An Auto-Local-Global Ensemble Correlation Model for Long-term Runoff Forecasting

Teng Zhang   Zhongjing Wang   Zixiong Zhang

Department of Hydraulic Engineering, Tsinghua University, Beijing, 100084, China

EGU General Assembly 2020

EGU2020-1369

1. Introduction

Long-term high-precision runoff prediction is of considerable significance to the water resources planning and management and benefit to regional sustainable development. With the global climate change, the water cycle is not only influenced by local meteorological elements but also global climate factors increasingly. To screen out the sensitive forecasting factors from various potential influences both in statistical issues and physical bodies, and to introduce the appropriate regression method, both traditional regression and data mining are the challenges.

Owing to the effect of climate change, the trend and periodic features of water cycle elements such as precipitation and runoff have changed inconsistent with historical series, becoming non-linear and non-stationary variables. The runoff is estimated generally based on local hydrological and/or meteorological variables, such as precipitation and potential evapotranspiration (Archer and Fowler 2008, Singh and Sankarasubramanian 2014). Researchers also predicted the runoff using the historical runoff by autoregression relationship (Ghorbani et al. 2016, Tan et al. 2018). However, either the historical runoff series or local meteorological data could not explain the non-linearity or non-stationarity. Many researches proved that climate indicators were global-correlated to hydrological variables strongly and coherently, and climate signals can improve the forecast skill of the model (Lee and Julien 2016). Thus, the influence of climate indicators can not be understated.

The pairwise combination forecasting of the above three methods were applied in runoff predicting. However, runoff as a periodic natural event, the historical flow will affect the future drainage obviously, and the autoregression should be considered in the forecasting. Thus, a principle of ALGEC was put forward, and the delay effect of predictors in the lag of 0-11 months is considered, which is the focus and innovation of this paper. Additionally, considering different key physical mechanisms of runoff in different months, especially in the arid region, we predict the runoff month by month separately.

3. Methodology

3.1 Machine Learning Model

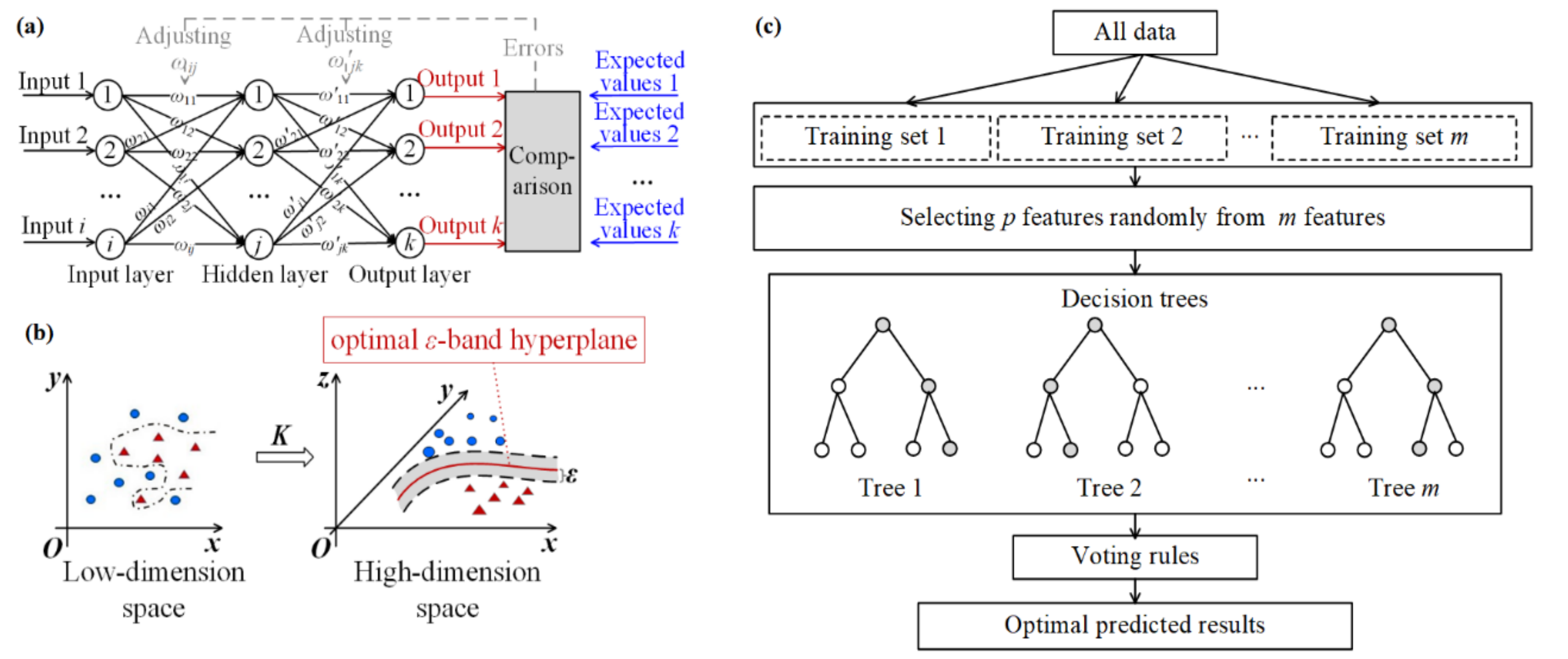


Fig. The schematic diagrams of BP (a), SVM (b), and RF (c)

3.2 Criteria of prediction accuracy

In this paper, correlation coefficient (R) and mean absolute percentage error (MAPE) are used to evaluate the model performance.

$$R=\frac{\sum_{i=1}^{n}(O_i^{'}-O_{ave}^{'})(O_i-O_{ave})}{\sqrt{\sum_{i=1}^{n}(O_i^{'}-O_{ave}^{'})^2\sum_{i=1}^{n}(O_i-O_{ave})^2}}$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{O_i^{'}-O_i}{O_i}\right|\times 100\%$$

where O_i is the measured value and $O^{'}_{ave}$ is the average of O_i; $O^{'}_i$ is the predicted value and $O^{'}_{ave}$ is the average of $O^{'}_i$; subscript 1 and 2 are the training and validation set.

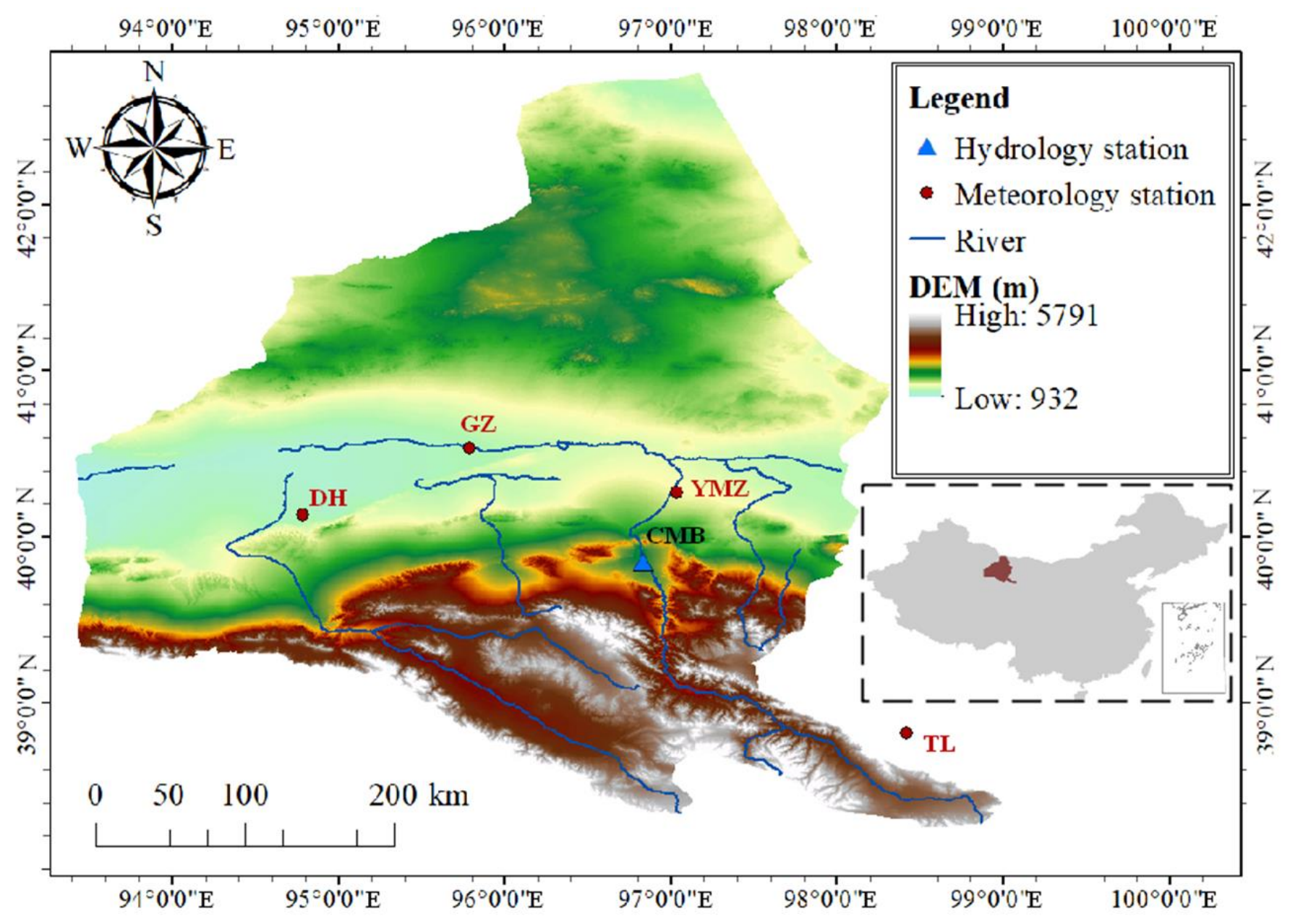
In order to judge the performance of the model more conveniently and intuitively when considering two or more indexes comprehensively, an indicator β is constructed to convert the objectives to an integration indicator

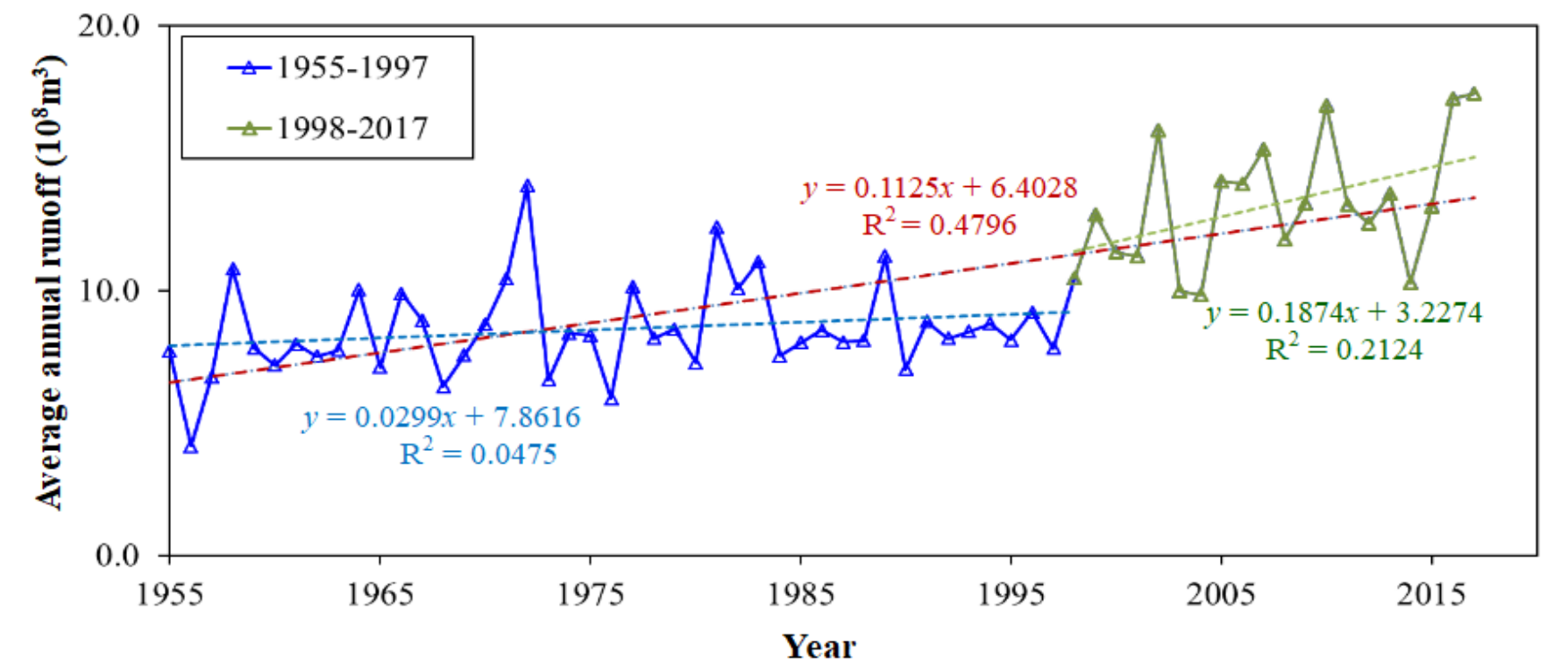
$$\beta=\frac{1}{2}(R_1^2(1-MAPE_1)+R_2^2(1-MAPE_2))$$

The R is ranged [-1,1] and $MAPE$ is [0，∞). The perfect absolute value of R is 1 and of $MAPE$ is 0, so the optimal value of β is 1. The closer the β is to 1, the better the model predicts.

2. Study area and data

2.1 Case area and data source





Shule River Basin

- length of 670 km
- area of 39,497 km^2
- altitude from 932 - 5,791 m
- annual precipitation is 30-60 mm
- annual evaporation is 1,500-2,700 mm
- Runoff replenishment
  - precipitation (39-55%)
  - shallow groundwater (19-29%)
  - glacial and snowmelt water (26-32%)

Monthly runoff data

- Jan. 1955- Dec. 2017
- CMB station
- Target & auto-correlation analysis

Local meteorological data

- Jan. 1954- Dec. 2017
- Guazhou (GZ), Tuole (TL), Yumenzhen (YMZ), and Dunhuang (DH) station
- Inputs & local-correlation analysis

Global climate data

- Jan. 1954- Dec. 2017
- Climate indexes
- Inputs & global-correlation analysis

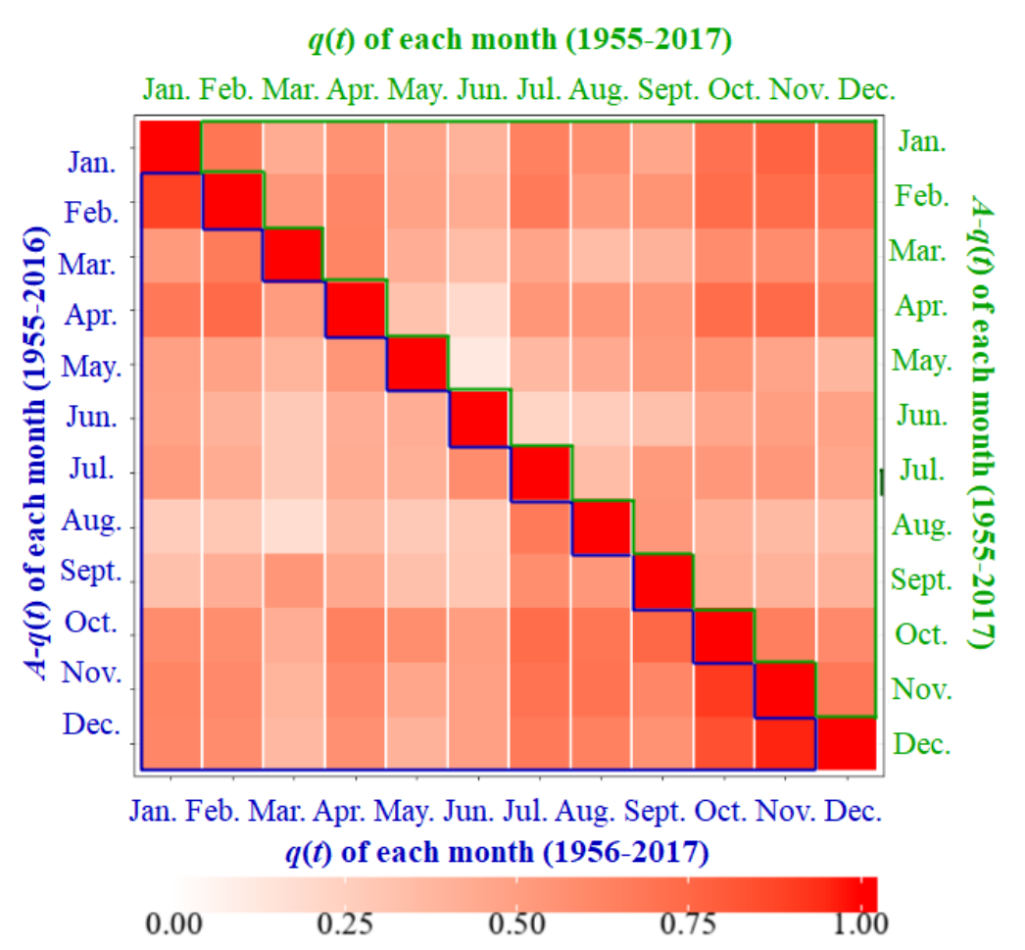
2.2 Ensemble and Screening Impact factors

***q*(*t*) of the dry season**

- highly correlated with *A-q*(*t-i*)
- *R* reaches 0.9 in Nov. & Dec.

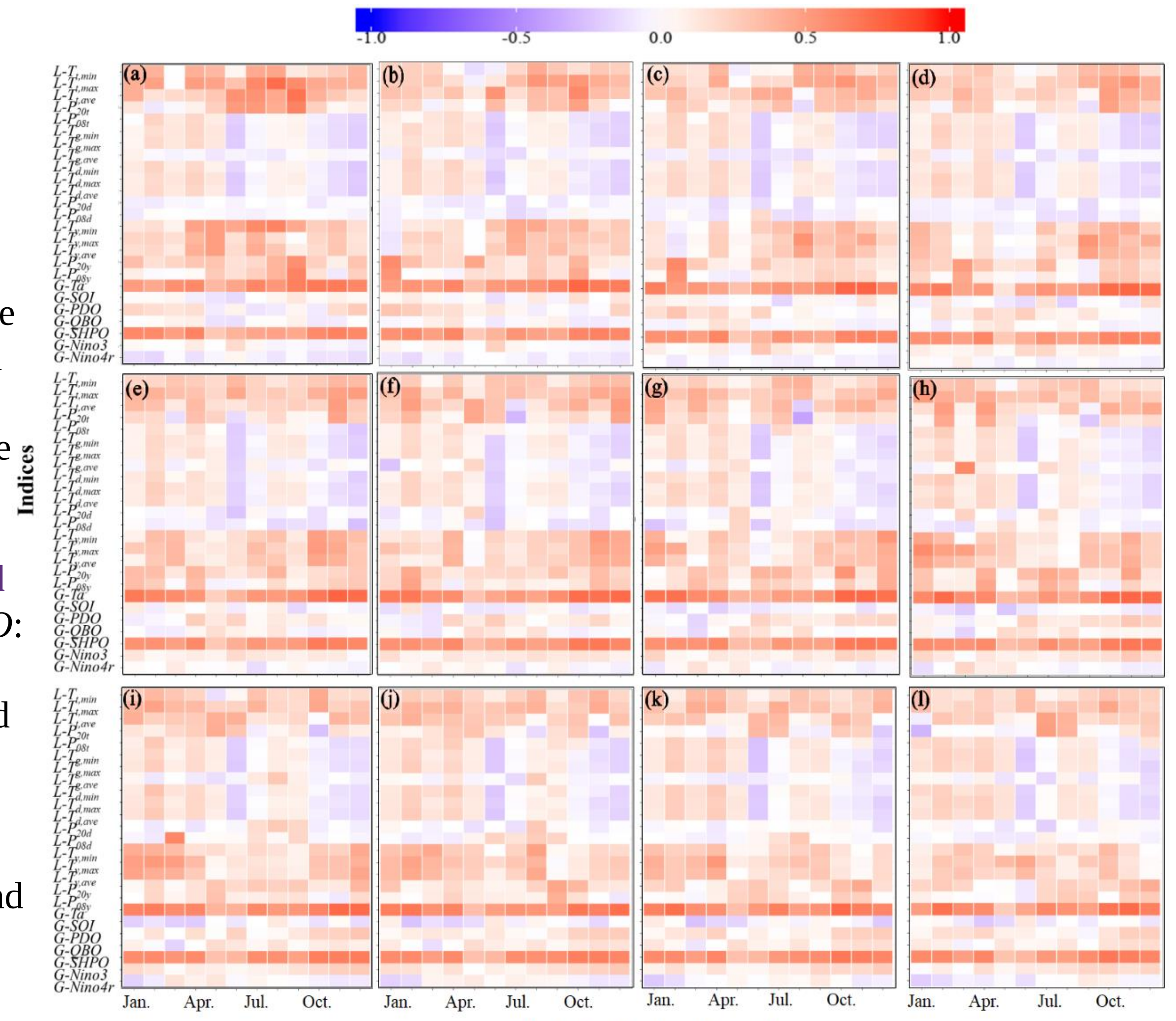
***q*(*t*) of the wet season**

- weakly affected by *A-q*(*t-i*)
- *R*=~0.5
- *q*(*t*) in the early months largely determines that in the later months, especially in months with low external replenishment
- *q*(*t*) has memory on the soil and groundwater storage



***R* between *q*(*t*) and**

- *G-Ta* or *G-SHPO*: > 0.5 (high)
- *T* and *P* of TL and YMZ: >0.5 in some months (considerable)
- *T* and *P* of GZ and DH: <0.4 (low)

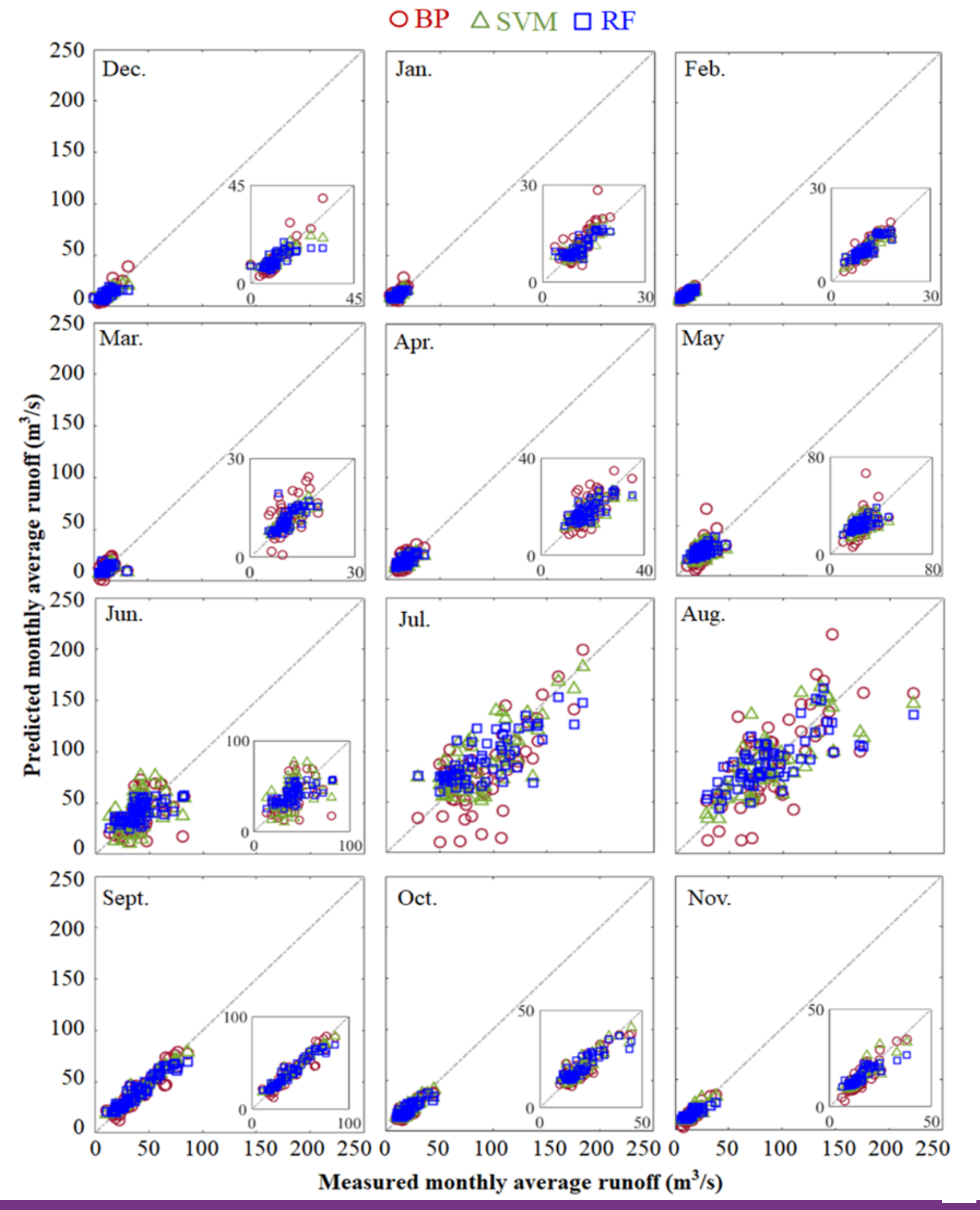


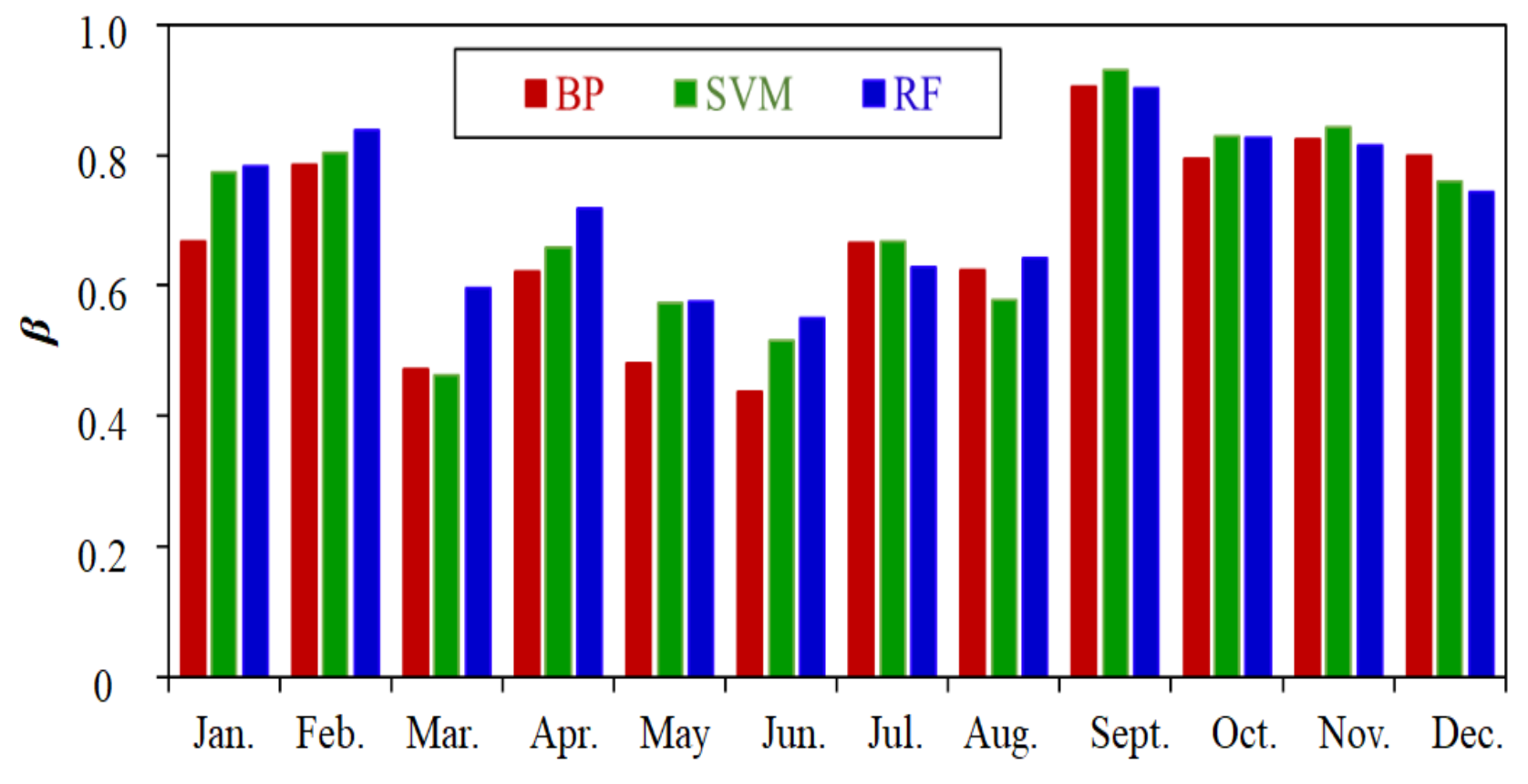
4. Result and Discussion

4.1 Result and comparison

The monthly runoff series of CMB is forecasted by the ALGEC model with BP, SVM, and RF algorithms using different prediction factor sets for each month, respectively. When establishing the model, k-fold cross-validation method is used, and the model performs well when k equals to 5. Thus, all samples are divided into three sets randomly, and each time, only one set is selected for verification, and the other two for model training. The model is trained and validated for five times.

In general, RF looks better than BP and SVM. The results suggest that the prediction models have a good simulation on the whole, especially in winter (Dec. to Feb.) and autumn (Sept. to Nov.). In other months, the simulation has a little worse.

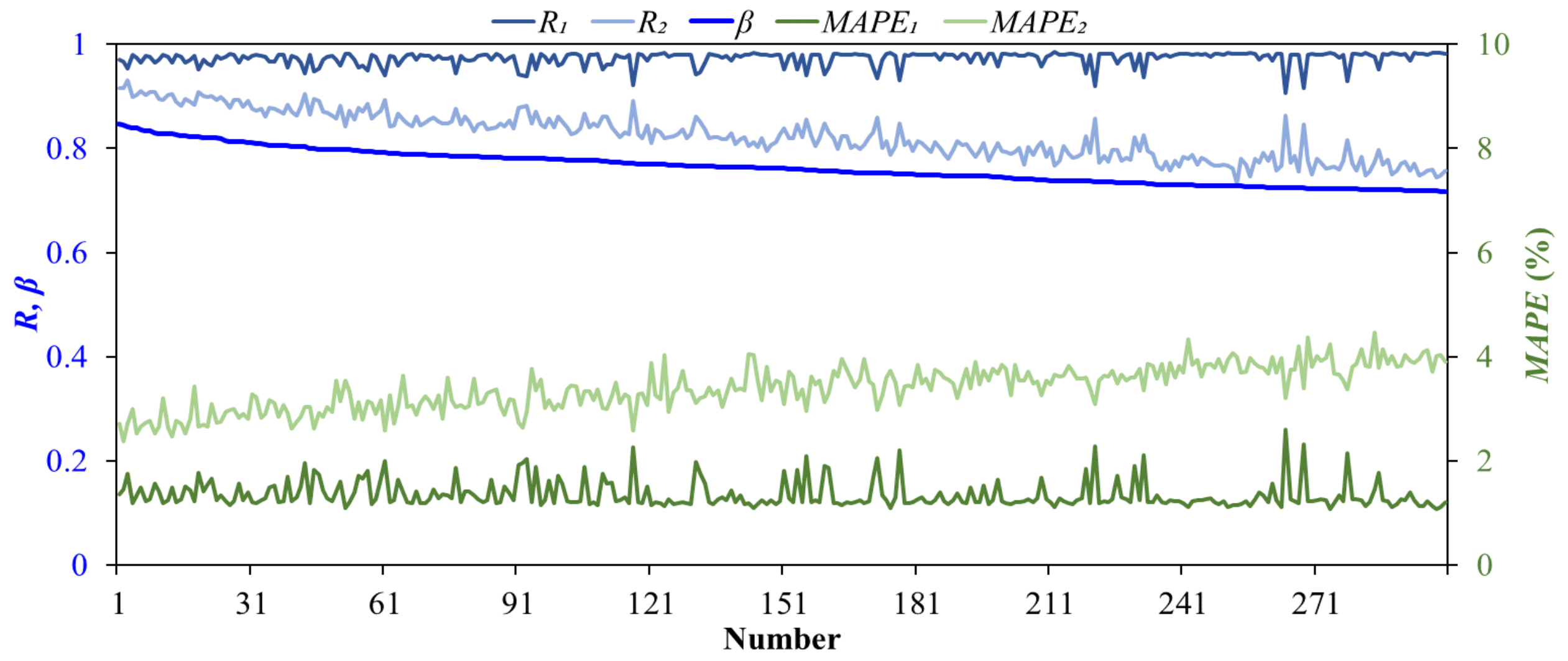




Among the three algorithms, BP has the lowest accuracy, and RF has the highest. Although machine learning algorithms are good at solving non-linear problems, they are deeply condition depended. BP is good at big data problems, but easy falls into local optimum. SVM is good at high-dimension but small-sample problem, so modeling is simple and fast. RF is simple, computes fast, has strong generalization ability and has advantages in processing high-dimension data and feature-missing data.

For monthly forecasting, the results of the dry season are better and more accurate than those of the wet season, especially May and June. In May and early June, the glacier begins to melt and the agriculture begins to irrigate, causing a large artificial influence increasing the uncertainty and instability to the runoff. The runoff of July and August is mainly affected by the precipitation, and the rainfall brings more uncertainty to runoff. For the dry season, the uncertainty of the runoff is low, so the prediction accuracy is high. In December, the runoff almost reduces to 0, so the MAPE of runoff is large, so the β is lower than other months of dry season.

4.2 Performance of the integration index β



- When R is closer to 1 and $MAPE$ is closer to 0 → β is higher
- When R_1 and R_2 are both high →$MAPE_1$ and $MAPE_2$ is low → β is high
- When R_1 is constant:
  - larger R_2 → larger β, better model
  - The smaller the difference of R_1 and R_2 → more stable and reliable the model

***β*: Reasonable to judge the model prediction ability and stability quickly and quantitively instead of selecting the best model subjectively**

4.3 Model generalization and improvement

- When the lag$_{min}$=0→ interpolation and extension of missed historical data
- When the lag$_{min}$=i，i≠0→ leading time = i year
- For other bigger region, more data of different meteorological stations and global climate factors could be analyzed to choose the prediction factors.
- Not enough discussion about the model performance and optimum effect of increasing the number of inputs → the relationship of the model improvement and the computational time and resource consume could be studied
- Further study for the reasons of the difference of the prediction factors of the runoff of each month and the whole runoff series

Archer, D.R. and Fowler, H.J. (2008) Using meteorological data to forecast seasonal runoff on the River Jhelum, Pakistan. Journal of Hydrology 361(1-2), 10-23.
Ghorbani, M.A., Zadeh, H.A., Isazadeh, M. and Terzi, O. (2016) A comparative study of artificial neural network (MLP, RBF) and support vector machine models for river flow prediction. Environmental Earth Sciences 75(6).
Lee, J.H. and Julien, P.Y. (2016) Teleconnections of the ENSO and South Korean precipitation patterns. Journal of Hydrology 534, 237-250
Singh, H. and Sankarasubramanian, A. (2014) Systematic uncertainty reduction strategies for developing streamflow forecasts utilizing multiple climate models and hydrologic models. Water Resources Research 50(2), 1288-1307
Tan, Q.-F., Lei, X.-H., Wang, X., Wang, H., Wen, X., Ji, Y. and Kang, A.-Q. (2018) An adaptive middle and long-term runoff forecast model using EEMD-ANN hybrid approach. Journal of Hydrology 567, 767-780.

More information: Please contact Teng Zhang
E-mail: t-zhang18@mails.tsinghua.edu.cn