considered as an effective parameter search method, which is widely used (Fienen et al., 2018). We have performed some numerical experiments to compare grid search and randomized search and grid search can obtain more reasonable and stable hyperparameter combination. In addition, more wide range of parameters has been performed according to Referee (#3)'s suggestion. For artificial neural networks-maximal information coefficient (ANN-MIC), a range of 2-20 neurons and four activation functions are selected by grid searching. Table 2 shows results of parameter optimization of ANN-MIC. Table 3 and Table 4 show results of parameter optimization of support vector regression-maximal information coefficient (SVR-MIC) and gradient boosting regression trees-maximal information coefficient (GBRT-MIC). More details will be given in the revised version.

Table 2: Tuning parameters of ANN-MIC

Model	Tuning parameter	1	2	3	4	5	6	7	8	9	10
ANN-MIC	Structure Activate function	19-5-1 tanh	19-2-1 logistic	19-3-1 tanh	19-2-1 logistic	19-2-1 logistic	19-2-1 logistic	19-2-1 logistic	19-2-1 logistic	19-2-1 logistic	19-2-1 tanh

Table 3: Tuning parameters of SVR-MIC.

Model	Tuning parameter	Tuning range	1	2	3	4	5	6	7	8	9	10
SVR-MIC	C	**(1, 100, 20)**	8.7368	50.0000	29.3684	19.0526	21.6316	6.1579	16.4737	6.1579	8.7368	3.5789
	epsilon	**(0.001, 0.1, 20)**	0.00048	0.00953	0.00032	0.00011	0.00001	0.00001	0.00032	0.00022	0.00058	0.00001
	gamma	**(0.0001, 0.01, 20)**	0.0100	0.0095	0.0019	0.0072	0.0019	0.0067	0.0038	0.0057	0.0072	0.0067

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

Table 4: Tuning parameters of GBRT-MIC

Tuning parameter	Tuning range	Optimal parameters (the lead times of 1-10 days)	
		GBRT	GBRT-MIC
max_leaf_nodes	[2, 3, …, 20]	7,3,3,3,3,2,3,3,3,3	10,11,16,11,12,10,6,6,4,4
min_samples_leaf	[1, 2, …, 20]	2,3,1,1,10,1,2,2,4,1	5,9,1,5,6,9,4,6,6,10
max_depth	[1, 2, ..., 20]	3,2,2,4,2,1,2,2,2,10	4,4,8,7,10,6,6,4,5,3
min_samples_split	[2, 3, …, 20]	9,14,13,20,11,3,6,2,3,4	16,20,19,16,20,20,17,17,20,3
n_estimators	[500,550, …, 4000]	1000,1000,1000,1500,1500,2500,1500,3500,2500,2500	2500,1000,1500,2500,1500,2500,1000,1000,1500,300
learning_rate	[0.001,0.0025,0.005,0.0075,0.01,0.025,0.05,0.075,0.1]	0.01,0.01,0.01,0.005,0.005,0.005,0.005,0.0025,0.0025,0.0025	0.0075,0.01,0.01,0.0025,0.01,0.0025,0.01,0.01,0.01,0.01

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. Bayesian optimization (Snoek et al., 2012) is proved as an effective parameter search method, especially for wide domain space. According to your suggestion, grid search has been used to replace Bayesian optimization for optimize parameters of SVR-MIC. The grid search spends much more time for hyperparameter optimization but and the result of optimization are more stable. And thus, grid search has been used for optimize parameters of SVR-MIC. In addition, for time consuming of grid search, will it is our next research direction to use heuristic algorithms, such as particle swarm optimization, genetic algorithm and gray wolf algorithm to optimize model parameters. More details will be given in the revised version.