Machine learning to optimize climate projection over China with multi-model ensemble simulations

Tong Li[1], Zhihong Jiang[5,2], Hervé Le Treut[3], Laurent Li[3], Lilong Zhao[1] and Lingling Ge[4]

Published 27 August 2021 •

Environmental Research Letters, Volume 16, Number 9

Citation Tong Li *et al* 2021 *Environ. Res. Lett.* **16** 094028

DOI 10.1088/1748-9326/ac1d0c

Multi-model ensemble is considered as the best way to explore the advantage and to avoid the weakness of individual models, and ultimately to achieve the best climate simulation. But the **design of an optimal strategy** and its practical **implementation** are both a challenging issue.

Laurent Li
Laboratoire de Météorologie Dynamique (LMD)
IPSL/CNRS, Sorbonne Université, Paris, France
Ecole Normale Supérieure, Ecole Polytechnique

Collaborative work with NUIST (Nanjing Univ. of Information Sci. and Tech.): Jiang Zhihong, Li Tong

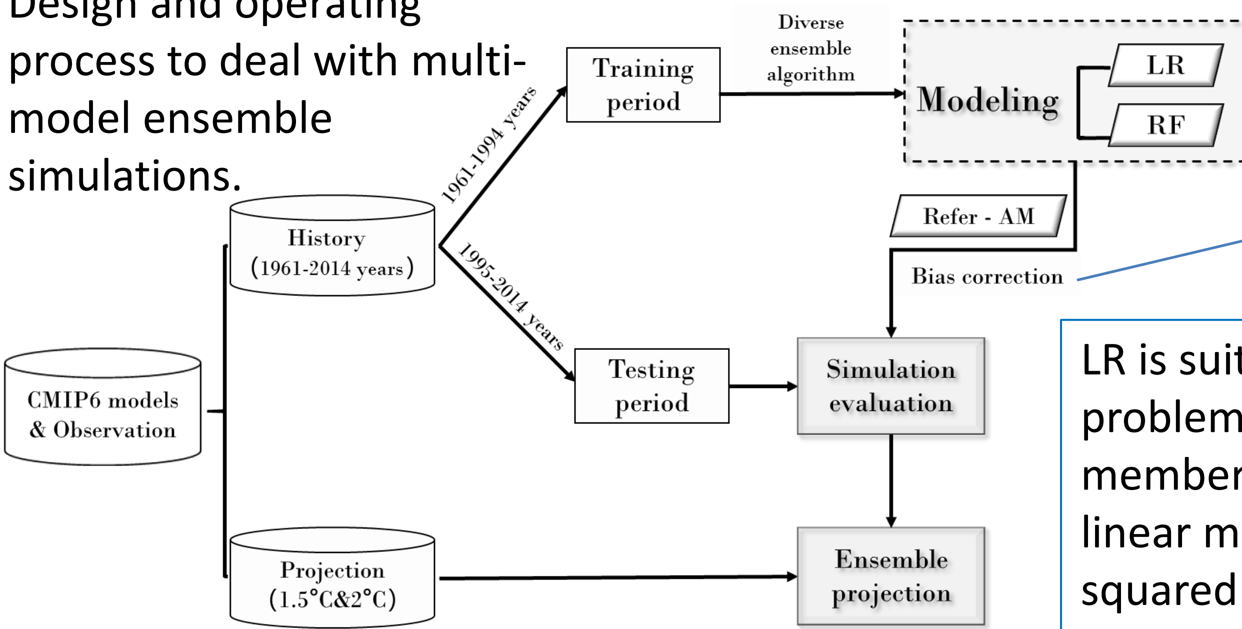
- ✓ We use the **Random Forest (RF)** algorithm to explore the information offered by the multi-model ensemble simulations of CMIP6. Our objective is to achieve a more reliable climate projection (mean climate and extremes) over China.
- ✓ RF is furthermore compared to two other ensemble-processing strategies of different nature, one is the basic **arithmetic mean (AM)**, and another is the **linear regression (LR)** across the ensemble members.

Number	Model Name	Modeling Center/ Country	Reso (lat×lon)
1 2	ACCESS-CM2 ACCESS-ESM1-5	Commonwealth Scientific and Industrial Research Organisation /Australia	1.25°×1.875° 1.25°×1.875°
3	BCC-CSM2-MR	Beijing Climate Center China Meteo. Administration /China	1.125°×1.125°
4	CanESM5	Canadian Centre for Climate Modelling and Analysis /Canada	2.8°×2.8°
5 6	CNRM-CM6-1 CNRM-ESM2-1	Centre National de Recherches Météorologiques–Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique /France	1.4°×1.4° 1.4°×1.4°
7 8	EC-Earth3 EC-Earth3-Veg	EC-EARTH consortium	0.7°×0.7° 0.7°×0.7°
9	FGOALS-g3	Chinese Academy of Sciences /China	2.25°×2°
10 11	GFDL-CM4 GFDL-ESM4	NOAA Geophysical Fluid Dynamics Laboratory /USA	1°×1.25° 1°×1.25°
12	HadGEM3-GC31-LL	Met Office Hadley Centre /UK	1.25°×1.875°
13 14	INM-CM4-8 INM-CM5-0	Institute for Numerical Mathematics, Russian Academy of Science /Russia	1.5°×2° 1.5°×2°
15	IPSL-CM6A-LR	Institut Pierre-Simon Laplace /France	1.26°×2.5°
16 17	MIROC6 MIROC-ES2L	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute, The University of Tokyo, National Institute for Environmental Studies, and RIKEN Center for Computational Science /Japan	1.4°×1.4° 2.8°×2.8°
18 19	MPI-ESM-1-2-HR MPI-ESM-1-2-LR	Max Planck Institute for Meteorology /Germany	0.9375°×0.9375° 1.875°×1.875°
20	MRI-ESM2-0	Meteorological Research Institute /Japan	1.125°×1.125°
21	NESM3	Nanjing Univ. of Information Sci. and Technology /China	1.875°×1.875°
22 23	NorESM2-LM NorESM2-MM	Norwegian Climate Centre /Norway	1.875°×2.5° 0.9375°×1.25°
24	UKESM1-0-LL	Met Office Hadley Centre /UK	1.25°×1.875°

We use six quantitative indices, including mean temperature (**TAS**), annual maximum (hottest daytime) temperature (**TXx**), annual minimum (coldest nighttime) temperature (**TNn**), total precipitation in wet days (**PRCPTOT**), annual maximum consecutive 5-day precipitation amount (**RX5DAY**) and annual total precipitation for events exceeding the 95th percentile (**R95P**).

Indices from models and observation were firstly calculated at their original grid and then interpolated, using bilinear interpolation, onto a common 1° × 1° grid comprising 928 geographic locations across China. The three ensemble-processing strategies, AM, LR and RF, were then practiced on this common grid

Design and operating process to deal with multi-model ensemble simulations.



$$T_{cor} = T_{ori} + \mu(T_{obs}) - \mu(T_{ori})$$

$$P_{cor} = P_{ori} \times \frac{\mu(P_{obs})}{\mu(P_{ori})}$$

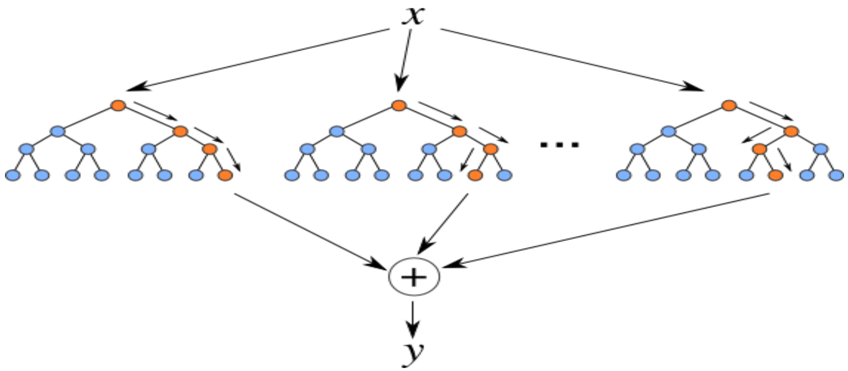
LR is suitable for resolving regression problems across multiple models or members in an ensemble. It fits a linear model to minimize the sum of squared errors. Its general form is:

Y = a0 + A·X

where x(i, k) is the input spatial field (i = 1, ..., 928) from the 24 models (k=1, ..., 24) and y(i) is the output spatial field. The regression coefficients a0 and A were fitted with data in the training period.

Function “LinearRegression” in the module “sklearn.linear_model” in python 3.8 (**https://scikit-learn.org**)

In our work, RF uses the function “RandomForestRegressor” from the python package “sklearn.ensemble” (https://scikit-learn.org). For the training, we have data covering 34 years, from 1961 to 1994, and 928 spatial points. The total number of samples into our RF training is thus 34×928=31552. **Each of the 24 climate models is treated as a feature** in our RF implementation.



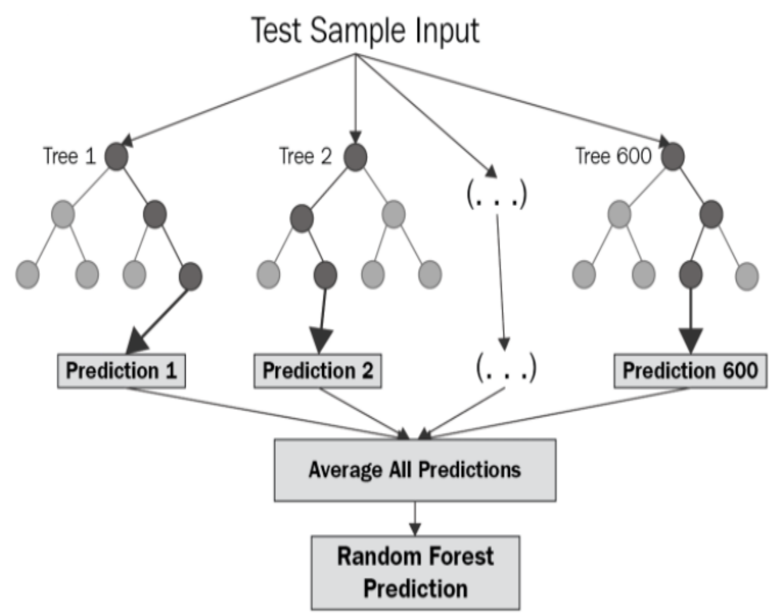
Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

A Random Forest Regression model is **powerful and accurate**. It usually performs great on many problems, including features with **non-linear relationships**.

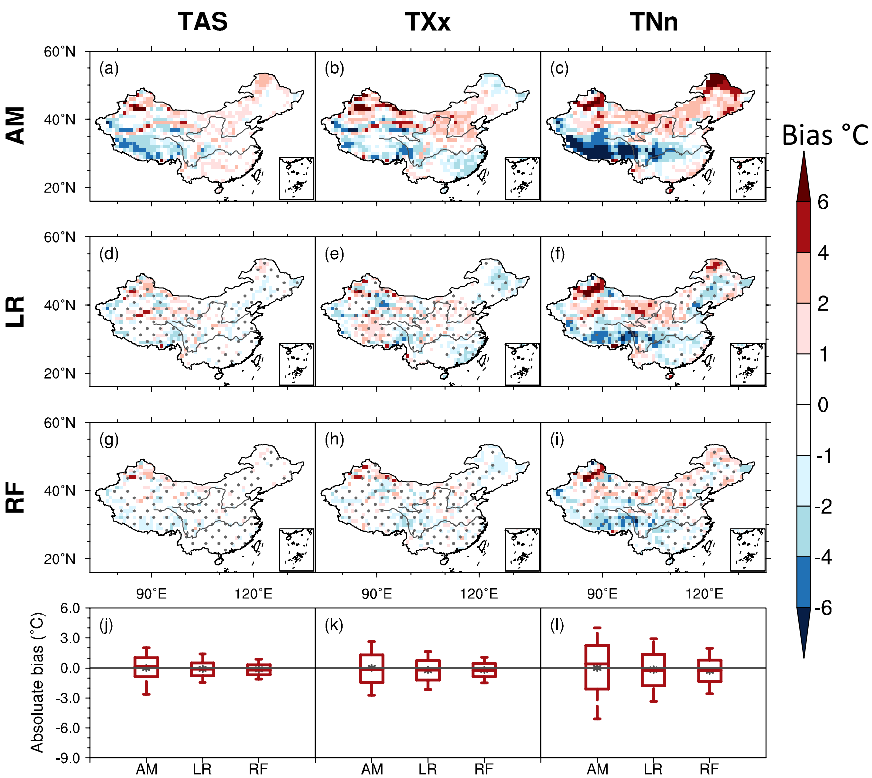
Disadvantages, however, include the following: there is **no interpretability**, **overfitting** may easily occur, we must choose the number of trees to include in the model.

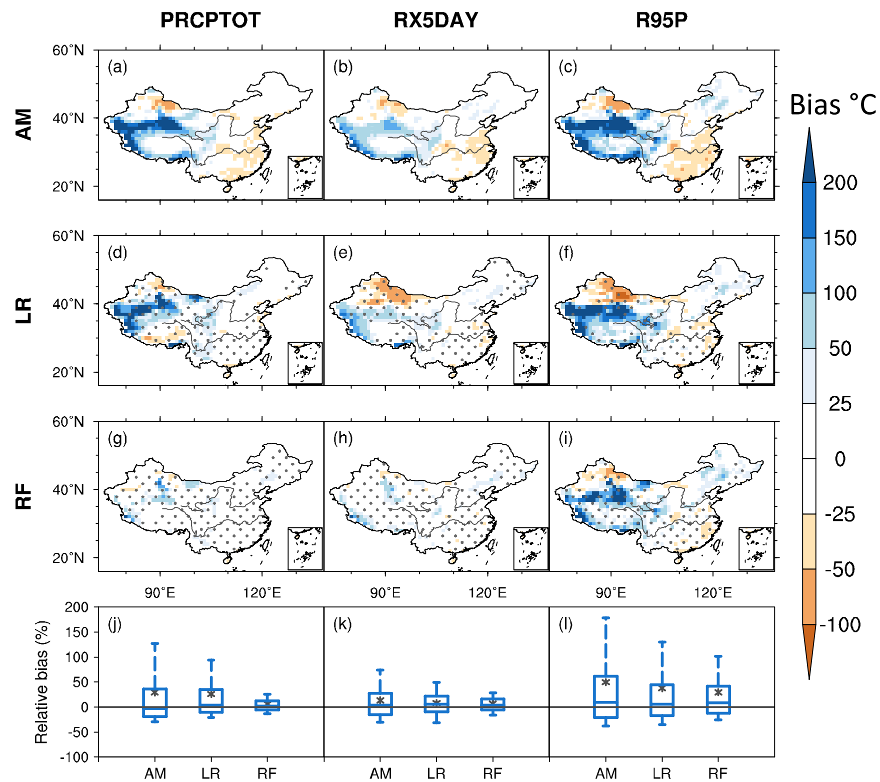
Trees run in **parallel** with no interaction amongst them. A Random Forest operates by **constructing several decision trees** during training time and outputting the mean of the classes as the prediction of all the trees. Steps:

1. Pick at random k data points from the training set.
2. Build a decision tree associated to these k data points.
3. Choose the number N of trees you want to build and repeat steps 1 and 2.
4. For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all of the predicted y values.

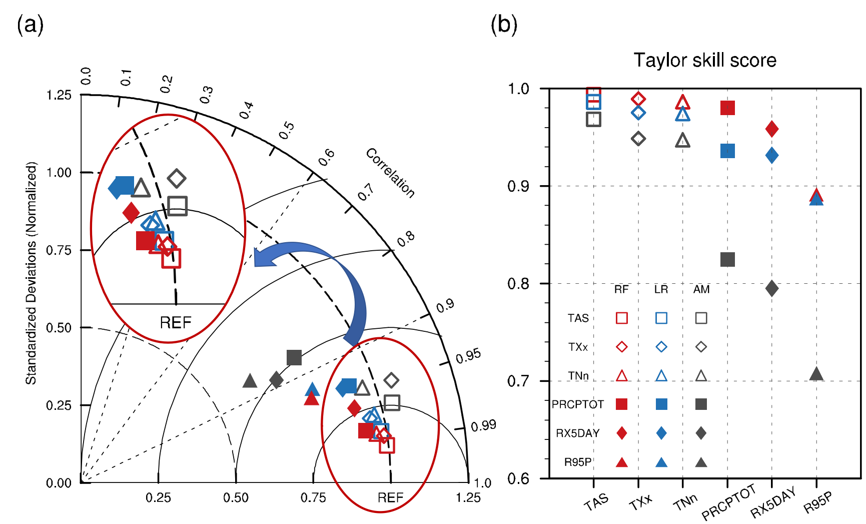


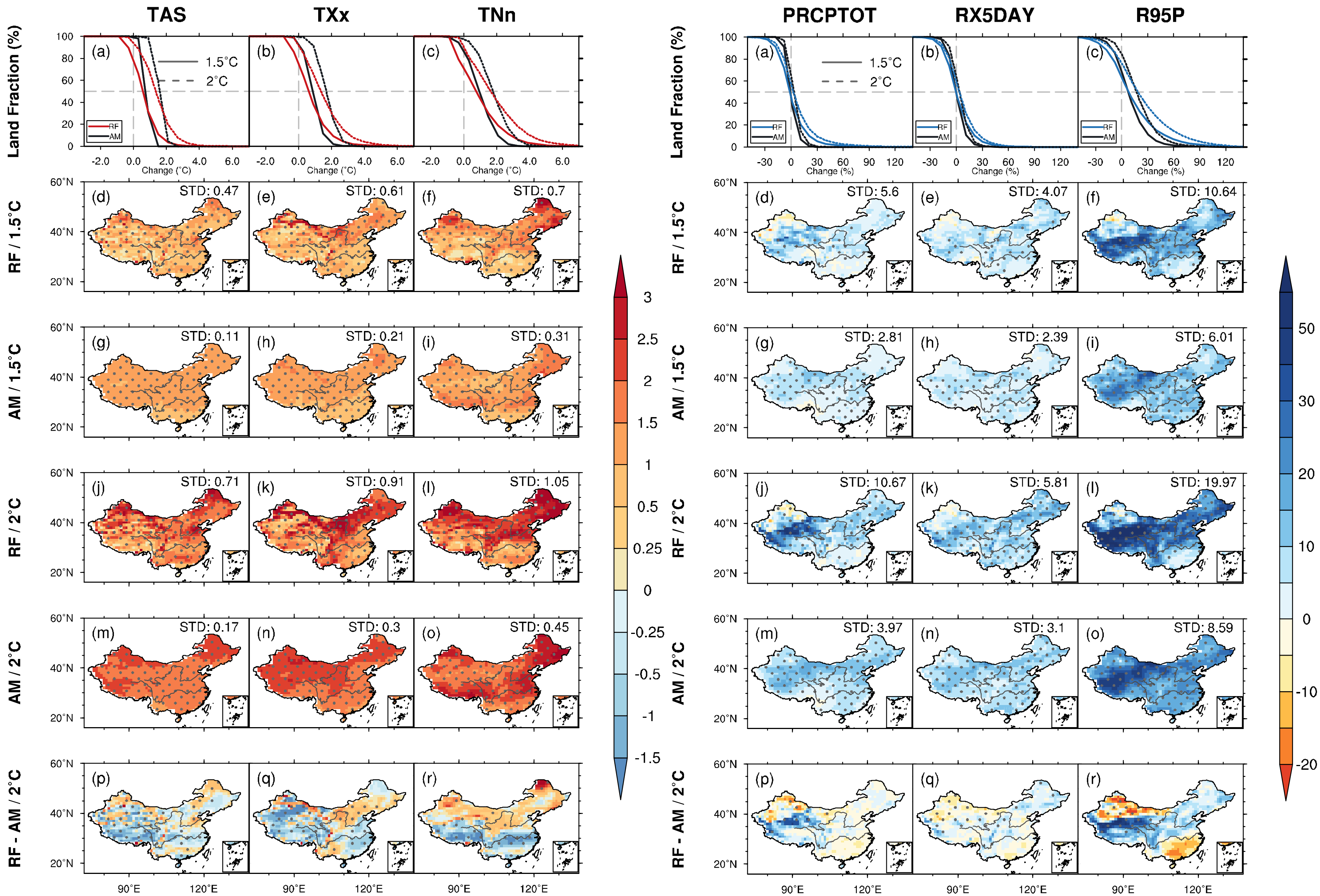
Courtesy Chaya https://levelup.gitconnected.com/random-forest-regression-209c0f354c84





Spatial distributions and corresponding boxplots of **biases** (°C for temperature, % for precipitation) from AM, LR, and RF algorithms in the validation period.





Projection of future climate change (°C for temperature, % for precipitation)

Conclusions

- In this work, three different ensemble-processing strategies, AM (arithmetic mean), LR (linear regression), and RF (Random Forest, machine learning decision tree algorithm), are used to explore information offered by the multi-model ensemble climate simulations of CMIP6. The main idea was to find the best way of processing the ensemble simulations to mimic observational climatic properties and to give a more reliable projection of future climate.
- AM is the simplest and most intuitive strategy. LR advocates the vision of a linear-regression approach to establish the relationship between simulations and observations, but it cannot necessarily represent any physical rules governing the climate system. RF is one of the most advanced machine-learning algorithms. It can extract non-linear and complex relations among climate models, instead of making a simple evaluation of models' apparent performance as in other ensemble-processing strategies.
- This leads to a hybrid approach that we advocate for climate change issues, which combines physical modelling and machine learning strengths, thus giving confidence in retrieving more valuable information.