



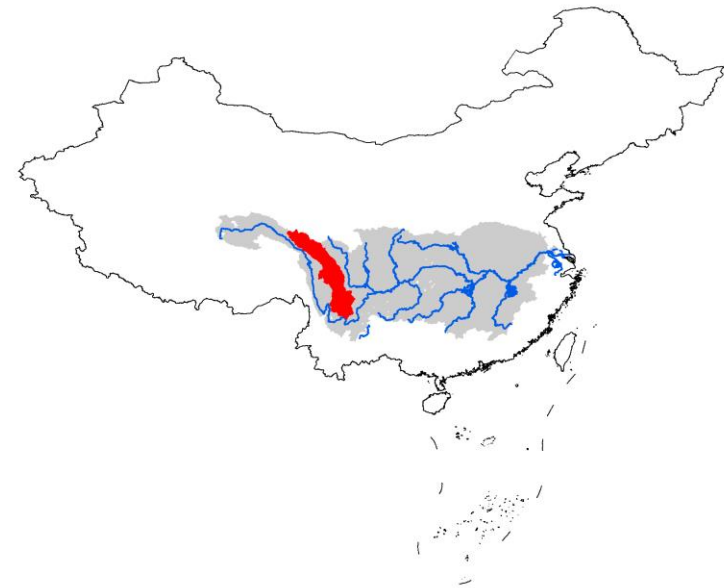
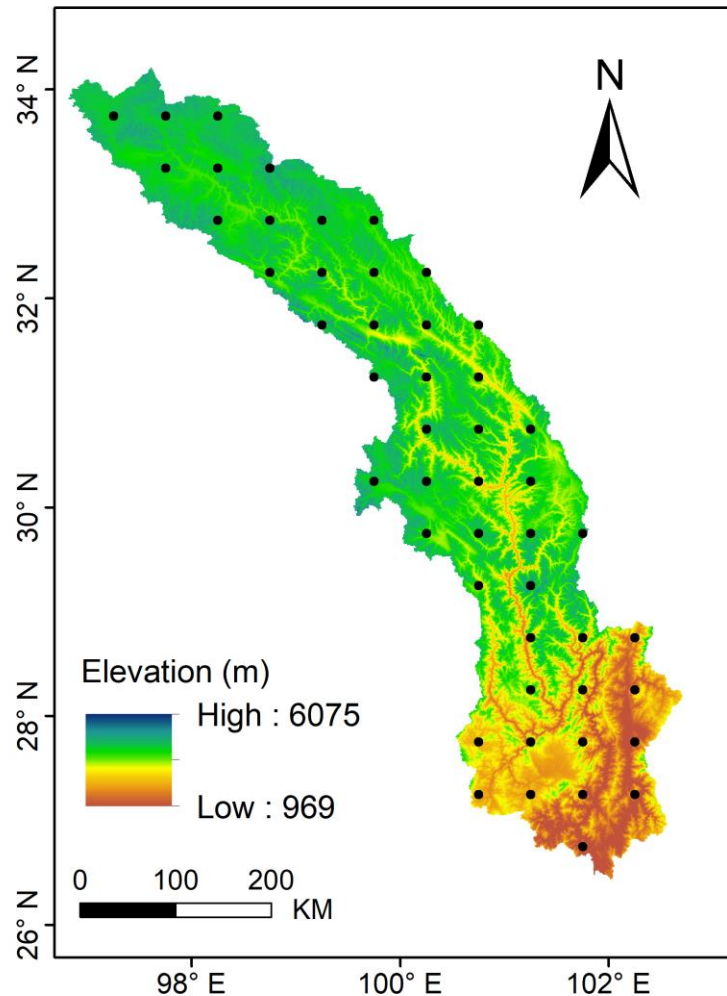
Machine learning for precipitation forecasts post-processing — Multi-model comparison and experimental investigation

Yuhang zhang
(zhangyh@19mail.bnu.edu.cn)
Beijing Normal University

Objectives:

- 1). Design a pipeline of machine learning algorithms for precipitation post-processing forecasting,
- 2). Test the applicability of 21 existing machine learning algorithms for precipitation post-processing forecasting,
- 3). Optimize the model configuration to obtain more accurate precipitation forecasts.

Study area and data



Legend

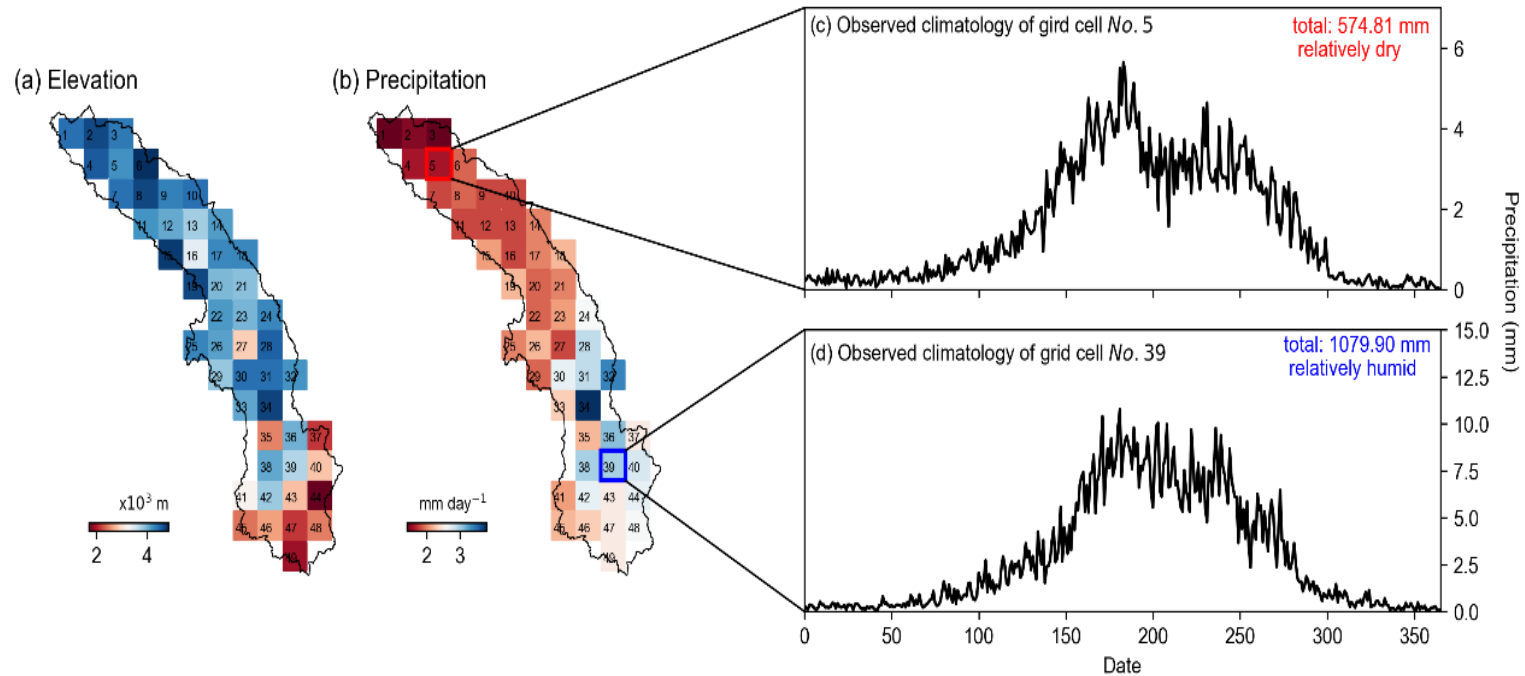
- Grid Center
- Chinese boundary
- Yangtze River
- Yangtze River Basin
- Yalong River Basin

Observation (reference) data: CMA 0.5 degree daily precipitation

(Re)Forecast data (hindcast): NCEP GEFS reforecast version 2 with lead time 8 day, 1 control, 10 perturbed

Time period: 1985-2018

Study area attributions



Typical sites

relative dry: Grid cell No.5; Continental climate, which is relatively cold and dry

relative wet: Grid cell No.39. Subtropical climate, relatively high temperature and large precipitation

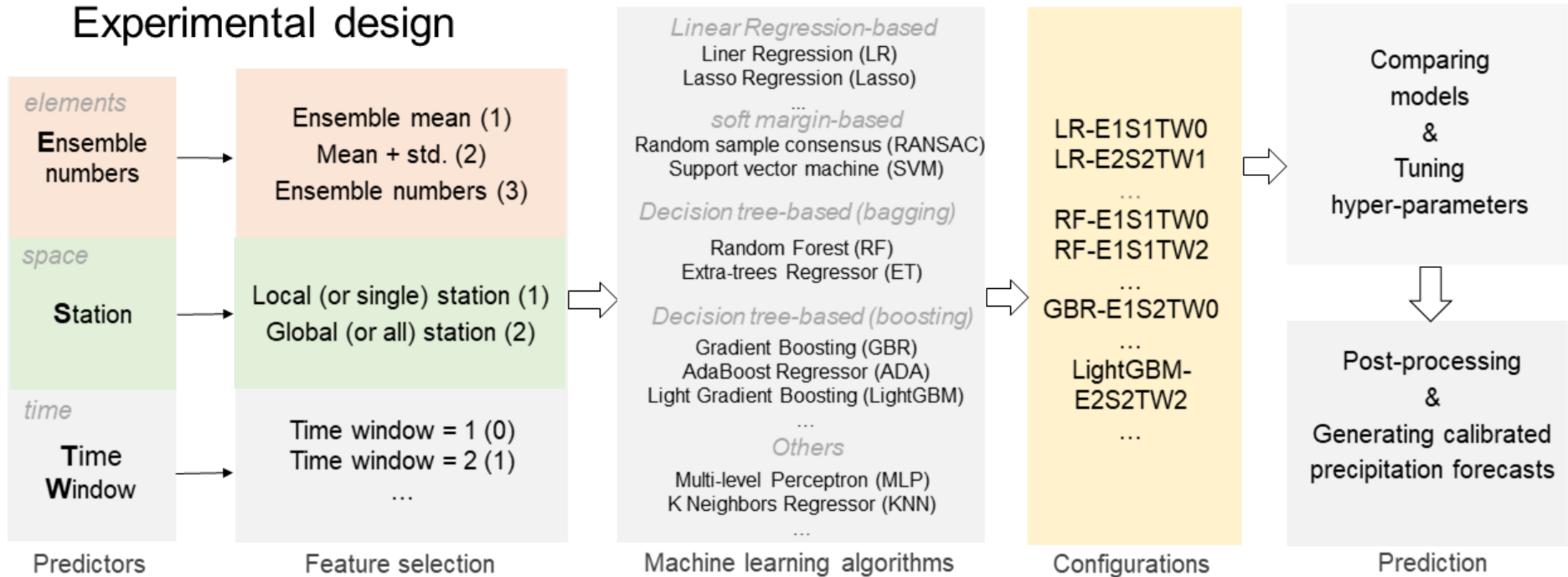
Aims:

1. Compare the differences between **regional and local models**
2. Compare the differences between **climate types**

Workflow

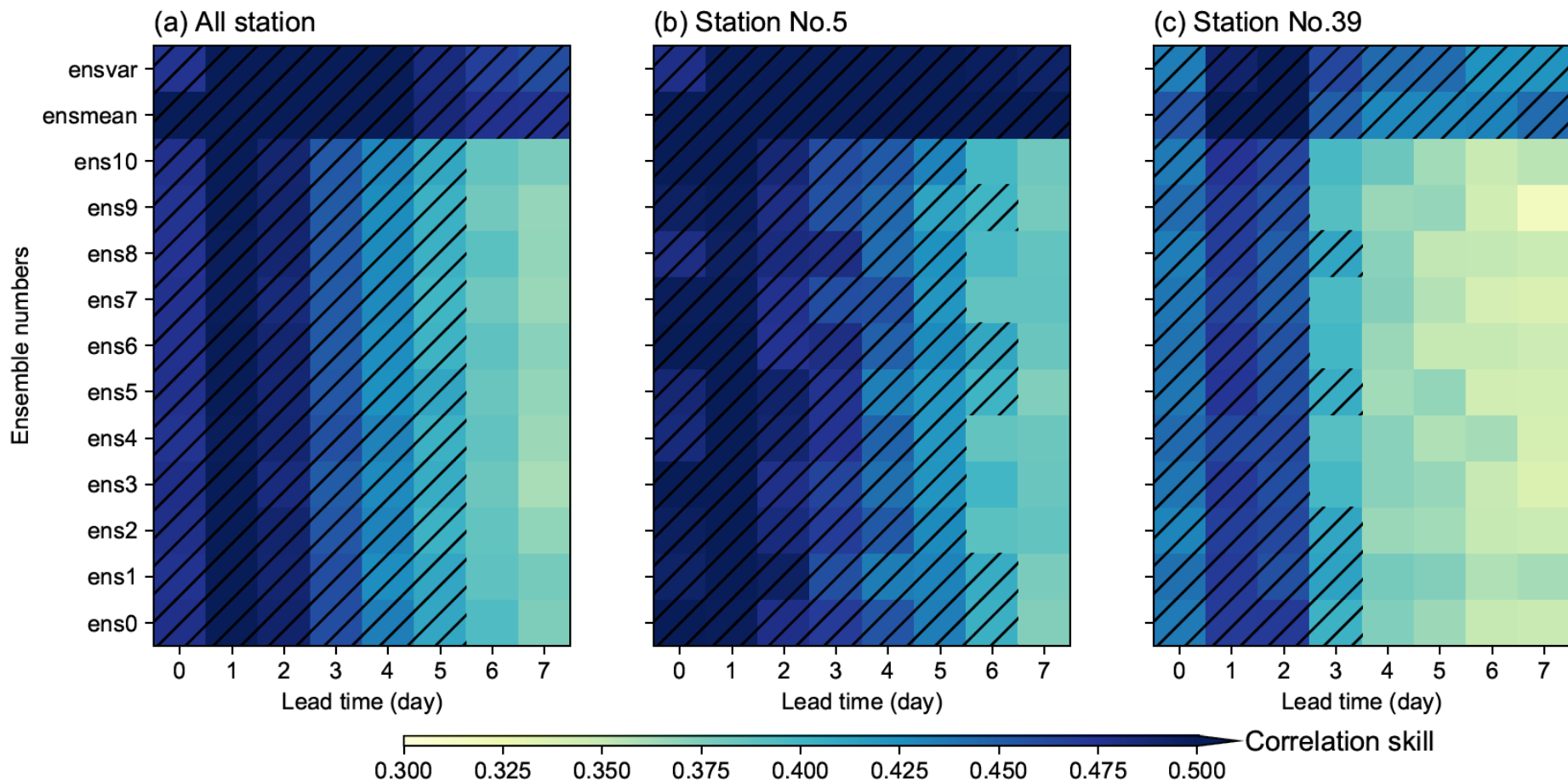
e.g.

Experimental design

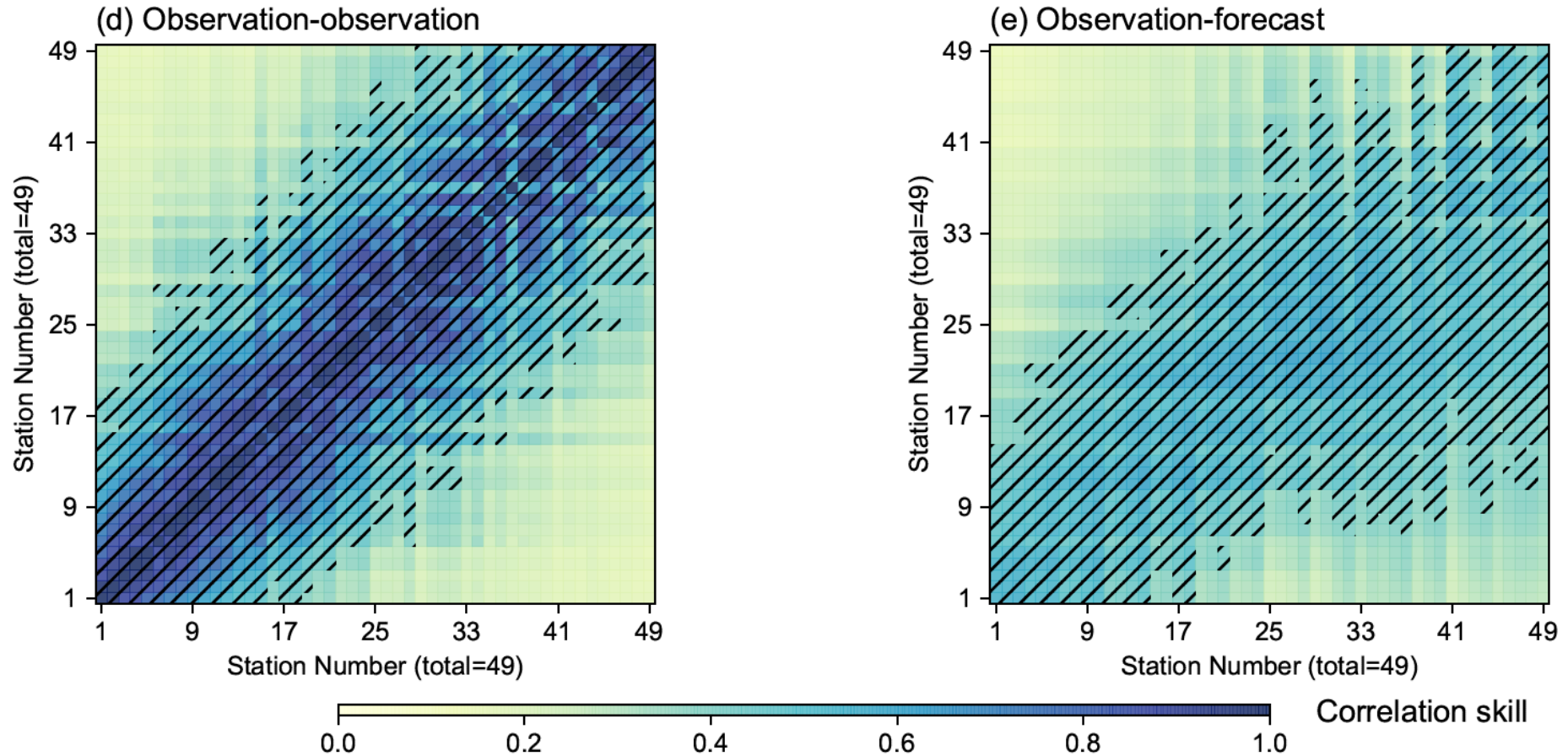


1. Data processing : data extraction, combination, interpolation, and transformation.
2. Predictors selection : ensemble numbers, space, time
3. Machine learning methods: (1) linear regression-based models; (2) decision tree-based models; (3) ensemble learning models; (4) other models, such as KNN, MLP.

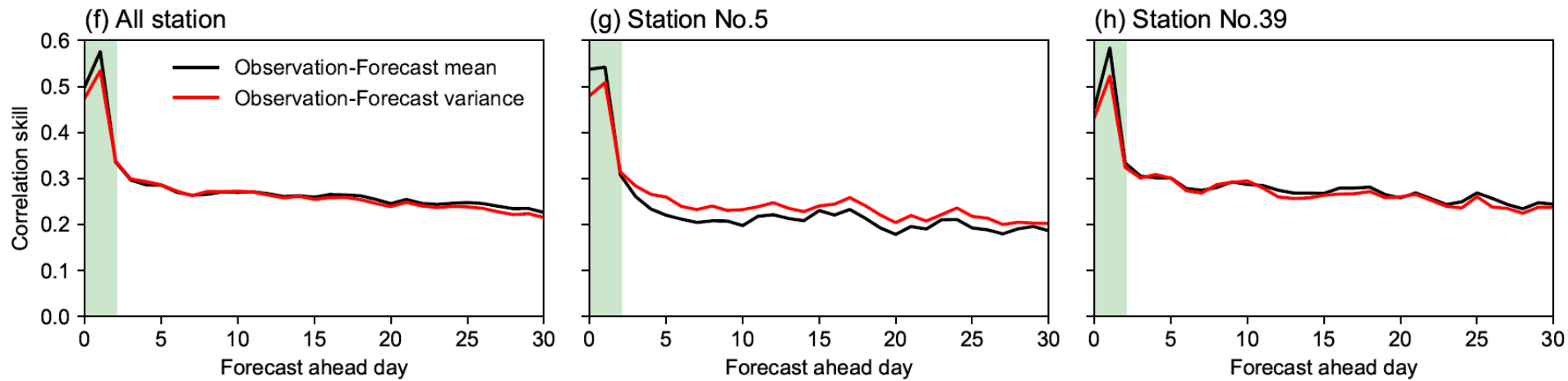
Feature selection (forecast predictor)—ensemble numbers



Feature selection (forecast predictor)—spatial correlation



Feature selection (forecast predictor)—time window



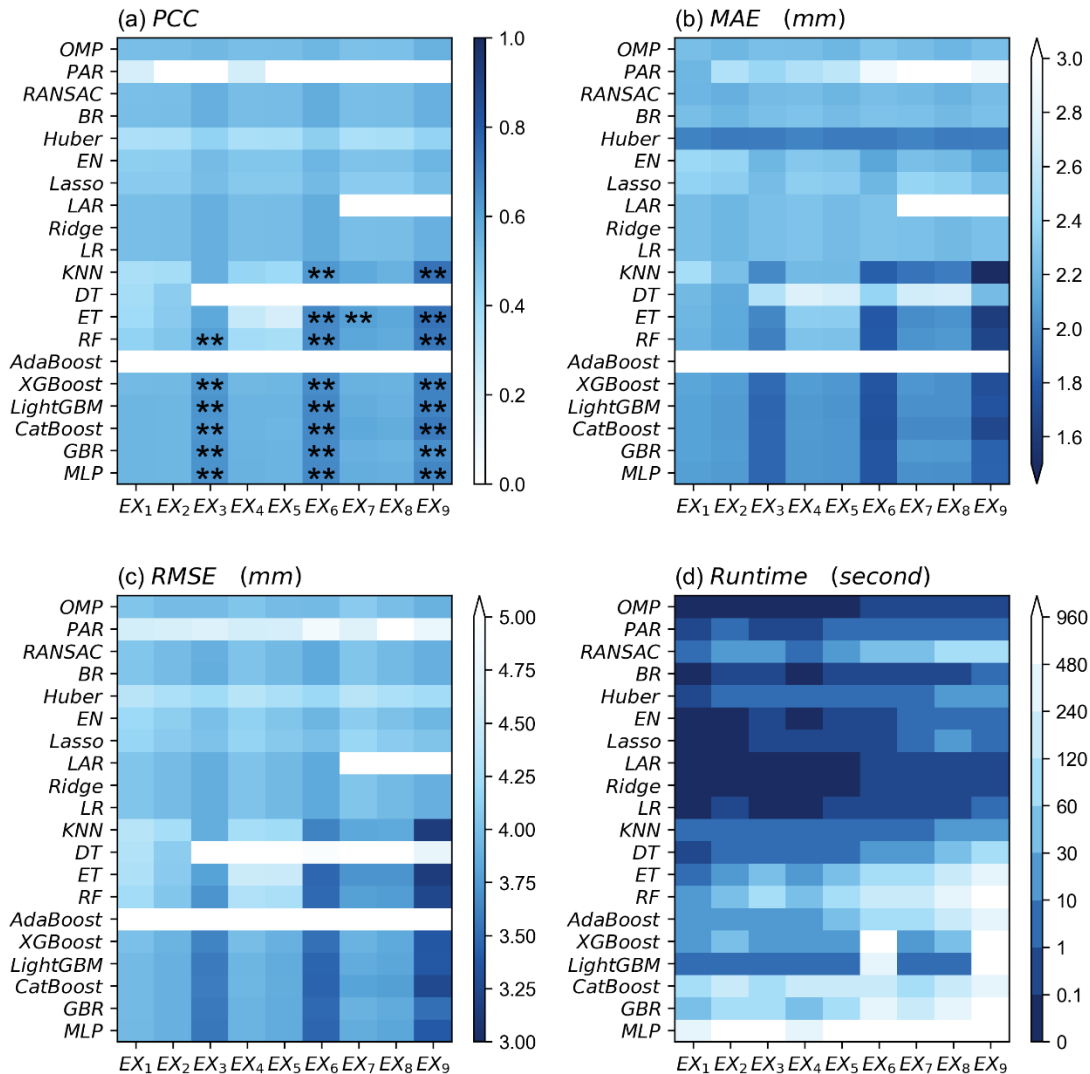
ML methods

ID	Full name and abbreviation	Brief description	Reference
M1	Multi-layer Perceptron (MLP)	Artificial Neural Network a.k.a.	Rumelhart, 1986
M2	Gradient Boosting Regressor (GBR)	An ensemble model based on Gradient Boosting Machine	Friedman, 2001
M3	CatBoost Regressor (CatBoost)	Similar to the GBR, but optimized for category features	Prokhorenkova, 2017
M4	Light Gradient Boosting (LightGBM)	Similar to XGBoost, but introduced the histogram algorithm	Ke et al., 2017
M5	Extreme Gradient Boosting (XGBoost)	Similar to the GBR, but optimized efficiency, generalization, and robustness	Chen, 2015
M6	AdaBoost Regressor (AdaBoost)	An ensemble model based on Gradient Boosting Machine and weight of sample points	Freund & Schapire, 1996
M7	Random Forest (RF)	An ensemble bagging model based on DT	Breiman, 2001
M8	Extra Trees Regressor (ET)	Similar to the RF, but randomness is added	Geurts, 2006
M9	Decision Tree (DT)	A binary tree with probability	Xu et al., 2005
M10	K Neighbors Regressor (KNN)	K nearest neighbor with distance	Altman, 1992
M11	Linear Regression (LR)	Ordinary least squares Linear Regression	-
M12	Ridge Regression (Ridge)	Linear least squares with L2 regularization	Hoerl, 1970
M13	Least Angle Regression (LAR)	Similar to forward stepwise regression	Efron et al., 2004
M14	Lasso Regression (Lasso)	Linear Model trained with L1 regularization	Kwangmoo, 2007
M15	Elastic Net (EN)	Linear regression with combined L1 and L2 regularization	Owen, 2006
M16	Huber Regressor (Huber)	A linear regression model that is robust to outliers.	Huber, 2004
M17	Bayesian Ridge (BR)	Similar to Ridge, introduced Bayesian inference	Bishop, 2006
M18	Random Sample Consensus (RANSAC)	Similar to LR, but ignored the outliers	Fischler & Bolles, 1981
M19	Passive Aggressive Regressor (PAR)	An incremental learning method	Crammer et al., 2006
M20	Orthogonal Matching Pursuit (OMP)	A greedy iteration method	Pati et al., 1993
M21	Support Vector machine(SVM)	A soft margin-based model with kernel transformation	Schölkopf et al., 2002

Experimental design

ID	Time period	ensemble number	Time window	Sample shape
EX1	2009-2018	ensemble mean	0	161063,1
EX2	1999-2018	ensemble mean	0	357945,1
EX3	2009-2018	ensemble mean	2	161063,2
EX4	2009-2018	ensemble mean, ensemble standard deviation	0	161063,2
EX5	1999-2018	ensemble mean, ensemble standard deviation	0	357945,2
EX6	2009-2018	ensemble mean, ensemble standard deviation	2	357945,4
EX7	2009-2018	11 ensemble numbers	0	161063,11
EX8	1999-2018	11 ensemble numbers	0	357945,11
EX9	2009-2018	11 ensemble numbers	2	161063,33

Baseline results

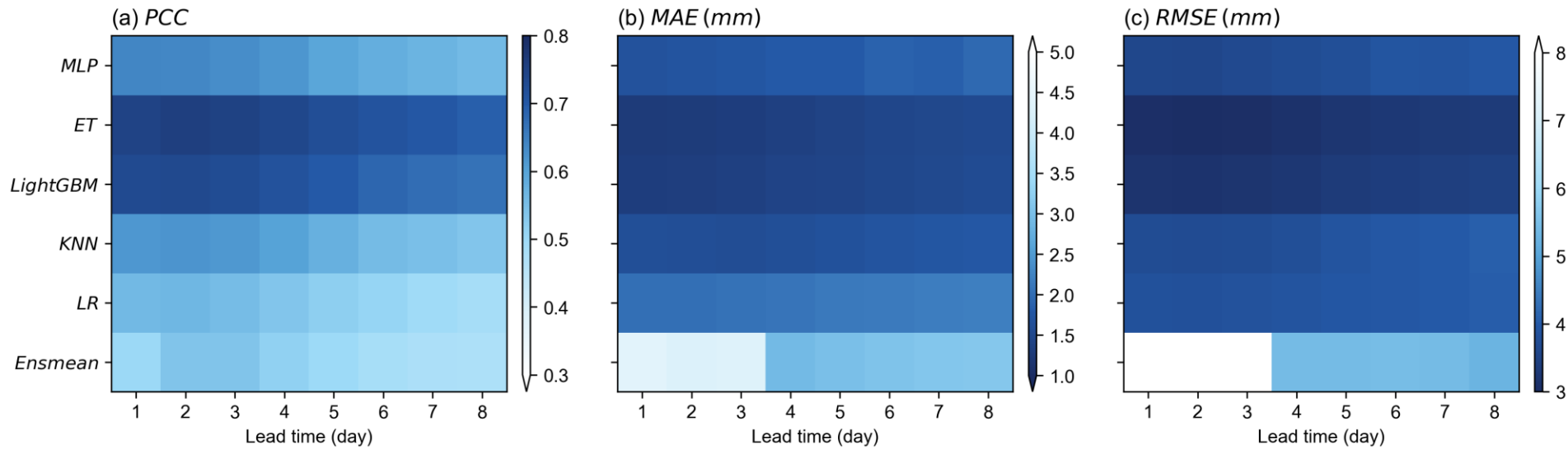


1. Experiments
2. Models
3. runtime

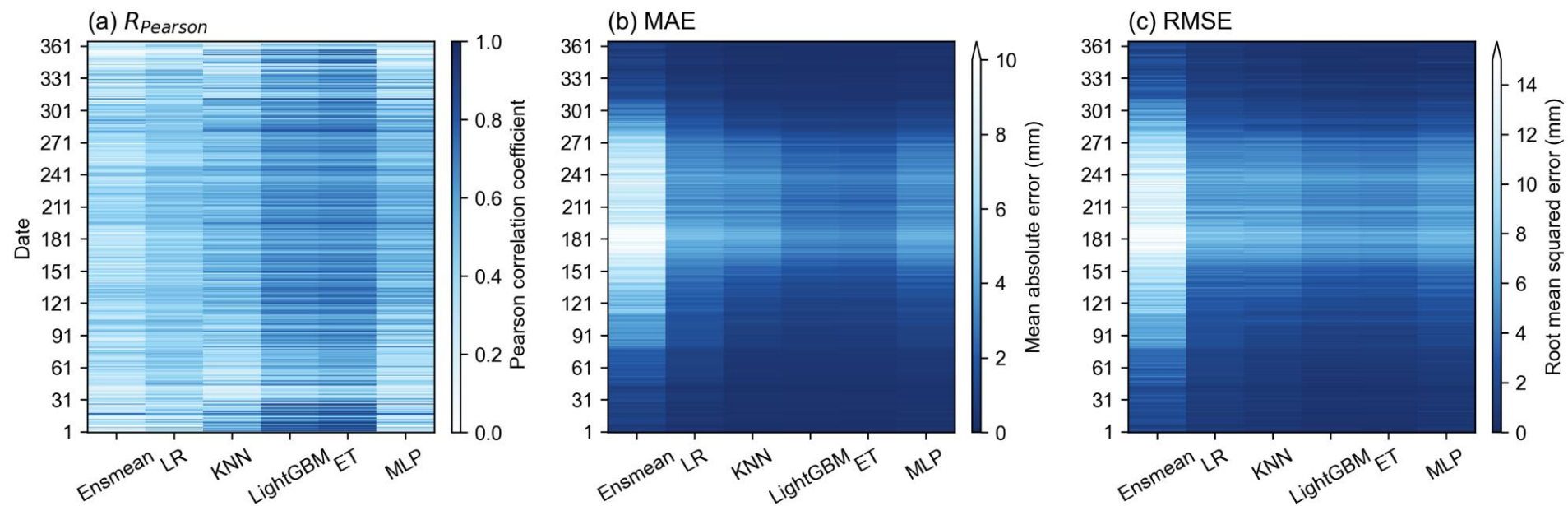
By comparing the results of the benchmark experiments, we selected five models from different categories for further analysis, namely, LR, ET, KNN, LightGBM, and MLP.

Meanwhile, we followed the configuration of EX9, with a 10-year training sample size (2009-2018), a 2-day time window, and 11 ensemble members as predictors.

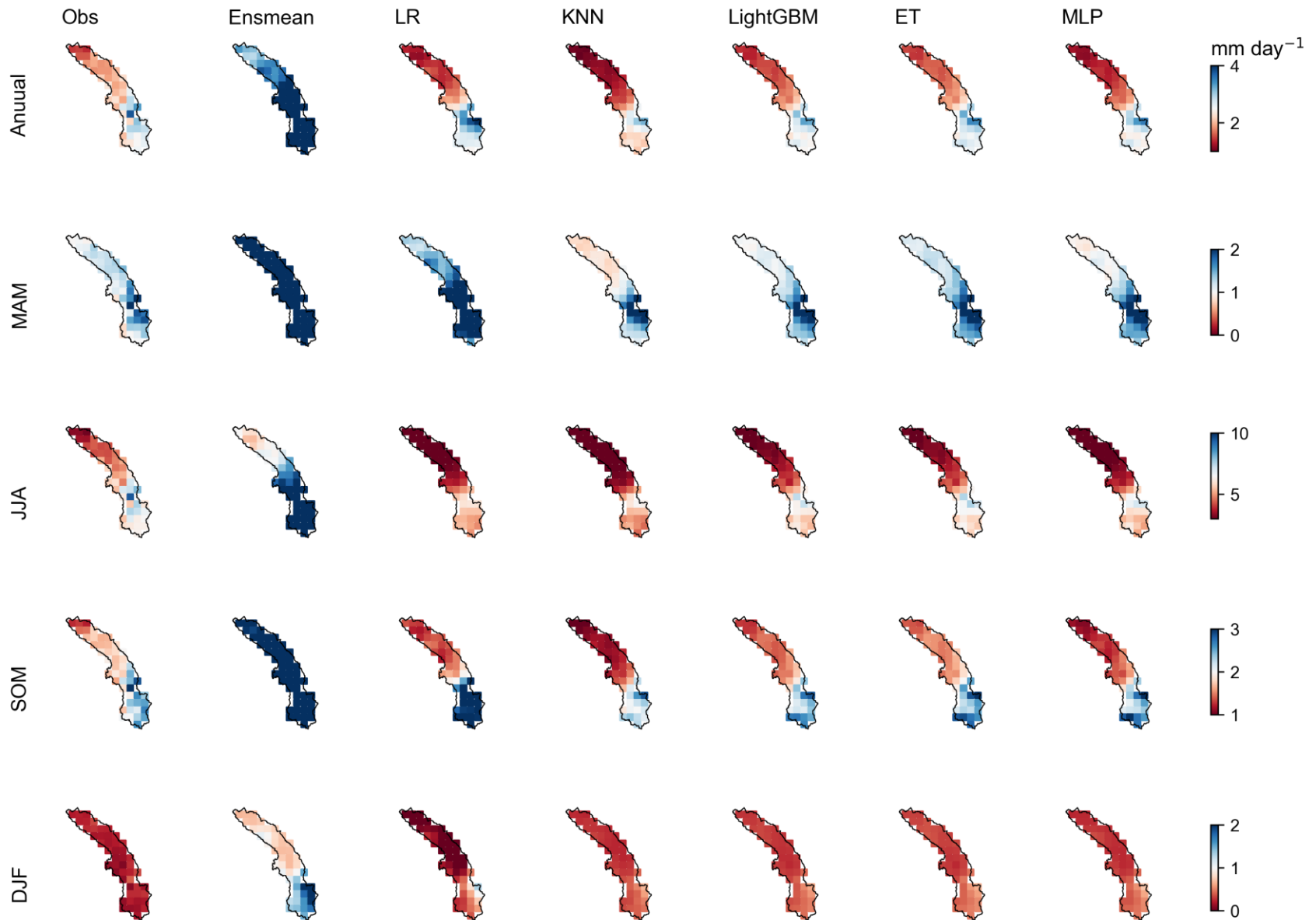
Results—Regional averaged model performance with lead time up to 8 day.



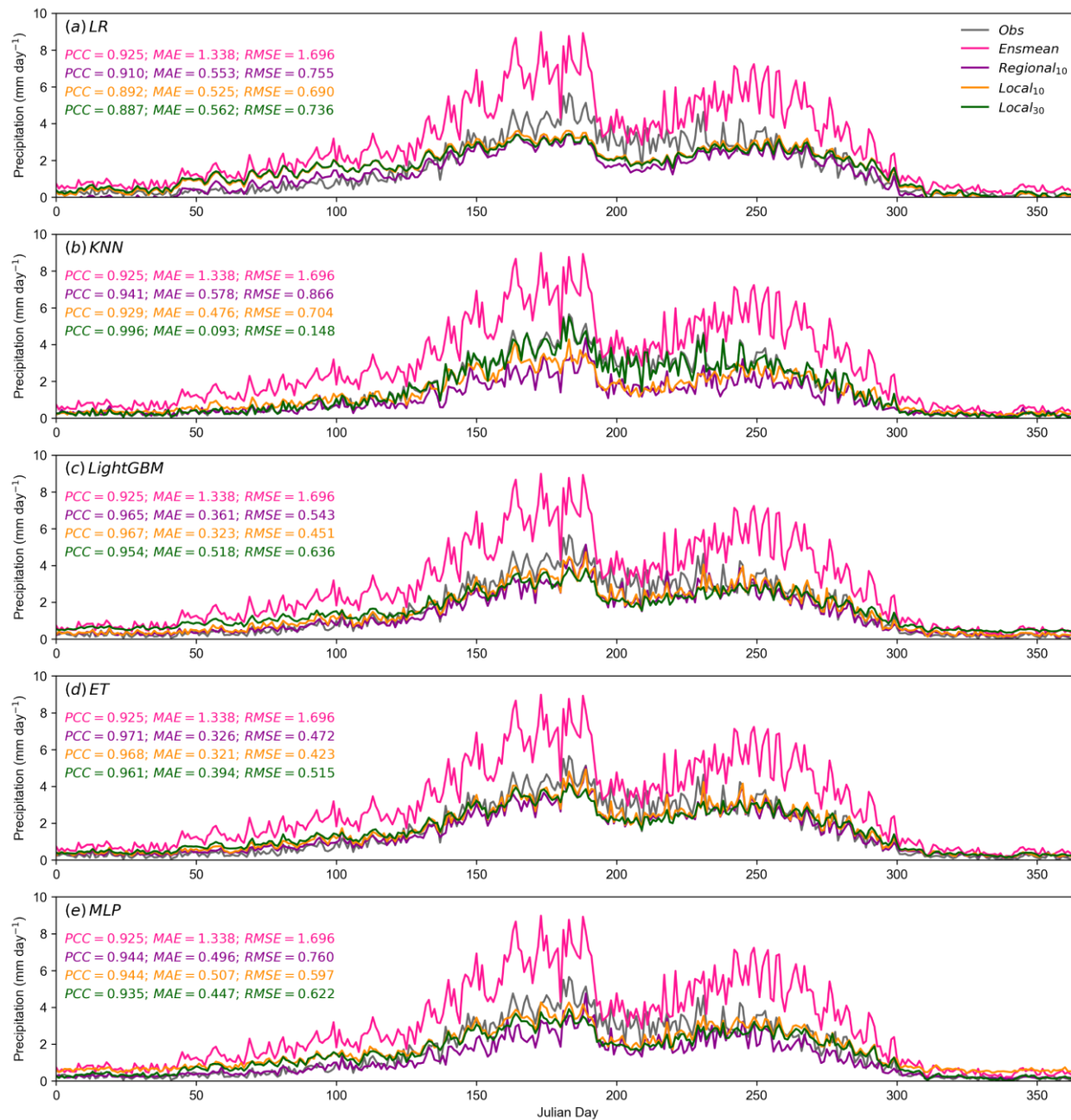
Results—Regional averaged model performance with different ML methods on each Julian Day



Results—Spatial distribution of precipitation forecasts from different models



Results—Performance comparison between regional and local model



Take home message

- 1). Twenty-one machine learning algorithms were compared as post-processors for precipitation forecasts.
- 2). The optimized extra-trees regressor exhibited the best performance among all ML algorithms.
- 3). The best model was obtained using 11 ensemble members and a 2-day time window.
- 4). The optimized models can also be applied as regional models at basin scales.

Performance metrics

$$PCC = \frac{\sum_{i=1}^n (O_i - \bar{O})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^n (S_i - \bar{S})^2}}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |S_i - O_i|$$

$$RMSE = \left(\frac{1}{n} \sum_{i=1}^n (S_i - O_i)^2 \right)^{\frac{1}{2}}$$