Study on the 3DVar emission inversion method combined with machine learning in CMAQ

Reporter: Congwu Huang

Congwu Huang, Tijian Wang, Tao Niu

Hubei University
Nanjing University
Chinese Academy of Meteorological Sciences

Contents

- Introduction
- Data and Methods
- Results
- Conclusion

Major sources of uncertainty in air quality models

Initial and Boundary conditions

- global distribution
- nested simulation results
- previous simulation results

Emissions

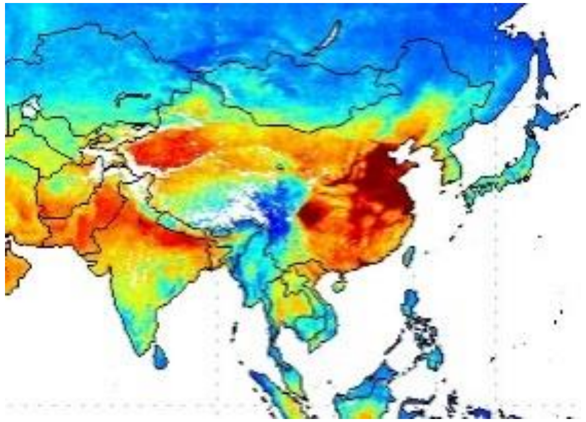
- anthropogenic emissions
- natural emissions

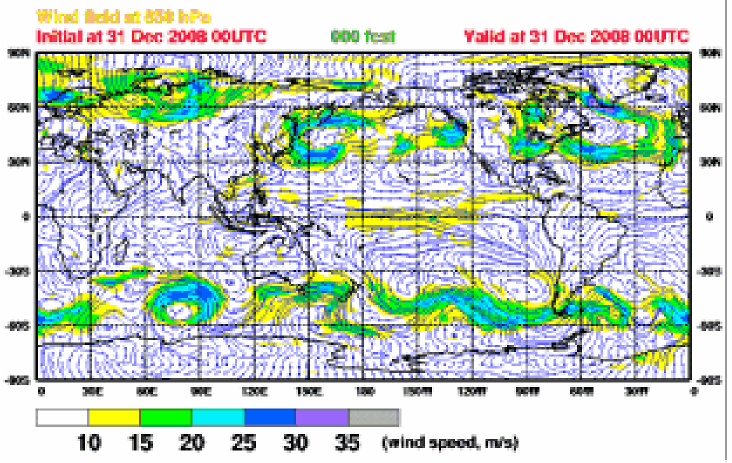
Meteorology

- Clouds
- Wind, temperature, humidity, pressure
- Planetary boundary layer height, local circulations

Processes

- chemistry
- dry deposition





Constantinescu et al., 2007a,2007b; Bocquet et al., 2015

Emission inversion

- 4DVAR and EnKF are two main methods that usually used to adjust emissions.

- Nudging is a relatively simpler method in emission data assimilation, but it can not deal with nonlinear problem or lack of observation

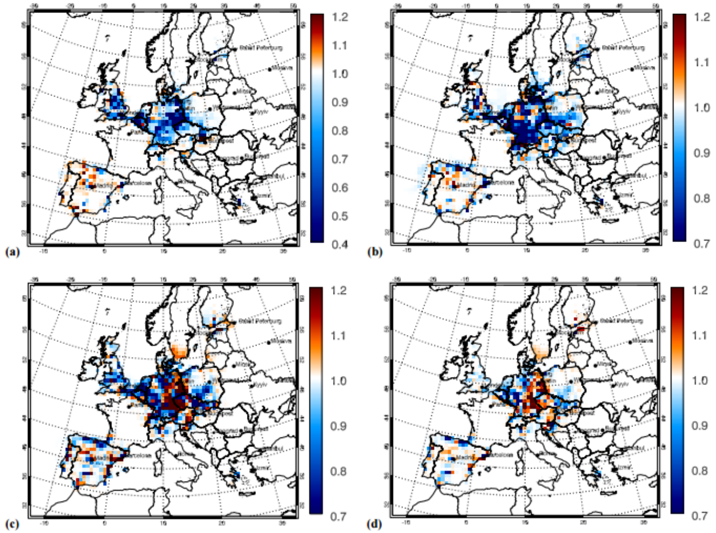


Fig. 10. Emission correction factors for **(a)** sulfur dioxide, **(b)** nitrogen dioxide, **(c)** terminal alkenes, and **(d)** isoprene at the surface layer, analysed by joint initial value/emission rate optimisation with 24 h assimilation interval placed at 17 August 1997.

4DVAR

Elbern et al,(2007)

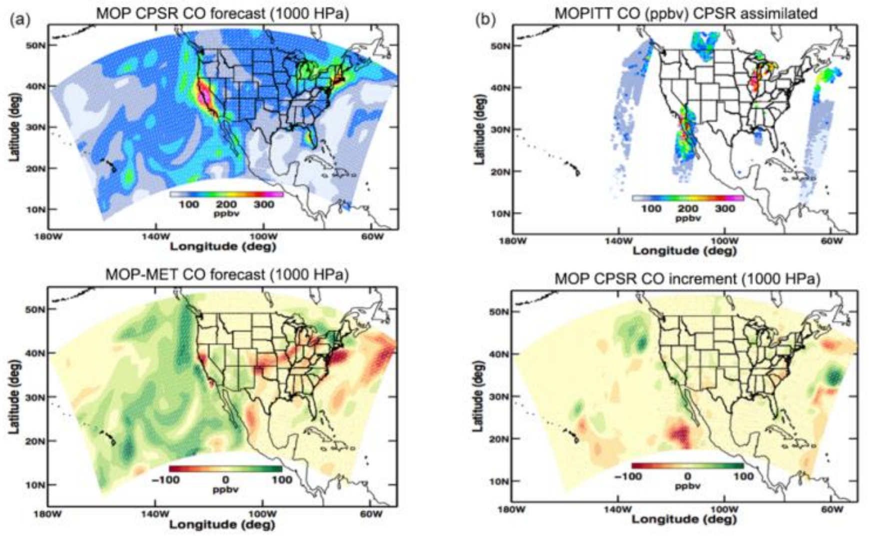


Figure 5. (a) Same as Fig. 1a except for the MOP CPSR experiment and the middle panel from Fig. 1a, the MET DA experiment is not plotted. **(b)** Same as Fig. 1b except for the MOP CPSR experiment.

EnKF

Mizzi et al,(2016)

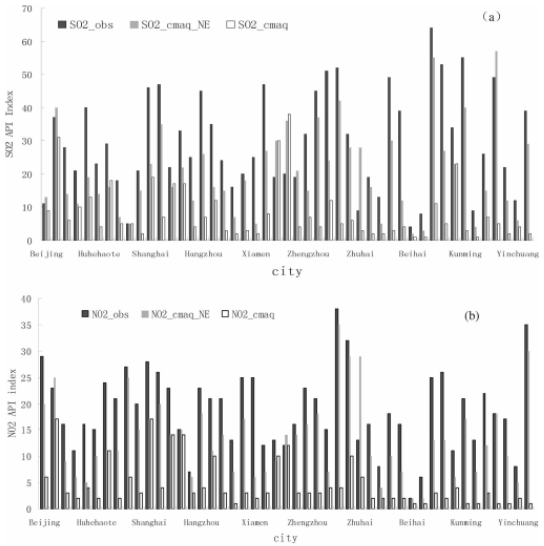


FIG. 2. Comparison of the 1–25 August mean SO_2 and NO_2 API values among the control experiment, nudging experiment, and the observations for the (a) SO_2 API and (b) NO_2 API. The APIs are computed for all 47 cities. It is obvious that the NO_2 and the SO_2 API values from the nudging experiment (light shade) are closer to the observations (dark shade) than are those from the control experiment (white) in nearly all cities.

Nudging

Cheng et al. (2008)

Machine learning

- Machine learning can improve the air quality forecasting accuracy significantly.
- Machine learning can deal with the nonlinear problem.
- In most cases, the effect of machine learning is increased with the growth of database.

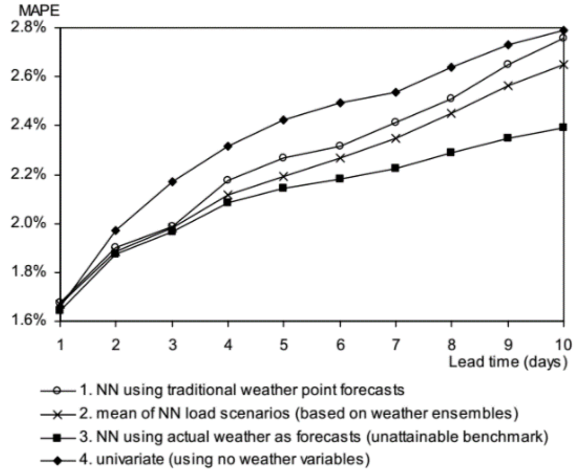


Fig. 3. *MAPE* for load point forecasts for post-sample period 1 November 1998 to 30 June 2000.

Taylor et al.,(2002). IEEE Transactions on Power Systems.

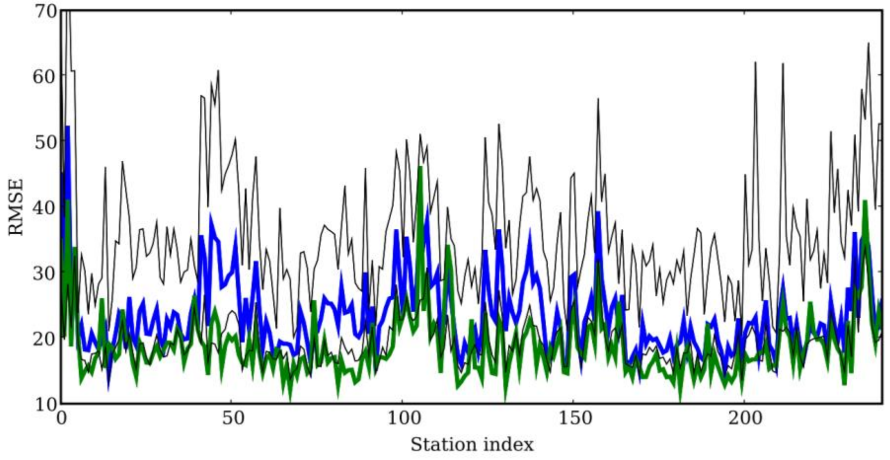


Figure 7. RMSE against the station index (for 241 stations). In green, $\mathcal{R}^{\beta}_{1000}$; in blue, the best model (over all stations); in black, the best model and the worst model for the station.

Mallet, V. et al.,(2009). Journal of Geophysical Research: Atmospheres

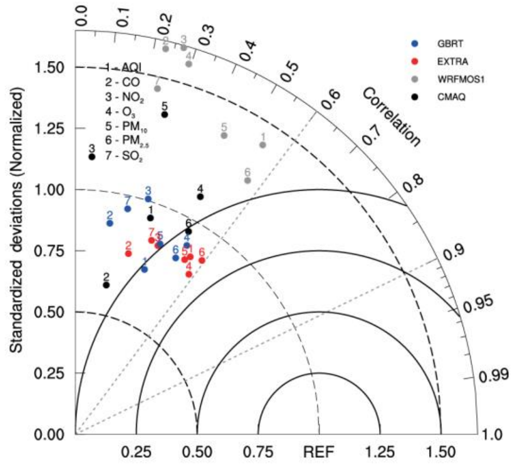


Fig. 4 Taylor diagram of concentrations of six pollutants and AQI from observations and outputs of CMAQ, WRFMOS1, EXTRA and GBRT during 13—24 September, 2016 in HHXC, Xuzhou

Huang. et al.,(2018). Acta Meteorologica Sinca

Contents

- Introduction
- Data and Methods
- Results
- Conclusion

Data and Methods

WRFv3.7.1

CMAQv5.3.2

Simulation period: 2 to 30 January in 2019

Resolution: 81km*81km, 27km*27km, 9km*9km

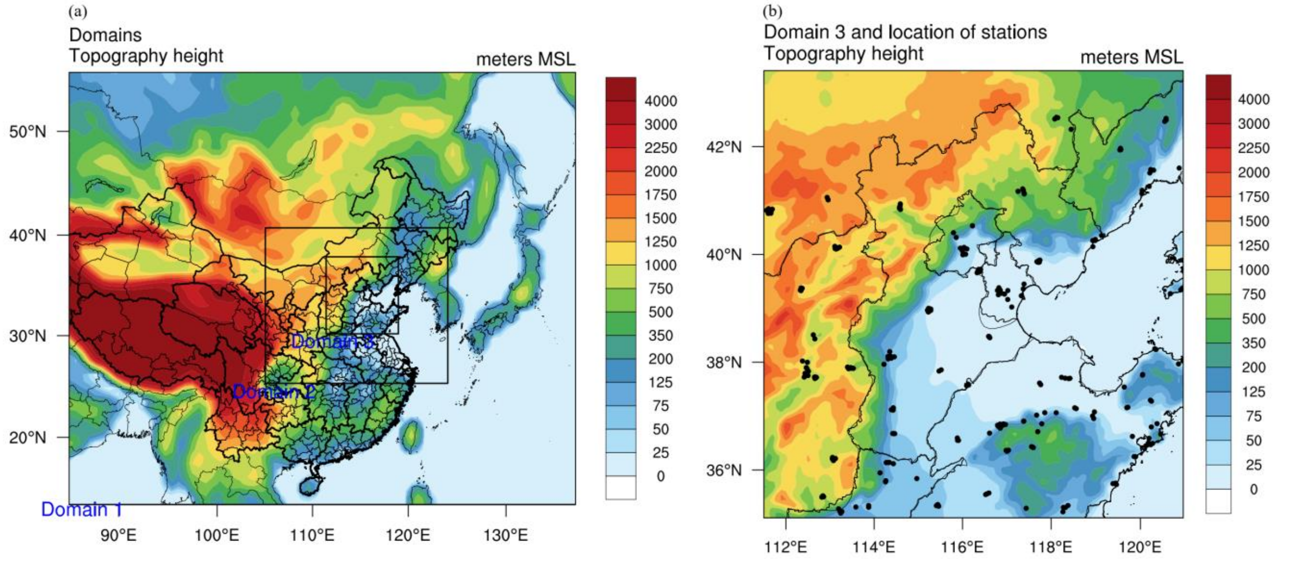
Pollutants: $PM_{2.5}$, O_3

Observation: $PM_{2.5}$,O_3 and NO_2 from 255 air quality stations in BTH region

Emission inventory: MEIC 2017

a. WRFv3.7.1			
Simulation period	3–30 January 2019		
Vertical resolution	33 vertical levels		
Microphysics scheme	WSM 3-class simple ice scheme		
Boundary layer scheme	YSU scheme		
Surface layer scheme	MM5 scheme		
Land-surface scheme	Unified Noah land-surface model		
Longwave radiation scheme	rrtm scheme		
Shortwave radiation scheme	Dudhia scheme		
Grid-nudging fdda	on		
Domain center	39.1248°N, 116.5657°E		
Domain id	1	2	3
Domain size	64×75	69×81	102×96
Starting IJ-indices from the parent domain	×	(30, 19)	(38, 23)
Horizontal resolution	81km	27 km	9km

a. CMAQv5.3.2			
Horizontal advection	Yamo		
Vertical advection	WRF		
Horizontal diffusion	Multiscale		
Vertical diffusion	ACM2		
Deposition	M3Dry		
Chemistry solver	EBI		
Aerosol module	AERO7		
Cloud module	ACM		
Mechanism	cb6r3_ae7_aq		
Domain id	1	2	3
Domain size	62×73	67×79	100×94



3DVar data assimilation method for initial conditions

$$J(x)=\frac{1}{2}(x-x_b)^T B^{-1}(x-x_b)+\frac{1}{2}(Hx-y)^T R^{-1}(Hx-y) \quad (1)$$

$$\downarrow$$

$$J(\delta x)=\frac{1}{2}(\delta x)^T B^{-1}(\delta x)+\frac{1}{2}(H\delta x-d)^T R^{-1}(H\delta x-d) \quad (2)$$

innovation $\delta x = x - x_b$

R: The given measurement instrument error and representative error obtained through spatial allocation

① $\varepsilon_R=\gamma\varepsilon_0\sqrt{\frac{\Delta l}{L}}$

B:*NMC* （Error between 24 and 48 hour simulations）

② $B=DCD^T$； $C=C_x\otimes C_y\otimes C_z$； $\delta x=DC^{\frac{1}{2}}\delta p$

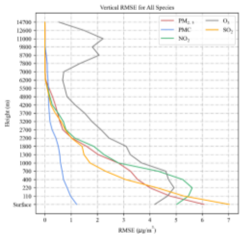
$$J(\delta p)=\frac{1}{2}(\delta p)^T\delta p+\frac{1}{2}\left(HDC^{\frac{1}{2}}\delta p-d\right)^T R^{-1}\left(HDC^{\frac{1}{2}}\delta p-d\right)$$

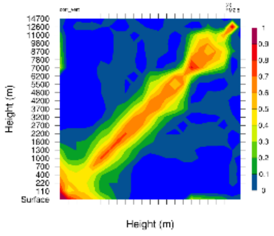
③ *H*： transformation matrix for linear interpolation of the model field to the location of the observation points

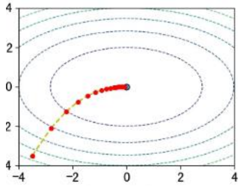
gradient descent method to solve the minimization process

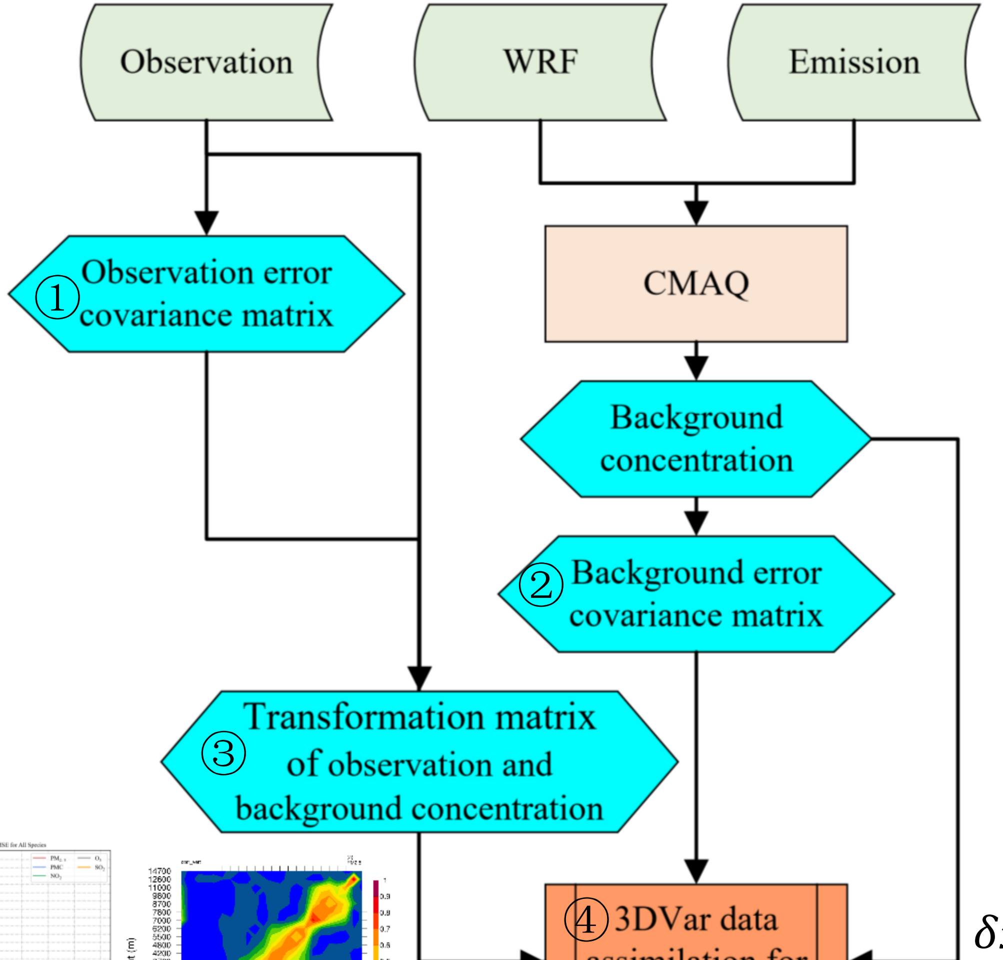
gradient： $\nabla J(\delta x)=B^{-1}(\delta x)+H^T R^{-1}(H\delta x-d)$

④
$$\delta x_{k+1}=\delta x_k+\rho_k s_k$$
$$s_k=-\nabla J(\delta x_k)$$

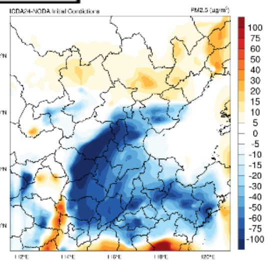








δx_a



$x_a = x_b + \delta x_a$

Extended 3DVar to emission inversion method based on machine learning

From (1)： $J(e_t)=\frac{1}{2}(e_t-e_{tb})^T B^{-1}(e_t-e_{tb})+\frac{1}{2}(H_t^e e_t-y_t)^T R^{-1}(H_t^e e_t-y_t)$ (3)

$H_t^e=H_t^o S$； $\delta c_t=H_t^e e_t-H_{tb}^e e_{tb}\approx H_t^{e\prime}(e_t-e_{tb})=H_t^{e\prime}\delta e_t$； $d_t=y_t-H_{tb}^e e_{tb}$

$$J(\delta e_t)=\frac{1}{2}(\delta e_t)^T B^{-1}(\delta e_t)+\frac{1}{2}\left(H_t^{e\prime}\delta e_t-d_t\right)^T R^{-1}\left(H_t^{e\prime}\delta e_t-d_t\right) \qquad (4)$$

In order to solving the Minimization of $J(\delta e_t)$, S'^T, the tangent linear adjoint mode of air quality model need to be developed,We use machine learning to replace S'^T.

$$H_t^{e\prime T}=S'^T H_t^{o\prime T}=M_{\delta c_t} H_t^{o\prime T}$$

$$\nabla J(\delta e_t)=B^{-1}(\delta e_t)+H_t^{e\prime T} R^{-1}\left(H_t^{e\prime}\delta e_t-d_t\right)=B^{-1}(\delta e_t)+M_{\delta c_t} H_t^{o\prime T} R^{-1}\left(H_t^{o\prime}\delta c_t-d_t\right) \qquad (5)$$

S:air quality model ;H_t^e : the transformation matrix between the emission intensity and the observation

Emission E $\xrightarrow{\text{air quality model } S}$ Simulated concentration C

Emission innovation δE $\xleftarrow[\text{machine learning model } M]{}$ Concentration innovation δC

$$\delta C_A=[\delta c_1 \quad \cdots \quad \delta c_t]$$
$$\delta E_A=[\delta e_1 \quad \cdots \quad \delta e_t]$$

$\delta E_A=M(\delta C_A)$

S'^T *means* δC *to* δE

$M\approx S'^T$

$$\boxed{H_t^{e\prime T}\approx M_{\delta c_t} H_t^{o\prime T}}$$

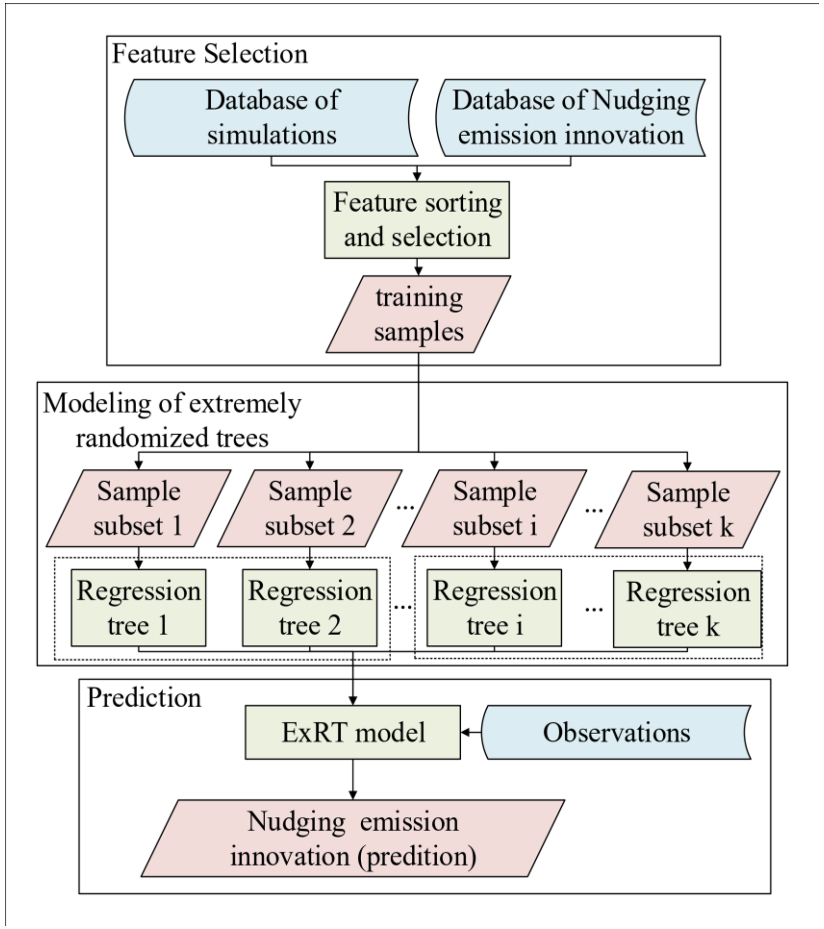
M: *machine learning model*

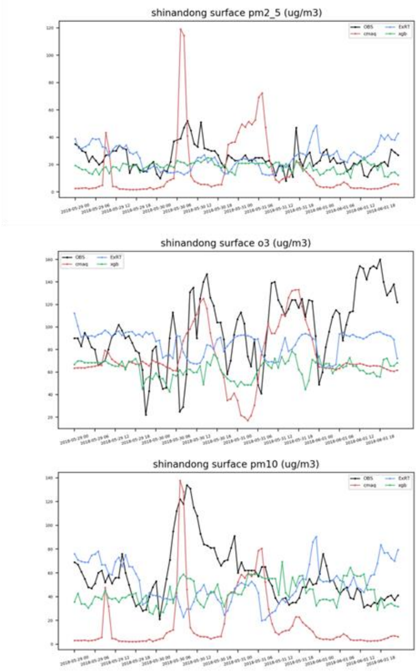
The H^{T} for δx in 3DVar initial data assimilation Changed to $M_{\delta c_t} H_t^{o\prime T}$ for δe_t in 3DVar emission inversion combined with machine learning.

δC_A： All the simulated concentration innovation in database

δE_A： Emission innovation in one-to-one correspondence with δC_A

Why Extremly Random Trees?

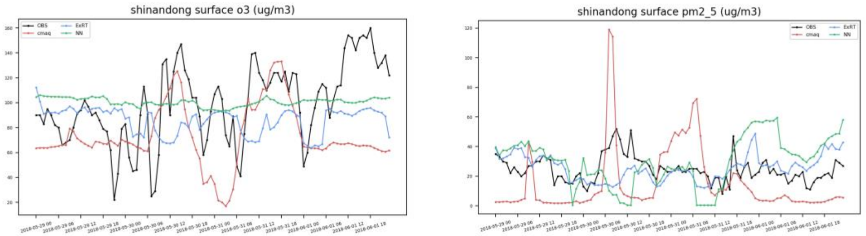




1.The comparison of ExRT and XGBoost in WRFCHEM-MOS in Qingdao

2.The comparison of ExRT、GBRT and Adaboost in CMAQ-MOS in Qingdao

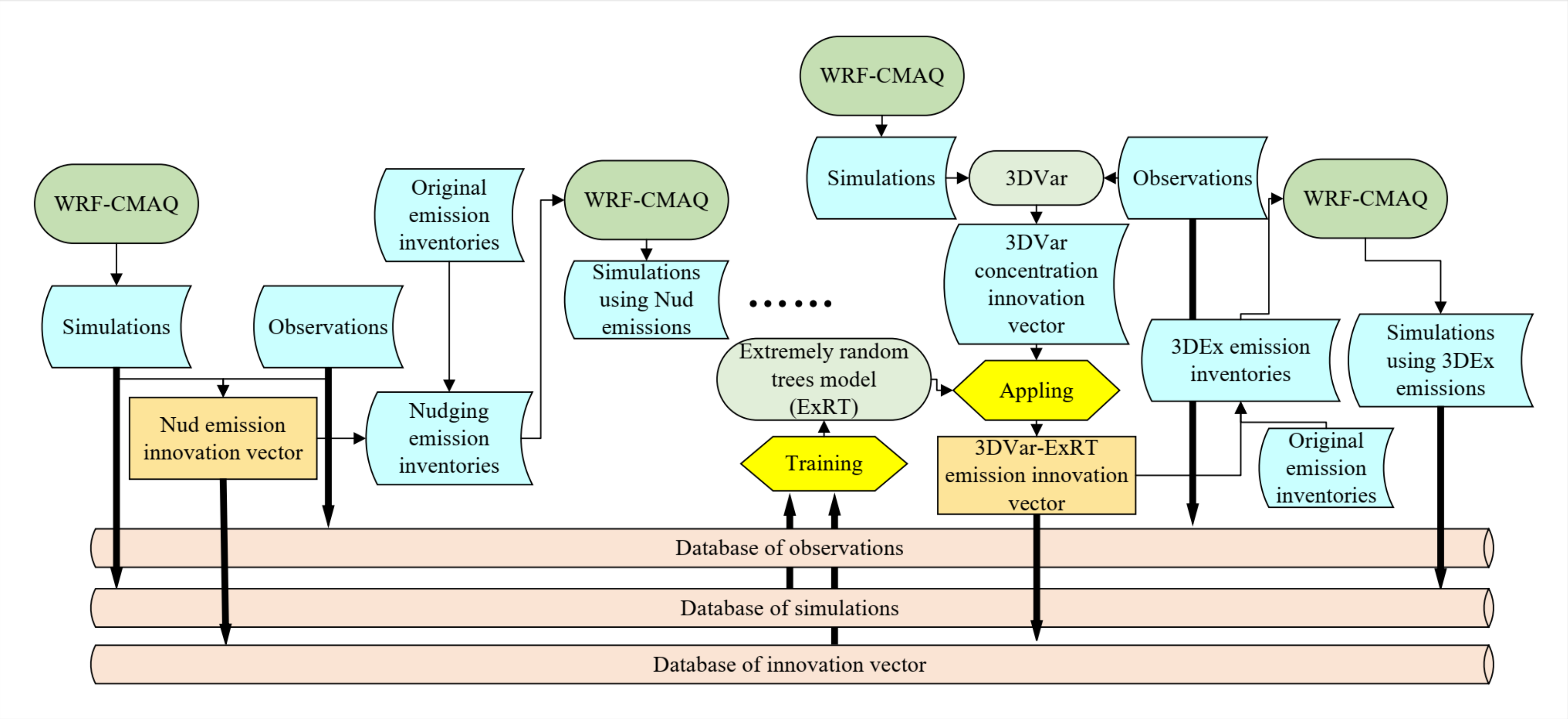
RMSE	SO_2	NO_2	PM_{10}	CO	O_3	$PM_{2.5}$
CMAQ	106.23	59.84	137.85	0.36	34.28	105.45
MOS_EXTRA	14.04	19.06	48.06	0.21	44.65	23.00
MOS_GBRT	15.85	19.26	48.85	0.21	44.45	21.58
MOS_ADA	13.91	19.81	50.24	0.21	47.55	23.03



3.The comparison of ExRT and BPNN in CMAQ-MOS in Qingdao

ExRT has been proved to be a stable and efficient method comparing to other methods in our previous studies in improving air quality forecasting accuracy using model output statistics method.

Framework of the 3DVar-ExRT method



- Nudging method was used to create the basic database of simulations and innovations, using simulations of CMAQ 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 innovation vector and simulations in the trees.

The emission data assimilation in operational forecast of CMAQ

- Using Nudging method to calculate the hourly emission innovation vector from 2 to 14 January, 2019 and create the database of observation, simulation and innovation vector . The model restart every 24 hours and store the data of observation, simulation and innovation vector in the database.

- Using Nudging (Nud) and 3DVar-ExRT (3DEx) method to adjust anthropogenic emissions of PM2.5, VOCs and NO_x from 15 to 30 January,2019.

- Every 24 hours, all the simulation of $PM_{2.5}$, O_3 and NO_2 and innovation vector of $PM_{2.5}$, VOC and NO_x data in the database was used to train the $PM_{2.5}$ and VOCs machine learning model.

- We use the machine learning model and 72h observation before model restart to get the 3DEx innovation vector for the next day.

- Notice:NO_2 observations and simulation was used to calculate the emission of NO_x

O_3 observations and simulation was used to calculate the emission of VOCs in Nud

O_3 and NO_2 observations and simulation was used to calculate the emission of VOCs in 3DEx

Contents

- Introduction
- Data and Methods
- Results
- Conclusion

Daily change of $PM_{2.5}$, VOC and NO_x emission inventories

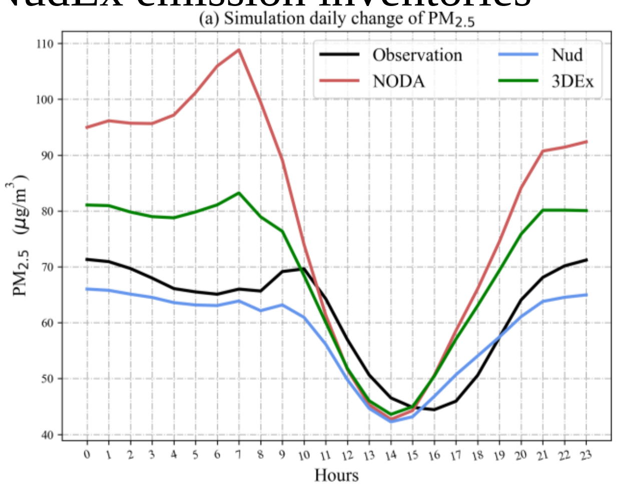
NO_2 simulations, observations and NO_x emission inventories was used in the ExRT model to calculate the 3DEx innovation vector of NO_x.

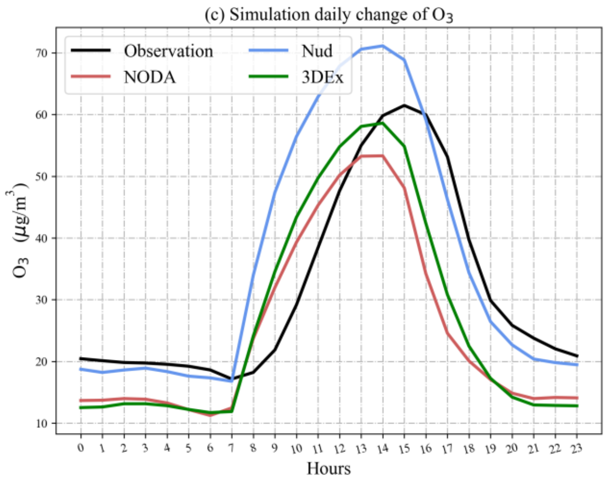
O_3 and NO_2 simulations, observations and VOC emission inventories was used in the ExRT model to calculate the 3DEx innovation vector of VOCs.

Nud method can only consider direct reaction between O3 and VOCs
3DEx can consider all direct and indirect reactions into the nonlinear ExRT model

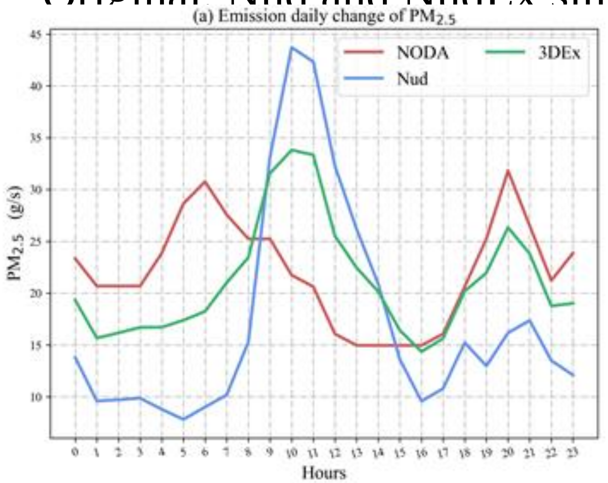
1	$NO_2 + hv + O_2 \rightarrow O_3 + NO$
2	$NO + O_3 \rightarrow NO + O_2$
3	$RH + OH + O_2 \rightarrow RO_2 + H_2O$
4	$HO_2 + NO \rightarrow OH + NO_2$
5	$RO_2 + NO + O_2 \rightarrow \varphi \cdot HO_2 + \varphi \cdot R'CHO + \varphi \cdot NO_2 + (1-\varphi) \cdot RONO_2$
6	$O_3 + hv + H_2O \rightarrow 2 \cdot OH + O_2$
7	$O_3 + OLE \rightarrow$ products
8	$O_3 + OH \rightarrow HO_2 + O_2$
9	$O_3 + HO_2 \rightarrow OH + 2 \cdot O_2$
10	$NO_2 + OH \rightarrow HNO_3$

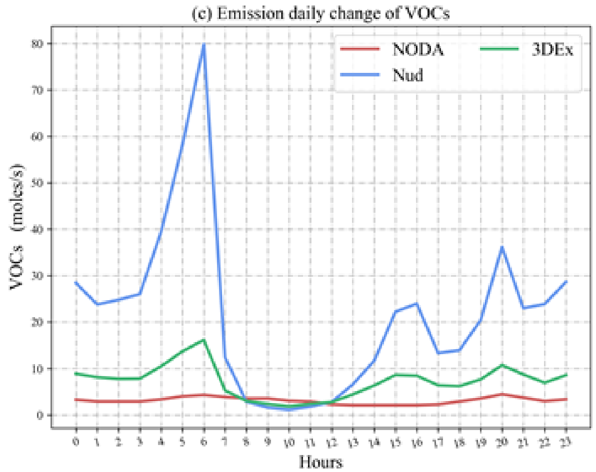
Simulation Daily Change of $PM_{2.5}$, O_3 using Original, Nud and NudEx emission inventories



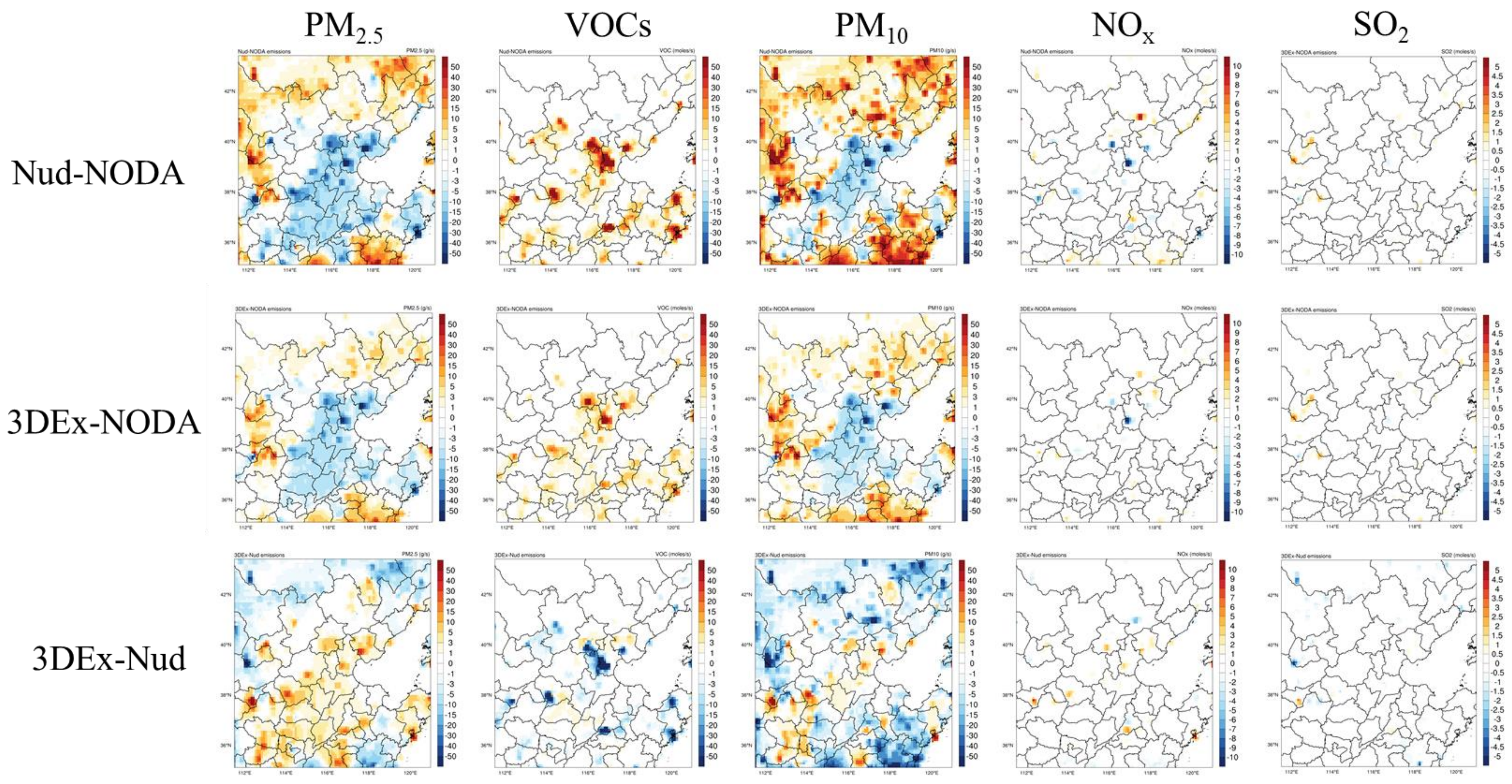


Emission Daily Change of $PM_{2.5}$, VOCs emission inventories in Original, Nud and NudEx simulations





The emission changes before and after emission inversion

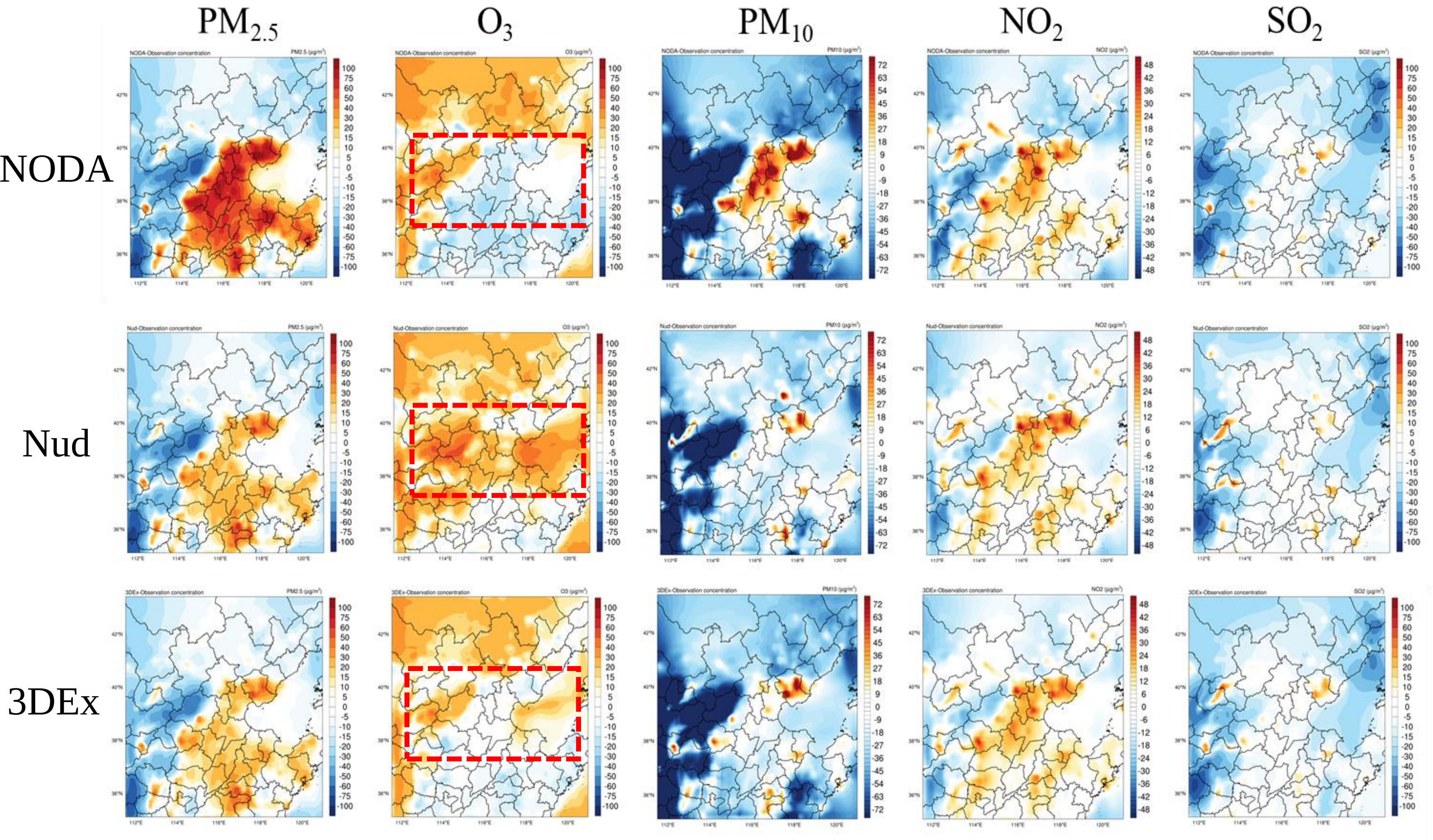


Simulation time:
2019.1.15 to 2019.1.30

- MEIC has time lag and cannot reflect the current emissions.
- There has been improvement in the inversion, with a significant decrease in PM2.5 and a significant increase in VOCs.

The 2019 emission inventory= original emission inventory (MEIC 2017)+emission innovation(inversed by using the 2019 observations)

Results of emission inversion



The Nudging method cannot handle nonlinear problems, attributing all O_3 errors to the underestimation of VOCs, and the inversion of O_3 actually deteriorates.

3DEx： O_3-NO_x-VOCs nonlinear processes≈ O_3,NO_2 concentration innovation+VOCs,NO_x emission innovation+ExRT model

$$\delta E_{NOx} = M(\delta C_{O3}, \delta C_{NO2})$$
$$\delta E_{VOCs} = M(\delta C_{O3}, \delta C_{NO2}, \delta E_{NOx})$$

3DEx can significantly improve the simulation of O_3.

Assessment of the 3DEx emission inversion method

As for $PM_{2.5}$, Nud can significantly improve the forecasting accuracy and 3DEx was better.

As for O_3, Nud reckon without nonlinear reactions of O3-Nox-VOC made the forecasting worse. But 3DEx can partly take the place of the nonlinear reactions and had a better performance.

	$PM_{2.5}$			**O_3**		
	NODA	Nud	3DEx	NODA	Nud	3DEx
$\overline{Rs}$	0.47	0.54	**0.54**	**0.33**	0.16	0.32
$\overline{RMSEs}$ **($\mu g/m^3$)**	56.44	38.23	**33.82**	24.60	36.61	**17.58**
Ra	0.85	0.91	**0.93**	0.75	0.81	**0.81**
$RMSEa$ **($\mu g/m^3$)**	24.41	**10.59**	12.45	13.91	14.86	**12.49**

- 3DEx-NODA $PM_{2.5}$: **RMSEs 40%** **RMSEa 49%**

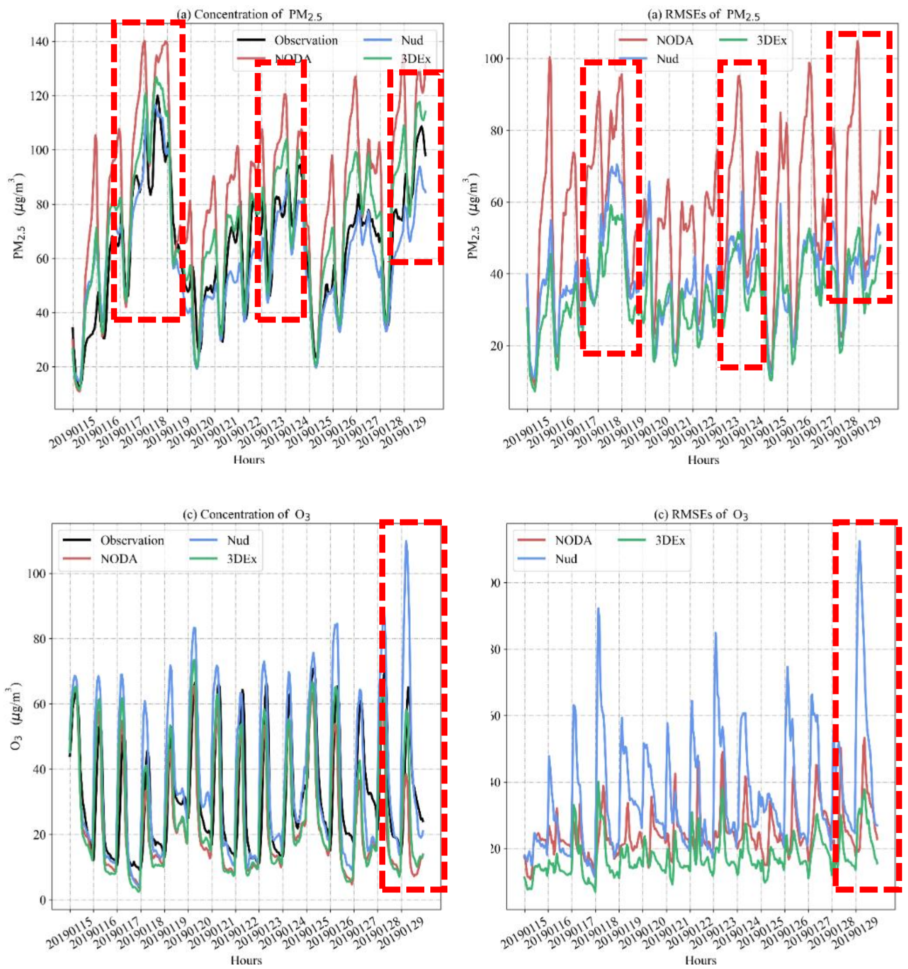
  O_3: **RMSEs 29%** **RMSEa 10%**

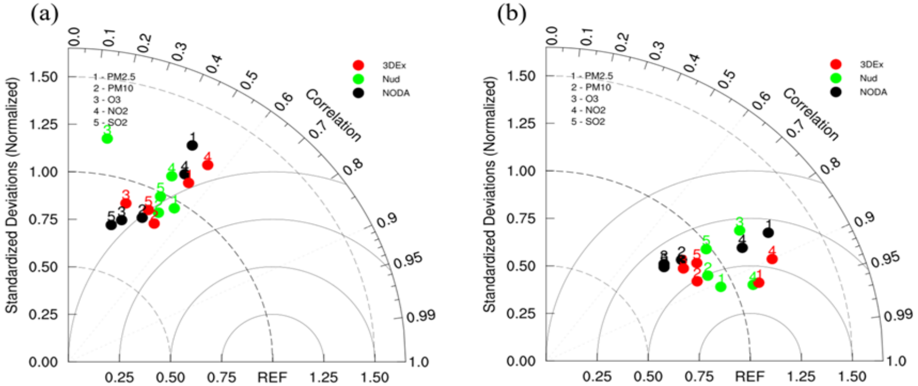
$\overline{Rs}$:the hourly averaged spatial correlation coefficient

$\overline{RMSEs}$:the hourly averaged spatial root mean squared error

Ra: the correlation coefficient of the site averaged concentration

RMSEa: the root mean squared error of the site averaged concentration





Contents

- Introduction
- Data and Methods
- Results
- Conclusion

Conclusion

- This efficiently and extensibility framework of 3DVar-ExRT method has been proved to be a good way to adjust anthropogenic emissions.

- 3DVar-ExRT method can improve the $PM_{2.5}$ and O_3 forecasting accuracy and the optimization was better with the growth of database.

- The iterations can be done with the operational forecast, which means the computing resources can be greatly reduced using this method.

- Both linear and nonlinear emission sources can be optimized using 3DVar-ExRT methods.

Thank You!

Email：congwuhuang@hubu.edu.cn

WeChat