Reply to Anonymous Referee #3

We thank you very much for reviewing the manuscript. The following are our supplementary reply for a lot of research has been done to your comments.

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 three significant measures. Firstly, ERA-Interim reanalysis data are 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 author to grasp the innovation of this paper, we will modify the "Abstract" and "Introduction" carefully to make the innovation more prominent. More details will be given in the revised version.

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. 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 your suggestion. Figure 1 shows the PACF, CCF and the corresponding 95% confidence bands from lag 1 to lag 10.

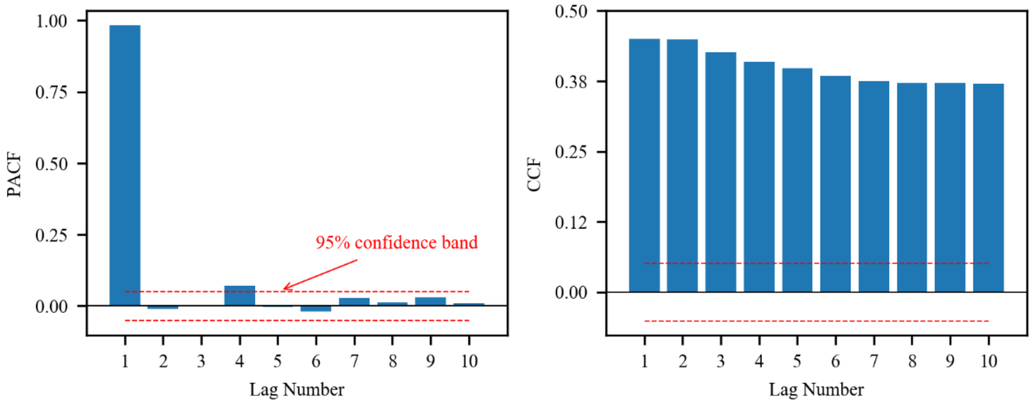

Figure 1: PACF of Xiaowan daily inflow and CCF of rainfall (2011-2014).

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.

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. Revised input structures are used to compare with developed model with ERA-Interim dataset. Table 1 shows performance indices of model in the test set. The experimental results indicate that the developed method generally performs better than other models and improves the accuracy of inflow forecasting about 1% and 14% in two-day and 10-day ahead forecasting. Especially for 5-10 day lead times, GBRT-MIC could be used for more accurate and reliable inflow forecasting. More details will be given in the revised version.

Table 1: Performance indices of the test set.

Indice	Model	1	2	3	4	5	6	7	8	9	10
MAE	GBRT-MIC	141	**158**	**174**	**181**	**188**	**191**	**194**	**201**	**208**	**210**
	SVR-MIC	132	161	183	191	212	219	226	231	235	239
	ANN-MIC	**132**	163	184	197	213	224	228	233	240	244
	MLR-MIC	136	170	191	209	226	240	243	249	254	258
RMSE	GBRT-MIC	219	**245**	**271**	**284**	**297**	**304**	**311**	**321**	**332**	**335**
	SVR-MIC	200	256	305	332	362	375	392	400	409	413
	ANN-MIC	**198**	254	292	317	338	357	371	385	398	403
	MLR-MIC	203	263	306	338	361	382	394	402	412	417
CORR	GBRT-MIC	0.9701	**0.9620**	**0.9539**	**0.9492**	**0.9445**	**0.9416**	**0.9386**	**0.9344**	**0.9302**	**0.9284**
	SVR-MIC	0.9753	0.9596	0.9438	0.9340	0.9214	0.9136	0.9055	0.9014	0.8973	0.8959
	ANN-MIC	**0.9756**	0.9596	0.9462	0.9363	0.9272	0.9184	0.9114	0.9045	0.8975	0.8948
	MLR-MIC	0.9744	0.9564	0.9407	0.9273	0.9164	0.9061	0.8998	0.8953	0.8900	0.8870
KGE	GBRT-MIC	0.9493	**0.9382**	**0.9270**	**0.9230**	**0.9189**	**0.9155**	**0.9121**	**0.9101**	**0.9081**	**0.9045**
	SVR-MIC	0.9506	0.9190	0.8700	0.8434	0.8165	0.8201	0.8053	0.7955	0.7817	0.7751
	ANN-MIC	**0.9631**	0.9375	0.9154	0.9021	0.8891	0.8791	0.8703	0.8635	0.8599	0.8611

	MLR-MIC	0.9619	0.9318	0.9062	0.8872	0.8701	0.8551	0.8430	0.8344	0.8260	0.8188
BHV	GBRT-MIC	**-0.1909**	**-0.1909**	**-0.1909**	**-0.1909**	**-0.1909**	**0.0886**	**0.3681**	**0.2720**	**0.1759**	**0.2720**
	SVR-MIC	-1.6396	-3.3890	-6.4785	-7.7173	-9.3569	-8.1066	-9.0180	-9.4042	-9.9785	-10.7111
	ANN-MIC	-0.2509	-0.7876	-1.2337	-1.5023	-1.6509	-1.8062	-2.7456	-3.1596	-3.1661	-2.8261
	MLR-MIC	-0.6867	-2.0428	-2.9254	-3.8346	-4.1555	-4.4089	-5.7323	-5.7912	-5.8660	-6.4303
ia	GBRT-MIC	0.9844	**0.9800**	**0.9756**	**0.9731**	**0.9705**	**0.9688**	**0.9671**	**0.9651**	**0.9631**	**0.9619**
	SVR-MIC	0.9870	0.9780	0.9670	0.9598	0.9506	0.9473	0.9414	0.9381	0.9342	0.9325
	ANN-MIC	**0.9874**	0.9788	0.9713	0.9657	0.9603	0.9552	0.9511	0.9473	0.9435	0.9424
	MLR-MIC	0.9868	0.9770	0.9680	0.9603	0.9537	0.9474	0.9432	0.9402	0.9367	0.9345

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

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 (see Question 2). More details will be given in the revised version.

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

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

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

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

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

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).

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

More details will be given in the revised version.

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. grid search is considered as an effective parameter search method, which is widely used (Fienen et al., 2018). In addition, more wide range of parameters has been performed according to your 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). However, comparing with selected ranges of the model parameters in the original manuscript, the new optimization of max_leaf_nodes, min_samples_leaf, max_depth and min_samples_split of GBRT-MIC generates 36100 models which spends about 200 minutes using 12 cores for parallel computing in each leadtime. 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 and scikit-learn package (Pedregosa et al., 2011). How to decrease time consuming of grid search will be our next research direction.

Table 2: Tuning parameters of ANN-MIC

Model	Tuning parameter	1	2	3	4	5	6	7	8	9	10
ANN-MIC	Structure	19-5-1	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
	Activate function	tanh	logistic	tanh	logistic	logistic	logistic	logistic	logistic	logistic	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 $[min + \frac{max-min}{step-1} \times 0, min + \frac{max-min}{step-1} \times 1, \ldots, min + \frac{max-min}{step-1} \times (step-1)]$.

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, …, 10]	2,3,1,1,10,1,2,2,4,1	5,9,1,5,6,9,4,6,6,10
max_depth	[1, 2, ..., 10]	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

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, 3 or 5 (see Table 4). More details will be given in the revised version.

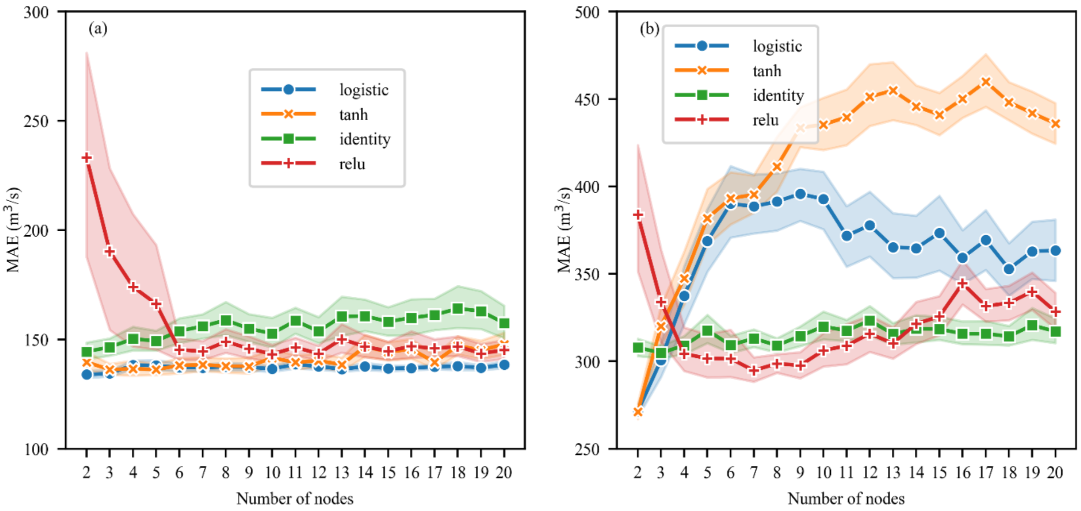

Figure 2: Sensitivity of the number of nodes and activation function in the hidden layer on the MAE of ANN-MIC, the shadow part is 95% confidence interval obtained by bootstrap of 50 trials. (a) One-day-ahead (b) Ten-day-ahead.

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. According to Referee (#2)’s and your suggestions, NSE is removed and KGE, BHV and IA are introduced as supplements. Table 1 and Figure 3 shows performance indices of model in the test set. More details will be given in the revised version.

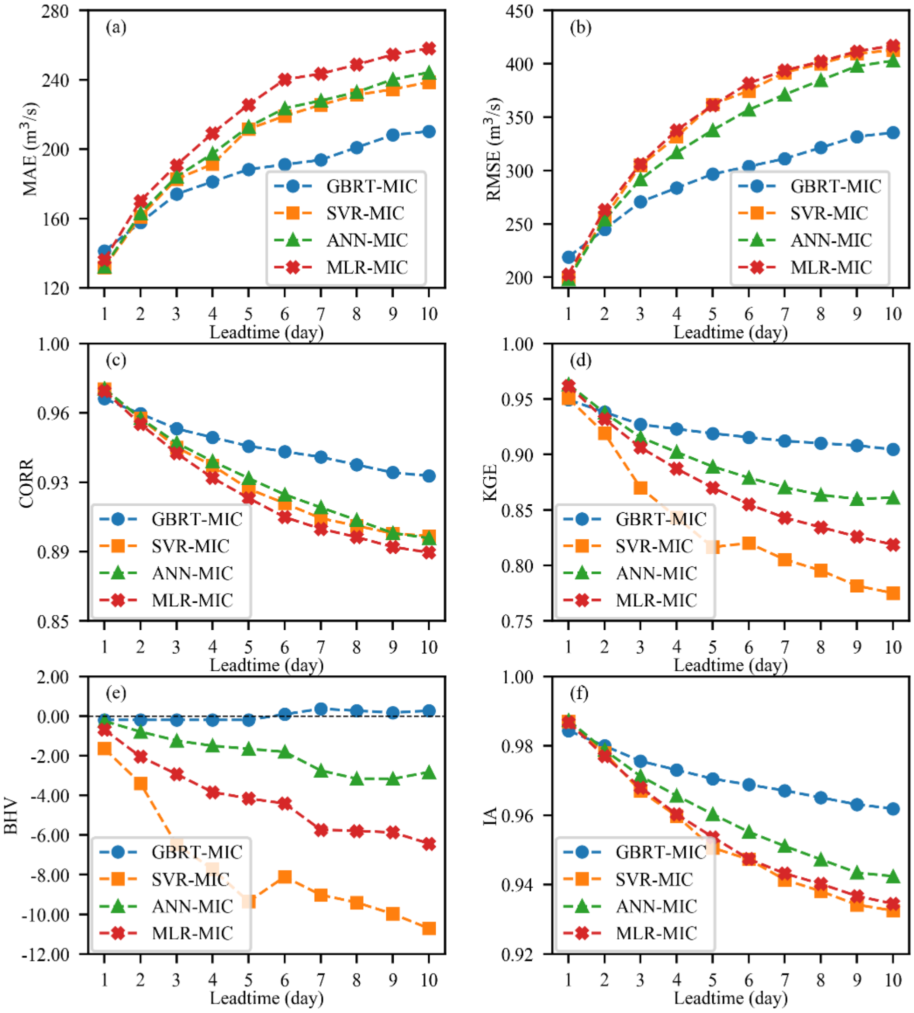

Figure 3: Performance of GBRT-MIC, SVR-MIC, ANN-MIC and MLR-MIC for the test set (2017-2018). (a) MAE (b) RMSE (c) CORR (d) KGE (e) BHV(f)IA.

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 (See Question 1 for details). 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 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 5, No.6-19). More details will be given in the revised version.

Table 5: List of inputs of GBRT-MIC. There are of two types, observed and reanalysis variables. The reanalysis variables are available four time a day at 00:00 UTC, 06:00 UTC, 12:00 UTC and 18:00 UTC. The cumulative variable (e.g., Total column water) is the sum of four periods and the instantaneous variable (e.g. 2 meter dewpoint temperature) is the mean of four periods.

No.	Variable	Index	Unit	MIC	Type
1	Inflow at day t – 1	Q_{t-1}	$m^3 \cdot s^{-1}$	-	Obs.
2	Inflow at day t – 2	Q_{t-4}	$m^3 \cdot s^{-1}$	-	Obs.
3	Rainfall at day t - 1	R_{t-1}	mm	-	Obs.
4	Rainfall at day t – 2	R_{t-2}	mm	-	Obs.
5	Rainfall at day t – 3	R_{t-3}	mm	-	Obs.
6	Forecast albedo	fal_t	-	0.865	ERA-I
7	Soil temperature level 3	$stl3_t$	K	0.846	ERA-I
8	2 meter dewpoint temperature	$d2m_t$	K	0.781	ERA-I
9	Total column water vapour	$tcwv_t$	$kg \cdot m^{-2}$	0.699	ERA-I
10	Total column water	tcw_t	$kg \cdot m^{-2}$	0.699	ERA-I
11	Soil temperature level 2	$stl2_t$	K	0.689	ERA-I
12	Minimum temperature at 2 meters	$mn2t_t$	K	0.683	ERA-I
13	Surface thermal radiation downwards	$strd_t$	$J \cdot m^{-2}$	0.669	ERA-I
14	Temperature of snow layer	tsn_t	K	0.664	ERA-I
15	Soil temperature level 4	$stl4_t$	K	0.642	ERA-I
16	Soil temperature level 1	$stl1_t$	K	0.631	ERA-I
17	Surface net thermal radiation, clear sky	$strc_t$	$J \cdot m^{-2}$	0.620	ERA-I
18	Runoff	ro_t	m	0.619	ERA-I
19	Volumetric soil water layer 1	$swvl1_t$	$m^3 \cdot m^{-3}$	0.614	ERA-I