Machine learning-based dynamical seasonal prediction of summer rainfall in China

Jialin WANG[1] Jing YANG*[1] Hongli REN[2] Jinxiao LI[3] Qing BAO[3] Miaoni GAO[4]

[1] State Key Laboratory of Earth Surface Process and Resource Ecology / Key Laboratory of Environmental Change and Natural Disaster, Faculty of Geographical Science, Beijing Normal University, Beijing, China

[2] Laboratory for Climate Studies and CMA-NJU Joint Laboratory for Climate Prediction Studies, National Climate Center, China Meteorological Administration, Beijing, China

[3] State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

[4] Institute for Disaster Risk Management, School of Geographic Sciences, Nanjing University of Information Science and Technology, Nanjing, China

- Contact: yangjing@bnu.edu.cn

1. Introduction

- **Seasonal prediction of precipitation is rather hard in current GCMs**
- **Statistical correction methods including machine learning have severe over-fitting problem**

⇒

Objectives

- **To develop a Machine-Learning Dynamical(MLD) method considering over-fitting**
- **To select optimum ML method with optimum hyper-parameter to predict seasonal rainfall *independently***

2. Data and Methods

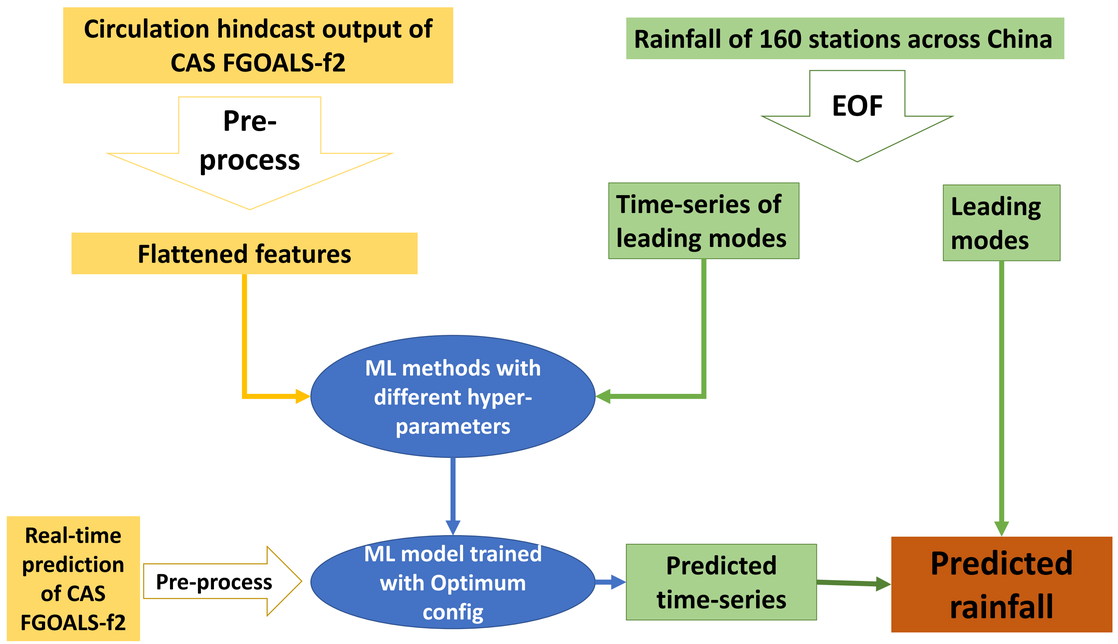

Fig1. Basic roadmap of MLD modelling and independent prediction

◆ **Selected Output variables of CAS FGOALS-f2**
- **SLP, U850, Z500, U200, and T200**
- **Variables are not used at same time, combination of them are tested**

◆ **ML methods tested**
- **ε-sensitive Support Vector Regression(SVR)**
- **Random Forest(RF)**
- **Gradient Boosting Regression Trees(GBRT)**

◆ **Cross validation and independent prediction**
- **For time-series validation, R^2 scores are used as metric**
- **To validate every ML model, K-Fold cross validation is used**
- **For rainfall validation, Pattern Correlation Coefficient of 160 stations is used**
- **1981-2010 used for EOF analysis, training and cross validation; 2011-2019 used for independent prediction**

3. Results

For each method tested, models with combinations of 31 combinations of input variables and thousands of combinations of hyper-parameters (grid searching) are trained and cross-validated via R^2 scores. Config with the best fitting and least over-fitting are selected as optimum.

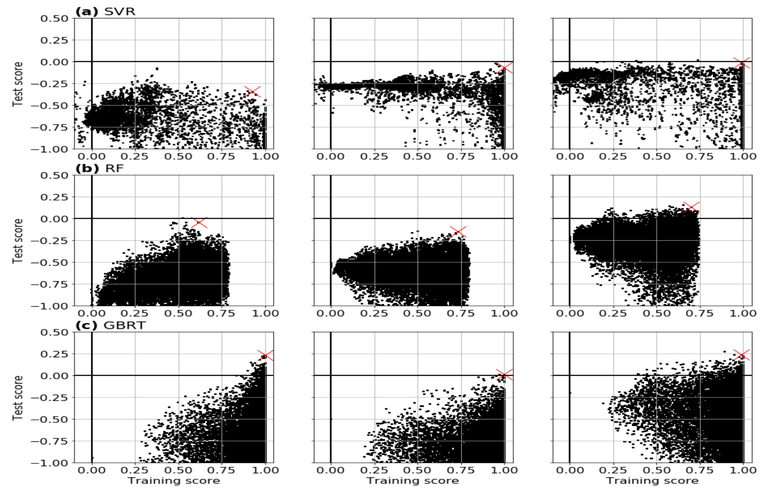

Fig 2. Mean training scores and test scores during cross validation with different hyperparameters and different circulation variables (31 kinds of variable combinations) for (a) SVR, (b) RF and (c) GBRT.

Rainfall prediction scores with optimum config of 3 methods comparing with dynamical model ensemble and multimodel ensemble. For independent years, GBRT+FGOALS-f2 has a great lead comparing to dynamical prediction and other 2 MLD methods

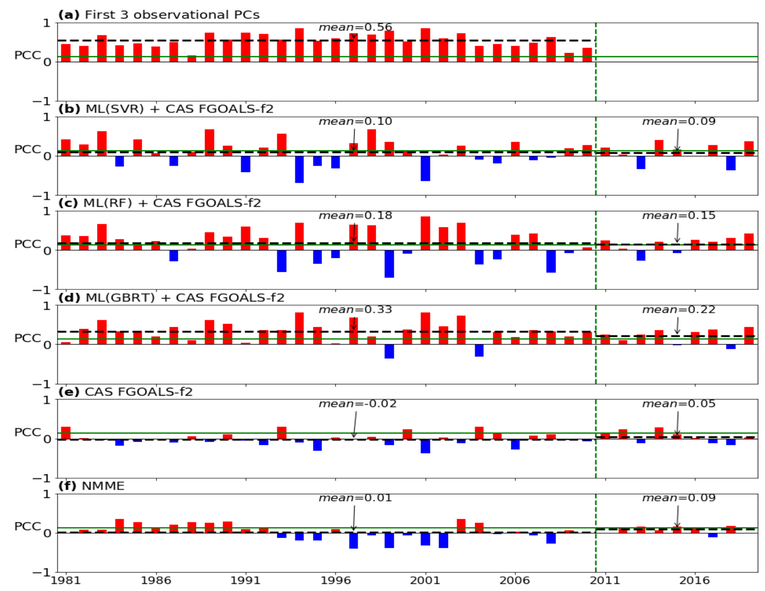

Fig 3. Pattern correlation coefficient (PCC) between (a) reconstructed precipitation anomaly using the first 3 observational PCs, (b-d) the reconstructed precipitation anomaly using three types of MLD predicted PCs, (e) the CAS FGOALS-f2 ensemble precipitation prediction, and (f) the NMME ensemble precipitation prediction and the observed summer precipitation anomaly in China. The horizontal solid green line denotes the 90% confidence level.

4. Discussions and conclusions

To determine if reducing the overfitting to the largest extent possible is required for MLD, GBRT is used as example. Group 1 (less overfitting) and Group 2 (more over-fitting) are compared in Fig 4.

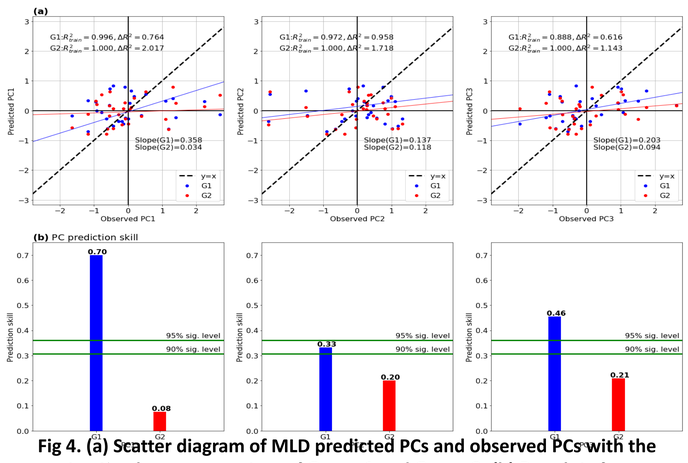

Fig 4. (a) Scatter diagram of MLD predicted PCs and observed PCs with the contrasting hyperparameters of Group 1 and Group 2, (b) Bar plots for prediction skill (correlation coefficient) in the historical reference period for the first three PCs in Groups 1 and 2.

To investigate whether dynamical prediction skill influence the MLD prediction results in individual years, the relationship of the prediction skills between the MLD and its corresponding dynamical circulation variable in 21 rolling 9-year epoch windows are evaluated

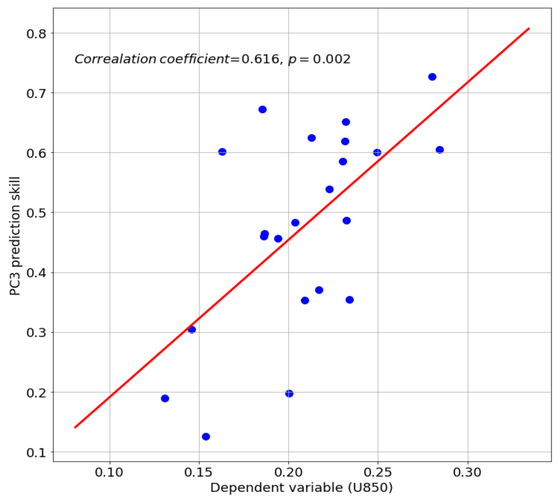

Fig 5. Scatter diagram representing the relationship between the prediction skill (measured by TCC) of the MLD-predicted PC (PC3) and the dynamical prediction skill of the dependent variable (U850, measured by the multiyear mean PCC). The rolling window is 9 years, and the total number of rolling windows is 21.

Conclusions:

- **MLD could be an efficient method to improve the current dynamical prediction.**
- **Reducing overfitting and using the best dynamical prediction are imperative in MLD application prospects**