The Sixth Session of the East Asia Winter Climate Outlook Forum
(EASCOF-6)

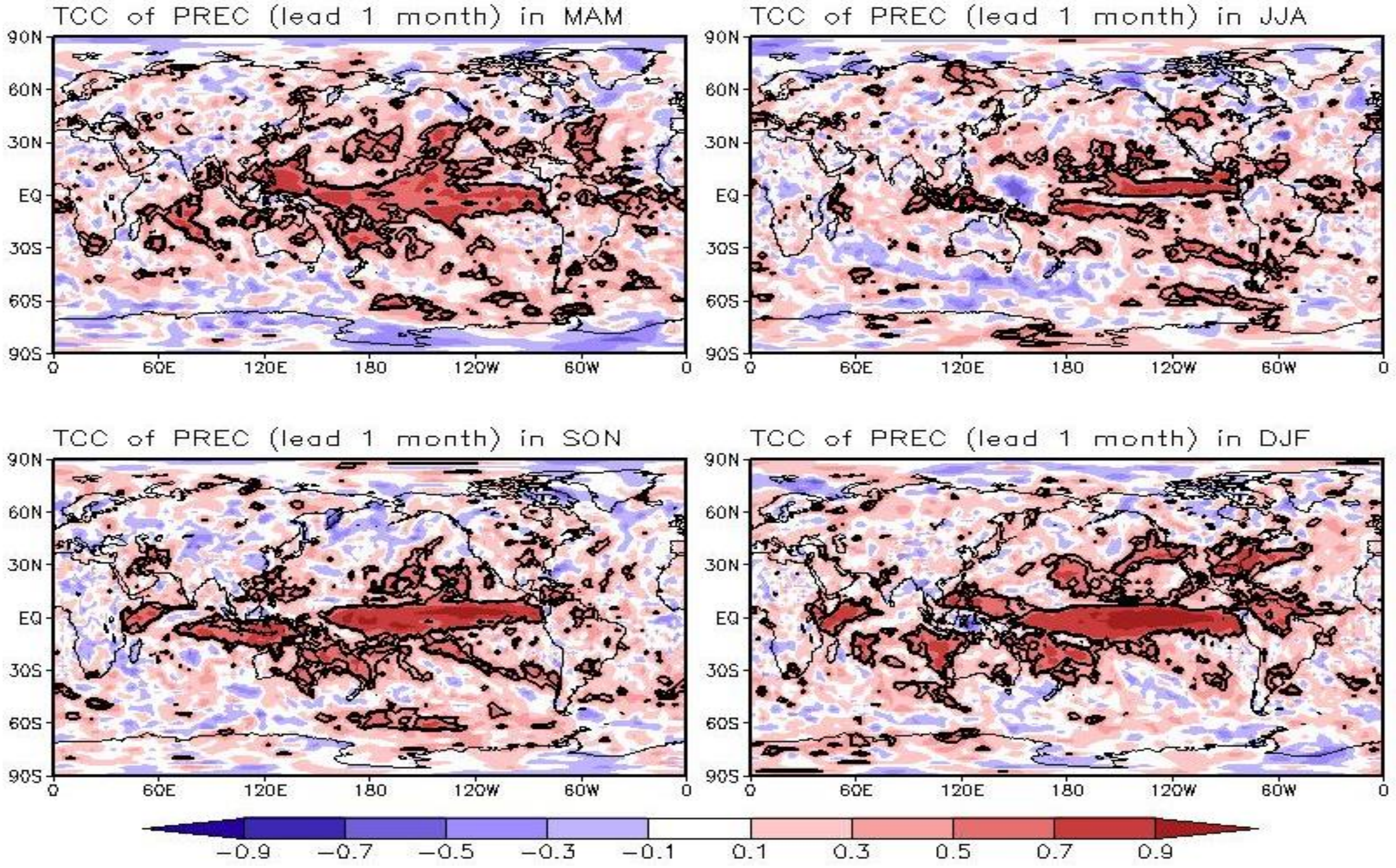
Statistical downscaling and Machine Learning in predicting monthly precipitation over China

Dr. Liu Changzheng
National Climate Center/CMA
November 7, 2018

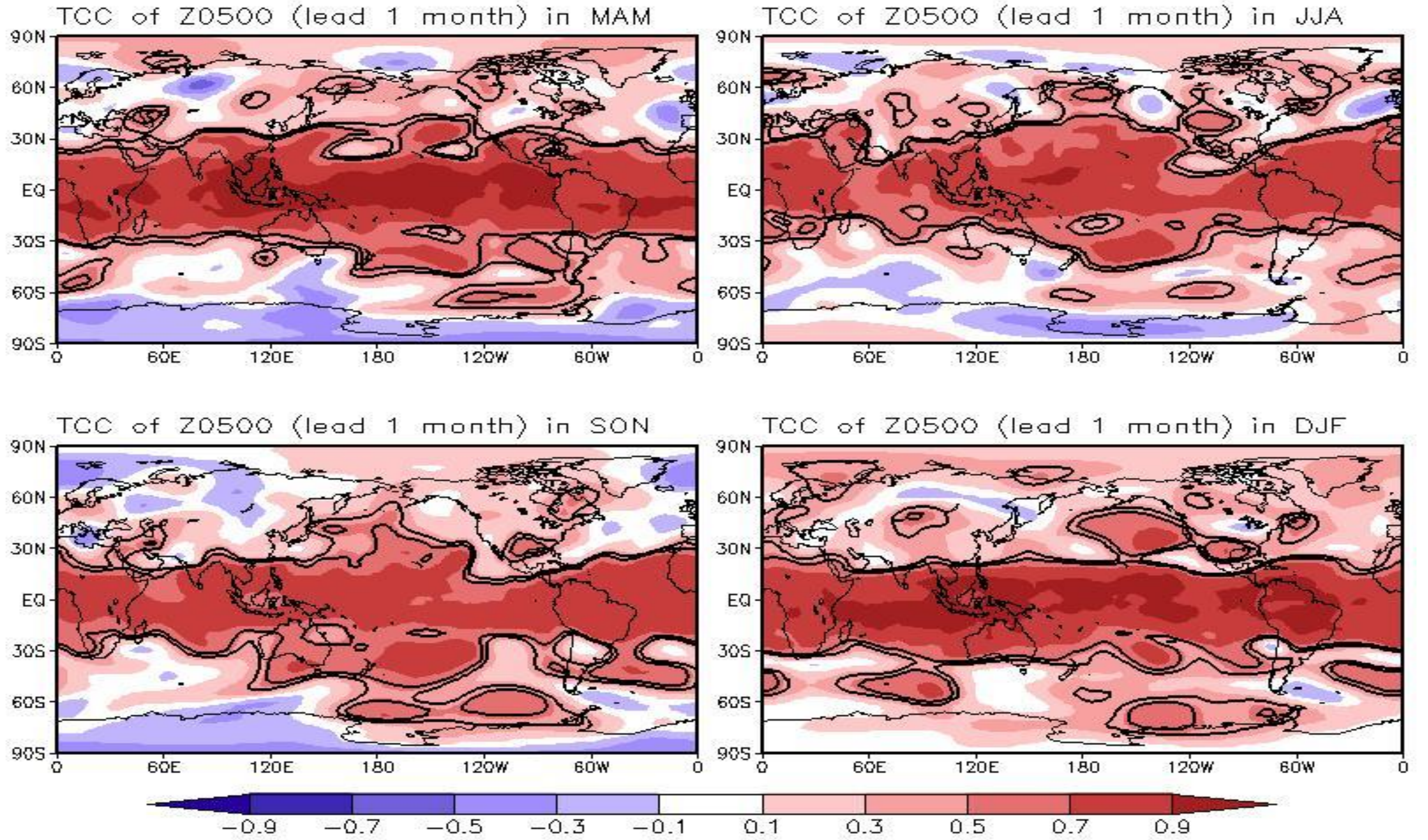
Outline

- **Introduction to statistical downscaling**
- **Introduction to machine learning**
- **Construction of a new statistical downscaling method**
- **Summary**

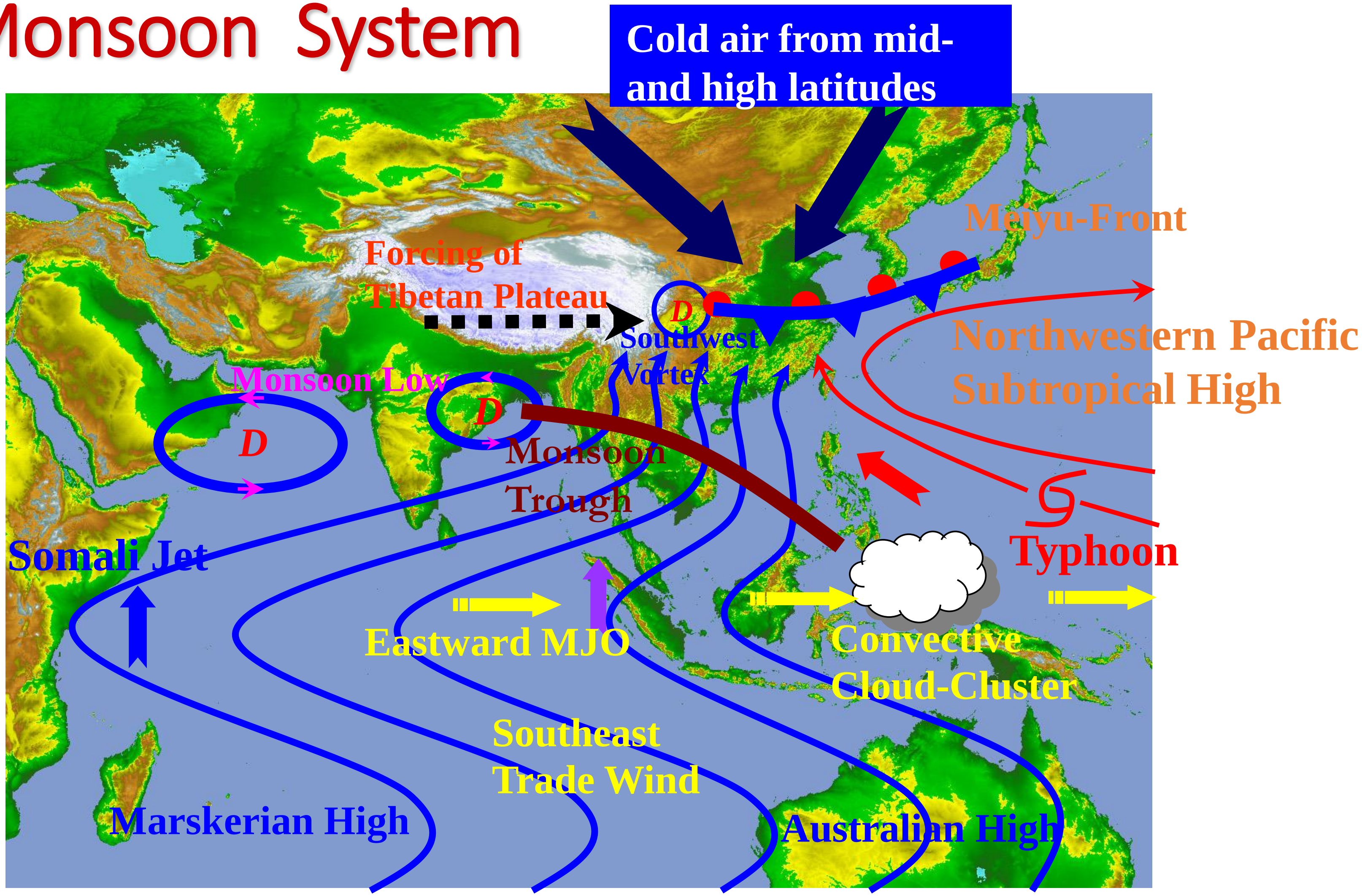
Prediction skills of precipitation of BCC CSM1.1



Prediction skills of H500 of BCC CSM1.1

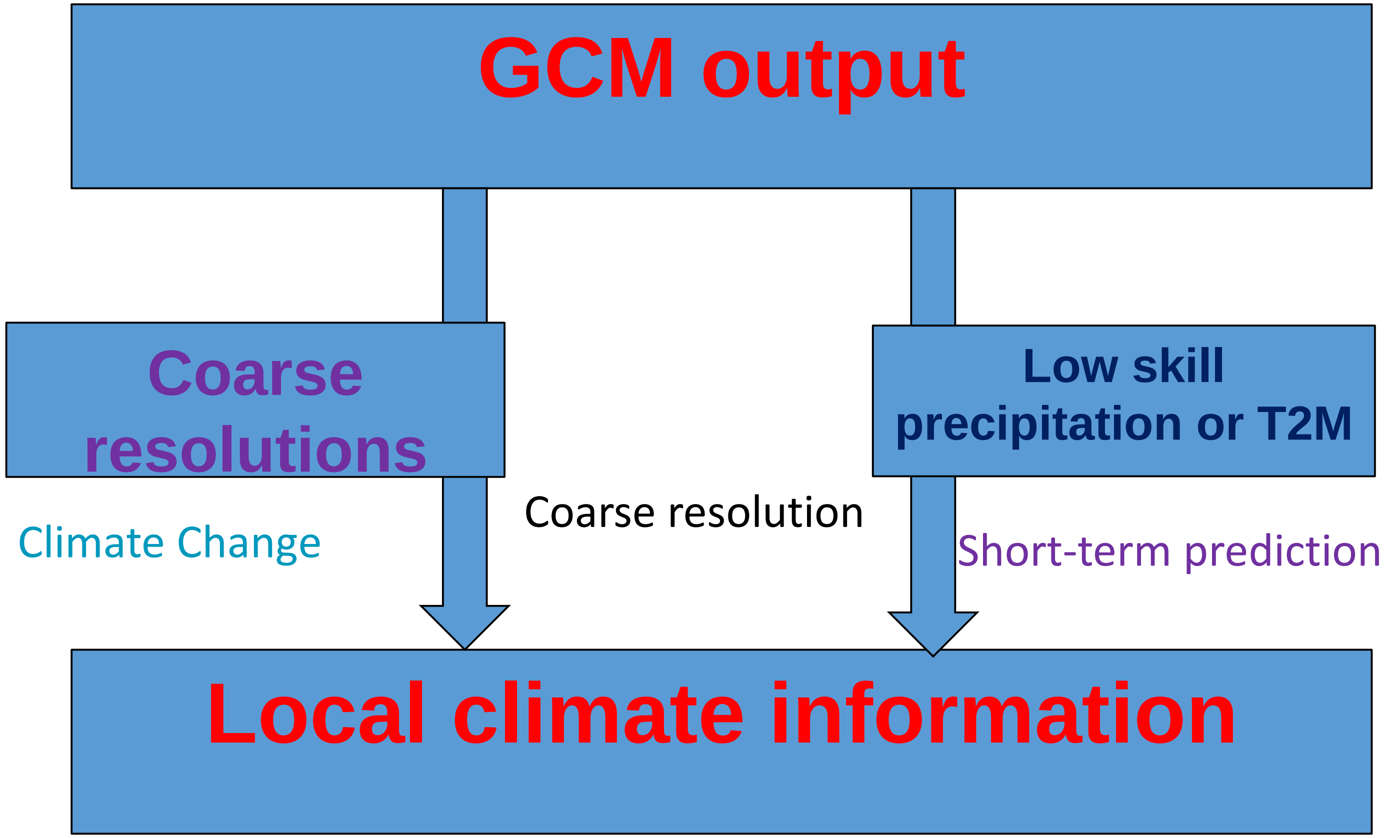


East Asian Monsoon System



rainfall : **<low frequency >** + <high-frequency>

Downscaling technique



Key points of statistical downscaling: predictor

- The statistical relationship between the predictor and predictand does not change over time.
- The predictor carries the climate signal.
- There is a strong relationship between the predictor and predictand.
- GCMs accurately simulate the predictor.

From < a review of downscaling methods for climate change projections> (USAID report)

Key points of statistical downscaling: modelling method

Linear methods

- Multiple linear regression
- Canonic correlation analysis(BP-CCA)
- Singular value decomposition (SVD)
- Optimal subset regression(OSR)
- CPPM
- …

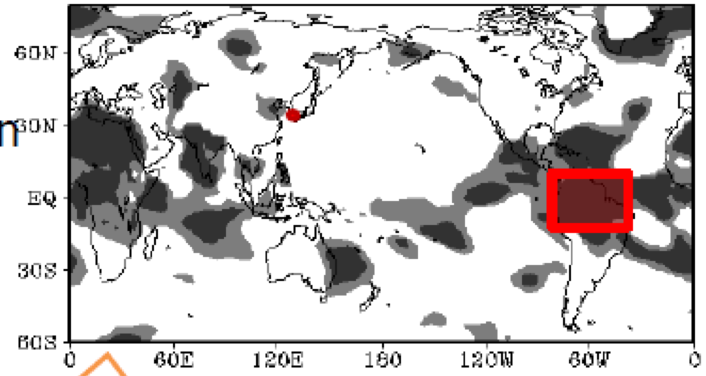
Nonlinear methods

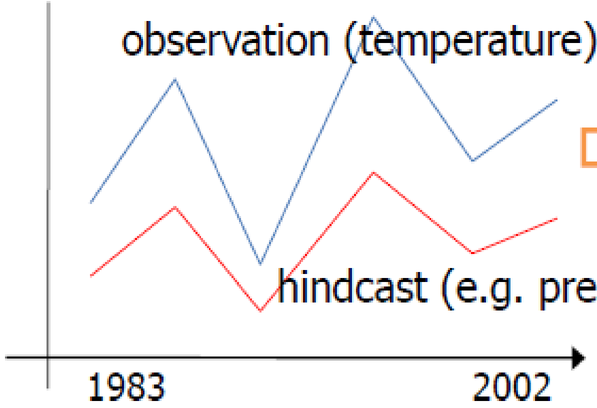
- Analog Method
- Self-organizing map (SOM)
- Artificial Neural Network（ANN）
-

Statistical downscaling method at APCC (Kang et al 2007)

step 1: to search coupled pattern between historical observation and hindcast

step 2: to estimate regression coefficient for downscaling





$$y = a + b\,x$$

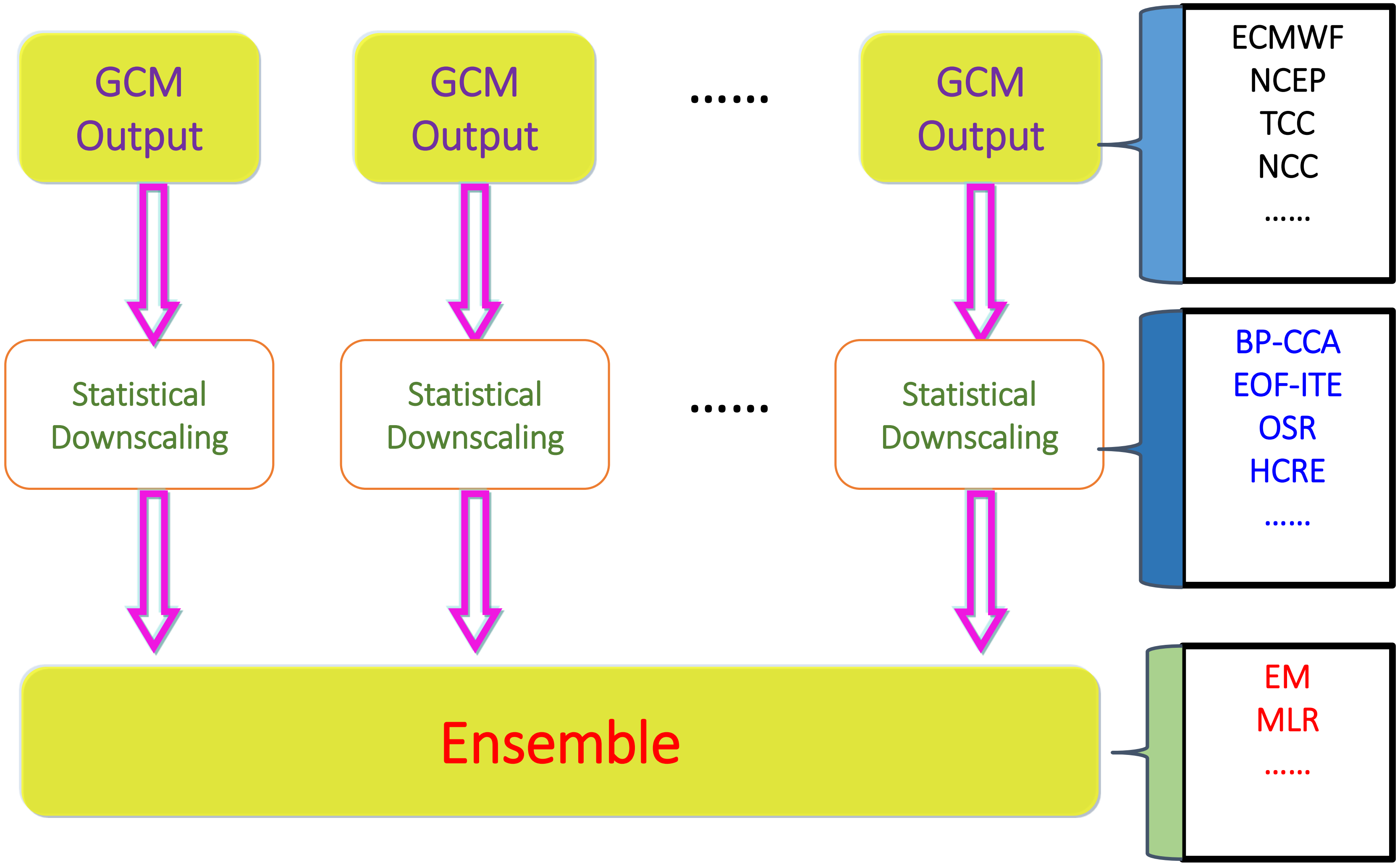
fcst	predictor
y_1	temperature
y_2	U wind
y_3	V wind
	:
y_i	pressure

predictor selection

step 3: to predict seasonal climate in each predictor

step 4: to select the best predictor based on cross-validation

Multi-Model Downscaling Ensemble Prediction System (MODES) at BCC



Key problems in statistical downscaling

Small examples

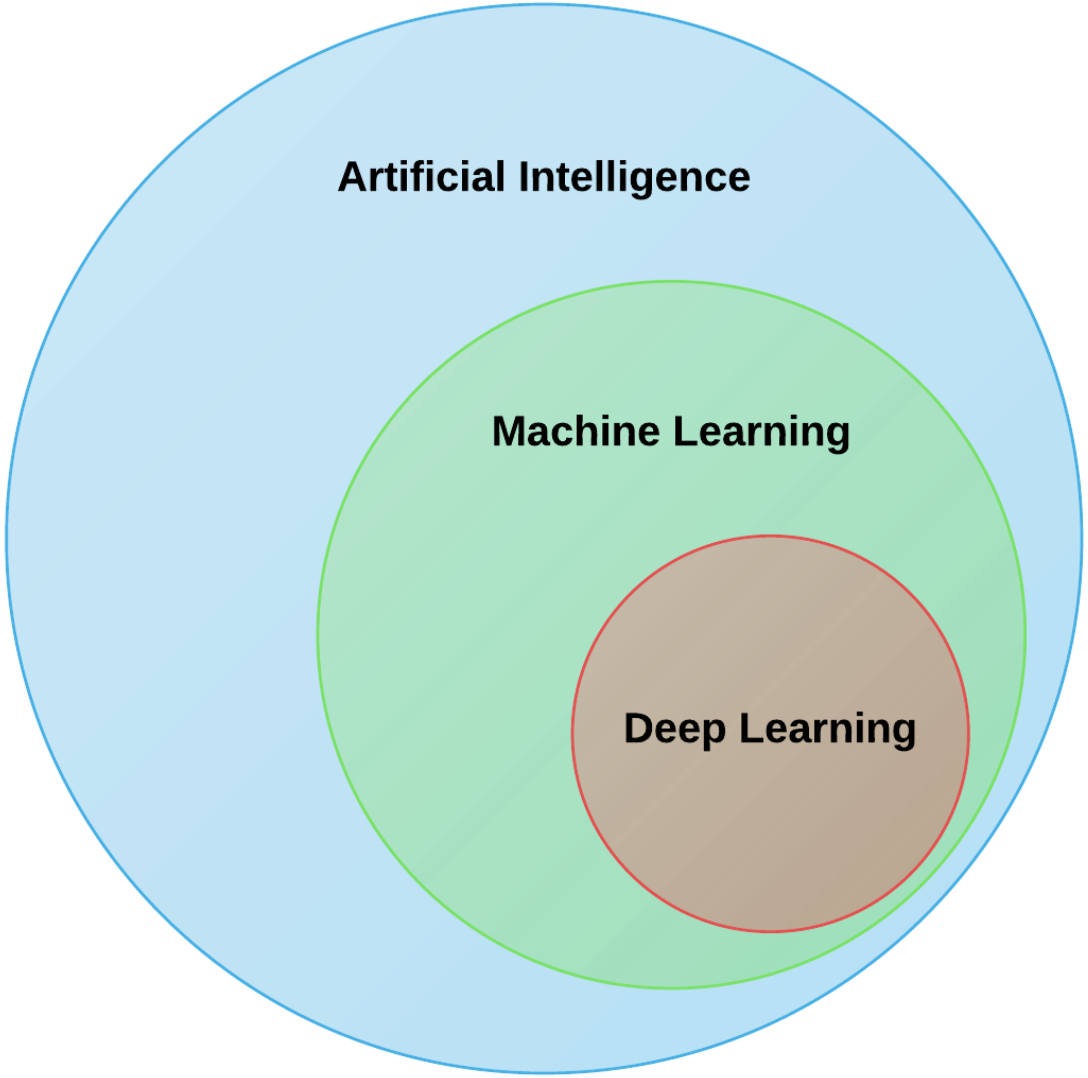
Precipitation time series:

- Non-stationary
- Non-normal distributed
- Outlies

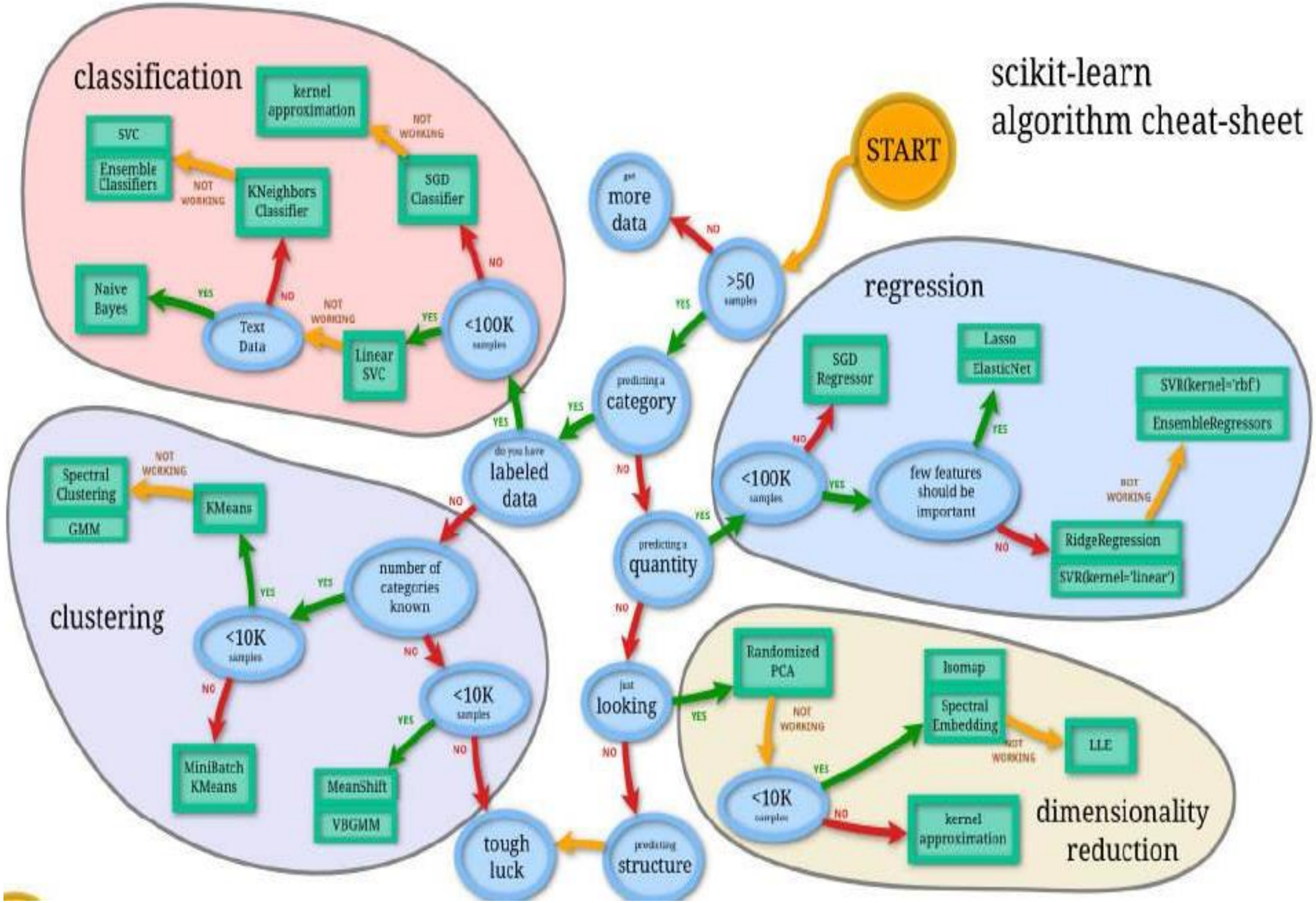
Factors:

- **many**
- **of multicollinearity**

Machine learning(ML)



	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction


scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
<100K samples
SGD Classifier
kernel approximation
Linear SVC
Text Data
Naive Bayes
KNeighbors Classifier
SVC
Ensemble Classifiers
clustering
number of categories known
<10K samples
KMeans
Spectral Clustering
GMM
MiniBatch KMeans
MeanShift
VBGMM
regression
SGD Regressor
few features should be important
Lasso
ElasticNet
RidgeRegression
SVR(kernel='linear')
SVR(kernel='rbf')
EnsembleRegressors
dimensionality reduction
Randomized PCA
Isomap
Spectral Embedding
LLE
YES
NO
NOT WORKING

Regression methods in ML

$$\hat{y}(w, x) = w_0 + w_1 x_1 + ... + w_p x_p$$

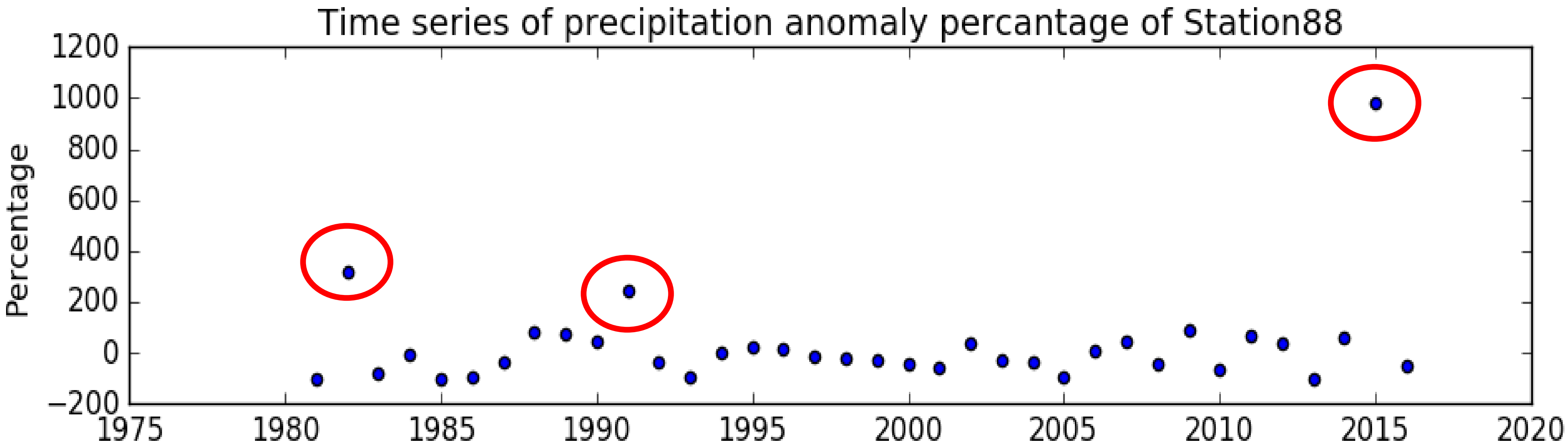
- **Ordinary Least Squares** $\underset{w}{min} \, ||Xw - y||_2{}^2$
- Ridge regression $\underset{w}{min} \, ||Xw - y||_2{}^2 + \alpha ||w||_2{}^2$
- LASSO regression $\underset{w}{min} \frac{1}{2n_{samples}} ||Xw - y||_2^2 + \alpha ||w||_1$
- Elastic Net $\underset{w}{min} \frac{1}{2n_{samples}} ||Xw - y||_2^2 + \alpha\rho ||w||_1 + \frac{\alpha(1-\rho)}{2} ||w||_2^2$
- Least Angle regression

- Robustness regression

- Support vector machines
- Polynomial regression

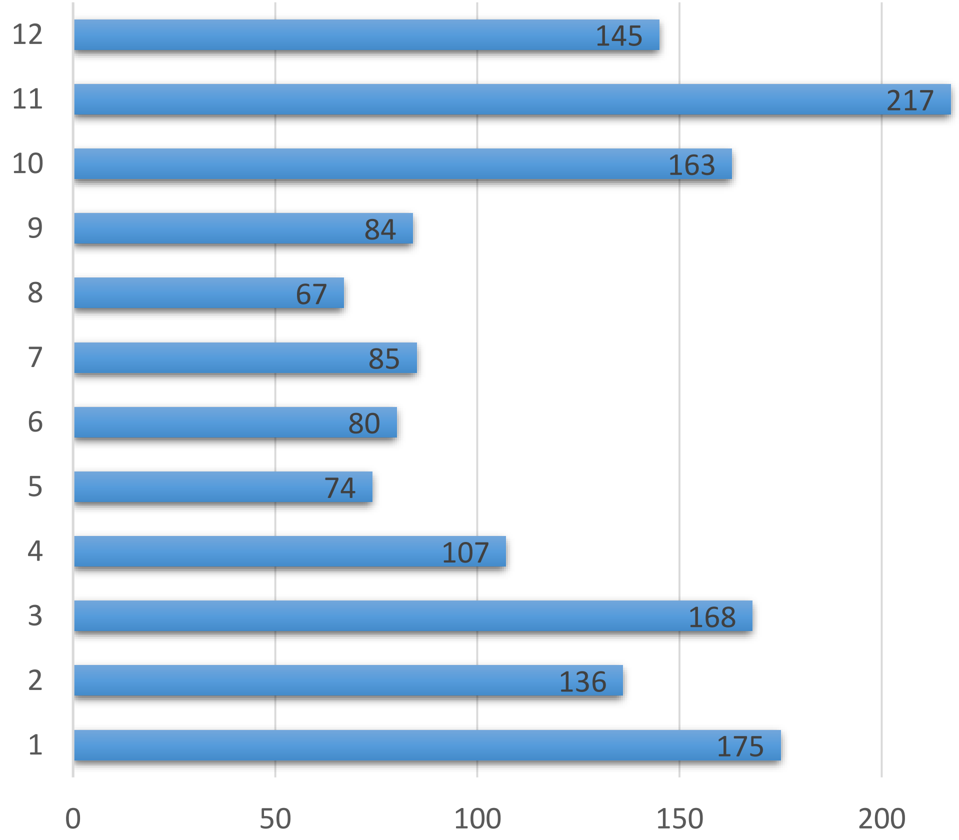
- decision trees and ensemble（random forest， Gradient Boosting Regression, xgboost）
- Neural network models (Multi-layer Perceptron)

Outliers in monthly precipitation

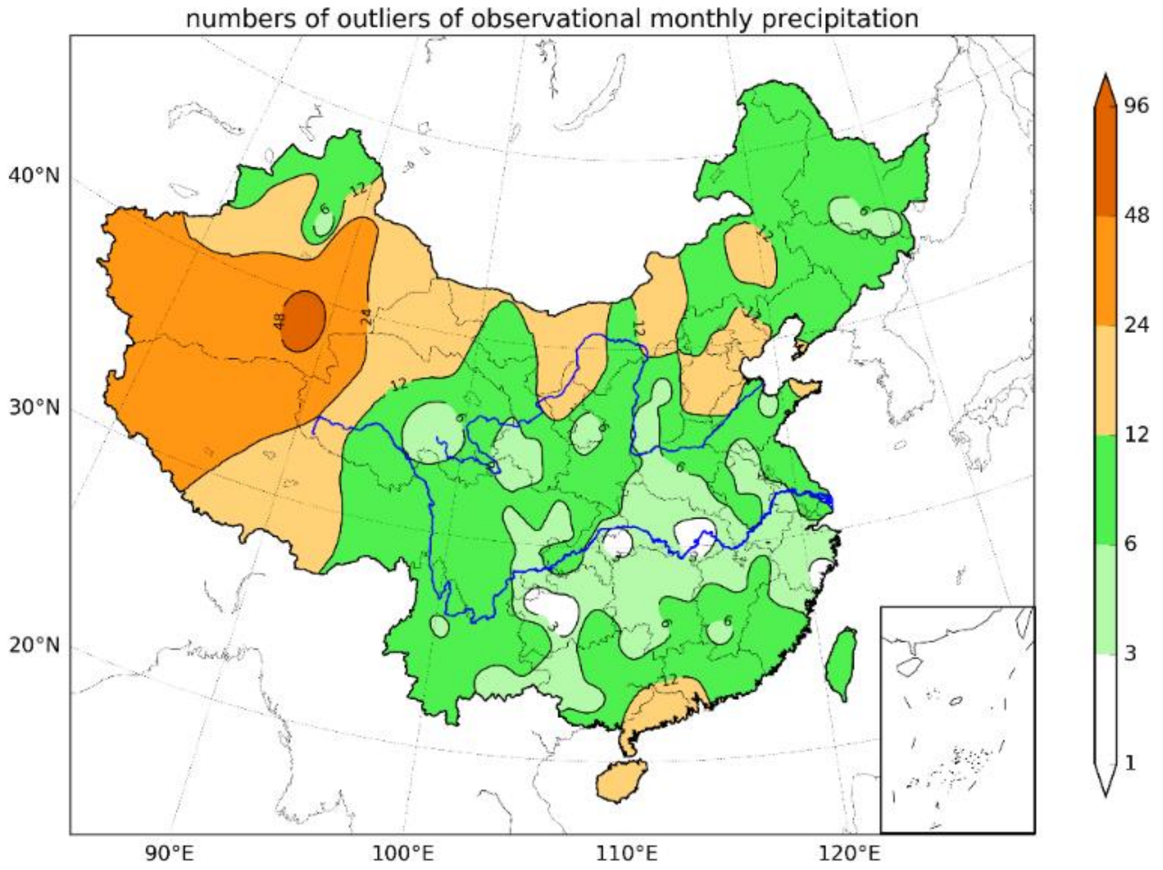


Outliers in monthly precipitation

Numbers varying with month



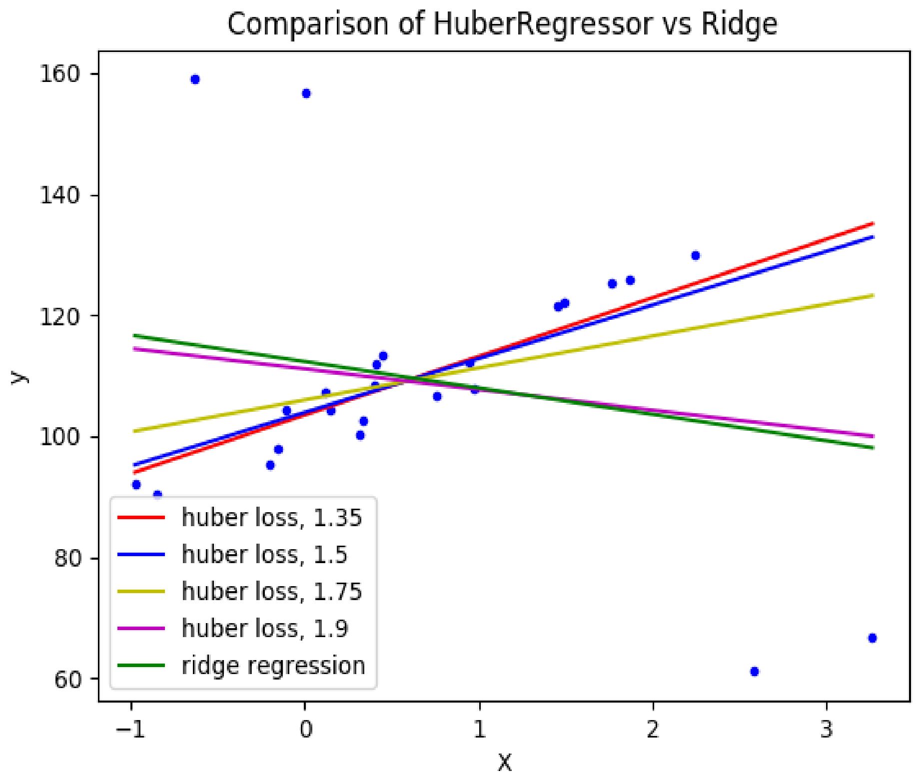
Spatial distribution



注: 1981-2016,160站

ML methods dealing with outliers

- RANSAC regression: RANdom SAmple Consensus
- Theil-Sen regression: generalized-median-based estimator
- Huber regression: it does not ignore the effect of the outliers but gives a lesser weight to them.



Multicollinearity of Factors

ENSO : Nino4, Nino3, Nino3.4， NinoZ, NinoA

Subtropical High： h1000, h500

ML methods:

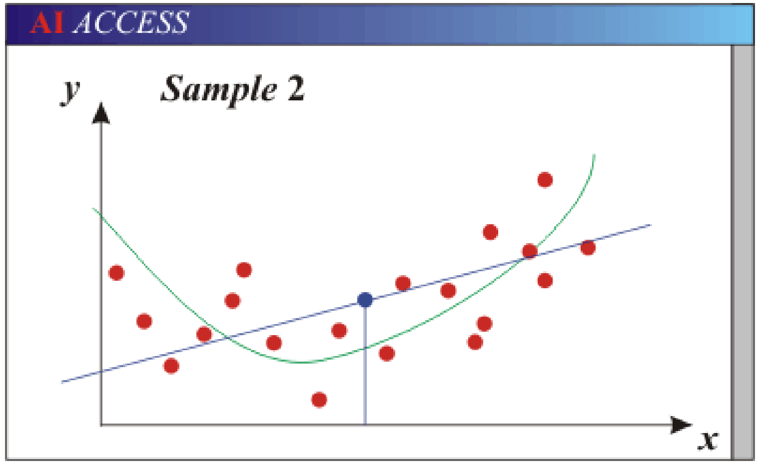
PCA: dimension reduction

LASSO, Ridge regression

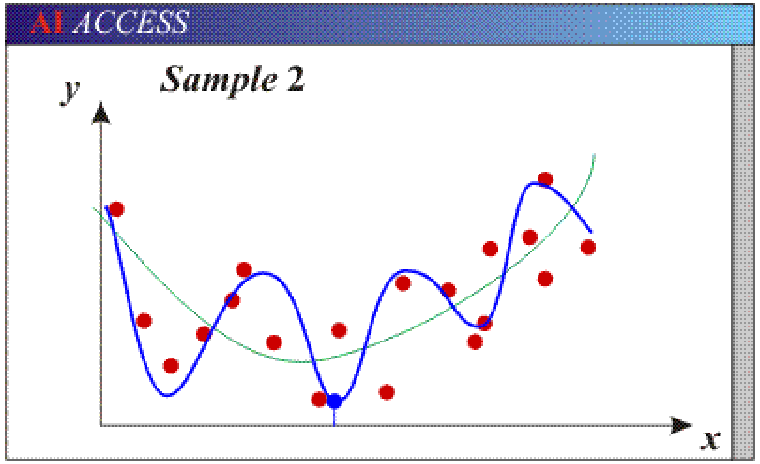
Number of factors and model's generalization

- Components of generalization error
  - **Bias:** how much the average model over all training sets differ from the true model?
    - Error due to inaccurate assumptions/simplifications made by the model
  - **Variance:** how much models estimated from different training sets differ from each other
- **Underfitting:** model is too "simple" to represent all the relevant class characteristics
  - High bias and low variance
  - High training error and high test error
- **Overfitting:** model is too "complex" and fits irrelevant characteristics (noise) in the data
  - Low bias and high variance
  - Low training error and high test error

Bias-Variance Trade-off



- Models with too few parameters are inaccurate because of a large bias (not enough flexibility).



- Models with too many parameters are inaccurate because of a large variance (too much sensitivity to the sample).

Slide credit: D. Hoiem

ML methods dealing with many factors

$$\hat{y}(w, x) = w_0 + w_1 x_1 + ... + w_p x_p$$

- **Ordinary Least Squares** $\min_{w} ||Xw - y||_2{}^2$
- Ridge regression $\min_{w} ||Xw - y||_2{}^2 + \alpha ||w||_2{}^2$
- LASSO regression $\min_{w} \frac{1}{2n_{samples}} ||Xw - y||_2^2 + \alpha ||w||_1$
- Elastic Net $\min_{w} \frac{1}{2n_{samples}} ||Xw - y||_2^2 + \alpha\rho ||w||_1 + \frac{\alpha(1-\rho)}{2} ||w||_2^2$
- Least Angle regression

- Robustness regression

- Support vector machines
- Polynomial regression

- decision trees and ensemble（random forest， Gradient Boosting Regression, xgboost）
- Neural network models (Multi-layer Perceptron)

Huber regression with penalized loss function

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p$$

The loss function that `HuberRegressor` minimizes is given by

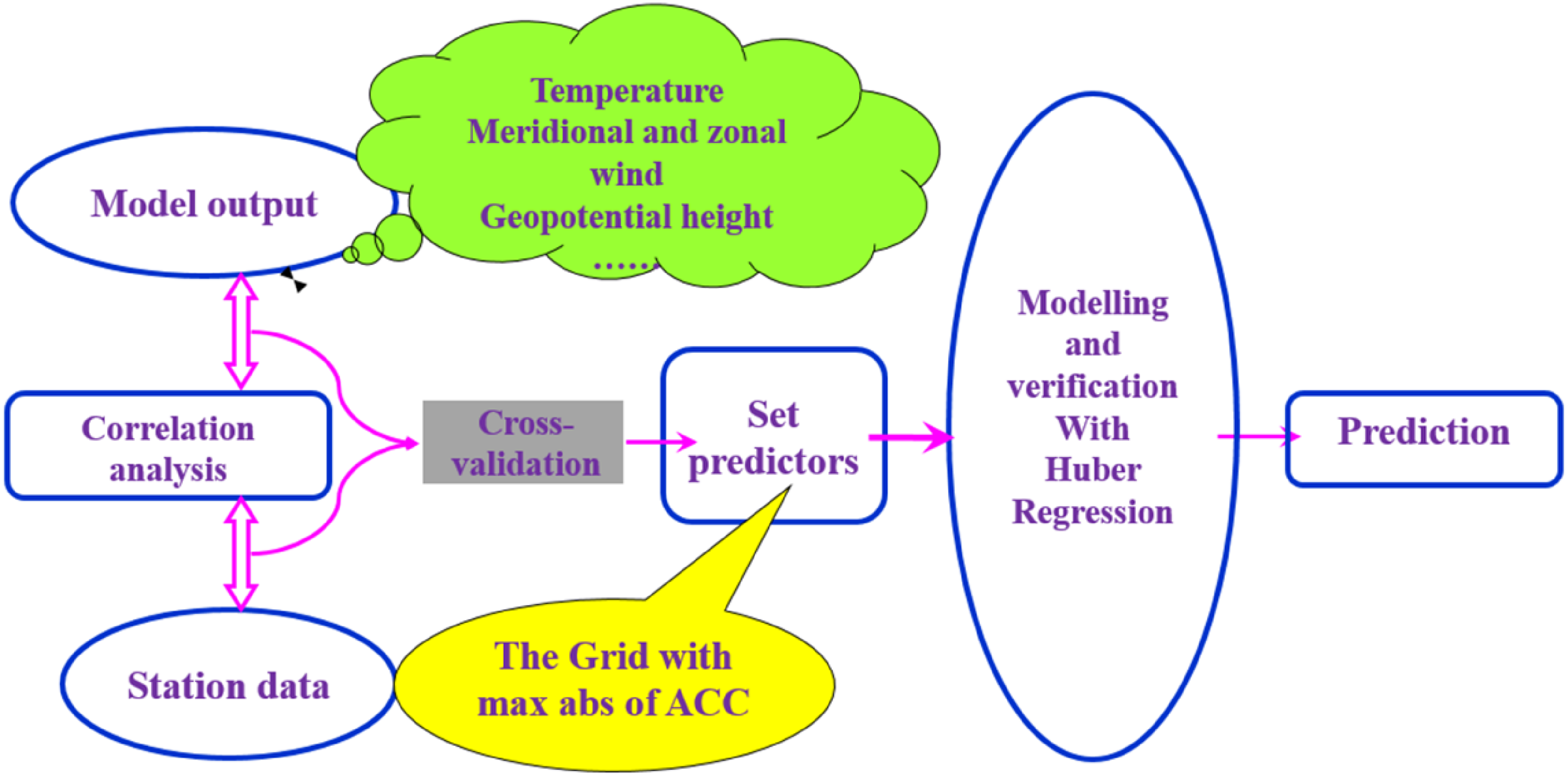
$$\min_{w,\sigma} \sum_{i=1}^{n} \left(\sigma + H_m\left(\frac{X_i w - y_i}{\sigma} \right) \sigma \right) + \alpha ||w||_2{}^2$$

where

$$H_m(z) = \begin{cases} z^2, & \text{if } |z| < \epsilon, \\ 2\epsilon|z| - \epsilon^2, & \text{otherwise} \end{cases}$$

It is advised to set the parameter `epsilon` to 1.35 to achieve 95% statistical efficiency.

Chart of the new downscaling method



Data

- Observation：

  Chinese 160 stations monthly rainfall data

  1981-2016

- Model Output：

  Seasonal prediction data of ECMWF System4 (monthly ）

  1981-2016

Two cases

Case 1:

choose the best factor of all

univariate linear regression

Case 2:

eight factors,

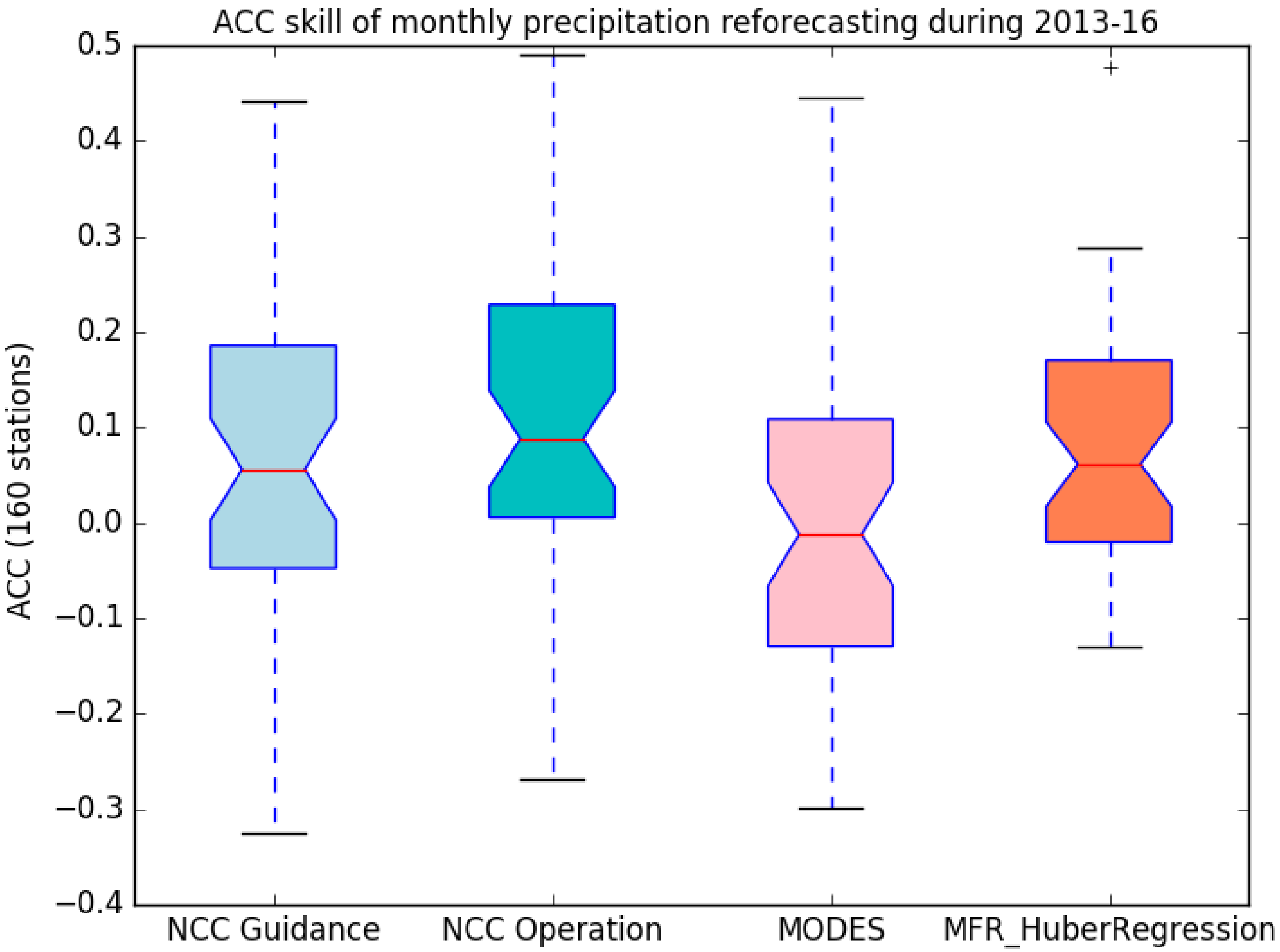
Huber regression

Reforecast skills

	ACC（160 stations)
NCC-Guidance product	**0.07**
NCC-Formal product	**0.11**
Case 1： univariate regression	**0.05**
Case2: Huber regression	**0.08**

(2013.1-16.12,Chinese 160 stations, monthly precipitation

Reforecast skills



2013.1-16.12,Chinese 160 stations, monthly precipitation

Summary

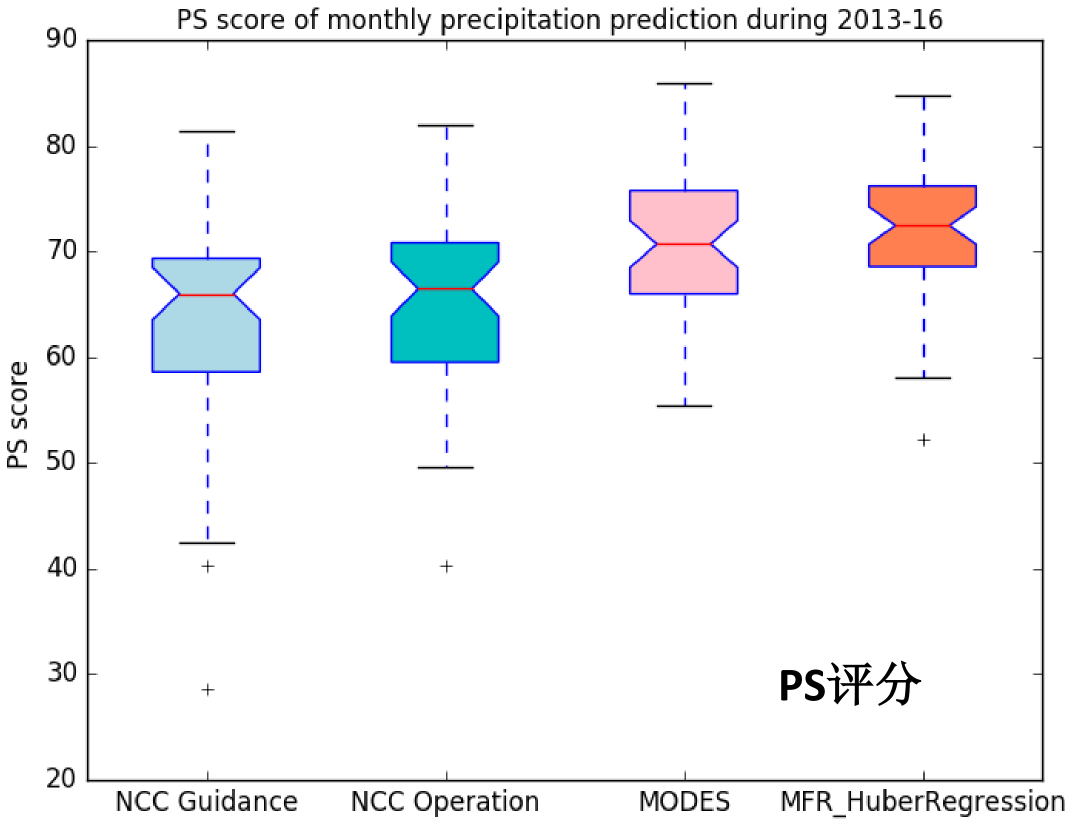
- Machine learning technique could be used to modified and improve statistical downscaling methods.

- As a new statistical downscaling method, Huber regression does well in predicting monthly precipitation over China.

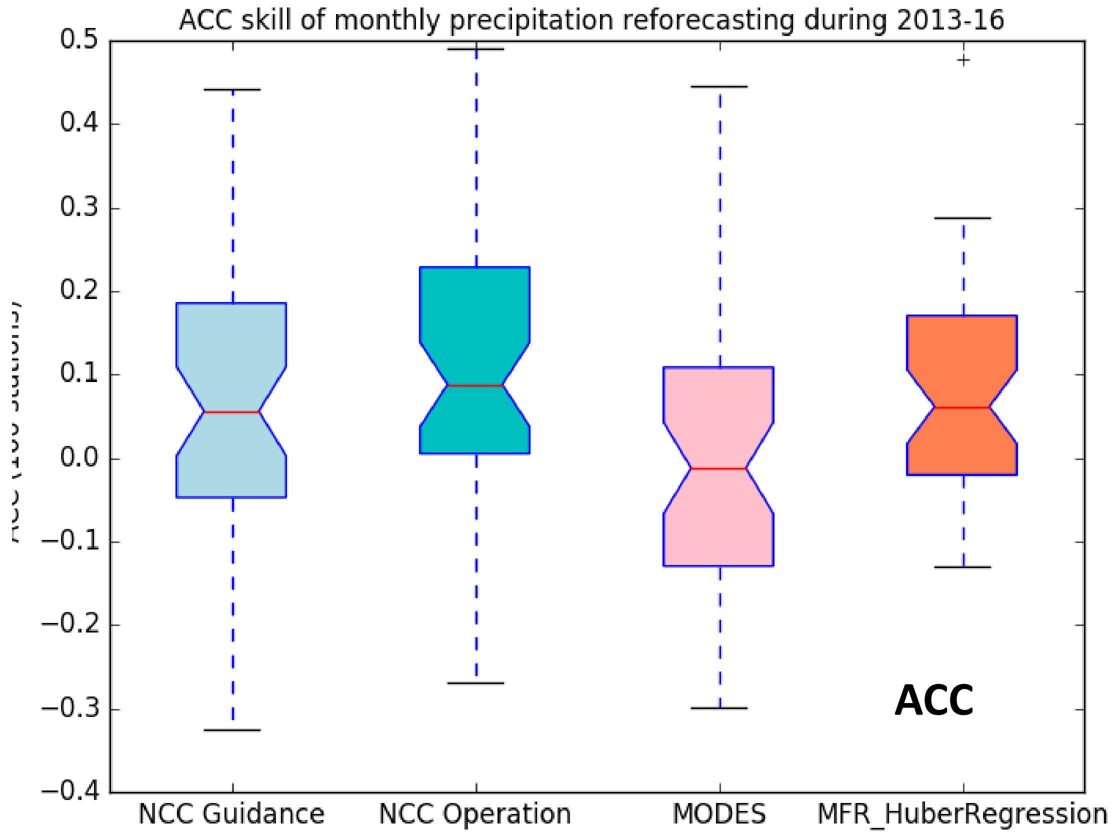
Thanks

Huber回归方法技巧评估

（2013.1–16.12我国160站月降水后报）



指导预报　发布预报　MODES　新方法



指导预报　发布预报　MODES　新方法

效果：***空间分布预测技巧显著提高***　预测技巧的下限有所提高