

## Reply to Referee #2 (ACPD-11-C2743-2746)

*This manuscript presents a set of experiments on the assimilation with the EnKF of ozone over Beijing. In my opinion few aspects and points of the presentation and analysis need to be clarified/expanded before this work can be published.*

Reply: The authors very much appreciate the valuable comments on this manuscript made by the reviewer. We have carefully considered the comments and have revised the manuscript accordingly. A point-by-point response is given in the following part.

### MAJOR COMMENTS

*Below I have listed few points that the authors need to address to make the manuscript suitable for publication:*

- 1) *It is not clear to me what is the added value of this contribution when compared to previous works. An effort should be made to stress out new results and what we have learn from this work that we did not already knew;*

Reply: Thank you very much for raising this important issue. According to your comments, we have specified the contributions of this study in the revised manuscript. Please refer to the revised abstract in Appendix A and the revised introduction in Appendix B. The results and discussions in Section 3 have also been revised accordingly in the new manuscript.

- 2) *There were 11 urban and 6 suburban stations; as stated at p. 7822 lines 25-28 only 2 urban and 1 suburban sites were not assimilated and used for verification. This choice is questionable and it may seriously limits the statistical significance of the results (given also that the assimilation experiments were less than 2 days long);*

Reply: Thanks for pointing out the issue of validation. In order to compensate for the limitation of this choice in the statistical significance, we have employed a cross validation method to improve the validation in the revised manuscript. The 17 monitoring stations are split into two subsets, 11 urban sites and 6 suburban sites. One experiment is to assimilate ozone observations at the 6 suburban sites. Ozone observations of the 11 urban sites are withheld for validation as independent data. In the other experiment, the 11 urban sites are used for assimilation and the 6 urban sites serve as validation sites. We have discussed the cross validation results in the revised manuscript. Please refer to Appendix C.

Based on the results of the cross validation data assimilation experiments, we found

out that withdrawing the urban site of any of the surrounding cities from data assimilation made it difficult to improve ozone forecast over this city. This deficiency of data assimilation in ozone forecast over surrounding cities when local observations are not assimilated is due to the long distance of these surrounding cities from other observation sites. The difficulties with the validation of this study are in the lack of abundant observations and in the unbalanced spatial distribution of the current observation sites over Beijing and its surrounding areas. For the current observation network, five urban sites and six suburban sites are located in or close to Beijing, while only one urban site is available at each of the six surrounding cities of Beijing.

- 3) *The analysis should be extended at least up to 6 hour forecasts; in data assimilation an important question is how long an observation has an impact on the forecast; as showed by the authors, in few stations it vanishes after 1 hour, in other seems to last longer. How long? Why?*

Reply: It is a good suggestion. According to your suggestion, we will add the results of several supplementary data assimilation experiments to investigate the effects of data assimilation on 6-12 hour ozone forecast in the revised manuscript. The forecast window after assimilating ozone observations will be extended to 6-12 hour. The focus is on the impacts of assimilating ozone observations in the 6-12 hour ozone forecast over Beijing and its surrounding areas. A section (Section 3.6) will be added to discuss the results in the revised manuscript.

- 4) *The section describing model error is not clear; given the importance of it, the authors should expand it and make it a better effort in explaining exactly the approach they have used;*

Reply: We agree. We have made a substantial revision of the description of model error in the revised manuscript. Please refer to the revised methodology description in Appendix D.

- 5) *- Below I have listed few places whit suggestions where the written English could be improved. However, there are many other sentences that are not clear, and if possible the authors should have the manuscript throughout reviewed by a native English speaker.*

Reply: Thanks for your remark. We have made a great effort to improve the English in

the revised manuscript.

*MINOR COMMENTS*

- *“on the other hand” and “however” often are not used properly in the text. The authors should revise any occurrence of those.*

- *P7812, L14: “However, adjustment” → “Adjustment”*

- *P7812, L21: “by implementing” → “via”*

- *P7813, L4: “serious problems” → “serious”*

- *Define “CAREBeijing” and add a reference if it is available*

- *P7813, L15: “is needed” → “was needed”*

- *P7815, L4: “It employs a Monte”*

- *P7815, L4: “variable with a large stochastic ensemble” → “variables with an ensemble”*

- *P7815, L5: “In this way, the” → “The”*

- *P7815, L6: “ensemble” → “the ensemble”*

- *P7816, L18: “promised” → “Promising”*

Reply: Accepted, we have revised these.

- *P7816, L22: “3 model grids”: : how high is that, roughly or on average?*

Reply: A clarification has been added in the revised manuscript. *“3 model grids (about 200 m)”*

- *P7817, L5-6: “It may lead”: : why? Please explain*

Reply: We have clarified it in the revised manuscript. *“It is worth noting that ensemble simulations can reconstruct the correlations between conventional state variables and emissions but not the correlations between emissions. Therefore, it is possible that some correlations between emissions are underestimated as a result of this configuration, which in turn influences the correlations between conventional state variables closely related to these emissions.”*

- *P7819: Is the assimilation performed in every domain?*

Reply: A description has been added to clarify this in the revised manuscript. *“In order to reduce the computational cost, data assimilation is only conducted in the third model domain.”*

- *P7819, L5: show some of the results of these sensitivity tests (e.g., a figure or a paragraph commenting on it). This is true also for other sensitivity experiments mentioned in the paper*

Reply: We agree. We have added some comments in the revised manuscript. *“The*

*spatial correlation scale of initial perturbation fields is set as 54 km in horizontal and 3 model grids (about 200 m) in vertical after several sensitivity tests. The correlation scale in horizontal or vertical is obtained independently from the best performed data assimilation experiment with the smallest RMSE of analyzed ozone. It should be noted that correlation scale is a tune parameter which can vary with species and space.”*

- P7819: *say more on how model error is treated and why you chose this approach*

Reply: Accepted, we have clarified this in the revised manuscript. Please refer to the equations (2.3)-(2.4)-(2.5)-(2.6)-(2.7) in Appendix D.

- P7821, L25: *“real-time way” → “real-time”*

- P7823, L10-11: *“not as valuable as much longer forecasts such as 24-hour and 48-hour forecast in application” → “not be as valuable as longer forecasts such as 24- or 48-hour forecasts”*

- P7824, L16: *new line after “areas.”*

- P7824, L20L *no new line after “Fig. 4.”*

Reply: Accepted, we have corrected these mistakes.

- P7824, L24-25-26: *this sentence is not clear*

Reply: Thanks for pointing it out. We have rewritten the sentence in the revised manuscript. *“Assimilating local observations in EXP1 strongly adjust the analysis and improve the forecast greatly with the analyzed values and forecasted values quite close to the observations.”... “At the two suburban sites, the forecast in EXP1 relaxes a lot to the reference simulation during the beginning 1-2 hours and gradually approaches the observation with the observational information integrated into model hour by hour.”*

- P7824, L27: *“a quit” → “a”*

Reply: We have fixed it.

- P7827, L5-6-7-8: *how do you justify the similar differences in daytime?*

Reply: We have added sentences to justify the similar differences during daytime in the revised manuscript. *“The similar great improvements of daytime ozone forecast from adjusting NOx initial values and VOCs initial values suggest that both NOx and VOCs initial conditions have an important role in the forecasted daytime ozone levels. It may relate to the rapid photochemical reactions between ozone, NOx and VOCs*

*during daytime.”*

- P7827, L24: “*furious*” → “*large*”
- P7828, L15: “*precious*” → “*previous*”
- P7828, L15: “*(Hanea*” → “*(e.g., Hanea*”
- P7828, L16: “*root mean square error*” → “*RMSE*”. Also, define *RMSE* the first time you use it in the text
- P7828, L27: “*zone*” → “*ozone*”

Reply: Thanks. We have corrected these mistakes and spelling errors.

- P7828, L29: *ozone is not emitted*

Reply: Thanks. We use the words “precursor emissions” to avoid this confusion in the revised manuscript.

- P7829, L21: “*a day in European and a*” → “*over Europe with a*”
- P7832, L21: “*as a powerful*” → “*is a powerful*”
- P7832, L22: “*emissions can*” → “*emissions to*”
- P7833, L25: take out “*while*”

Reply: We have corrected these grammar mistakes.

- P7833-7834: *are other errors coming for the portion of model error that is not accounted for, and for the errors coming from the assumptions built in the EnKF design (e.g., Gaussinity, linearity)?*

Reply: Yes. We have addressed it in the revised manuscript. “*Model error in this study is assumed to be mainly from the error in few model parameters including precursor emissions, photolysis rates and vertical diffusion coefficients. Actually, there are still other sources of model error such as the errors in other model parameters, the uncertainty from the Gaussian assumption of error and linearity of the ozone-precursor relationship in EnKF. Better representation of these errors in simulating model error might improve the performance of EnKF.*”

---

## **Appendix A: Revised Abstract**

Data assimilation approach is firstly employed to improve the surface ozone forecast over Beijing and its surrounding areas in this study. Several advanced data assimilation strategies based on ensemble Kalman filter are designed to adjust ozone

initial conditions, precursor initial conditions and precursor emission rates separately or jointly through assimilating ozone observations. The results suggest that adjusting ozone initial conditions, precursor initial conditions and precursor emission rates either separately or jointly can improve the ozone forecast over Beijing and its surrounding areas to different degrees. Adjusting precursor initial conditions demonstrates a potential for improving the short-term ozone forecast almost as great as shown by adjusting precursor emissions. However, either adjusting precursor initial values or emissions show a deficiency in improving the short-term ozone forecast at suburban areas. Optimizing ozone initial values brings significant improvement to the short-term ozone forecast during both daytime and nighttime. Its limitation lies in the difficulty in improving the ozone forecast at some urban sites. Simultaneous adjustment of ozone initial conditions, precursor initial conditions and precursor emission rates can overcome these limitations and displays overall better performances in improving the ozone forecast over Beijing and its surrounding areas. The root mean square errors of 1-hour ozone forecast at urban sites and suburban sites decreased by 54% and 59% respectively compared with those in free run. One more important finding is that assimilating local ozone observations is decisive in the behavior of ozone forecast over the observational area, while assimilating remote ozone observations is unnegligible for its role in reducing the uncertainty of regional transport.

---

## **Appendix B: Revised Introduction**

As one of the typical city clusters in China, Beijing and its surrounding areas are facing serious challenge in surface ozone pollutions within urbanization and motorization processes (Chan and Yao, 2008; Shao et al., 2006). Exposure to high ozone concentrations leads to heavy damages on both human health and plant life (Anderson et al., 1996; Burnett et al., 1997). Providing ozone forecast is undoubtedly quite important, not only for the public, but also for the decision makers. However,

ozone forecast is not included in the current operational air quality forecast over these areas, and only quite few previous studies focus on the issues of ozone forecast over these areas.

In previous studies on ozone forecast over Beijing, An et al. (2010) and Yu et al. (2011) developed statistical forecast models to forecast ozone concentrations based on several statistical techniques including multiple linear regressions, principal component analysis and neural network methods, while Tang et al. (2010a) and Zhang et al. (2010) employed ensemble forecast methods based on chemical transport model (CTM) to forecast ozone. A main drawback with the statistical forecast model is the difficulty in describing non-local influences such as emission changes, transport processes and complex chemical reactions (Flemming et al., 2001). The ensemble forecasting methods with CTM contains the influences from the complex chemical and dynamical processes and do not have the conceptual limitations with statistical forecast method. Its limitations lie in that large uncertainty in CTM is still a great challenge and forecast with ensemble mean brings limited improvement of forecast skill (Mallet et al., 2009; von Loon et al., 2007). Linear combination of each ensemble member based on past observations and past forecasts can produce better forecast performances (Mallet et al., 2009; Zhang et al., 2010). However, the effectiveness of the combining weights is limited to the locations and variables with available observations, and the errors in observation are normally not taken into account (Mallet et al., 2010).

In this paper, advanced data assimilation method, as an alternative approach, is firstly employed to improve the ozone forecast of CTM over Beijing and its surrounding areas. Data assimilation method integrates the observational information into a numerical model in order to obtain the estimate of the model state that minimizes the error variance. The attractive feature of data assimilation method is in their ability to improve the estimations of the physical properties that are not observed directly. Furthermore, the observation errors are taken into account adequately. Several applications of data assimilation method in ozone modeling brought out relevant findings for ozone forecast improvement in many locations (Chai et al., 2006; Elbern

et al., 2007; van Loon et al., 2000; Hanea et al., 2004; Constantinescu et al., 2007). However, how to adopt high performance data assimilation method and design well-fitting data assimilation strategy for improving ozone forecast over Beijing and its surrounding areas have not yet been addressed in previous publications.

The objective of this study is to investigate the performances of several data assimilation strategies and their implications for designing a suitable ozone data assimilation strategy over Beijing and its surrounding areas. The focus is on using data assimilation method to adjust ozone initial conditions, precursor initial conditions and precursor emissions, which are convenient to be taken as control variables in data assimilation and among which ozone initial conditions and precursor emissions have been identified as the most important uncertainty sources for short-term (less than 24 hours) ozone forecast over these areas by Tang et al. (2010b). Ensemble Kalman filter (EnKF) is employed for its strong attractive features in application for complex models, which has also been pointed out by previous literatures (Carmichael, et al., 2008; Constantinescu et al., 2006; Evensen, 2009). It supports fully nonlinear evolution of the error statistics through the highly nonlinear model and is convenient to deal with model error. Furthermore, its implementation is very simple and suitable for parallel computation with no need for tangent linear or adjoint model. Section 2 describes the adopted data assimilation method, regional air quality model, regional air quality observation network and the designed experiments. Results and discussions are presented in Section 3 and conclusions are given in Section 4.

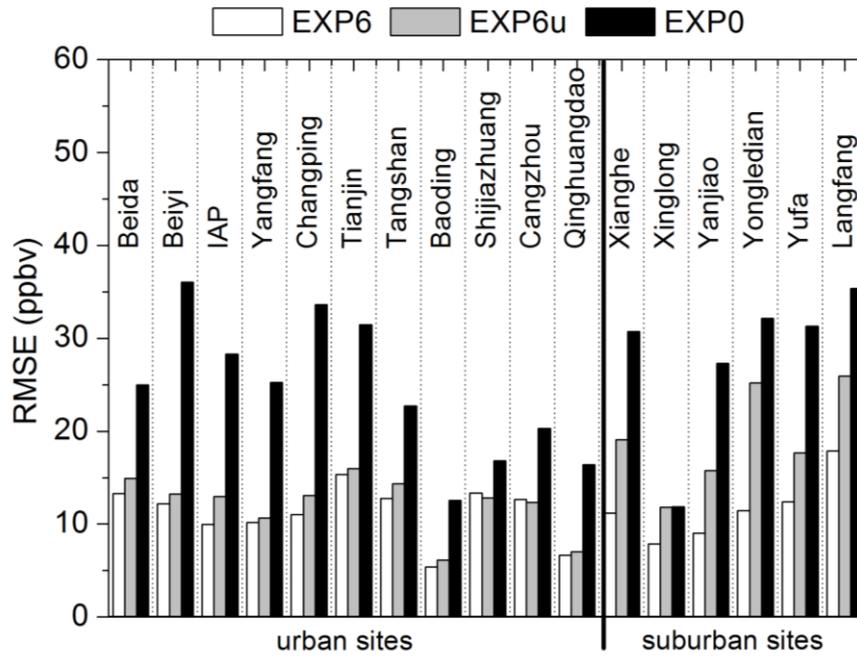
---

### **Appendix C: Cross validation data assimilation experiment**

In order to further evaluate the effects of data assimilation on ozone forecast at the areas without ozone observation, we design two cross validation data assimilation experiments. The 17 monitoring stations are split into two subsets, 11 urban sites and 6 suburban sites. The experiment EXP6u is to assimilate ozone observations at the 11 urban sites with the simultaneous adjustment strategy of EXP6. Ozone observations of

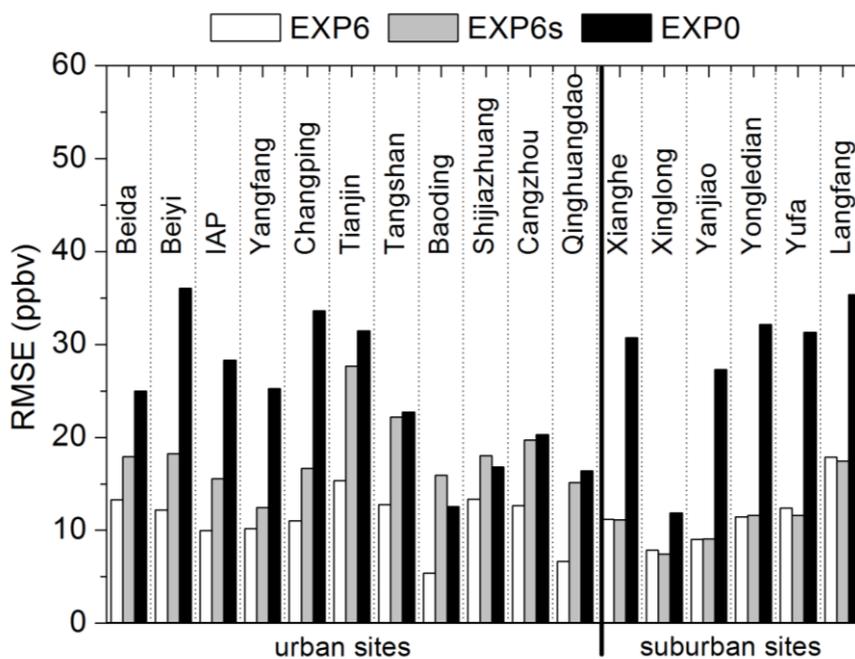
the 6 suburban sites are withheld for validation as independent data. In the other experiment EXP6s, the 6 suburban sites are used for assimilation with the same data assimilation strategy and the 11 urban sites serve as validation sites.

In Fig. 11a, the RMSEs of 1-hour ozone forecast at the 17 sites in EXP6u are compared with those in reference experiment and those in EXP6. It can be seen that data assimilation in EXP6u can reduce the RMSEs at both assimilation sites and independent sites (except for Xinglong). On the other hand, however, the reduction rates of RMSE at independent sites in EXP6u are not as high as those at the same sites in EXP6. It highlights the importance of assimilating local observations in improving ozone forecast of this area. Another interested phenomenon in Fig. 11a is that the RMSEs at several assimilation sites such as Beida and IAP in EXP6u are a little higher than those in EXP6. This phenomenon is probably related to the role of the 6 independent suburban sites in reducing the uncertainty of regional-transport ozone, because these sites are located at the suburban areas between three megacities (Beijing, Tianjin and Tangshan). Assimilating their ozone observations in EXP6 can improve the ozone initial values over these areas and further improve the ozone forecast at their downwind areas. This result also implies that the influence of regional transport should be taken into account for ozone forecast over Beijing and its surrounding cities.



**Fig. 11a.** A comparison of the RMSEs of 1-hour ozone forecast at the 17 sites in EXP6u against those in reference experiment and those in EXP6.

In Fig. 11b, a comparison is made of the RMSEs of 1-hour ozone forecast at the 17 sites in EXP6s against those in reference experiment and those in EXP6. At the six assimilation sites of EXP6s, data assimilation exhibits almost the same performances as in EXP6. At the 11 validation sites, quite different responses of ozone forecast to data assimilation in EXP6s are observed at different sites. Among these validation sites, the RMSEs at five urban sites of Beijing (Beiyi, Beida, IAP, Yangfang and Changping) and one urban sites of Tianjin are significantly reduced by data assimilation in EXP6s. It validates the effectiveness of data assimilation on ozone forecast at the areas without ozone observation. However, at the urban sites of the other cities, data assimilation in EXP6s has not brought marked improvement of ozone forecast. The RMSEs at these sites in EXP6s are a little higher or lower than those in reference experiment, which may be induced by the noise from perturbations or assimilation. The deficiency of data assimilation over these areas is probably caused by the lack of enough observations over these areas. At the five cities of Baoding, Cangzhou, Shijiazhuang, Tangshang and Qinghuangdao, only one urban site has been established for each city in the current regional observation network.



**Fig. 11b.** A comparison of the RMSEs of 1-hour ozone forecast at the 17 sites in EXP6s against those in reference experiment and those in EXP6.

The previous results from the two validation experiment suggest that the current data assimilation strategy with EnKF can improve the ozone forecast not only at assimilation sites, but also at validation sites. And what is more, assimilating local observations is necessary to obtain a best performance of data assimilation over the observational area, especially over the cities with only one monitoring sites.

---

#### **Appendix D: Revised description of methodology**

EnKF, proposed by Evensen (1994), is an approximate version or extension of Kalman filter (Kalman, 1960). It uses a group of stochastic ensemble samples to obtain error statistics of model state variable or parameter. The ensemble mean and ensemble spread of samples are assumed to be the best estimate of state variable or parameter and the error respectively. Error statistics can be propagated with linear or nonlinear dynamic model through simply implementing ensemble simulations of the dynamic model. There are several variants of EnKF (Anderson, 2001; Houtekamer and Mitchell, 2001; Keppenne, 2000; Sakov and Oke, 2008) suitable for applying in

large geophysical system. In this study, we adopt the sequential algorithm proposed by Houtekamer and Mitchell (2001) to implement EnKF for its efficiency in computation. Its implementation and detailed setup are as following.

#### (1) Definition of state vector

In CTM, the state vector  $x$  of model system evolves from time  $k-1$  to time  $k$  can be represented in discrete form:

$$x_k^f = M_{k-1}(x_{k-1}^b, \theta_{k-1}^b) \quad (2.1)$$

where the superscripts  $f$  and  $b$  denote forecast and background (or first guess) respectively,  $M_{k-1}$  denotes the model dynamic operator.  $\theta$  represents model inputs such as meteorological and chemical reaction parameters. The state vector  $x$  defined in this study contains not only conventional state variables (concentrations of the species), but also some parameters. Ozone initial conditions, VOCs and NO<sub>x</sub> initial conditions, and VOCs and NO<sub>x</sub> emission rates consist of the state vector (or control variables) in EnKF. It should be noted that the state vector is able to be extended to include more other variables.

#### (2) Initial perturbation of state vector

A key step of EnKF is generating an initial set of samples of state vector to provide initial conditions for Monte Carlo ensemble simulations. The initial ensemble samples are obtained through perturbing the background values of state vector:

$$x'_{k-1}(i) = x_{k-1}^b + \delta x(i), \quad i = 1, 2, \dots, N \quad (2.2)$$

where  $\delta x$  and  $i$  represents the random perturbation samples added to the background value at the initial time ( $k=1$ ) and its index in ensemble.  $x'_{k-1}$  is the obtained initial ensemble sample of state vector. The ensemble size  $N$  is set as 50, which has been proved to be credible for application in ozone data assimilation by previous publications (Carmichael et al., 2008; Constantinescu et al., 2007).

Ideally, initial perturbation should keep the statistic characterization of the error in background values (Evensen, 2003). For application in CTM, initial perturbation

seems not as important as in application for meteorological and ocean model. Wu et al. (2008) shows credible results of ozone data assimilation without initial perturbation in EnKF. In this study, we employ the method suggested by Evensen (1994) to generate a pseudo smooth random perturbation field in three dimensions. This method is convenient to set the amplitude, horizontal and vertical scales of perturbations. With reference to the uncertainty analysis result of Tang et al. (2010b), the perturbation magnitudes of NO<sub>x</sub> and VOCs emissions are restricted to be within 60% and 80% of the first guess emission rates respectively. The perturbation ranges of initial conditions of O<sub>3</sub>, NO<sub>x</sub> and VOCs are assumed to be 50% of the background values in reference simulation. Initial perturbations on different variables in state vector are assumed to be independent with each other due to the difficulty to obtain their correlations directly. It is worth noting that ensemble simulations can reconstruct the correlations between conventional state variables and emissions but not the correlations between emissions. Therefore, it is possible that some correlations between emissions are underestimated as a result of this configuration, which in turn influences the correlations between conventional state variables closely related to these emissions. The spatial correlation scale of initial perturbation fields is set as 54 km in horizontal and 3 model grids (about 200 m) in vertical after several sensitivity tests. The correlation scale in horizontal or vertical is obtained independently from the best performed data assimilation experiment with the smallest RMSE of analyzed ozone. It should be noted that correlation scale is a tune parameter which can vary with species and space.

### (3) Ensemble forecast of state vector

Propagation of initial error of state vector in model is conducting ensemble runs of original CTM. Each initial ensemble sample obtained in equation (2.2) serves as an initial condition of state vector in each ensemble run:

$$x_k^f(i) = M_{k-1}(x_{k-1}^i(i), \theta_{k-1}^b), \quad i = 1, 2, \dots, N \quad (2.3)$$

In this way, the initial error of state vector is propagated to the forecasted ensemble samples of state vector  $x_k^f(i)$ . It should be noted that the error of the forecasted state vector comes not only from the initial error of state vector, but from error in other sources such as model parameter or numerical technique. We define the latter error as model error in this research.

In data assimilation, how to deal with model error is a very important and difficult issue. Disregarding model error results in an underestimation of background error. In EnKF, it can lead to a serious problem of filter divergence which is characterized by too small ensemble spread and disregard of observation during analysis. A simple method to compensate the missed model errors is inflating background error covariance (Constantinescu et al., 2007). A main drawback of inflation method is lack of physical basis and leading to spurious linear increase of background error at the area far away from observation sites. In this study, we adopt an alternative approach to deal with model error in EnKF. An assumption is made that model error is mainly from the error in few model parameters. An ensemble of model parameter is generated to approximate the error in parameter through perturbing the first guess value of parameter at each integration step:

$$\theta_{k-1}^i(i) = \theta_{k-1}^b + \delta\theta_{k-1}(i) \quad (2.4)$$

where  $\delta\theta$  is the random perturbation sample obtained from Gaussian distribution. Photolysis rates and vertical diffusion coefficients, beside precursor emissions, are assumed to be dominant error sources for the model error and are perturbed in equation (2.4). The perturbation magnitudes of photolysis rate of  $\text{NO}_2$  and vertical diffusion coefficient are restricted to be within 30% and 35% of the first guess values respectively as suggested by Tang et al. (2010b).

In order to integrate model error into the ensemble runs in a smooth way and prevent rapid fluctuations of model error, we adopt a time-correlated noise to generate random perturbation samples of parameter, as suggested by Segers (2002) and von Loon et al. (2000).

The colored noise is simulated by:

$$\delta\boldsymbol{\theta}_k = \alpha\delta\boldsymbol{\theta}_{k-1} + \sqrt{1-\alpha^2}\boldsymbol{\sigma}\mathbf{w}_{k-1}$$

$$\mathbf{w}_k \in N(0, 1)$$

$$\delta\boldsymbol{\theta}_{k-1} = [\delta\theta_{k-1}(1), \delta\theta_{k-1}(2), \dots, \delta\theta_{k-1}(N)] \quad (2.5)$$

where  $\mathbf{w}_{k-1}$  is a sequential of white noise,  $\boldsymbol{\sigma}$  denotes the standard deviation of error in parameter.  $\alpha$  represents the smooth coefficient dependent on time decorrelation scale ( $\tau$ ):

$$\alpha = \exp(-1/\tau) \quad (2.6)$$

We use 24 hours as the first guess value of the time decorrelation scale. The same decorrelation length has been used by Segers (2002) to simulate the uncertainty in emissions, photolysis rates and deposition parameters. It is worth noting that this assumption may not be true. Other options might improve the performance of EnKF.

After integrating the model error in equation (2.4) into the ensemble model runs, the equation (2.3) is transformed to:

$$\mathbf{x}_k^f(i) = M_{k-1}(\mathbf{x}_{k-1}^f(i), \boldsymbol{\theta}_{k-1}^f(i)), \quad i = 1, 2, \dots, N \quad (2.7)$$

#### (4) Update of state vector

After obtaining a group of ensemble samples of state vector, a key step of EnKF is updating the forecasted state vector with assimilating observation data. The forecast error covariance  $\mathbf{P}^f$  of state vector is estimated based on the forecast ensemble samples of equation (2.7):

$$\mathbf{P}_k^f = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{x}_k^f(i) - \overline{\mathbf{x}_k^f})(\mathbf{x}_k^f(i) - \overline{\mathbf{x}_k^f})^T \quad (2.8)$$

where the overline denotes the ensemble mean of samples.

The observational error is assumed to be Gaussian with mean zero and covariance  $\mathbf{R}$ .

An ensemble of observation samples is generated accordingly:

$$y_k^f(i) = y_k + \eta_k(i), \quad i = 1, 2, \dots, N$$

$$\eta_k \in N(0, \mathbf{R}_k) \quad (2.9)$$

where  $y_k$  is the original observation value and  $\eta_k$  is the random perturbation sample from Gaussian distribution. The observation error, including both representative error and measurement error, is assumed to be uncorrelated in time and space. The amplitude of ozone observation is set as 10% of the original observation value with reference to von Loon et al. (2000).

Based on the error statistics of the forecast and observational state vector, the state vector is updated:

$$x_k^a(i) = x_k^f(i) + \mathbf{K}_k(y_k' - \mathbf{H}x_k^f(i)), i = 1, 2, \dots, N$$

$$\mathbf{K}_k = \mathbf{P}_k^f \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k^T + \mathbf{R}_k)^{-1} \quad (2.10)$$

$\mathbf{H}$  denotes the observation operator mapping the state vector in model space to the expected value in observation space.  $\mathbf{K}$  represents the Kalman gain dependent on forecast error covariance and observation error covariance.  $x_k^a$  is the updated state vector, or analysis of state vector. In order to assimilate the observation data in an efficient way, we assimilate the ozone observations at different sites in a sequential way. Only observation at one site is assimilated at each analysis step in equation (2.10), and then the updated state vector is used as the background for assimilating observation at next sites. The sequential way is suggested to be better than the way with observations of all sites assimilated once (Houtekamer and Mitchell, 2001).

A major limitation of EnKF is using finite ensemble size which leads to spurious correlation between two independent variables in background error covariance, underestimation of analysis error covariance, and spurious increment of state vector during analysis (Evensen, 2009). In this study, a local analysis scheme is employed to reduce the spurious influence of remote observation during analysis which is caused by the finite ensemble size. Updating state vector of one grid only uses the observations within a certain distance (localization scale) of this grid. An optimal localization scale can efficiently eliminate the influence of spurious correlation without excluding useful observations and bring better forecast skill than other scales. The optimal localization scale can vary with ensemble size, dynamic system and life cycle of chemical species. The localization scale is set as 54 km for updating ozone

initial conditions and 45 km for updating precursor initial conditions and emissions after several sensitivity tests.