Response to Referee #1 (ACPD-11-C2563–C2566)

Tang et al. describe an application of data assimilation to improve Ozone forecasts in the Beijing area. From the methodology point of view the work is not on the cutting edge. The Ensemble Kalman filter algorithm used, the model error term introduced and the localization method used have all been employed before for Ozone prediction problems (see also the references mentioned by the authors). The interesting part might be the application to the Beijing area. But then it should become more clear what we learn from this specific application. The presentation of the methodology is a weak point. It isn't sufficiently related to other work and the justification is iffy. In my view the paper is not yet ready. I suggest a major revision.

Reply: The authors very much appreciate the reviewer for the constructive and up-to-point comments. We have carefully considered the comments and have revised the manuscript accordingly.

In order to specify the contributions of this study, we have made a substantial revision to the abstract and introduction 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 revised manuscript.

A substantial revision to the description of the methodology has also been made in the revised manuscript. Please refer to Appendix C.

A point-by-point response to the review's comments is given in the following part.

Major comments

1) *The notation used for the state space formulation of the model is not clear. To achieve a better understanding of the paper an established notation should be used. For applications of data assimilation in the field of weather prediction, oceanography and air quality there is a standard notation that is often used: See "Unified Notation for Data Assimilation: Operational, Sequential and Variational.", K. Ide et al, 1997.*

Reply: Thanks for pointing out this issue. According to your comments, a standard

notation has been used to descript the state space of the model in the revised manuscript. Please refer to the revised methodology description in Appendix C.

2) *An important issue in data assimilation is the representation of the model error. If one starts with reading the equations (1) and (2) and then read the equation (7) for the model error it is not clear how the model error is related to the state vector and how the model error is actually implemented. The use of an established notation could help to solve this problem.*

Reply: Thank you very much for raising these important issues. We have made a substantial revision to the description of the model error. Please refer to the equations (2.3)-(2.4)-(2.5)-(2.6)-(2.7) in Appendix C.

3) *The choice of the Ensemble Kalman filter algorithm used is not motivated. There are many options here. The scheme used is generally not the best one. A number of very attractive schemes have been introduced recently, see e.g. "Implications of the form of the ensemble transformation in the ensemble square root filters", Sakov P, Oke PR, MONTHLY WEATHER REVIEW, 2008. At least a discussion should be added to convince the reader that the chosen algorithm is a good option for this application.*

Reply: We agree. We have specified the motivation for employing the ensemble Kalman filter algorithm in the revised manuscript.

"*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.*"

"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."

4) *Regarding the colored noise introduced in equation (7) the authors should motivate the specific choice of the equation: qt is now a stationary process with variance sigma/(1-alpha^2). Also they should motivate the choice of the decorrelation length. Has a sensitivity study been carried out to come up with the value used in this study?*

Reply: Sorry, there are typos in equation (7) in the previous manuscript. We have corrected these errors in the revised manuscript. Please refer to the equation (2.5) and (2.6) in Appendix C. The reasons for choosing this equation have been specified in the revised manuscript. "*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 configuration of the decorrelation length of model error in this study is not based on a sensitivity study. It is just a first guess value to obtain smooth model error. The same decorrelation length has been used by Segers (2002) to simulate the uncertainty in emissions, photolysis rates and deposition parameters. We have clarified it and also pointed out the possible uncertainty as a result of this choice in the revised manuscript. "*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.*"

5) *As one can see in figure 1 the model covers 3 different domains: The largest domain D1 and the nested domains D2 and D3. It is not clear from the text how the boundary conditions have been taken into account in case of the nested models. It is well known that there might be difficulties in the treatment of the*

boundary conditions in these kinds of problems.

Reply: Thanks for pointing out this issue. The employed chemical transport model in this study supports both one-way and two-way nesting techniques at a given nesting ratio. Wang et al. (2001) provide a detailed description of the nesting technique in this model: "*NAQPMS contains a capability of nesting with one-way and two-way interacting at given nesting ratio. "Two-way interaction" means that the nest's input from the coarse mesh comes via its boundaries, while the feedback to the coarser mesh occurs over the nested interior. The "one-way" nesting differs from the two-way nesting in having no feedback and coarse temporal resolution at the boundaries.*"

According to your comment, we have clarified it in the revised manuscript. "*In this study, the model is configured with three nested model domains (displayed in Fig. 1a). The coarse domain (D1 and D2) is to provide boundary conditions for its nested domain with one-way nesting technique. The boundary conditions of the largest domain (D1) are provided by a global transport model CHASER (Sudo et al., 2002) at three hour intervals.*"

- *It is important to validate the results presented, e.g. by considering validation data that is not used in the assimilation experiments. Validation issues are hardly discussed in the paper.*

Reply: Thanks for pointing out the issue of validation. 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 D.

Minor remarks

- *page 7813, line 18, there is a reference to "30 ppbv" without defining what the*

abbreviation "ppbv" means.

line 25, where the authors define data assimilation: "... into three-dimensional model...". DA is not restricted to three-dimensional models.

Reply: We agree. We have revised these in the revised manuscript. "*Part Per Billion by Volume (ppbv)*" "*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.*"

- *page 7314, equation (2): It seems to me that the noise gamma should also depend on i.*

Reply: We agree. We have revised this in the revised manuscript. Please refer to the equation (2.4) in Appendix C.

- *page 7815. Concepts as 4D variational methods and the adjoint are presented here without introduction and explanation, and without references.*

  line 17, The following sentence needs rewriting, because it is not clear what the authors would like to say here: "The system is employed for constraining input uncertainties of ozone modeling including ozone initial conditions, precursors (NOx and VOC) initial conditions and emissions".

Reply: Thanks for your comment. These sentences have been deleted through the major revision of the methodology description in the revised manuscript.

- *page 7816. After equation (2) the choice of 50 ensemble is made. References to others are added, but how similar are the referred applications with the application studied in the submitted paper? Why would 50 also be good for the application presented here?*

Reply: Thanks for pointing out this issue. The two studies cited have applied EnKF for ozone data assimilation in chemical transport model as our present study. Furthermore, we have verified the effectiveness of this ensemble size in application for the current research through sensitivity experiments which are configured with the

same model setting and observation network as this study. The detailed results as in Fig. 1 were not presented in the manuscript due to limit of space. The result shown in Fig. 1 suggests that 50 samples can keep a good balance between computational efficiency and assimilation performance. According to your comment, we have specified the reasons for the choice of ensemble size in the revised manuscript. "*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).*"

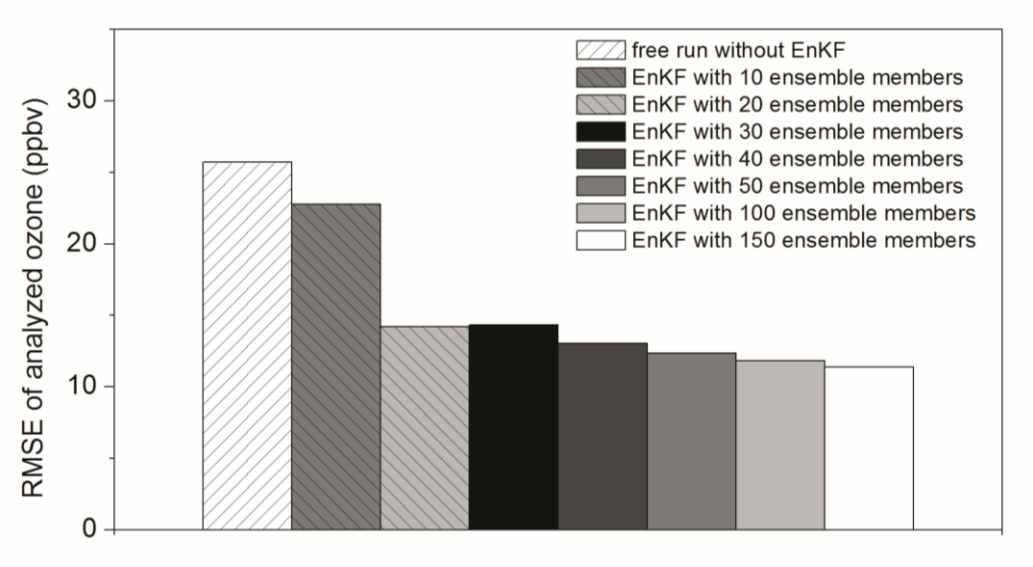

Fig. 1. Sensitivity of root mean square error (RMSE) of the analyzed ozone over Beijing and its surrounding areas to ensemble size in data assimilation experiments with ensemble Kalman filter (EnKF).

- *page 7817, line 11. The operator M cannot be the same as the operator M in equation (1) since the dimension of V is larger than the dimension of x. M should be redefined here.*

  line 16: It seems to me that tau1, tau2, ... should depend on t+1 too.

Reply: We agree. We have revised these. Please refer to Appendix C.

- *page 7818, line 19. The authors discuss the " : : : underestimation of analysis error covariance". A reference to this result should be added or an additional explanation is needed.*

  line 23. The authors discuss the localization used. They are referring to an

"optimal" one. The localization option is optimal with respect with what? What optimality criterion is used here?

Reply: We have added a reference to clarify this ambiguity. "*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)*"

An optimal localization scale is defined in this manuscript based on the criterion that this scale can bring the best forecast skill compared with other scales. We have clarified this in the revised manuscript. "*An optimal localization scale can efficiently eliminate the influence of spurious correlation without excluding useful observations and bring better forecast skill than other scales.*"

References

Segers, A.: Data assimilation in atmospheric chemistry models using Kalman filtering, Ph.D. thesis, Delft Univ., Delft, Netherlands. 2002.

Wang, Z., Maeda, T., Hayashi, M., Hsiao, L.F., and Liu, K.Y.: A nested air quality prediction modeling system for urban and regional scales: application for high-ozone episode in Taiwan, Water, Air, and Soil Pollution, 130, 391-296, 2001.

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 conceptional 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: 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. θ 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 NOx initial conditions, and VOCs and NOx 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:

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

where δx 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 NOx 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_3, NOx 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:

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

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:

$$\theta'_{k-1}(i) = \theta^{b}_{k-1} + \delta\theta_{k-1}(i) \tag{2.4}$$

where δθ 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 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:

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

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

$$\delta\mathbf{\theta}_{k-1} = [\delta\theta_{k-1}(1), \delta\theta_{k-1}(2), ..., \delta\theta_{k-1}(N)] \tag{2.5}$$

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

$$\alpha = \exp(-1/\tau) \tag{2.6}$$

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:

$$x_k^f(i) = M_{k-1}(x_{k-1}^{'}(i), \theta_{k-1}^{'}(i)),\ i = 1,2,...,N \tag{2.7}$$

(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 $\mathbf{P}^{\mathrm{f}}$ of state vector is estimated based on the forecast ensemble samples of equation (2.7):

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

where the overline denotes the ensemble mean of samples.

The observational error is assumed to be Gaussian with mean zero and covariance **R.** An ensemble of observation samples is generated accordingly:

$$y_k^{'}(i) = y_k + \eta_k(i),\ i = 1,2,...,N$$

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

where y_k is the original observation value and η_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:

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

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

H denotes the observation operator mapping the state vector in model space to the expected value in observation space. **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.

Appendix D: 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.

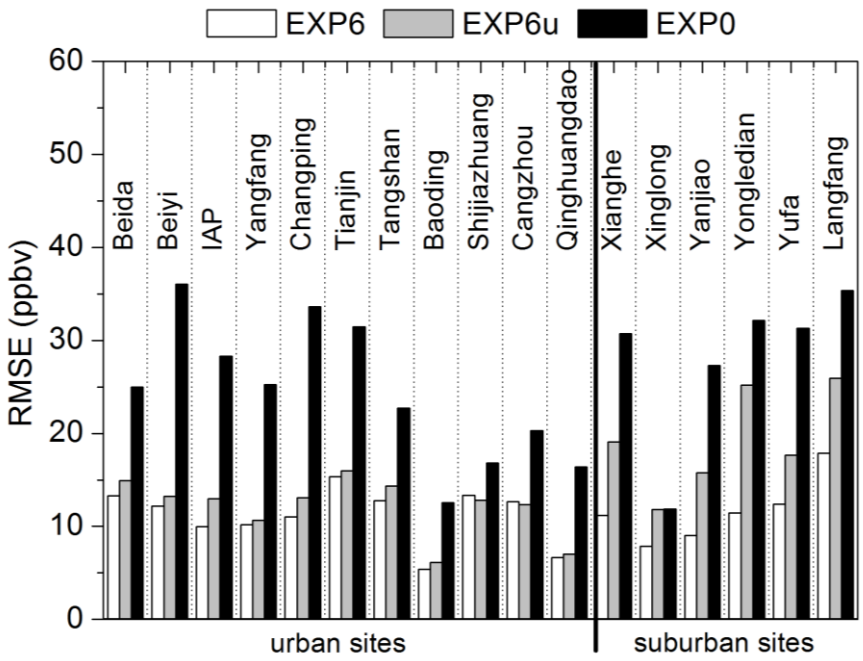

Fig. 11a. A comparison of the RMSEs of 1-hour ozone forecast at the 17 sites in EXP6u against those in reference experiment and those in EXP6.

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.

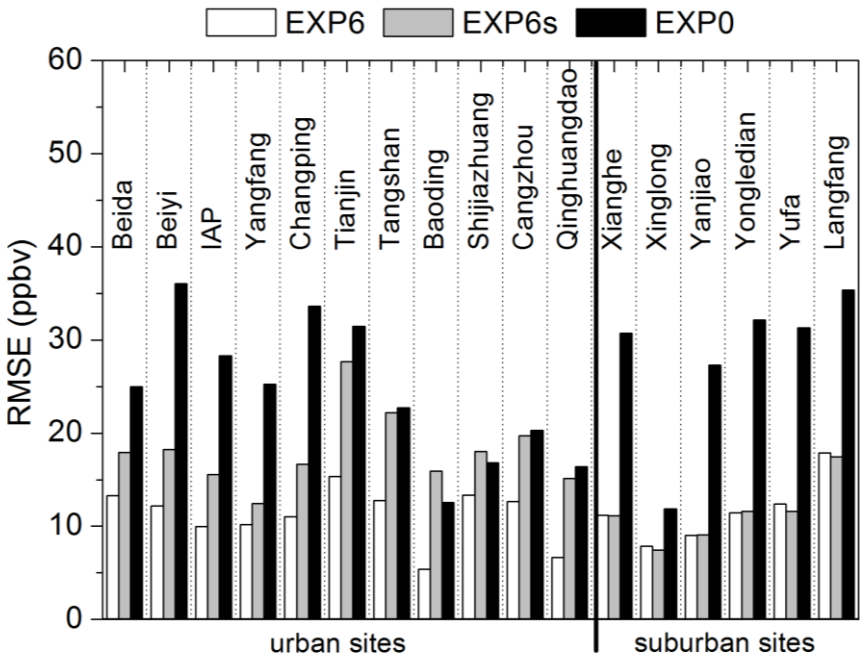

Fig. 11b. A comparison of the RMSEs of 1-hour ozone forecast at the 17 sites in EXP6s against those in reference experiment and those in EXP6.

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.