A data assimilation method combined with machine learning and its application to anthropogenic emission adjustment in CUACE model

Reporter: Congwu Huang

Congwu Huang, Tijian Wang, Tao Niu, Mengmeng Li, Hongli Liu, Chaoqun Ma

Nanjing University, China
Chinese Academy of Meteorological Sciences, China

Introduction & Methods

In order to improve the pollution forecasting accuracy in Chinese unified atmospheric chemistry environment (CUACE) model, a data assimilation method combined with machine learning has been developed and applied to adjust anthropogenic emissions.

- **Data assimilation method: Nudging (Kistler，1974；Hoke and Anthes，1976)**

$$\frac{\partial \alpha}{\partial t} = F(\alpha, X, t) + G_\alpha \frac{\sum_{i=1}^{N} W_i^2(X,t)\gamma_i(\alpha_0 - \hat{\alpha})_i}{\sum_{i=1}^{N} W_i(X,t)}.$$

- **Machine learning method: Extremely random trees (Geurts et al., 2006)**

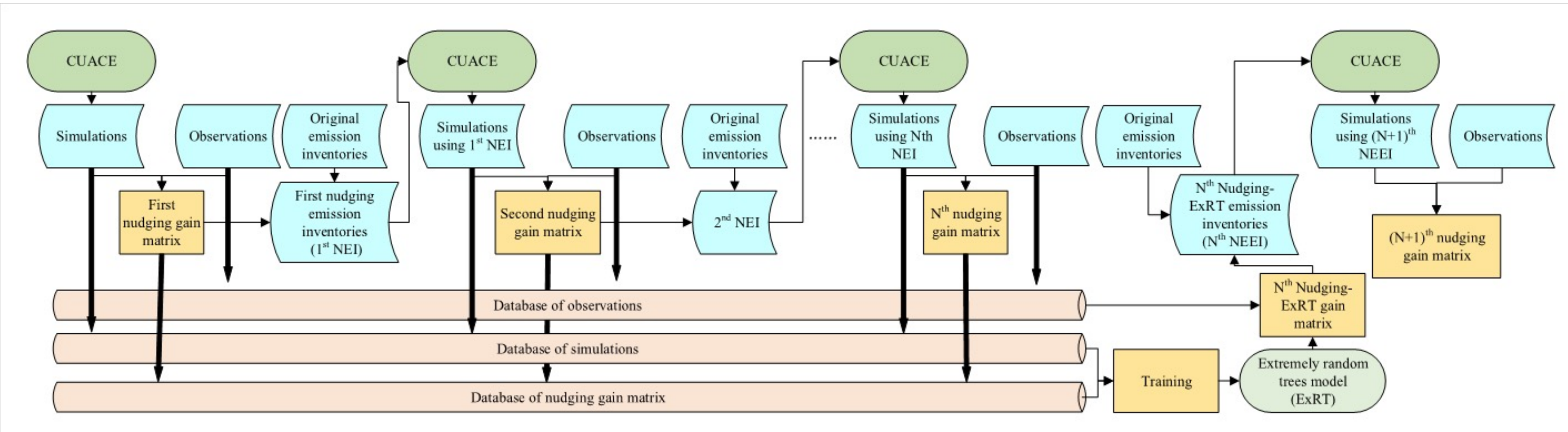
Table 1 Extra-Trees splitting algorithm (for numerical attributes)

Split_a_node(S)
Input: the local learning subset S corresponding to the node we want to split
Output: a split $[a < a_c]$ or nothing
– If **Stop_split**(S) is TRUE then return nothing.
– Otherwise select K attributes $\{a_1, \ldots, a_K\}$ among all non constant (in S) candidate attributes;
– Draw K splits $\{s_1, \ldots, s_K\}$, where $s_i =$ **Pick_a_random_split**(S, a_i), $\forall i = 1, \ldots, K$;
– Return a split s_* such that $\text{Score}(s_*, S) = \max_{i=1,\ldots,K} \text{Score}(s_i, S)$.

Pick_a_random_split(S,a)
Inputs: a subset S and an attribute a
Output: a split
– Let $a_{\max}^S$ and $a_{\min}^S$ denote the maximal and minimal value of a in S;
– Draw a random cut-point a_c uniformly in $[a_{\min}^S, a_{\max}^S]$;
– Return the split $[a < a_c]$.

Stop_split(S)
Input: a subset S
Output: a boolean
– If $|S| < n_{\min}$, then return TRUE;
– If all attributes are constant in S, then return TRUE;
– If the output is constant in S, then return TRUE;
– Otherwise, return FALSE.

Framework of the Nudging-ExRT method in CUACE



- Nudging method was used to create the database of nudging gain matrixes, using simulations of CUACE and the ground-based observations.
- These data are employed to train a machine learning model using extremely random trees method (ExRT), and to store the relations between nudging gain matrixes and simulations in the trees.

The experiment and assessment of the emission data assimilation in CUACE

Observations: ground-base stations in China

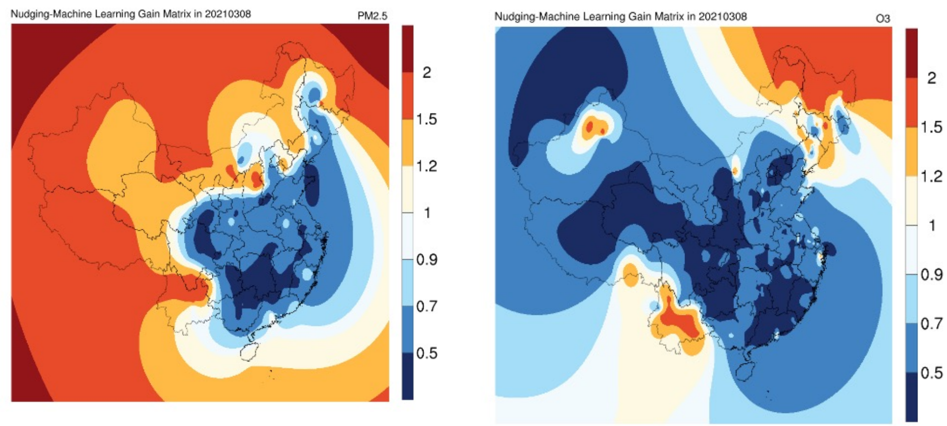
Configurations: Iterative the Nudging gain matrix in 20 o'clock everyday from March 1,2021 to March 8,2021, each iteration used 72 hours' observations and simulations.

Forecasting time periods: March 9,2021 ~ March 14,2021

Domain: China

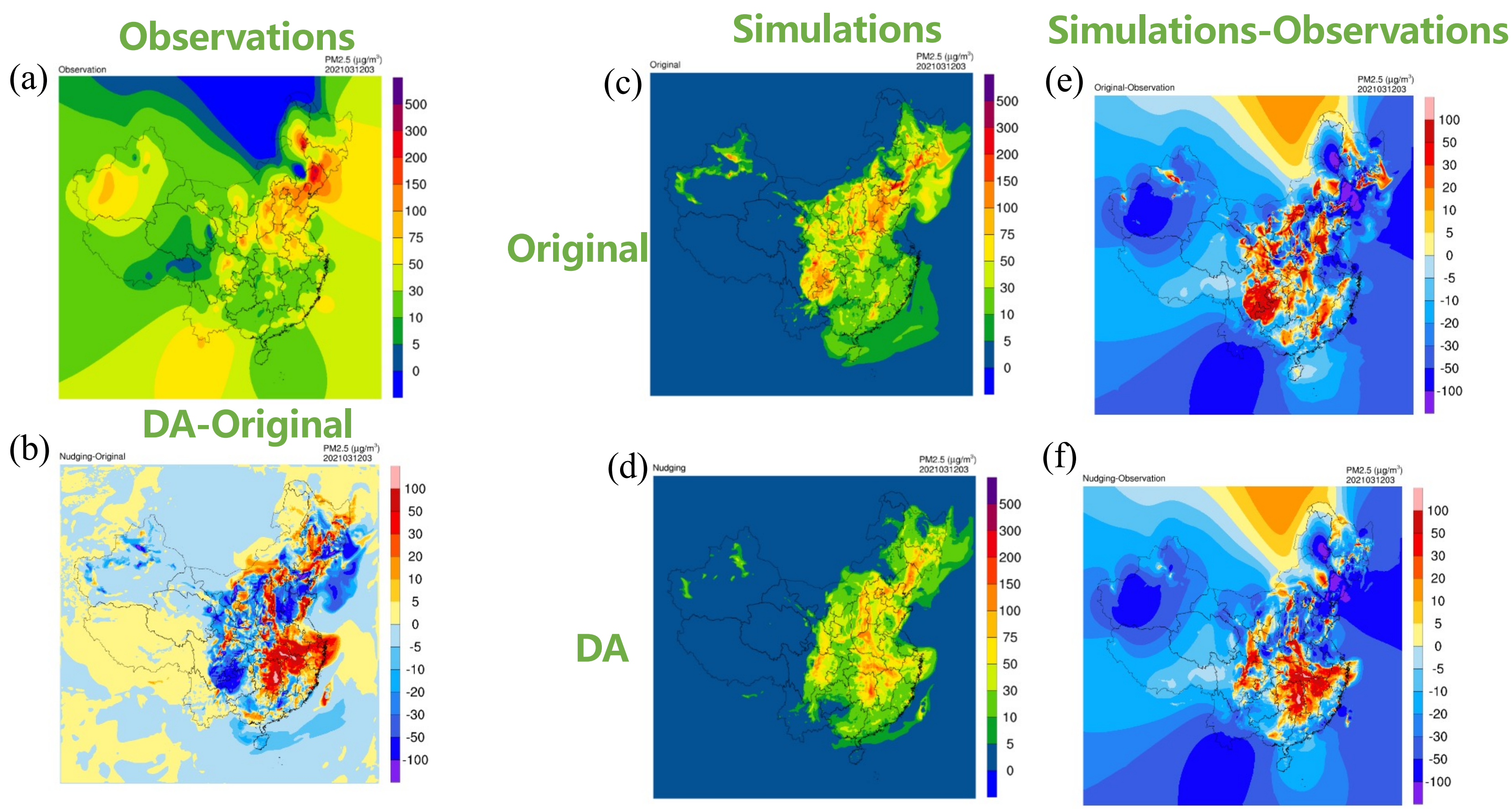
Pollutants: $PM_{2.5}$, O_3

The $PM_{2.5}$ and O_3 Nugding-ExRT gain matrix in CUCE in March 8, 2021



Nudging-ExRT emission inventories=Original emission inventories •Gain matrix

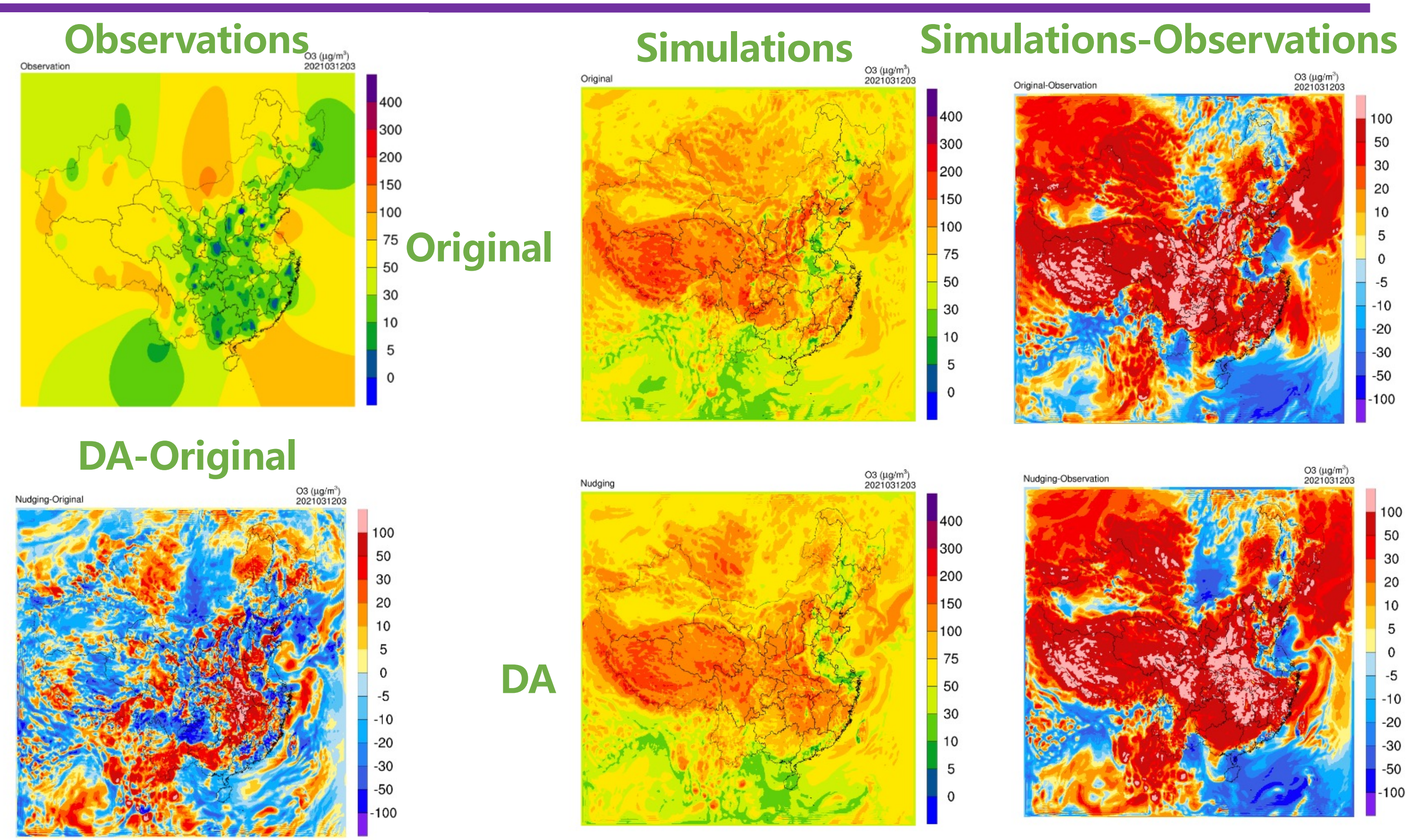
Results of emission data assimilation-$PM_{2.5}$



Results of emission data assimilation-O_3

The gain matrix of NO_2 was used to train the emissions of NO_2 and NO.

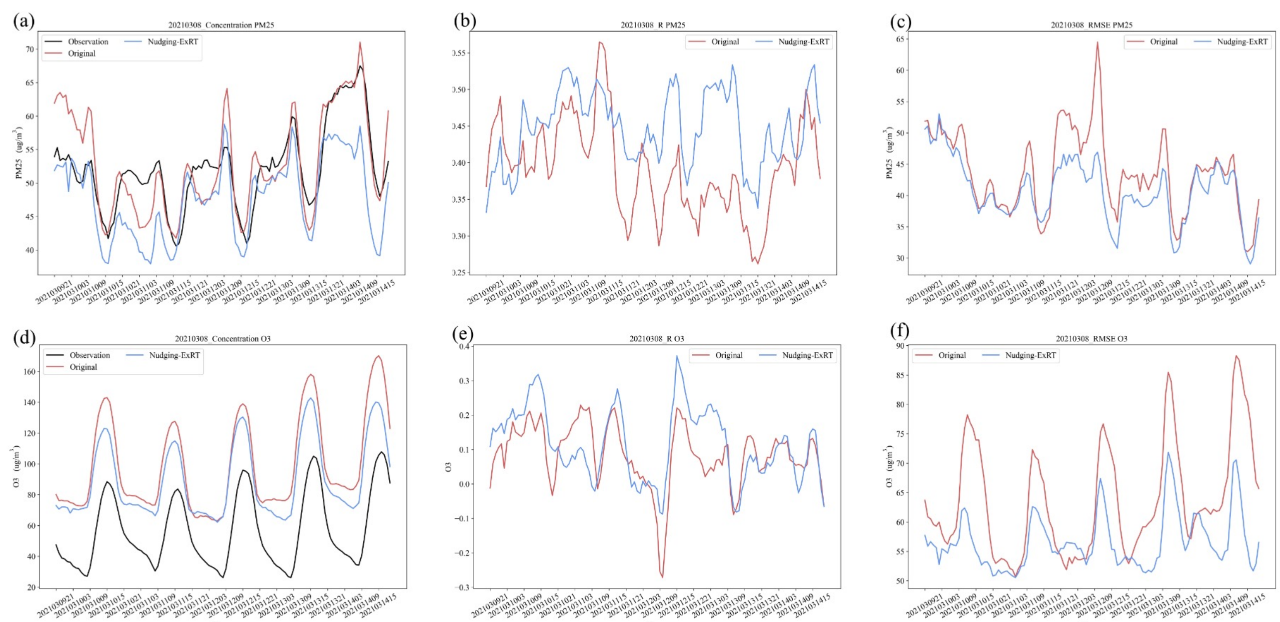
The gain matrix of NO_2 and O_3 was used to train the emissions of VOCs.



Assessment of the Nudging-ExRT emission data assimilation method

$PM_{2.5}$:
R increased from 0.39 to 0.45
RMSE decreased from 43.59 $\mu g/m^3$ to 40.71 $\mu g/m^3$.

O_3:
R increased from 0.08 to 0.11,
RMSE decreased from 63.66 $\mu g/m^3$ to 56.70 $\mu g/m^3$.



R:the hourly average spatial correlation coefficient
RMSE:the hourly average spatial root mean square error

Conclusion

- This simplicity, efficiently and extensibility framework of Nudging-ExRT method has been proved to be a good way to adjust anthropogenic emissions in CUACE.

- Nudging-ExRT method can improve the $PM_{2.5}$ and O_3 forecasting accuracy in CUACE and the optimization did not weaken even if the iterations were not in one simulation, which means the computing resources is greatly reduced using this method in the operational forecasting.

Thank You!

Email：congwuhuang@smail.nju.edu.cn

WeChat