doi:10.11676/qxxb2018.036

基于极端随机树方法的 WRF-CMAQ-MOS 模型研究*

黄丛吾[1]　陈报章[2]　马超群[1]　王体健[1]

HUANG Congwu[1]　CHEN Baozhang[2]　MA Chaoqun[1]　WANG Tijian[1]

1. 南京大学大气科学学院,南京,210023

2. 中国矿业大学环境与测绘学院,徐州,221116

1. *School of Atmospheric Sciences, Nanjing University, Nanjing* 210023, *China*

2. *School of Environment Science and Spatial Informations, China University of Mining and Technology, Xuzhou* 221116, *China*

2017-10-25 收稿,2018-04-10 改回.

黄丛吾, 陈报章, 马超群, 王体健. 2018. 基于极端随机树方法的 WRF-CMAQ-MOS 模型研究. 气象学报, 76(5):779-789

Huang Congwu, Chen Baozhang, Ma Chaoqun, Wang Tijian. 2018. WRF-CMAQ-MOS studies based on extremely randomized trees. *Acta Meteorologica Sinica*, 76(5):779-789

Abstract With the rapid development of urbanization and industrialization, air pollution has become one of the problems which are most concerned with. In this paper, the CMAQ (Community Multi-Scale Air Quality) model is combined with the WRF (Weather Research and Forecast Model) forecast of meteorological data, observed meteorological data and pollutants concentration data to build the WRF-CMAQ-MOS statistical correction model based on the method of Extremely Randomized Trees. The output of CMAQ model like AQI and pollutants concentrations of $PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3 and CO etc. are optimized. In the experiment that combines WRF meteorological elements with the CMAQ-MOS model, it is found that the linear regression method cannot well reflect the effect of optimization. Thereby the Extremely Randomized Trees method and Gradient Boosted Regression Trees were selected to improve the model performance. The investigations show that the Extremely Randomized Trees method has greatly improved the performance of CMAQ-MOS model combined with WRF meteorological elements. Finally, the model and method are evaluated using the data of January, February and March 2016. The results show that the most obvious optimization effect can be found in two pollutants, i.e., NO_2 and O_3. The correlation coefficient of NO_2 in January, February and March 2016 is increased from 0.3542 to 0.6274, the correlation coefficient of O_3 is increased from 0.3883 to 0.7886, the root-mean-square error of NO_2 is decreased from 0.0346 to 0.0243 mg/m^3, and the root-mean-square error of O_3 is decreased from 0.0447 to 0.0367 mg/m^3. The WRF-CMAQ-MOS statistical correction model developed in this study can effectively improve the prediction accuracy of various pollutants, and thus has a great application prospect in operational air quality prediction.

Key words CMAQ, WRF, Extremely randomized trees, CMAQ-MOS model

摘　要　随着城市化、工业化的快速发展,空气污染已经成为了公众最关注的问题之一。为了提高空气质量预报的准确度,以多尺度空气质量模型(Community Multi-Scale Air Quality,CMAQ)为工具,结合中尺度 WRF(Weather Research and Forecast Model)气象预报数据、气象观测数据、污染物浓度观测数据,基于极端随机树方法建立了 WRF-CMAQ-MOS(Weather Research and Forecast Model-Community Multi-Scale Air Quality-Model Output Statistics)统计修正模型。结果表明,结合 WRF 气象预报的 CMAQ-MOS 方法明显修正了由于模型非客观性产生的模式预报偏差,提高了预报效果。使用线性回归方

* 资助课题:国家重大科技研发项目(2017YFC0209803、2016YFC0208504)和国家自然科学基金项目(91544230、41575145)。

作者简介:黄丛吾,主要从事大气污染数值模拟研究。E-mail:851554475@qq.com

通讯作者:王体健,主要从事大气环境和大气化学研究。E-mail:tjwang@nju.edu.cn

法不能获得较好的优化效果,选取极端随机树方法和梯度提升回归树方法对模型进行改进和比较,发现极端随机树方法对结合 WRF 气象要素的 CMAQ-MOS 模型有较大的提升。针对徐州地区空气质量预报,进一步使用基于极端随机树方法的 WRF-CMAQ-MOS 模型对 2016 年 1、2、3 月的空气质量指数(AQI)及 $PM_{2.5}$、PM_{10}、NO_2、SO_2、O_3、CO 六种污染物优化试验进行验证,发现优化效果最为明显的两种污染物分别是 NO_2 及 O_3,2016 年 1、2、3 月整体相关系数 NO_2 由 0.35 升至 0.63,O_3 由 0.39 升至 0.79,均方根误差 NO_2 由 0.0346 减至 0.0243 mg/m^3,O_3 由 0.0447 减至 0.0367 mg/m^3。文中发展的 WRF-CMAQ-MOS 统计修正模型可以有效提升预报精度,在空气质量预报中具有很好的应用前景。

关键词 多尺度空气质量模型(CMAQ),WRF,极端随机树,CMAQ-MOS 模型

中图法分类号 P456

1 引 言

多尺度空气质量模型(CMAQ)作为 MODES-3 的核心组成部分,可用于局地、城市、区域和大陆等多尺度的空气质量预报研究。为了提高模式预报精度,使模型能够真实反映研究区域的大气污染过程,中外已经有许多学者开展了相关研究,如许建明等(2005)进行的 CMAQ-MOS 区域空气质量统计修正模型预报途径研究,通过观测气象要素值以及观测污染物浓度值对模型模拟结果进行统计订正,成功将概率预报方式与数值预报方式结合起来;Peng 等(2014)基于 CMAQ 模型成功建立了东亚地区的碳同化反演系统;Chang 等(2016)使用了同化后的 MACC(Monitoring Atmospheric Composition and Climate)数据作为初始和边界条件,成功地对 CMAQ 模型进行优化;Gou 等(2011)使用四维变分方法优化了 CMAQ 模型的化学传输过程;Zubrow 等(2008)使用了集合调整卡尔曼滤波方法对 CMAQ 模型预报中 CO 的化学传输过程进行优化;针对 CMAQ 模型所使用的排放源清单,程兴宏等(2013,2016)提出了使用自适应偏最小二乘回归法以及张弛逼近源同化反演对排放源清单进行订正的方法。

为了提高多尺度空气质量模型(CMAQ)的模拟准确度,本研究引入了结合 WRF 气象要素的 CMAQ-MOS 模型。在模拟试验过程中,发现传统的多元线性回归方法限制了 MOS 模型的优化效果,采用非线性的极端随机树方法和梯度提升回归树方法对模式输出统计模型进行优化,将两种方法应用到徐州空气质量预报业务中,并进行性能比较。

2 模式和方法

2.1 WRF-CMAQ-MOS 模型

传统的多元线性回归方法限制了结合 WRF 气象要素的 CMAQ-MOS 模型的发挥,多元线性回归方法可较好地预测线性过程,但对于非线性过程,例如大气污染物预报过程,其效果显然会受到较大的限制。为了解决以上问题,提升结合 WRF 气象要素的 CMAQ-MOS 模型优化效果,文中选取较为新颖且易于实现的极端随机树(Extremely Randomized Trees)方法(Galelli, et al, 2013;Scalzo, et al, 2012)和梯度提升回归树(Gradient Boosted Regression Trees)方法(Carslaw, et al, 2009),对结合 WRF 气象要素的 CMAQ-MOS 模型进行改进。传统的使用多元线性回归方法、基于 WRF 预报气象要素的 MOS 模型计算式为

$$\boldsymbol{X}_{t|t-t_0} = \widetilde{\boldsymbol{G}}_{m_{t|t-t_0}} + \widetilde{\boldsymbol{S}}_{N+L_{t|t-t_0}} \times \widetilde{\boldsymbol{A}}^{N+L}_{m_{t|t-t_0}} \tag{1}$$

$$\boldsymbol{S}_{t+t_1|t} = \widetilde{\boldsymbol{G}}_{m_{t|t-t_0}} + \widetilde{\boldsymbol{S}}_{N+L_{t+t_1|t}} \times \widetilde{\boldsymbol{A}}^{N+L}_{m_{t|t-t_0}} \tag{2}$$

式中,$\boldsymbol{X}_{t|t-t_0}$ 代表 $t-t_0$ 时刻至 t 时刻某种污染物的浓度观测值矩阵,$\widetilde{\boldsymbol{G}}_{m_{t|t-t_0}}$ 代表 $t-t_0$ 时刻至 t 时刻常数矩阵,$\widetilde{\boldsymbol{S}}_{N+L_{t|t-t_0}}$ 代表 $t-t_0$ 时刻至 t 时刻 N 列 WRF 大气要素预报值、L 列空气质量模型污染物浓度模拟值,$\widetilde{\boldsymbol{A}}^{N+L}_{m_{t|t-t_0}}$ 为 $t-t_0$ 时刻至 t 时刻相对应系数项矩阵。根据式(1),求得 $\widetilde{\boldsymbol{G}}_{m_{t|t-t_0}}$ 和 $\widetilde{\boldsymbol{A}}^{N+L}_{m_{t|t-t_0}}$ 后代入式(2)。在式(2)中,$\boldsymbol{S}_{t+t_1|t}$ 为 t 时刻至 $t+t_1$ 时刻某种污染物的浓度预报值矩阵,本系统中,取 t_0 为 1×24 h,取 t_1 为 1×24 h,$\widetilde{\boldsymbol{S}}_{N+L_{t+t_1|t}}$ 代表 t 时刻至 $t+t_1$ 时刻 WRF 大气要素预报值、空气质量模型污染物浓度模拟值。

由式(1)、(2)可知,使用 MOS 模型综合考虑了污染物浓度观测数据、大气要素观测数据、WRF 预报气象数据和空气质量模型预报数据进行污染物浓度预报订正,考虑了动力作用、化学作用和污染物排放的影响。

图 1 为结合 WRF 预报气象要素的 CMAQ-

MOS 模型运行流程,分为模拟阶段和预报阶段。在模拟阶段,模型读取 $t-1$ 时段观测气象数据,t 时段 CMAQ 模型预报污染物浓度数据,t 时段 WRF 预报气象数据,t 时段观测污染物浓度数据,建立基于多元线性回归的矩阵方程,其中,观测气象要素数据为 4×25 的矩阵,4 列分别为观测温度、湿度、气压、风速,25 行为从 $t-1$ 天 00 时至 t 天 00 时的时间序列;WRF 气象要素数据为 4×25 的矩阵,4 列分别为 WRF 预报温度、湿度、气压、风速,25 行为从 t 天 00 时至 $t+1$ 天 00 时的时间序列;CMAQ 模型预报污染物浓度数据为 7×25 的矩阵,7 列分别为空气质量指数(AQI)、$PM_{2.5}$、PM_{10}、CO、NO_2、O_3、SO_2 的模型模拟值,25 行为从 t 天 00 时至 $t+1$ 天 00 时的时间序列;观测污染物浓度数据为 7×25 的矩阵,7 列分别为空气质量指数、$PM_{2.5}$、PM_{10}、CO、NO_2、O_3、SO_2 的观测值,25 行为从 t 天 00 时至 $t+1$ 天 00 时的时间序列。通过以上数据组成多元线性回归模型,将所得模型代入预报阶段中,得到 $t+1$ 时段 MOS 模型污染物预报浓度值,为 7×25 的矩阵,7 列分别为空气质量指数、$PM_{2.5}$、PM_{10}、CO、NO_2、O_3、SO_2 的污染物浓度预报值,25 行为从 t 天 00 时至 $t+1$ 天 00 时的时间序列。

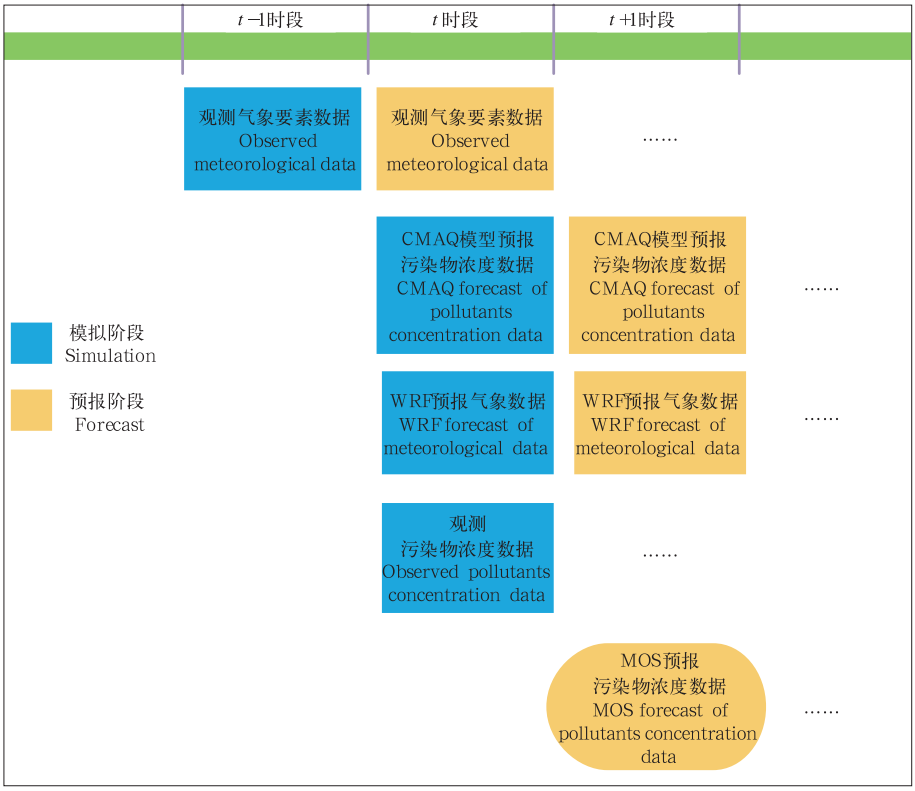


图 1 结合 WRF 预报气象要素的 MOS 模型运行流程

Fig. 1 Flow chart of the CMAQ-MOS model based on WRF forecast of meteorological data

2.1.1 极端随机树方法

Geurts 等(2006)提出了极端随机树方法,根据经典的自上而下的方法,极端随机树构建了一系列“自由生长”的回归树集合。与随机森林方法相似,极端随机树方法也是由多棵决策树构成,但与随机森林方法不同在于极端随机树方法是完全随机的得到分叉值,从而进行对回归树的分叉,不同于随机森林的在一个随机子集内得到最佳分叉属性。此外,极端随机树方法中的每一棵回归树用的都是全部训练样本。该方法的实现流程为:

(1) 根据观测污染物浓度数据建立随机拆分 (S,a)

计算 S 的最大值和最小值,分别记为 $a_{\min}^{S}$ 和 $a_{\max}^{S}$;选择均匀切点为 $[a_{\min}^{S},a_{\max}^{S}]$ 中的 a_c;返回拆分

$[a<a_c]$。

(2) 结合 WRF 预报气象数据、模型预报数据建立一棵极端随机树 t

在以下 3 种情况时,返回叶子结点:(ⅰ)$|S|<n_{\min}$,即叶子节点的最小样例;(ⅱ)S 中包含了所有的候选属性;(ⅲ)S 中包含了输出变量。运行流程如下:在所有的属性中选择互不相同的随机 K 属性 $\{a_1,\cdots,a_K\}$;利用随机拆分(S,a_i),构建 K 拆分 $\{S_1,\cdots,S_K\}$;选择拆分 S_*,其中,$\mathrm{Score}(S_*,S)=\max_{i=1,\cdots,K}\mathrm{Score}(S_i,S)$,Score 为评分函数,文中,$\mathrm{Score}_R(S_i,S)=$

$$\frac{\mathrm{var}\{y\mid S\}-\frac{|S_l|}{|S|}\mathrm{var}\{y\mid S_l\}-\frac{|S_r|}{|S|}\mathrm{var}\{y\mid S_r\}}{\mathrm{var}\{y\mid S\}},$$

$\mathrm{var}\{y|S\}$为 y 在 S 中的方差;根据 S_*,在分叉 S_l 和 S_r 中进行切片 S;分叉 S_l 和 S_r 中分别形成下一步的回归树 t_l 和 t_r;创建分叉点 S_* 将 t_l 和 t_r 作为左右子树并返回树 t。

(3) 建立极端随机树集合

根据(2)构建极端回归树集合 $T=\{t_1,\cdots,t_M\}$。

2.1.2 梯度提升回归树方法

梯度提升回归树是根据梯度提升算法构建的回归树模型,梯度提升是一种组合算法,它的基分类器是决策树,既可以用来回归,也可以用作分类。与随机森林相比,梯度提升算法在分类性能上能够与其媲美,甚至有可能表现得更好。梯度提升回归树具有以下三点优势:自然而然地处理混合类型的数据、预测能力强及在输出空间通过强大的损失函数对于异常值的鲁棒性强。

Friedman(2001)提出梯度提升算法,其核心就在于每棵树是从先前所有树的残差中来学习。利用的是当前模型中损失函数的负梯度值作为提升树算法中的残差的近似值,进而拟合一棵回归树。其运行流程为:

(1) 初始化,结合 WRF 预报气象数据、模型预报数据和观测污染物浓度数据通过函数

$$f_0(x)=\mathrm{argmin}_c\sum_{i=1}^{N}L(y_i,c) \tag{3}$$

估计一个使损失函数极小化的常数值,式中,y_i 为 WRF 预报气象数据和模型预报数据,c 为观测污染物浓度数据,此时它是只有一个节点的树。

(2) 迭代法建立 M 棵提升树

在第 1 层循环中 $m=1,2,\cdots,M$;在第 2 层循环中 $i=1,2,\cdots,N$,计算损失函数的负梯度在当前模型的值,并将其作为残差的估计值。

$$r_{m,i}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]f(x)=f_{m-1}(x) \tag{4}$$

对于 $r_{m,i}$ 拟合一棵回归树,得到第 m 棵树的叶节点区域 $R_{m,j}$,$j=1,2,\cdots,J$;在第 2 层循环中 $j=1,2,\cdots,J$,计算

$$C_{m,j}=\mathrm{argmin}_c\sum_{x_i\in R_{m,j}}L(y_i,f_{m-1}(x_i)+c) \tag{5}$$

利用线性搜索估计叶节点区域的值,使损失函数极小化;之后,更新

$$f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J}C_{m,j}I \tag{6}$$

式中,指示函数 I 在 $x\in R_{m,j}$ 时,$I=1$,否则 $I=0$。

(3)最后得到的 $f_m(x)$就是最终的梯度提升回归树模型

$$\tilde{f}(x)=f_M(x)=\sum_{m=1}^{M}\sum_{j=1}^{J}C_{m,j}I \tag{7}$$

2.2 研究区域设置

以六大污染物(NO_2、$PM_{2.5}$、PM_{10}、SO_2、O_3、CO)及空气质量指数(AQI)为研究对象,利用空气质量模型 CMAQ 及 Python、Fortran 语言,shell 脚本语言,进行 CMAQ-MOS 统计修正模型研究,用于徐州空气质量预报。

模型使用 3 层嵌套网格(图 2),水平分辨率分别为 27、9 和 3 km,第 1 层区域覆盖整个东亚、大部分东南亚和部分南亚地区,第2层区域覆盖大部分

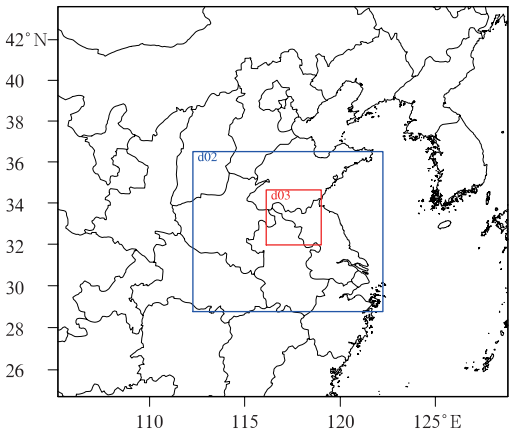


图 2 研究区域

Fig. 2 Model domain for simulation

华东地区及部分华中地区,第 3 层区域覆盖徐州市全域,网格数分别为 88×76,112×94 和 97×97;垂直方向采用 η 坐标系,共 29 层,模式层顶气压为 50 hPa,模式层顶距地面 20 km,边界层(约 2 km)内分为 10 层。对于污染源排放数据,区域内层主要采用 SMOKE 模型输出,外层嵌套采用清华大学 MEIC 清单数据。

2.3 误差分析方法

采用的误差分析方法主要有 4 种,分别是均方根误差(RMSE)、相关系数(R)、平均相对误差(MFE)、平均相对偏差(MFB)(Boylan, et al, 2006)。当平均相对误差在±30%内,且平均相对偏差小于 50%时,可认为预报是准确的。

$$\mathrm{RMSE}=\left[\frac{1}{n}\sum_{i=1}^{n}(S(i)-O(i))^2\right]^{\frac{1}{2}} \tag{8}$$

$$R=\frac{\mathrm{Cov}(S-O)}{\sqrt{D(S)}\ \sqrt{D(O)}} \tag{9}$$

$$\mathrm{MFB}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{S(i)-O(i)}{O(i)+S(i)/2}\right) \tag{10}$$

$$\mathrm{MFE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{S(i)-O(i)}{O(i)+S(i)/2}\right| \tag{11}$$

3 WRF-CMAQ-MOS 统计修正模型的应用

3.1 极端随机树方法和梯度提升回归树方法比较

选取 2016 年 9 月 13—24 日徐州市黄河新村站点逐时污染物浓度观测数据、WRF 气象预报数据、CMAQ 模型污染物浓度预报数据,使用极端随机树方法和梯度提升回归树方法并结合 WRF 气象要素进行 CMAQ-MOS 模型试验。分别以 WRFMOS1 代表使用多元线性回归方法、结合 WRF 气象要素预报的 CMAQ-MOS 模型优化结果,EXTRA 代表使用极端随机树方法、结合 WRF 气象要素预报的 CMAQ-MOS 模型优化结果,GBRT 代表使用梯度提升回归树方法、结合 WRF 气象要素预报的 CMAQ-MOS 模型优化结果,试验结果如下:

由表 1 可以看出,EXTRA 方案优化效果最佳。结合观测污染物浓度值(OBV)、WRFMOS1 方案结果、EXTRA 方案结果、GBRT 方案结果,绘制出 2016 年 9 月 13—24 日徐州市空气质量指数、CO、NO_2、O_3、PM_{10}、$PM_{2.5}$、SO_2 浓度观测、WRFMOS1 方案、EXTRA 方案、GBRT 方案时间变化和泰勒图(图 3、4)。

表 1 徐州市黄河新村站点 3 种 CMAQ-MOS 方法比较结果

Table 1 Comparison of results obtained by three methods in HHXC, Xuzhou

	R	RMSE(mg/m³)	MFB(%)	MFE(%)
AQI-WRFMOS1	0.4915	0.0309	47.17	-18.02
AQI-EXTRA	0.5453	0.0199	22.29	5.12
AQI-GBRT	0.3881	0.0187	22.47	6.28
CO-WRFMOS1	0.2250	0.7175	28.41	5.99
CO-EXTRA	0.2827	0.4228	27.37	12.40
CO-GBRT	0.1622	0.4455	30.45	13.04
NO_2-WRFMOS1	0.2798	0.0357	51.83	4.05
NO_2-EXTRA	0.4020	0.0220	30.66	14.16
NO_2-GBRT	0.2971	0.0241	32.73	15.92
O_3-WRFMOS1	0.2728	0.0575	173.61	-95.34
O_3-EXTRA	0.5796	0.0355	40.86	17.92
O_3-GBRT	0.5097	0.0373	43.82	21.45
PM_{10}-WRFMOS1	0.4472	0.0307	23.12	-0.32
PM_{10}-EXTRA	0.5324	0.0223	28.04	7.98
PM_{10}-GBRT	0.4079	0.0224	30.69	7.19
$PM_{2.5}$-WRFMOS1	0.5680	0.0249	46.81	-24.80
$PM_{2.5}$-EXTRA	0.5904	0.0180	28.76	2.88
$PM_{2.5}$-GBRT	0.4958	0.0176	29.64	1.50
SO_2-WRFMOS1	0.2178	0.0278	21.57	9.24
SO_2-EXTRA	0.3665	0.0179	23.11	10.43
SO_2-GBRT	0.2270	0.0188	24.94	12.87

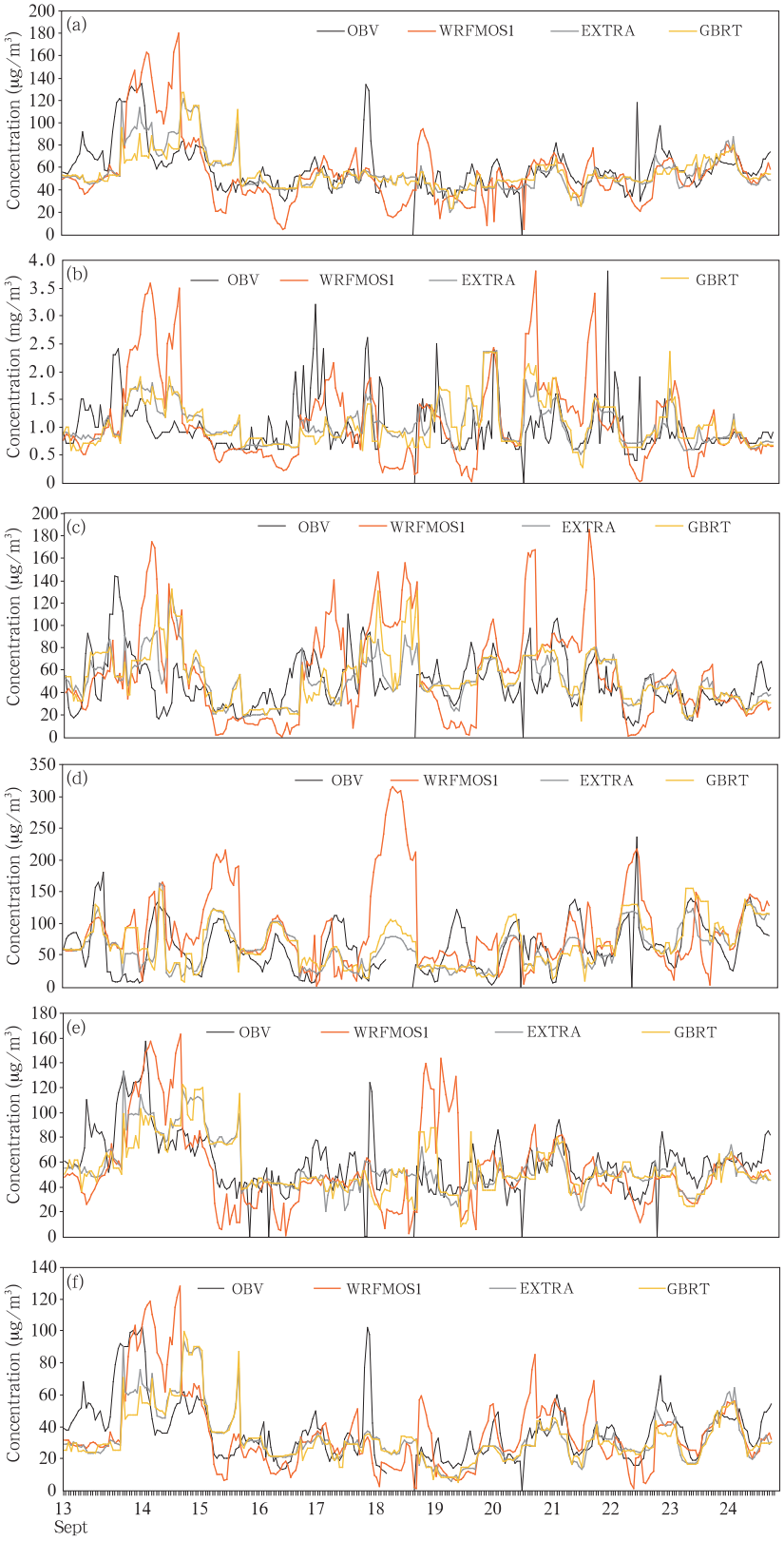
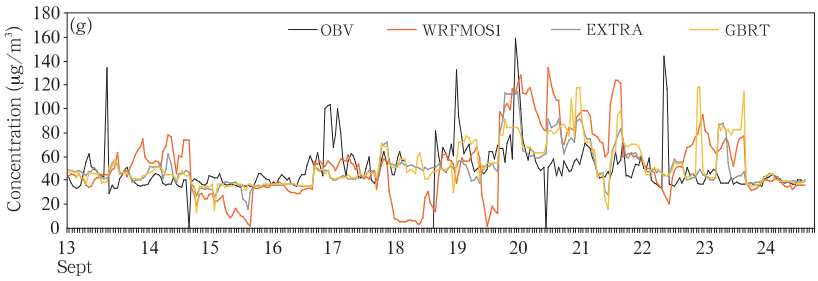


图 3 2016 年 9 月 13—24 日徐州市黄河新村站点 AQI(a)、CO(b)、NO_2(c)、O_3(d)、PM_{10}(e)、$PM_{2.5}$(f)、SO_2(g)浓度观测、WRFMOS1 方案、EXTRA 方案、GBRT 方案时间变化

Fig. 3 Hourly variations of concentrations of AQI (a), CO (b), NO_2(c), O_3(d), PM_{10}(e), $PM_{2.5}$(f) and SO_2(g) from observations and outputs of WRFMOS1, EXTRA and GBRT during 13 – 24 September, 2016 in HHXC, Xuzhou



续图 3

Fig. 3 Continued

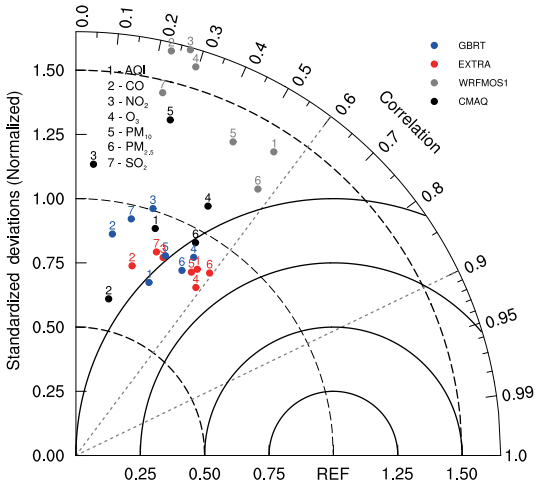


图 4 2016 年 9 月 13—24 日徐州市黄河新村站点六大污染物及 AQI 浓度观测、CMAQ 模拟值、WRFMOS1 方案、EXTRA 方案、GBRT 方案泰勒图

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

针对 CMAQ 模型,使用多元线性回归方法,结合 WRF 预报气象要素的 CMAQ-MOS 模型,其优化效果仍未达到预期。在选取极端随机树方法和梯度提升回归树方法,分别对结合 WRF 气象要素的 CMAQ-MOS 模型进行改进之后,经由试验结果发现,使用极端随机树方法比梯度提升回归树方法的改进效果更显著,其原因主要是因为其每一棵树都是用了全部的样本,并进行了完全随机的拆分。

3.2 极端随机树方法的应用

为了验证使用极端随机树方法并结合 WRF 气象要素的 CMAQ-MOS 模型优化效果并进行应用,进行了针对徐州地区 2016 年 1—3 月的逐时预报应用试验,选取黄河新村站点(1177A)观测数据进行评估,并针对 7 个站点的优化效果进行统计分析。模型运行效率较高,进行单站点一个月逐时预报试验耗时约 15 min,业务化运行中每日对 7 个站点进行未来 3 d 逐时预报,耗时约 3 min。

针对空气质量指数及六大污染物,使用极端随机树方法结合 WRF 气象要素预报的 CMAQ-MOS 模型进行试验,7 个站点试验结果见表 2。可见,经过优化后的六大污染物及空气质量指数的准确度均得到了极大的提高。根据试验结果,针对黄河新村站点的空气质量指数与六种污染物分别绘制 2016 年 1—3 月徐州市空气质量指数、CO、NO_2、O_3、PM_{10}、$PM_{2.5}$、SO_2 污染物浓度预报和订正误差时间变化、散点分布和泰勒图。从图 5—7 可明显看出,针对空气质量指数、CO、NO_2、O_3、PM_{10}、$PM_{2.5}$、SO_2,使用极端随机树方法结合 WRF 气象要素预报的 CMAQ-MOS 模型具有良好的优化效果。

4 总 结

为了提高 CMAQ 空气质量模型的预报精度,基于多元线性回归方法,引入了结合 WRF 气象要素的 CMAQ-MOS 模型,通过试验发现其优化效果一般,主要原因是多元线性回归方法限制了结合 WRF 气象要素的 CMAQ-MOS 模型的发挥。为了提升模型优化效果,选取极端随机树方法和梯度提升回归树方法,对结合 WRF 气象要素的 CMAQ-MOS 模型进行改进,通过试验发现,相对梯度提升回归树方法,极端随机树方法的改进效果更为显著。

表 2 2016 年 1—3 月徐州市 7 站点六大污染物及空气质量指数浓度预报和订正误差统计结果

(a. 相关系数,b. 均方根误差(mg/m^3))

Table 2 Relative coefficients and biases between hourly forecast, and corrected and observed data of six pollutants and AQI during January－March 2016 in Xuzhou

(a. correlation coefficient, b. root mean square error (mg/m^3))

a. 相关系数		AQI	CO	NO_2	O_3	PM_{10}	$PM_{2.5}$	SO_2
黄河新村	CMAQ	0.40	0.44	0.35	0.39	0.34	0.47	0.12
	MOS	0.47	0.53	0.63	0.73	0.49	0.55	0.53
淮塔	CMAQ	0.47	0.47	0.43	0.39	0.47	0.50	0.15
	MOS	0.48	0.52	0.61	0.57	0.50	0.51	0.22
铜山兽医院	CMAQ	0.19	0.35	0.22	0.51	0.16	0.29	0.06
	MOS	0.49	0.52	0.68	0.54	0.52	0.50	0.20
新城区	CMAQ	0.26	0.42	0.41	0.49	0.22	0.41	−0.04
	MOS	0.39	0.60	0.55	0.71	0.42	0.57	0.20
桃园路	CMAQ	0.41	0.44	0.33	0.40	0.38	0.46	0.16
	MOS	0.49	0.49	0.65	0.62	0.52	0.52	0.38
农科院	CMAQ	0.22	0.36	0.31	0.47	0.19	0.35	−0.01
	MOS	0.45	0.55	0.63	0.68	0.48	0.54	0.32
铜山区环保局	CMAQ	0.17	0.33	0.26	0.38	0.17	0.25	0.13
	MOS	0.43	0.45	0.66	0.59	0.49	0.48	0.40
b. 均方根误差		AQI	CO	NO_2	O_3	PM_{10}	$PM_{2.5}$	SO_2
黄河新村	CMAQ	0.072	1.076	0.045	0.056	0.120	0.051	0.051
	MOS	0.063	0.800	0.024	0.037	0.080	0.048	0.030
淮塔	CMAQ	0.077	1.123	0.049	0.069	0.117	0.059	0.065
	MOS	0.068	0.889	0.021	0.019	0.087	0.056	0.031
铜山兽医院	CMAQ	0.077	1.264	0.049	0.053	0.124	0.054	0.050
	MOS	0.076	0.921	0.027	0.042	0.101	0.059	0.040
新城区	CMAQ	0.068	0.898	0.041	0.047	0.128	0.041	0.035
	MOS	0.047	0.573	0.023	0.032	0.077	0.034	0.024
桃园路	CMAQ	0.076	1.172	0.048	0.061	0.118	0.056	0.057
	MOS	0.070	0.889	0.024	0.032	0.093	0.056	0.035
农科院	CMAQ	0.071	1.071	0.044	0.049	0.124	0.047	0.042
	MOS	0.062	0.748	0.025	0.037	0.088	0.047	0.032
铜山区环保局	CMAQ	0.074	1.199	0.048	0.053	0.122	0.052	0.053
	MOS	0.063	0.887	0.025	0.034	0.090	0.049	0.036

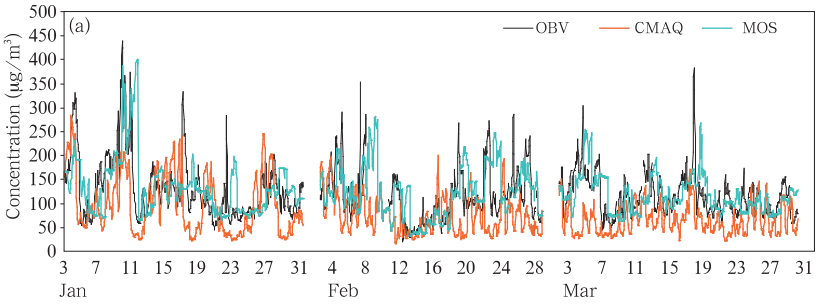


图 5 2016 年 1—3 月徐州市黄河新村站点空气质量指数(a)、CO(b)、NO_2(c)、O_3(d)、PM_{10}(e)、$PM_{2.5}$(f)、SO_2(g)污染物浓度预报和订正误差时间变化

Fig. 5 Temporal variations of hourly forecast, and corrected and observed concentrations of AQI (a), CO (b), NO_2(c), O_3(d), PM_{10}(e), $PM_{2.5}$(f) and SO_2(g) during January－March 2016 in HHXC, Xuzhou



续图 5

Fig. 5 Continued

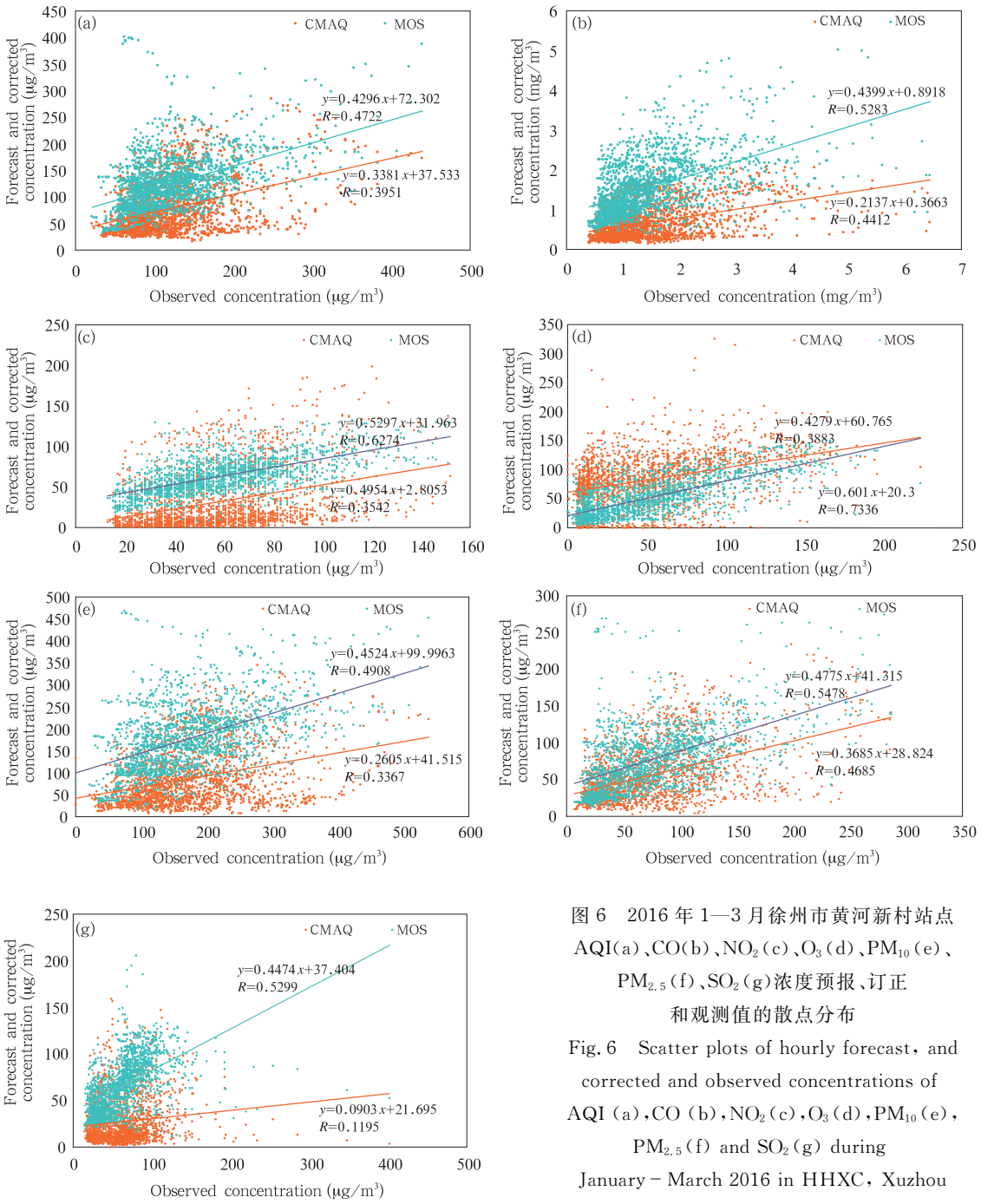


图 6 2016 年 1—3 月徐州市黄河新村站点 AQI(a)、CO(b)、NO_2(c)、O_3(d)、PM_{10}(e)、$PM_{2.5}$(f)、SO_2(g)浓度预报、订正和观测值的散点分布

Fig. 6 Scatter plots of hourly forecast, and corrected and observed concentrations of AQI (a), CO (b), NO_2 (c), O_3 (d), PM_{10} (e), $PM_{2.5}$ (f) and SO_2 (g) during January – March 2016 in HHXC, Xuzhou

采用极端随机树方法,并结合 WRF 预报气象要素的 CMAQ-MOS 模型在徐州地区进行 2016 年 1—3 月的模拟应用试验。结果表明,针对空气质量指数及六大污染物逐时预报,与未经优化的 CMAQ 模型相比,采用极端随机树方法并结合 WRF 预报气象要素的 CMAQ-MOS 模型在相关系数、均方根误差、平均相对误差、平均相对偏差方面都有较为明显的优化效果,NO_2 及 O_3 两种污染物优化效果最为明显,其原因可能是 O_3、NO_2 与 MOS 模型引入的温度气象要素的相关性较高,这是因为温度的变化往往与太阳辐射的变化相关,而太阳辐射的变化会导致 O_3 光化学反应速率的变化。与此同时,NO_2 和 O_3 的化学耦合较强,O_3 浓度的变化会导致 NO_2 浓度也发生变化,因此MOS模型针对NO_2及

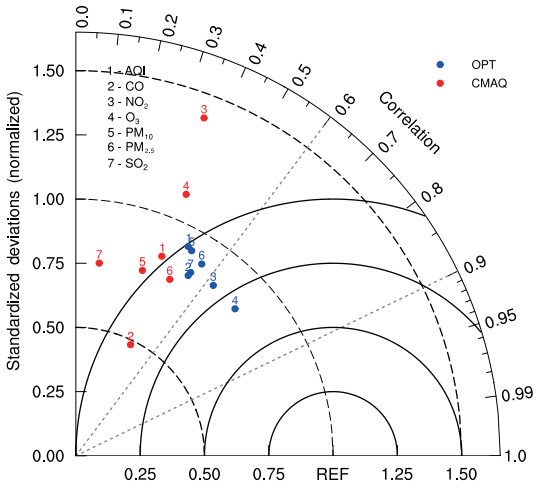


图 7　2016 年 1—3 月徐州市黄河新村站点六种污染物与 AQI 浓度预报、订正(OPT)和观测值的泰勒图

Fig. 7　Taylor diagram of hourly forecast, and corrected (OPT) and observed concentrations of six pollutants and AQI during January－March 2016 in HHXC, Xuzhou

O_3 两种污染物优化效果最为明显。

综上所述,采用极端随机树方法并结合 WRF 预报气象要素的 CMAQ-MOS 模型预报准确度得到较大的提升,未来可以投入实际业务应用。

此外,采用极端随机树方法并结合 WRF 预报气象要素的 CMAQ-MOS 模型仍存在有一定的局限性,有待进一步改进,主要包括以下两个方面:(1)本方法优化的仅为点位数据,并不能完成对整个预报场的优化;(2)本方法针对空气质量模型预报较差时优化效果较好,但针对空气质量模型预报较好时优化效果并不明显。

参考文献

程兴宏, 李德平, 徐祥德等. 2013. 北京地区 CMAQ 源同化模式预报 PM_{10} 产品订正方法研究//2013 中国环境科学学会学术年会浦华环保优秀论文集. 昆明: 中国环境科学学会. Cheng X H, Li D P, Xu X D, et al. 2013. Study on PM_{10} product correction method for CMAQ source assimilation model prediction in Beijing region//Proceedings of 2013 Chinese Academy of Environmental Sciences. Kunming: Chinese Society of Environmental Sciences (in Chinese)

程兴宏, 刁志刚, 胡江凯等. 2016. 基于 CMAQ 模式和自适应偏最小二乘回归法的中国地区 $PM_{2.5}$ 浓度动力-统计预报方法研究. 环境科学学报, 36(8): 2771-2782. Cheng X H, Diao Z G, Hu J K, et al. 2016. Dynamical-statistical forecasting of $PM_{2.5}$ concentration based on CMAQ model and adapting partial least square regression method in China. Acta Sci Circumst, 36(8): 2771-2782 (in Chinese)

许建明, 徐祥德, 刘煜等. 2005. CMAQ-MOS 区域空气质量统计修正模型预报途径研究. 中国科学 D 辑:地球科学, 35(S1): 131-144. Xu J M, Xu X D, Liu Y, et al. 2005. Study of statistically correcting model CMAQ-MOS for forecasting regional air quality. Sci China Ser D Earth Sci, 48(S2): 155-172

Boylan J W, Russell A G. 2006. PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models. Atmos Environ, 40(26): 4946-4959

Carslaw D C, Taylor P J. 2009. Analysis of air pollution data at a mixed source location using boosted regression trees. Atmos Environ, 43(22-23): 3563-3570

Chang L S, Cho A, Park H, et al. 2016. Human-model hybrid Korean air quality forecasting system. J Air Waste Manag Assoc, 66(9): 896-911

Friedman J H. 2001. Greedy function approximation: A gradient boosting machine. Ann Stat, 29(5): 1189-1232

Galelli S, Castelletti A. 2013. Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. Hydrol Earth Syst Sci, 17(7): 2669-2684

Geurts P, Ernst D, Wehenkel L. 2006. Extremely randomized trees. Mach Learn, 63(1): 3-42

Gou T Y, Sandu A. 2011. Continuous versus discrete advection adjoints in chemical data assimilation with CMAQ. Atmos Environ, 45(28): 4868-4881

Peng Z, Zhang M, Kou X, et al. 2014. A regional carbon data assimilation system and its preliminary evaluation in East Asia. Atmos Chem Phys Discuss, 14(14): 20345-20381

Scalzo F, Hamilton R, Asgari S, et al. 2012. Intracranial hypertension prediction using extremely randomized decision trees. Med Eng Phys, 34(8): 1058-1065

Zubrow A, Chen L, Kotamarthi V R. 2008. EAKF-CMAQ: Introduction and evaluation of a data assimilation for CMAQ based on the ensemble adjustment Kalman filter. J Geophys Res Atmos, 113(D9): D09302