Responses to the comments of Reviewer #1:

We would like to thank the anonymous referee for his/her comprehensive review and valuable suggestions. These suggestions help us to present our results more clearly. In response, we have made changes according to the referee's suggestions and replied to all comments point by point. All the page and line number for corrections are referred to the revised manuscript, while the page and line number from original reviews are kept intact.

General comments:

1. A major problem arises due to the choice of the Ensemble Kalman Filter (EnKF) to perform the emission inversion. A key assumption of the EnKF, and indeed of any sequential DA method, is that the model minus observation errors have mean zero bias, be random, and have a Gaussian distribution. The authors' own results show here that this assumption is invalidated. They report an a posteriori emission scalings of between 5-1045% for each of the pollutants studied – this strongly implies the presence of large biases between the observations and model when run with a priori emissions. Such large biases will affect the optimality of the EnKF solution, and it is essential that the authors consider this in a revised manuscript. For solutions the authors should look to other fields. The problems of bias with the EnKF have been well discussed in its application in the fields of land surface (Brandhorst et al., 2017; De Lannoy et al., 2007) and in oceanographic DA (e.g., Keppenne et al., 2005). Air quality/atmospheric chemistry cannot be exempt from this. Extensive published work exists trying to solve this problem in other fields by exploring various possibilities: bias estimation as an additional state term, removal of observation bias through statistical methods, and parameter estimation. The parameter estimation presented in Brandhorst et al. (2017) gives a clear pathway for the authors to resolve this problem. If emission inversion could be framed as a bias correction method via parameter estimation (citing the correct literature), rather than being an end and means unto itself alone, then I think this would be acceptable. Right now, emission estimation is present in the first sentence of the

abstract and appears to be the primary focus, but due to the limitations of the EnKF this emphasis combined with the current approach is problematic.

Response: We really appreciate these valuable comments and suggestions. Yes, we agree that optimal state estimated using a EnKF relies on the assumption of zero mean forecast error, which is not guaranteed in such highly nonlinear and large biases systems (e.g., Chen et al., 2019). In this study, some pollutants (e.g., CO, PMC) may have very large simulated biases, thus if a small uncertainty is adopted, the emission bias cannot been fully reduced, while if a very large uncertainty is adopted, then the degree of freedom of adjustment is too large, and the inverted daily emissions will fluctuate abnormally. Therefore, we only set a larger prior uncertainty in the first three windows, adopted a moderate uncertainty in the following windows, and used a "two-step" inversion scheme and cyclic iteration to gradually correct the emission errors. Figure R1(a) shows the time series of the relative differences between the prior and posterior emissions in each window. There are the relatively large adjustments for the emissions in the first three windows, especially for coarse PM_{10} (PMC), but the adjustment ranges of the five species after the first 3 windows are basically within the uncertainty range (e.g., $\pm 25\%$), indicating that with this scheme, the EnKF method used in this system still has a good performance in emission inversion. For PMC, because the posterior emission includes the contribution of natural dust, which dominates in PMC emissions in China (Li et al., 2010; Lu et al., 2020; Ma et al., 2019), but it was not included in the original emission inventory (MEIC 2016). The large increment (1045%) of PMC inverted in this study is consistent with the result of Ma et al (2019).

The model error exists objectively, and this study ignores the model error, which will indeed have a certain impact on the inversion results. Previous studies using parameter estimation or/and bias correction method to account for a certain degree of the model errors, including random and systematic errors (Brandhorst et al., 2017; De Lannoy et al., 2007; Evensen, 2009). The bias correction method is usually used for correcting the systematic error (Baek et al., 2006; Danforth et al., 2007; Ruiz and Pulido, 2015). De Lannoy et al. (2007) assumed that the bias evolves more slowly than the forecast. A

good bias forecast model is needed to represent the variation characteristics of model error especially when the bias evolves quickly (De Lannoy et al., 2007). However, the biggest dilemma is to assume that the covariance structure of the bias errors is a fraction of the state errors, which cannot be guaranteed if emission is framed as a bias correction term. In Brandhorst et al. (2017), they directly applied the bias correction method to the observations, which cannot give a bias correction for the emissions in the whole domain. Additionally, their studies interpret roughly constant mismatch between simulated and measured values as a bias. Therefore, it may be inflexible for correcting emission with accident or drastic changes (e.g., primary $PM_{2.5}$ during 16-20 December). Moreover, it is quite difficult to identify several key uncertain parameters of different species, which comes not only from the complexity of the atmospheric chemical model itself, but also from hundreds of other model input variables (Tang et al., 2013). Therefore, considering model errors is still a big challenge in emission inversions (e.g., Miyazaki et al., 2012; Peng et al., 2018; etc.). In general, there is currently no reasonable and convincing algorithm to describe the model error well for atmospheric data assimilation (Houtekamer and Zhang, 2016), and the emission is basically regarded as the state variable in current inversion studies (e.g., Ma et al., 2019; Miyazaki et al., 2015; etc.).

We have added related discussions in the revised paper. The changes are also listed as follows:

Lines 1082-1129, pages 57-59.

"Optimal state estimation using an EnKF relies on the assumption of unbiased Gaussian prior error, which is not guaranteed in such highly nonlinear and large biases systems In this study, some pollutants (e.g., CO, PMC) have very large simulated biases, thus if a small uncertainty is adopted, the emission bias cannot been fully reduced, while if a very large uncertainty is adopted, then the degree of freedom of adjustment is too large, and the inverted daily emissions will fluctuate abnormally. Therefore, we only set a larger prior uncertainty in the first three windows, adopted a moderate uncertainty in the following windows, and used a "two-step" inversion scheme and cyclic iteration to gradually correct the emission errors. Figure 10(a) shows the time series of the relative

differences between the prior and posterior emissions in each window. There are the relatively large adjustments for the emissions in the first three windows, especially for PMC, but the adjustment ranges of the five species after the first 3 windows are basically within the uncertainty range (e.g., ±25%), indicating that with this scheme, the EnKF method used in this system still has a good performance in emission inversion.

The model-data mismatch error not only comes from the emissions, but also from the inherent model errors arising from model structure, discretization, parameterizations and the biases in the simulated meteorological fields. Neglecting model errors would attribute all uncertainties to emissions, and lead to considerable biases in the estimated emissions. In the version of CMAQ model used in this study, there is no heterogeneous reactions (Quan et al., 2015; Wang et al., 2017), the parameterization scheme for the formation of secondary organic aerosol (SOA) is imperfect (Carlton et al., 2008; Jiang et al., 2012; Yang et al., 2019), no feedback between chemistry and meteorology is considered, and we used an idea profile for chemical lateral boundary conditions. All of the above problems can lead to underestimated concentrations of pollutants, which in turn require more emissions to compensate, leading to overestimations in emissions. In addition, previous studies have shown that the emission of ammonia in the MEIC inventory was underestimated (Kong et al., 2019b; Paulot et al., 2014; Zhang et al., 2018). Due to lack of ammonia observations, our system does not include emission estimates of ammonia, which means that the concentration of ammonium aerosol was underestimated in this system, also resulting in an overestimation in the $PPM_{2.5}$ emission. Wind-blown dust was also not simulated here, thus the PMC emission inverted in this system do not only come from anthropogenic activities, but also from natural sources. Although some of these shortcomings could be solved in the future by updating the CTM model, there will still be errors in each parameterization and each process. Generally, parameter estimation method was used to reduce the model errors, in which, some uncertain parameters were included in the augmented state vector and were optimized synchronously based on the available observations (Brandhorst et al., 2017; Evensen, 2009). However, it is still quite difficult to identify the key uncertain

parameters of different species in different models, which generally comes not only from the complex atmospheric chemical model, but also from hundreds of model inputs (Tang et al., 2013). Another method is bias correction, which treats the model error as a bias term, and includes it in the augmented state vector (Brandhorst et al., 2017; De Lannoy et al., 2007; Keppenne et al., 2005). In addition, the weak-constraint 4D-Var method can also be used to reduce the model errors, which adds a correction term in the model integration to account for the different sources of model error (Sasaki, 1970). Although reliable diagnosis of model error is still a challenge at present (Laloyaux et al., 2020), it should be considered in an assimilation system. We will consider model errors in our system in the future to obtain better emission estimates."
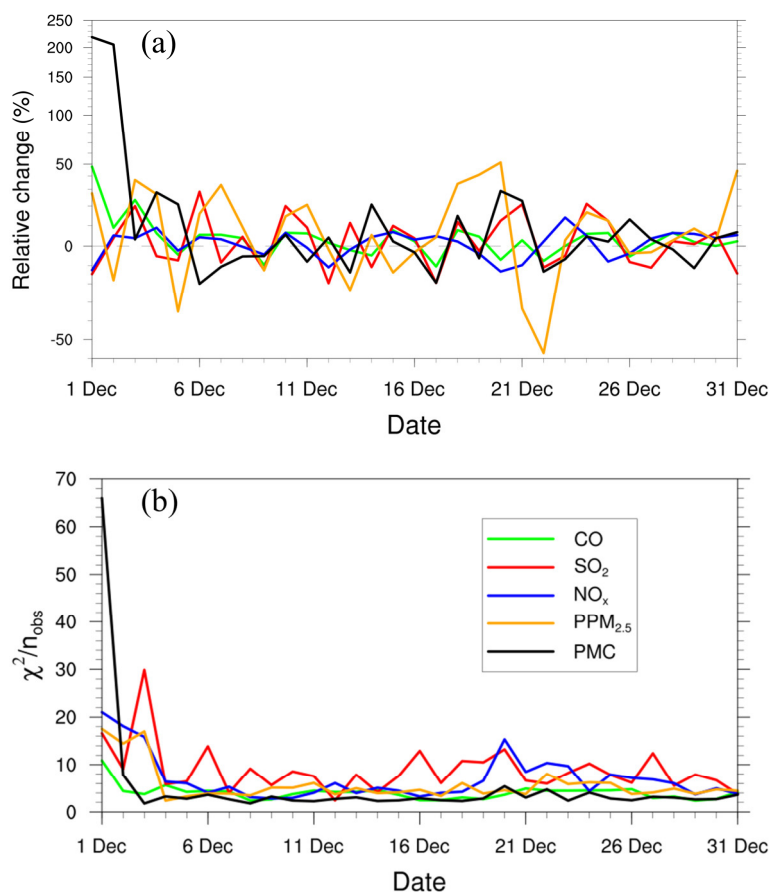

Figure R1. Relative changes (a) in a posteriori emission estimates of CO, SO_2, NO_x, $PPM_{2.5}$ and PMC, and χ^2 statistics (b) of these state vectors in each window. (Figure 10 in the revised manuscript)

Baek, S. J., Hunt, B. R., Kalnay, E., Ott, E., and Szunyogh, I.: Local ensemble Kalman filtering in the presence of model bias, Tellus Series a-Dynamic Meteorology and Oceanography, 58, 293-306, 2006.

Brandhorst, N., Erdal, D., and Neuweiler, I.: Soil moisture prediction with the ensemble Kalman filter: Handling uncertainty of soil hydraulic parameters, Advances in Water Resources, 110, 360-370, 2017.

Chen, D., Liu, Z., Ban, J., and Chen, M.: The 2015 and 2016 wintertime air pollution in China: SO2 emission changes derived from a WRF-Chem/EnKF coupled data assimilation system, Atmospheric Chemistry and Physics, 19, 8619-8650, 2019.

Danforth, C. M., Kalnay, E., and Miyoshi, T.: Estimating and correcting global weather model error, Monthly Weather Review, 135, 281-299, 2007.

De Lannoy, G. J. M., Reichle, R. H., Houser, P. R., Pauwels, V. R. N., and Verhoest, N. E. C.: "Correcting for forecast bias in soil moisture assimilation with the ensemble Kalman filter" (vol 43, artn W09410, 2007), Water Resources Research, 43, 2007.

Evensen, G.: The Ensemble Kalman Filter for Combined State and Parameter Estimation MONTE CARLO TECHNIQUES FOR DATA ASSIMILATION IN LARGE SYSTEMS, Ieee Control Systems Magazine, 29, 83-104, 2009.

Feng, S., Jiang, F., Wang, H., Wang, H., Ju, W., Shen, Y., Zheng, Y., Wu, Z., and Ding, A.: NOx Emission Changes Over China During the COVID-19 Epidemic Inferred From Surface NO2 Observations, Geophysical Research Letters, 47, 2020.

Houtekamer, P. L. and Zhang, F.: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation, Monthly Weather Review, 144, 4489-4532, 2016.

Li, J.-d., Deng, Q.-h., Lu, C., and Huang, B.-l.: Chemical compositions and source apportionment of atmospheric PM10 in suburban area of Changsha, China, Journal of Central South University of Technology, 17, 509-515, 2010.

Lu, P., Zhao, X.-Y., Yin, B.-H., Zhang, N., Wang, X.-H., Yu, H., Yang, W., and Wang, X.-L.: Distribution Characteristics and Source Apportionment of Elements Bonded with PM2.5 and PM10 in Linyi, Environment Science, 41, 2036-2043, 2020.

Ma, C., Wang, T., Mizzi, A. P., Anderson, J. L., Zhuang, B., Xie, M., and Wu, R.: Multiconstituent Data Assimilation With WRF-Chem/DART: Potential for Adjusting Anthropogenic Emissions and Improving Air Quality Forecasts Over Eastern China, 124, 7393-7412, 10.1029/2019jd030421, 2019.

Miyazaki, K., Eskes, H. J., and Sudo, K.: Global NOx emission estimates derived from an assimilation of OMI tropospheric NO2 columns, Atmospheric Chemistry And Physics, 12, 2263-2288, 2012.

Miyazaki, K., Eskes, H. J., and Sudo, K.: A tropospheric chemistry reanalysis for the years 2005-2012 based on an assimilation of OMI, MLS, TES, and MOPITT satellite data, Atmospheric Chemistry and Physics, 15, 8315-8348, 2015.

Peng, Z., Lei, L., Liu, Z., Su, J., Ding, A., Ban, J., Chen, D., Kou, X., and Chu, K.: The impact

of multi-species surface chemical observation assimilation on air quality forecasts in China, Atmospheric Chemistry and Physics, 18, 2018.

Ruiz, J. and Pulido, M.: Parameter Estimation Using Ensemble-Based Data Assimilation in the Presence of Model Error, Monthly Weather Review, 143, 1568-1582, 2015.

Tang, X., Zhu, J., Wang, Z. F., Wang, M., Gbaguidi, A., Li, J., Shao, M., Tang, G. Q., and Ji, D. S.: Inversion of CO emissions over Beijing and its surrounding areas with ensemble Kalman filter, Atmospheric Environment, 81, 676-686, 10.1016/j.atmosenv.2013.08.051, 2013.

2. Furthermore, the authors need to make use of the available diagnostics exist to help diagnose the optimality of the EnKF analysis, e.g., the chi-squared diagnostic. Based on reported examples in the literature, my guess is that with each sequential iteration of the EnKF the emission biases are slowly reduced (prior to the reported overestimation) and the EnKF will slowly reach a more optimal state. The authors should plot chi-squared along with the relative change in a posteriori emission estimates for each successive emission inversion cycle with the EnKF. Doing this should highlight whether the optimality improves over time. If it does not then it would imply other fundamental problems with the approach.

Response: Thank you very much for this suggestion. We have added a subsection and a figure (Figure R1 in this response manuscript, and Figure 10 in the revised manuscript) about the chi-squared statistics along with the relative changes of the posterior emission estimates. Following descriptions and figure are also added in the revised manuscripts. See lines 796-818, pages 40-41.

4.1.5 Evaluation using chi-squared statistics

To diagnose the performance of the EnKF analysis, the chi-squared (χ^2) statistics was calculated, which is generally used to test whether the prior ensemble mean RMSE with respect to the observations is consistent with the prior "total spread" (square root of the sum of ensemble variance and observation error variance). Following Zhang et al. (2015), for the tth window, χ^2 is defined as

$$\chi_t^2 = (y - H\overline{X^b})^T (HP^b H^T + R)^{-1} (y - H\overline{X^b}) \qquad (20)$$

Figure 10 shows the time series of the relative changes between the prior and posterior emissions and the χ^2 statistics. There are relatively large adjustments of emissions in the first three windows, especially for PMC. After that, the optimality of the five species reaches a more optimal state with successive emission inversion cycle. The χ^2 statistics shows a similar variation characteristics with the daily changes in the emissions. The χ^2 value is slightly greater than 1, indicating that the uncertainties from error covariance statistics do not fully account for the error in the ensemble simulations. A similar situation also appeared in Chen et al. (2019). Further investigations should be conducted to generate larger spreads by accounting for the influence of model errors. Since we imposed a same uncertainty of prior emission at each DA window to partially compensate for the influence of model errors, χ^2 statistics showed small fluctuations, indicating that the system updates emissions consistently and stably.

3. The combination of the 3D-Var and EnKF together could be more clearly motivated and the choice of combining these two algorithms seems somewhat expedient. There are various self-citations justifying this choice, but the authors should dedicate some text for justifying why the two algorithms combined represent something greater than the sum of their parts. I am not saying a justification cannot be made, but we are missing a fundamental justification of the method choice.

Response: Thank you for your suggestion. Previous studies have demonstrated that emissions would be adjusted unreasonably to compensate initial field error if the initial field is not optimized (Miyazaki et al., 2017; Tang et al., 2013). An EMDA experiment without the IA step also shows a great impact of the initial field optimization on the emission inversion (Comment 11). Therefore, it is essential to remove the bias in initial field for the subsequent emission inversion. To get a better initial condition during spin-up period, we directly use 3DVAR module developed before and did not need to extend

EnKF module to optimize initial field. Jiang et al., (2017) also used a combination of two methods to optimize a 15-year record of CO emissions. Additionally, it has been found that the 3DVAR method can obtain a better initial field than the EnKF method (Schwartz et al., 2014). We have added following words in the revised manuscript. See lines 197-200, page 7.

"The introduction of 3DVAR mainly considers its great performance based on our previous study and lower computational cost during spin-up period in optimizing ICs. Additionally, it has been found that the 3DVAR method can obtain a better initial field than the EnKF method (Schwartz et al., 2014)."

Jiang, Z., Worden, J. R., Worden, H., Deeter, M., Jones, D. B. A., Arellano, A. F., and Henze, D. K.: A 15-year record of CO emissions constrained by MOPITT CO observations, Atmospheric Chemistry And Physics, 17, 4565-4583, 10.5194/acp-17-4565-2017, 2017.

Miyazaki, K., Eskes, H., Sudo, K., Boersma, K. F., Bowman, K., and Kanaya, Y.: Decadal changes in global surface NOx emissions from multi-constituent satellite data assimilation, Atmospheric Chemistry and Physics, 17, 807-837, 2017.

Schwartz, C. S., Liu, Z., Lin, H.-C., and Cetola, J. D.: Assimilating aerosol observations with a "hybrid" variational-ensemble data assimilation system, Journal Of Geophysical Research-Atmospheres, 119, 4043-4069, 10.1002/2013jd020937, 2014.

Tang, X., Zhu, J., Wang, Z. F., Wang, M., Gbaguidi, A., Li, J., Shao, M., Tang, G. Q., and Ji, D. S.: Inversion of CO emissions over Beijing and its surrounding areas with ensemble Kalman filter, Atmospheric Environment, 81, 676-686, 10.1016/j.atmosenv.2013.08.051, 2013.

4. As an example of potential problems with their combination, it seemed odd to first do the 3D-Var IA step and then the EnKF EI step afterwards for a mixture of long- and short-lived gases. In fact, if you optimize the concentration of CO, which is long-lived relative to the DA windows and DA cycles, then this should remove some of the bias associated with low a priori emissions of CO. The authors say this removes some of the bias associated with errors from other issues other than emissions, but this raises other

issues – see next comment (5). I think this would be fine if optimizing the concentrations for forecasting purposes is the only goal, but if the emission estimation is a primary or secondary goal (as presented) then I think that this goal is undermined by the approach taken. At the very minimum the authors should include the changes in the emission estimates obtained from and experiment with the EI without first running the 3D-var. If the goal was emission estimation, then it would seem to make more sense to perform a long spin-up simulation prior to running the EI step rather than to partially fix the model concentration field prior to this. The authors cite concerns about other types of model bias as a justification for this step, but I do not find this fully convincing. Surely, if the emissions are biased to the extent shown, this would impact the a priori ICs as much as the model run after the IA step. The large biases discussed in sect. 4.1.2 between the ICDA and ICNO experiments seem to be consistent with this point. Indeed, the authors even find that the spatial patterns in the simulated concentrations using posteriori and a priori emissions are similar to the increments calculated in the 3D-var, which indicates a potentially similar cause in both cases. I think the authors need to make at least another experiment to test the effect of this in the EI step (see specific comment 11).

Response: Thank you for this comment. Our goal is to optimize the emission. The goal of 3D-Var in the IA step is used for getting a better IC for the subsequent EI step. Besides model errors, the model-data mismatch errors in the EI step mainly come from the IC and the emissions. Previous studies have demonstrated that there would be large emission discrepancies resulting from the initial field errors (Jiang et al., 2013; Miyazaki et al., 2017; Tang et al., 2013), Removing the bias in initial conditions as much as possible is essential for inverse analysis (Jiang et al., 2017). In the revised manuscript, we also added a new inversion experiment, in which the IA step was stopped. The results show that the optimized initial field has the greatest impact on the emission inversion of long-lived species (e.g., CO). The maximum difference in the first few days of posterior CO emissions can reach 26.1%. With the gradual disappearance of the benefit of initial field assimilation, the two experiments can reach

a unified state after some windows.

Jiang, Z., Jones, D. B. A., Worden, H. M., Deeter, M. N., Henze, D. K., Worden, J., Bowman, K. W., Brenninkmeijer, C. A. M., and Schuck, T. J.: Impact of model errors in convective transport on CO source estimates inferred from MOPITT CO retrievals, Journal Of Geophysical Research-Atmospheres, 118, 2073-2083, 2013.

Jiang, Z., Worden, J. R., Worden, H., Deeter, M., Jones, D. B. A., Arellano, A. F., and Henze, D. K.: A 15-year record of CO emissions constrained by MOPITT CO observations, Atmospheric Chemistry And Physics, 17, 4565-4583, 2017.

Miyazaki, K., Eskes, H., Sudo, K., Boersma, K. F., Bowman, K., and Kanaya, Y.: Decadal changes in global surface NOx emissions from multi-constituent satellite data assimilation, Atmospheric Chemistry and Physics, 17, 807-837, 2017.

Tang, X., Zhu, J., Wang, Z. F., Wang, M., Gbaguidi, A., Li, J., Shao, M., Tang, G. Q., and Ji, D. S.: Inversion of CO emissions over Beijing and its surrounding areas with ensemble Kalman filter, Atmospheric Environment, 81, 676-686, 2013.

5. The authors appear to claim that the IA step removes some biases and prevents biased emission estimates. My concern here though is that I see no way for the 3D-Var to distinguish model-observation biases arising due to emissions or biases due to some other source. If this is somehow possible then the authors should explain how. In either case I think the authors should be clearer, either about the limitations of partitioning different types of bias, or on how this part of the algorithm functions. Certainly the similarity in spatial patterns of the 3D-var IA increments and a priori vs a posteriori EI hints that both systems are addressing the same root cause of the biases but with different assumptions about the cause.

Response: Thank you for this comment. Yes, the 3D-Var cannot distinguish model-observation biases arising due to emissions or biases due to some other source. And we also do not need to distinguish the source types of model-data mismatch errors in the IA step. The goal of 3D-Var in this step is only to get a better IC for the subsequent EI step. Besides model errors, the model-data mismatch errors in the EI step mainly come from the IC and the emissions. As stated in General Comment 4, in the EI subsystem, we attribute all the model-data mismatch error to the bias in emission, thus we need to

minimize the IC errors to get better emissions estimates. The following descriptions are added in the revised manuscript. See lines 206-207, page 8.

"In the IA subsystem, we do not need to distinguish the type of sources of the model-observation mismatch error."

6. As a further comment on model error, bias, and emission estimation. I think the authors should be clearer that despite the IA step, from which the model is assumed perfect, inherent errors arising from discretization and model parameterizations will still exist meaning that the model will have other nonresolvable biases. This will naturally mean that other types of model bias will feed into the emission inversion step leading to biased emission estimates. Perhaps the authors could discuss the work on DA methods that consider model error, e.g., weak constraint 4D-var, that could be used to resolve such problems. This would help the discussion on the overestimate of emissions lines 740 onwards.

Response: Many thanks for this suggestion. We fully agree that the model-data mismatch error not only comes from emissions, but also from the inherent model errors arising from model structure, discretization and parameterizations, such as biases in meteorological fields, feedbacks of meteorology and chemistry, and chemical reactions and gas-particle transformations, etc. Neglecting model errors in the EnKF would attribute all uncertainties to emissions. For these random, systematic, or time-correlated model errors, they can be partly resolved using parameter estimation, bias correction and weak constraint 4D-var methods. We have added following related discussions in the revised manuscript. See lines 1096-1129, pages 58-59.

"The model-data mismatch error not only comes from the emissions, but also from the inherent model errors arising from model structure, discretization, parameterizations and the biases in the simulated meteorological fields. Neglecting model errors would attribute all uncertainties to emissions, and lead to considerable biases in the estimated emissions. In the version of CMAQ model used in this study, there is no heterogeneous

reactions (Quan et al., 2015; Wang et al., 2017), the parameterization scheme for the formation of secondary organic aerosol (SOA) is imperfect (Carlton et al., 2008; Jiang et al., 2012; Yang et al., 2019), no feedback between chemistry and meteorology is considered, and we used an idea profile for chemical lateral boundary conditions. All of the above problems can lead to underestimated concentrations of pollutants, which in turn require more emissions to compensate, leading to overestimations in emissions. In addition, previous studies have shown that the emission of ammonia in the MEIC inventory was underestimated (Kong et al., 2019; Paulot et al., 2014; Zhang et al., 2018). Due to lack of ammonia observations, our system does not include emission estimates of ammonia, which means that the concentration of ammonium aerosol was underestimated in this system, also resulting in an overestimation in the $PPM_{2.5}$ emission. Wind-blown dust was also not simulated here, thus the PMC emission inverted in this system do not only come from anthropogenic activities, but also from natural sources. Although some of these shortcomings could be solved in the future by updating the CTM model, there will still be errors in each parameterization and each process. Generally, parameter estimation method was used to reduce the model errors, in which, some uncertain parameters were included in the augmented state vector and were optimized synchronously based on the available observations (Brandhorst et al., 2017; Evensen, 2009). However, it is still quite difficult to identify the key uncertain parameters of different species in different models, which generally comes not only from the complex atmospheric chemical model, but also from hundreds of model inputs (Tang et al., 2013). Another method is bias correction, which treats the model error as a bias term, and includes it in the augmented state vector (Brandhorst et al., 2017; De Lannoy et al., 2007; Keppenne et al., 2005). In addition, the weak-constraint 4D-Var method can also be used to reduce the model errors, which adds a correction term in the model integration to account for the different sources of model error (Sasaki, 1970). Although reliable diagnosis of model error is still a challenge at present (Laloyaux et al., 2020), it should be considered in an assimilation system. We will consider model errors in our system in the future to obtain better emission estimates."

7. In addition to problems with the DA of long-lived gases in the IA step, I do question the DA of shortlived gases, i.e., NOx. I am aware of unpublished negative results showing that NOx concentration DA has limited benefit for the model forecast skill and in the worst case negatively impacts results by perturbing ozone chemistry in unrealistic ways. Hints of such problems can be found in ECMWF reports (Flemming et al., 2009; Inness et al., 2009, 2015) that clearly show the limited efficacy of NOx concentration assimilation due to the decay of DA increments. The authors should discuss these limitations in a meaningful way.

Response: Many thanks for this suggestion. We have added a discussion section in the revised manuscript, and added following discusses about the limited benefits or even negative effect of NO2 concentration initial field optimization on ozone chemistry and emission inversions. See page * and page *.

"Similarly, the model may not be able to resolve local-scale NO_2 well because of uniform distribution of concentration over the whole grid. Therefore, the model is shifted towards a NO_x (VOC)-limited regime in high (low) pollution regions, which negatively impacts results by perturbing ozone chemistry in unrealistic ways (Inness et al., 2015)."

8. Important details of the work could have been explained in a clearer and more concise way (specific comments to follow), e.g., why do the authors not include the length of DA window and analysis run the figure describing the architecture of their system? These details appear almost at the end of the descriptions, but such information, that can easily be compacted into concise form, could appear earlier either in figure 1 or in a suitable table. See specific comments below. There are also some contradictions in the explanations of the different experiments, which make it awkward to evaluate what has been done.

Response: Many thanks for this suggestion. We have added descriptions about the length of DA window and the architecture of our system (see Specific Comment 7) in

the revised manuscript. We also shorten some text to describe our system more clearly and concisely (e.g., Specific Comment 4). Additionally, the descriptions of EMS1 in Table 3 and lines 582-584 of the original manuscript were rephrased to avoid any misunderstanding. (see Specific Comment 10)

9. The DA experiments are only run for a single month. Given that this system will presumably be used in an operational capacity throughout the year it would have been nice to see statistics on the results of other experiments at different times of the year. I think this point is especially relevant for the pollutants with a strong seasonality in their emissions, e.g., PM2.5. Why was December 2016 chosen as the period of study? I could not see any discussion on this. Please can the authors add something to explain this.

Response: Many thanks for these suggestions. December is one of the most severely polluted and most concerned months in China every year. Therefore, we chose December as the period of study to reflect the capacity in optimizing anthropogenic emissions. According to this suggestion, we also performed a new inversion for July to further verify the system's capacity. Different from the inversion in winter, the localization scale of NO_2 was set to 80 km, because the lifetime of NO_2 is much shorter in summer (about 6 hours) due to the stronger atmospheric oxidation (Beirle et al., 2003). Figure R2 shows the prior and posterior emissions in July and December. It could be found that the emissions in December are always greater than those in July. Table S2 lists the statistical results of the evaluations in July averaged over the whole mainland of China. After assimilation, the evaluation with posterior emissions is significantly improved. The RMSEs decrease by 39.7, 63.5, 69.1, 34.1 and 27.0%, and the CORRs increase by 160.0, 491.7, 72.3, 49.0 and 60.5% for CO, SO_2, NO_2, $PM_{2.5}$ and PMC, respectively. Overall, the performance of the system in July is comparable to that in December. The following discussions are added in the revised manuscript.

See lines 549-551, page 24.

"December is one of the months with most severe air pollution, while July is one of the

least polluted months in China. Therefore, this study mainly tested the performance of the RAPAS system in these two months. For December, the IA …"

See lines 556-559, page 24.

"For July, the system also operated in the same way as for December. It needs to be noted that due to the stronger atmospheric oxidation, the lifetime of NO2 in July is significantly shorter than that in December, thus we adopted a smaller localization scale for NO2 (80 km)."

See lines 868-873, page 45.

"The evaluation results for July show that the emission uncertainty can still be significantly reduced, and the performance of the system in July is comparable to that in December (Table S2). Additionally, the seasonal variation of emissions can be well reflected (Figures S4 and S5), which means that our system can perform well at different times of the year."
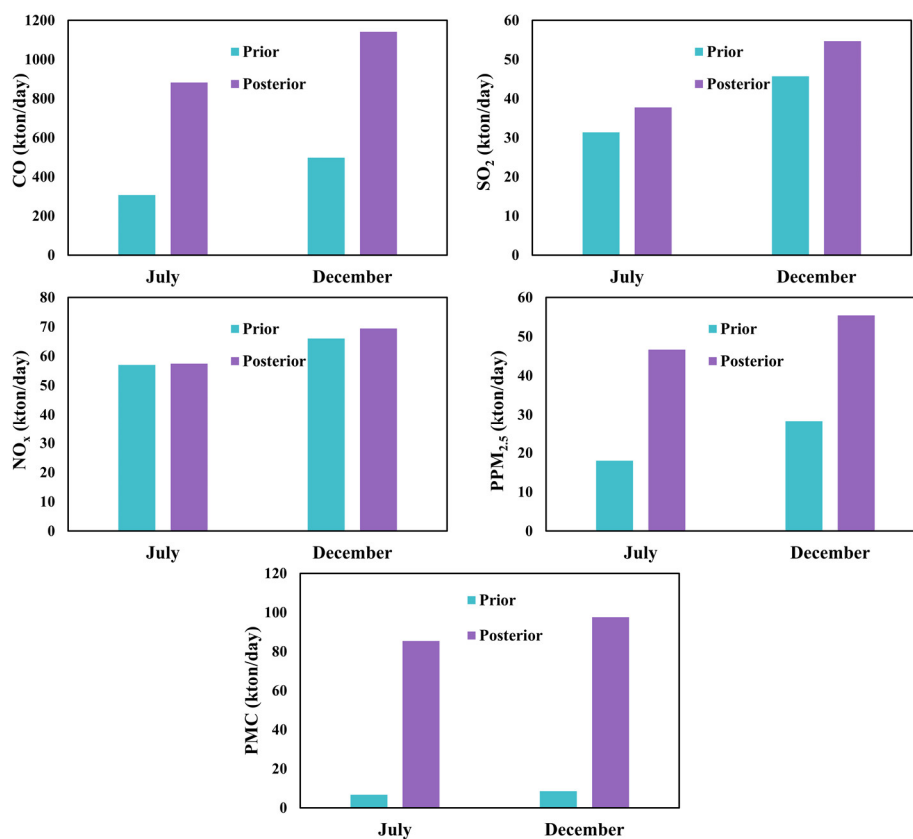
Figure R2. The prior and posterior emissions (kton/day) in July and December 2016 over the whole mainland China. (Figure S5 in the revised manuscript)
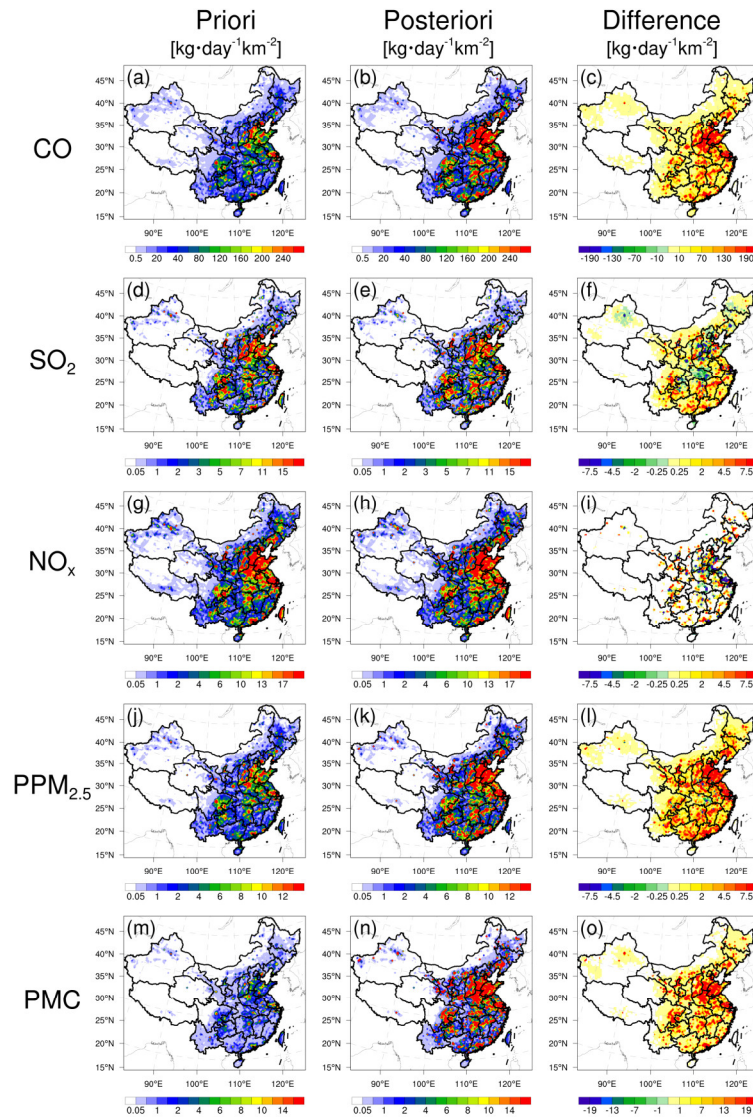

Figure R3. Spatial distribution of the time-averaged prior emissions (left column, MEIC 2016), posterior emissions (middle column), and differences (right column, posterior minus prior) in July 2016. (Figure S4 in the revised manuscript)

Table R1. Statistics comparing the pollution concentrations from the simulations with prior (CEP) and posterior (VEP) emissions against observations for July. CO unit: mg m^{-3}; others units: μg m^{-3}. (Table S2 in the revised manuscript)

Species	Mean Obs.	Mean Sim.		BIAS		RMSE		CORR	
		CEP8	VEP8	CEP8	VEP8	CEP8	VEP8	CEP8	VEP8
CO	0.79	0.33	0.63	-0.46	-0.16	0.58	0.35	0.25	0.65
SO$_2$	12.9	15.6	9.6	2.7	-3.3	19.7	7.2	0.12	0.71
NO$_2$	20.0	23.0	16.5	3.0	-3.5	22.0	6.8	0.47	0.81
PM$_{2.5}$	29.2	21.9	23.3	-7.3	-6.0	21.1	13.9	0.51	0.76
PMC	53.6	27.6	42.3	-26.0	-11.3	42.3	30.9	0.38	0.61

* BIAS, mean bias; RMSE, root mean square error; CORR, correlation coefficient

Beirle, S., Platt, U., Wenig, M., and Wagner, T.: Weekly cycle of NO2 by GOME measurements: a signature of anthropogenic sources, Atmospheric Chemistry and Physics, 3, 2225-2232, 2003.

10. Another point regarding the season chosen and the performance of the EI step. Since the EI system does not include emission estimates of ammonia, and ammonia can be very important to spring time loadings of PM2.5, could its absence from the system lead to compensation effects on the primary PM2.5 emission estimates? I think the authors should discuss this as a potential issue for the wider application of RAPAS throughout the year. This might help to identify the necessity for future development paths.

Response: Many thanks for this suggestion. Due to lack of ammonia observations, the current version of RAPAS does not include emission estimates of ammonia, which actually leads to compensation effects on the estimates of primary PM$_{2.5}$ emission. Kong et al. (2019) conducted emission inversion of NH3 with the same prior inventory as this study (MEIC), and found NH3 emissions increased heterogeneously almost over the whole mainland China, with an average increase of 27%. Therefore, this deficiency

may lead to a certain overestimation of posterior PPM2.5 emissions. We have added the following discussions in the revised paper, see lines 1107-1111, page 58.

"In addition, previous studies have shown that the emission of ammonia in the MEIC inventory was underestimated (Kong et al., 2019; Paulot et al., 2014; Zhang et al., 2018). Due to lack of ammonia observations, our system does not include emission estimates of ammonia, which means that the concentration of ammonium aerosol was underestimated in this system, also resulting in an overestimation in the PPM$_{2.5}$ emission."

Kong, L., Tang, X., Zhu, J., Wang, Z., Pan, Y., Wu, H., Wu, L., Wu, Q., He, Y., Tian, S., Xie, Y., Liu, Z., Sui, W., Han, L., and Carmichael, G.: Improved Inversion of Monthly Ammonia Emissions in China Based on the Chinese Ammonia Monitoring Network and Ensemble Kalman Filter, Environmental Science & Technology, 53, 12529-12538, 2019.

Specific Comments:

1. Can the authors briefly mention the ICNO experiment results in the abstract.

Response: Thanks for this suggestion. We have added following descriptions in the revised manuscripts. See lines 62-64, page 3.

"Results showed that the simulated concentrations of CO, NO2, SO2, PM2.5 and PM10 with the prior inventory have large systematic biases, with relative biases in the range of -48.2-54.2%. In the IA subsystem, after 3DVAR, the root mean squared error (RMSE) …"

2. Sentence running from line 138-line 141. Please can the authors rephrase this? I had difficulty making sense of this sentence.

Response: Thank you very much for your suggestion. We are so sorry for that the description is not clear enough, which increases reading difficulties for readers. The sentence (See lines 156-162, page 6) has been rephrased as follows:

"In addition, in this method, the constraints of the chemical ICs with observations in each assimilation window make the emission inversions are independent between assimilation windows, means if the emission in one window is overestimated or underestimated, it cannot be transferred to the next window for further correcting and be compensated in the following windows. This may result in a systematic bias in the inverted emissions (Jiang et al., 2021)."

3. Please reformulate sentence running from line 147 to 150.

Response: Thanks! We have revised that sentence as follows, which also could be found in lines 168-172, page 6 in the revised manuscript.

"Considering the possible shortcomings of the simultaneous optimization method (named as "one-step" method in this study) as metioned by Jiang et al. (2021), we adopted a "two-step" method (Sect. 3) in this system. Unlike the "one-step" method, the ICs of each DA window in the "two-step" method is simulated using the posterior emissions of the pervious DA window."

4. Line 172 onwards in section 2.1.1. Please can the authors find a way to shorten the summary text to avoid repetitions. There is already a lot of detail here some of which is repeated in the following sections.

Response: Many thanks for this suggestion. These sentences have been further summarized in the revised manuscript, including the composition and operation of IA subsystem, one-step introduction, etc. Additionally, we supplemented more technical details in the first step of EI subsystem according to the suggestion of another reviewer. See lines 201-228, pages 7-8.

"Based on above three components, the RAPAS is divided into two subsystems, namely the IC assimilation (IA) subsystem (CTM plus 3DVAR) and the emission inversion (EI) subsystem (CTM plus EnSRF). As shown in Figure 1, the IA subsystem is first run to optimize chemical ICs (Kleist et al., 2009; Wu et al., 2002) for the subsequent EI subsystem. In the IA subsystem, we do not need to distinguish the type of sources of the model-observation mismatch error. The EI subsystem runs cyclically with a "two-step" scheme. In the first step, the prior emissions (X^b) are perturbed and put into the CTM model to simulate chemical concentration ensembles. The simulated concentrations of the lowest model level are then interpolated to the observation space according to the locations and times of the observations using the nearest neighbor interpolation method. The prior emissions (X^b), simulated observations and real observations are entered into the EnSRF module to generate the optimized emissions (X^a). In the second step, the optimized emissions are entered into the CTM model again to generate the initial fields of the next DA window. Meanwhile, the optimized emissions are transferred to the next window as the prior emissions. Different from the "one-step" scheme, this "two-step" scheme needs to run the CTM model twice, which is time consuming, but it could transfer the potential errors of the inverted emissions in one DA window to the next for further correction. The benefit of this scheme will be further presented in Sect. 4.3."

5. I found no mention of the temporal variability of the emissions. The EI seems to be run for single days resolving an emission estimate for each day. Are the daily a posteriori emission estimates somehow applied to the hourly emission variabilities each day?

Response: Thanks for this comment. Yes, the EI resolves an emission estimate for each day. For both the prior and posterior emissions, we adopted a same scheme to convert the emissions from daily to hourly in each day and each grid. It is difficult to obtain the diurnal variations of each emission source. Therefore, in our scheme, for all the species emitted from area sources, we converted them to hourly emissions using a same diurnal

profile (Figure R4), and for the point sources, we assumed that there was no diurnal change. The reason for estimating daily emissions in the EI subsystem is that, first, due to the complex of hourly emissions, it is very difficult to simulate hourly concentrations that can match the observations well; second, since the observations are concentrated in cities, to retrieve the emission in each grid, the emitted pollutants need to be fully mixed and transmitted in the atmosphere.
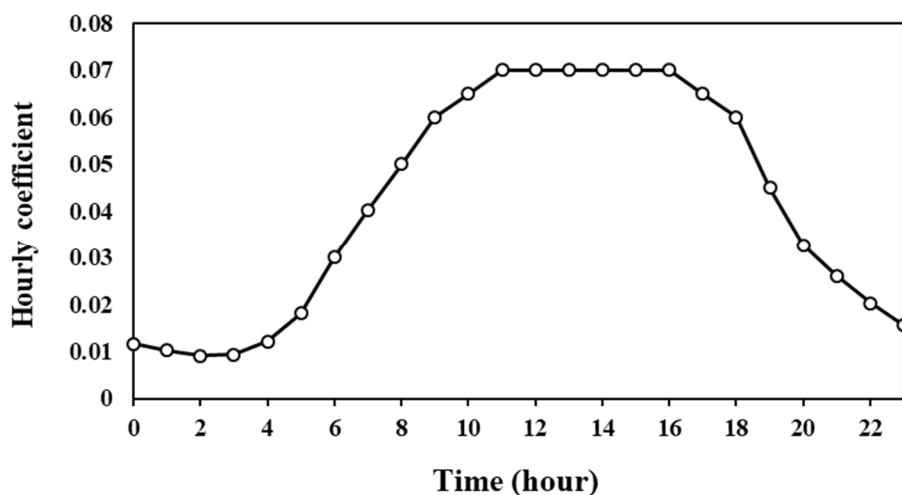

Figure R4. Hourly emission variabilities. (Figure S2 in the revised manuscript)

The following introductions have been added in the revised manuscript, see lines 476-479, page 21.

"During the simulations, the daily emissions were further converted to hourly emissions. For all the species emitted from area sources, we converted them to hourly using a same diurnal profile (Figure S2), and for the point source, we assumed that there was no diurnal change."

6. If the daily emission estimates from the EI are estimated for a single day and then applied to the following day as the updated a priori, how does this system behave due to day-of-the-week effects? Emissions from traffic and other sources are known to change at the weekend compared to the day, so how does this affect the day-to-day

performance of the DA system when transitioning Friday-Saturday or Sunday-Monday? Further to that, how does this affect the hourly emission variability following point 5 above?

Response: Thanks for this comment. Yes, the daily emissions are estimated and then applied to the following day as the updated a priori. NOx is mainly emitted by transportation (Li et al., 2017), which can better reflect the level of economic activities to a certain extent. Figure R5 shows the posterior NOx emission changes on different days of the week overlaid with concentration changes. The power emissions are not taken into account in the statistics because the production processes of resource-based power sector are little interrupted. During this study period, high emissions mainly occur from Thursday to Saturday, while low emissions occur in other days. It could be found that the posterior emission changes are in good agreement with the observations. The emission changes transitioning Friday-Saturday or Sunday-Monday are not obvious during this study. However, the emission changes transitioning Saturday-Sunday significantly decrease. Wang et al. (2014) showed that Monday and Sunday had lower levels of NO2 than the other days of the week in Beijing–Tianjin–Hebei, while Wang et al. (2015) showed that NO2 concentrations on weekends are higher than those on weekdays in Bijing. Deng and Jian (2020) showed that the "weekend effect" within the Yangtze River Economic Belt varies greatly. Additionally, Beirle et al. (2003) reported that there is no weekly pattern can be found in China. These results indicate that weekly emission variation may be related to local production and lifestyle in different regions and at different times.

As mentioned in response 5, we didn't optimize hourly emissions and applied a diurnal profile to convert emissions from daily to hourly on each day and at each grid.

We have added related discussions in the revised paper. Lines 1180-1194, page 61.

"NOx is mainly emitted by transportation (Li et al., 2017), which can better reflect the level of economic activities to a certain extent. Weekly emission changes were also explored to verify the performance of the system in depicting emission changes (Figure

S13). Although the "weekend effect" of emissions in China is not significant (Wang et al., 2014; Wang et al., 2015), the posterior NOx emission changes showed a good agreement with the observations. In our previous studies (Feng et al., 2020a; Feng et al., 2020b), the system was successfully applied to optimize NOx and CO emissions, respectively. The inverted emission changes were also in line with the time points of epidemic control. Additionally, the emission changes can well reflect the emission migration from developed regions or urban areas to developing regions or surrounding areas over recent years, which were consistent with the emission control strategies in China. Although the system does not consider the model error, resulting in a certain difference between the posterior emission and the actual emission, the spatiotemporal changes in posterior emissions are relatively reasonable, which can be used to monitor emission changes and make emission regulations."
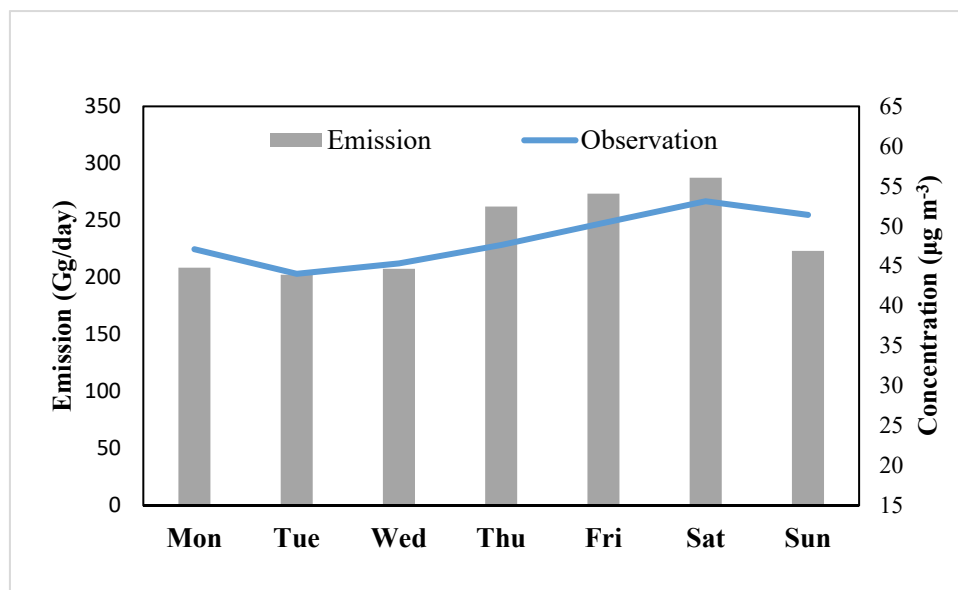

Figure R5 Weekly variation in posterior NOx emission and NO_2 observation.

Beirle, S., Platt, U., Wenig, M., and Wagner, T.: Weekly cycle of NO2 by GOME measurements: a signature of anthropogenic sources, Atmospheric Chemistry and Physics, 3, 2225-2232, 2003.

Deng, H. and Jian, X.: Study on weekend and holiday effects of urban haze pollution in the Yangtze River Economic Belt, Journal of Central China Normal University. Natural Sciences Edition, 54, 536-545, 2020

Li, M., Liu, H., Geng, G., Hong, C., Liu, F., Song, Y., Tong, D., Zheng, B., Cui, H., Man, H., Zhang, Q., and He, K.: Anthropogenic emission inventories in China: a review, National Science Review, 4, 834-866, 2017.

Wang, Y. H., Hu, B., Ji, D. S., Liu, Z. R., Tang, G. Q., Xin, J. Y., Zhang, H. X., Song, T., Wang, L. L., Gao, W. K., Wang, X. K., and Wang, Y. S.: Ozone weekend effects in the Beijing-Tianjin-Hebei metropolitan area, China, Atmospheric Chemistry and Physics, 14, 2419-2429, 2014.

Wang, Z., Li, Y., Dong, X., Sun, R., Sun, N., and Pan, L.: Analysis on weekend effect of air pollutants in urban atmosphere of Beijing, Journal of University of Chinese Academy of Sciences, 32, 843-850, 2015.

7. Figure 1. Please can the authors show the length of DA window in each step of this system.

Response: Thank you very much for your suggestion. We have added the length of DA window in Figure 1. Additionally, more details on the description of symbols, precise times and data update frequency are supplemented to the caption of Figure 1.

"**Figure 1**. The composition and flow chart of RAPAS. The x_a and x_b represent the prior and posterior emissions. The 3DVAR assimilation stage lasts 5 days with data input frequency of 6 hours, and the DA window in the EI subsystem is set to 1 day."

8. Line 210. Would "interpolated" be more accurate than "compressed"?

Response: Thanks for your suggestion. We have changed 'compressed' to 'interpolated'. See line 251, page 10.

9. I think the name of the ICNO experiment should be renamed to NODA or DANO to indicate more clearly that neither the IA or EI steps are performed. As it is, ICNO implies only not IA.

Response: Thanks for your suggestion. We have renamed 'ICDA' to 'NODA'.

10. The descriptions of EMS1 in Table 3 and on lines 498-501 seem to be in contradiction of one another, i.e., Table 3 says EMS1 is initialized with prior emissions from the previous window (no indication of 3DVar) and the text from line 498 claims that the ICs are from the 3D-Var experiment. The authors should make this clearer to avoid any confusion.

Response: Thanks! Yes, the description of EMS1 is not clear enough. In EMS1, the initial field of each DA window is first taken from a forward simulation with the prior emissions in the previous DA window, and then optimized using observations with the 3DVAR method. We have revised the sentence and reorganized Table 3 to make it clear.. See lines 580-584, page 25 in the revised manuscript.

"In EMS1, rather than forward simulated using the optimized emissions of the previous DA window in EMDA, the initial fields of each DA window were first taken from forward simulation with the prior emissions of the previous DA window, and then optimized using the 3DVAR algorithm and the observations at the corresponding moment as mentioned in Sect. 2.3."

11. It would have been good to see an EMDA experiment without any 3D-Var to see what affect the 3DVar has on the emission inversion step. I think this is an important test because of the potential overlap between emission error and initial condition error arising from emission errors.

Response: Thank you for your opinion. We conducted an EMDA experiment without 3D-Var, which is named as EMS8, to test the effect of the initial field optimization on the emission inversions. Figure R6 shows the time series of the relative differences in daily posterior emissions of the five species between the EMDA and EMS8 experiments. It can be found that the optimization of initial field has great impact on the emission inversion, especially for the long-lived species (e.g., CO). With the gradual disappearance of the benefit of initial field assimilation, the two experiments can reach

a unified state after some windows. The following changes has been added in the revised paper.

Table 3

Exp. Type	Exp. Name	Period	IC of the first DA window	IC of the first DA window	Emission
Sensitivity	EMS8	1-31 December	0000 UTC on December 1, taken from ICNO	The same as EMDA	The same as EMDA

Lines 593-595, page 26

"The last EMS8 experiment aims to evaluate the impact of IC optimization of the first window on emission estimates, in which the ICs were taken from a 5-day spin-up simulation."

Lines 1015-1035, page 55

4.3.4 Impact of the IC optimization of the first window

Many studies have shown that there would be large emission discrepancies resulting from the IC errors (Jiang et al., 2013a; Miyazaki et al., 2017; Tang et al., 2013), which means that if the IC is not optimized, the errors of concentrations would be compensated through the adjustment of emissions. To evaluate the impact of the IC optimization of the first window on the emission inversions, the EMS8 experiment without the IA step was conducted. Figure 16 shows the time series of the relative differences in daily posterior emissions of the five species between the EMDA and EMS8 experiments. It can be found that the optimization of IC has great impact on the emission inversions of long-lived species (i.e., CO). The overall difference in the inverted CO emissions between the two experiments is about 5.3%, and in the first few windows, the maximum difference can reach 26.1%. For the short-lived species, the IC optimization has little

impact on the emission, for example, the averaged emission differences of SO_2, NO_x and PMC in the two experiments are 0.3%, 0.3% and 0.9%, respectively. For PPM2.5, it is affected not only by the primary emission, but also by the complex chemistry of its precursors. Therefore, the difference between the two experiments fluctuates at a certain extent, with overall difference of 2%. It is worth noting that with the gradual disappearance of the benefit of IC assimilation, the two experiments can reach a unified state after some windows. For CO, the impact of IA on emission inversion lasts about half a month. These results indicate that removing the bias of IC of the first DA window is essential for subsequent inverse analysis (Jiang et al., 2017).
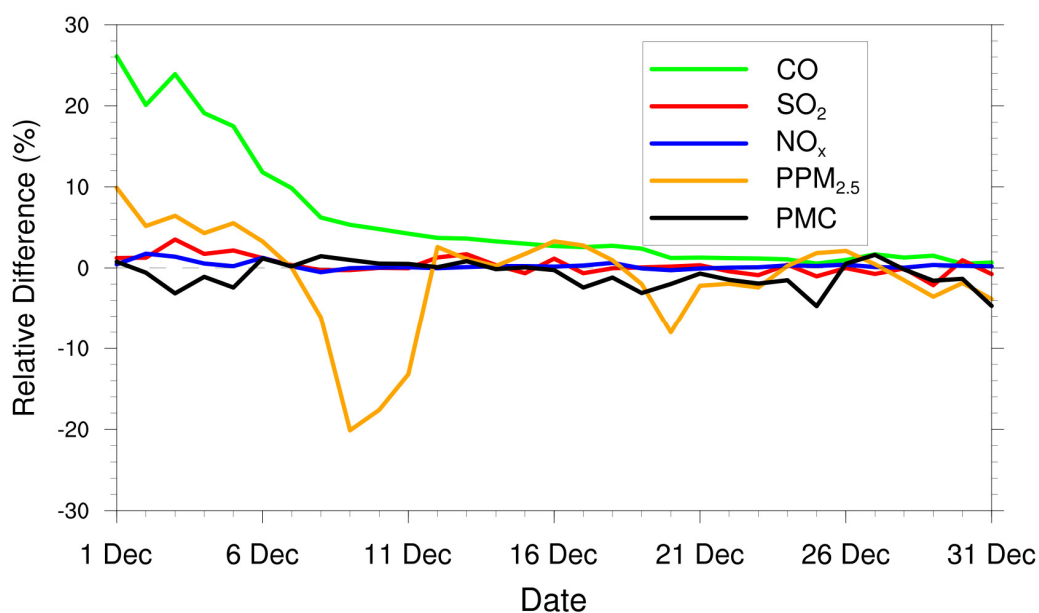

Figure R6. Relative differences in CO, SO2, NOx, PPM2.5 and PMC emissions (%, the ratio of absolute difference to EMDA) between the EMDA and EMS8. (Figure 16 in the revised manuscript)

12. I think Figure 10 should be expanded to show the changes in emissions beyond the borders of China. I am assuming of course that the EI system's state variables include emission terms beyond China's frontiers. If they do not, then I would ask the authors to

make this point clearer. But given the large emission region (north India) included in the western part of the modelling domain, it would make sense to show any emission changes in that region.

Response: Thanks for this comment. Yes, the EI system's state variables include emissions outside China, but in Figure 10 (Figure 12 in the revised manuscript), we only showed the emissions within China. Since the observation sites are all located in China, the changes of the emissions outside China are very small. The affected areas are related to the covariance localization scale of a chemical species, which are mainly located in the Indochina Peninsula, the Korean Peninsula and the northwest and northeast borders of China. Due to the sparse stations in Tibet province, there is basically no adjustment for the emissions in the north of India. The following description has been added on lines 834-836, page 43 in the revised paper.

 "It should be noted that the emissions outside China were masked, since the observation sites are all within China in this study, there is little change in the emissions outside China."
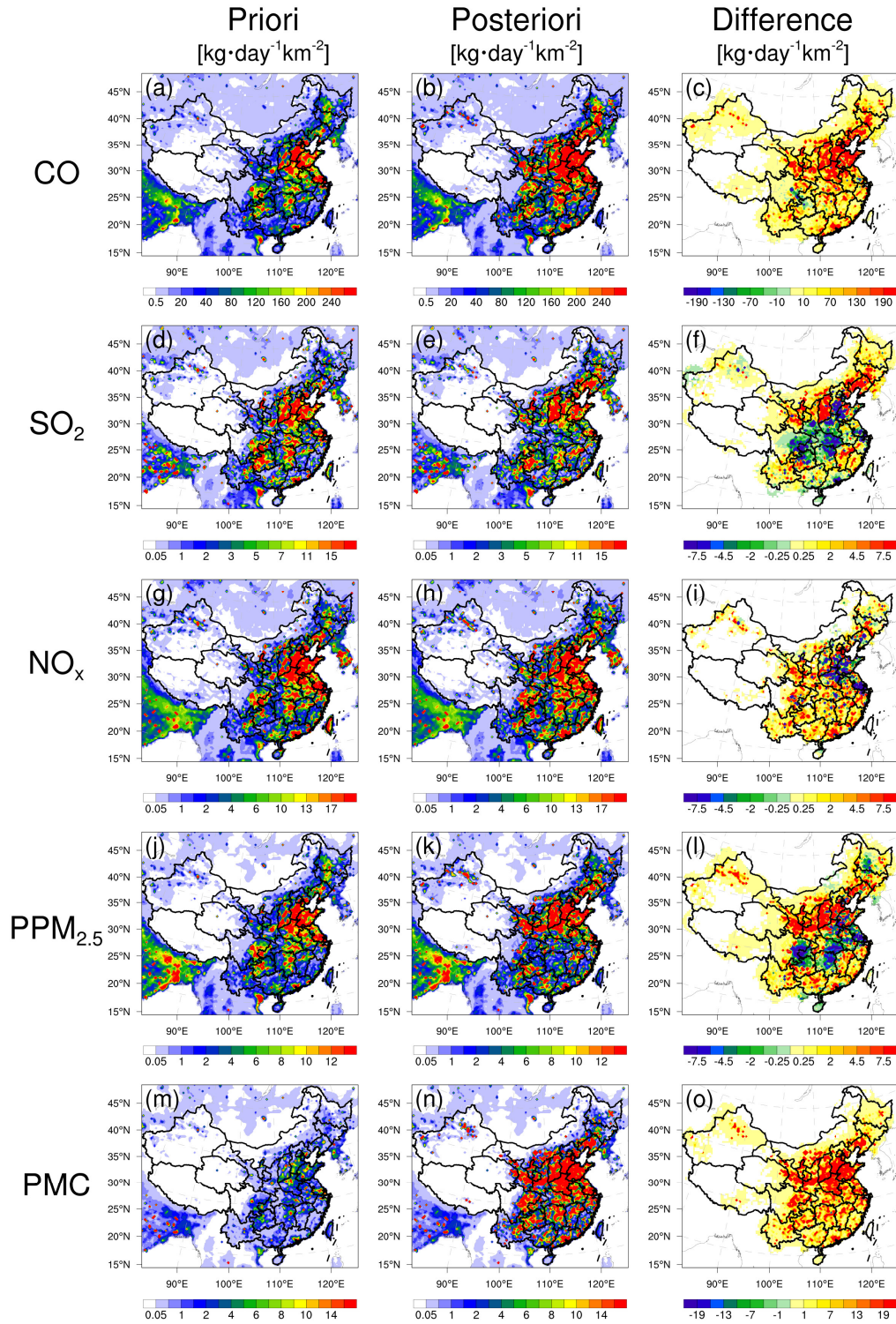
Figure R7 Spatial distribution of the time-averaged prior emissions (left column, MEIC 2016), posterior emissions (middle column), and differences (right column, posterior minus prior).

13. I assume that the authors have performed some offline testing on the number of ensemble members in the EnKF. It would be interesting to hear about these tests and what they showed with regard to selecting 40 over any other number of ensemble members.

Response: Thanks for this comment. Generally, in EnKF, the larger the number of ensembles, the smaller the impact of spurious correlations, but the more computationally expensive. The choice of ensemble size involves a tradeoff between the computation costs and the accuracy of the inversions. We have tested the impact of the number of ensemble members (e.g., 20, 30, 40, 50, and 70) on the inversion results. Figure R8 shows the emissions of the 5 species in China inferred using different ensemble numbers. It could be found that when the number of ensembles is less than 40, as the number of ensembles we use decreases, the inverted emissions will increase significantly, and when the ensemble number is greater than 40, as the number of ensembles increases, the change in the inverted emissions is very little. Therefore, we set the ensemble number to 40. The following description has been added on lines 377-378, page 16 in the revised paper.

"… N is the ensemble size, which was set to 40 in consideration of a tradeoff between the computation cost and inversion accuracy (Figure S1), and δX_i^b represents …"
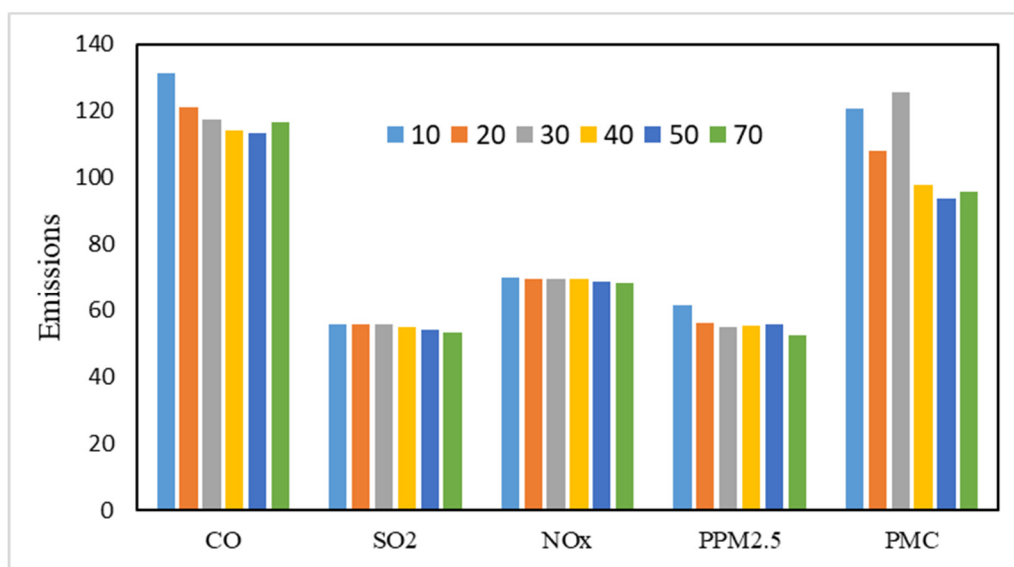
Figure R8 Comparison of emissions of different species inferred by selecting different ensemble numbers. CO unit:10 Gg; others units: Gg. (Figure S1 in the revised manuscript)

14. I think the application of the chi squared metric to the discussion in sections 4.3.3 and 4.3.4 would be very informative of how the optimality of the EnKF is changing in each case.

Response: We really appreciate this suggestion. We have added the Chi-square statistics in sections 4.3.2 and 4.3.3 and in Figure 14 and S10 in the revised manuscript. It could be found that a larger uncertainty usually leads to a faster convergence. After a few windows, there are no significant differences for the Chi-square statistics among the different experiments. However, the day-to-day changes in emissions can also cause slight fluctuations. Additionally, the total spread (including the ensemble variance and the observation error) is relatively small compared to the ensemble mean RMSE of surface SO2 in the NCP, but the opposite in the YRD, indicating that it may be necessary to apply different a priori uncertainties according to different regions in the future (Chen et al., 2019). Figure S10 shows the time series of the relative changes in a posteriori emission estimates and χ2 statistics in EMS7 experiment. Due to applying larger observation error, the day-to-day changes in emissions is small compared with that in EMDA, and may not reflect the actual situation. Although χ2 statistics also shows that the emission converges to a certain range after some windows, it is slightly larger than the statistics in EMDA, indicating that there may be a large discrepancies $(y_t - H_t\overline{X_t^b})$ between simulations with prior emissions and observations, which is more verified that the emissions cannot respond in a timely manner with too large observation errors. The changes in the revised manuscript are marked in red as follows:

See line 959, page 50.

"Figure 14 shows the time series of SO_2 emission changes, the Chi-square statistics and

the RMSEs …"

"… …and as shown in Figure 14a-d, larger uncertainty will lead to a faster convergence… …"

"… but in the later stage of the experiment, there are no significant differences for the RMSE and Chi-square statistics among the different experiments. However, the day-to-day changes in emissions can also cause slight fluctuations."

"Due to the spatial-temporal inhomogeneity of emissions, the differences of Chi-square statistics between the YRD and NCP show that it may be necessary to apply different a priori uncertainties according to different regions (Chen et al., 2019)."

"These results usually correspond to sluggish emission changes and large Chi-square statistics (Figure S10), suggesting that too large observation error may substantially impact the estimated emissions."
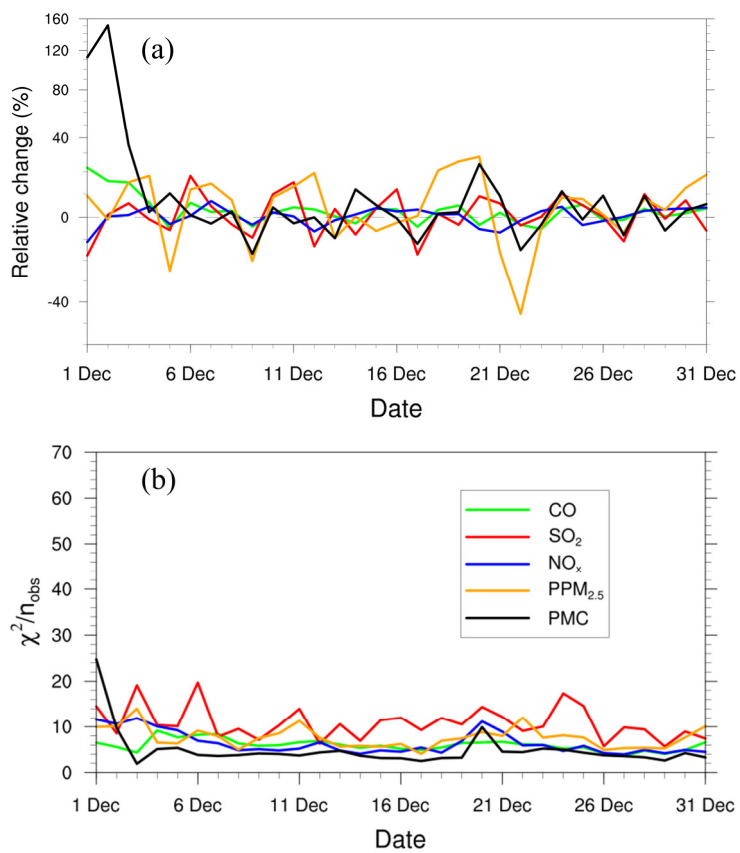
Figure R10. Relative changes (a) in a posteriori emission estimates of CO, SO_2, NO_x, $PPM_{2.5}$ and PMC, and χ^2 statistics (b) of these state vectors in EMS7 experiment. (Figure S10 in the revised manuscript)
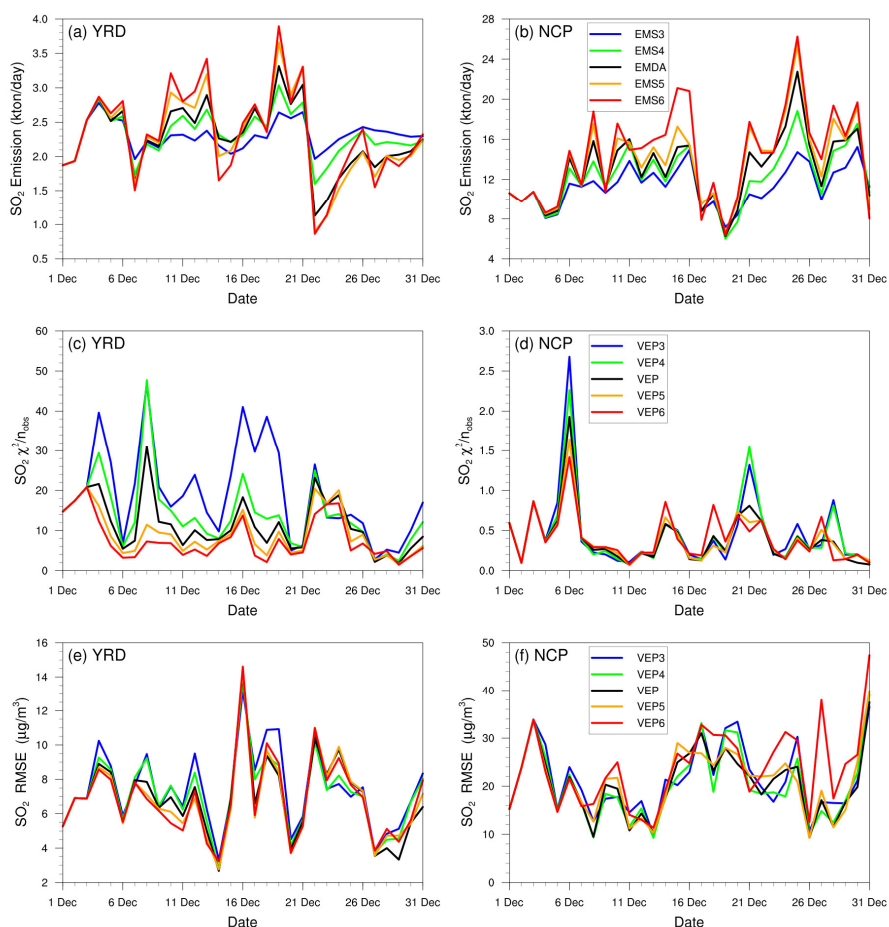
Figure R9. Time-series of SO_2 emission changes, the Chi-square statistics and the RMSE of simulated SO_2 with updated SO_2 emissions in the EMDA and EMS3-6 experiments over the Yangtze River Delta (YRD) and North China Plain (NCP). (Figure 14 in the revised manuscript)

Chen, D., Liu, Z., Ban, J., and Chen, M.: The 2015 and 2016 wintertime air pollution in China: SO2 emission changes derived from a WRF-Chem/EnKF coupled data assimilation system, Atmospheric Chemistry and Physics, 19, 8619-8650, 2019.

15. Line 854. When the authors speak about reduced emission uncertainties, I think some care is needed. In fact, the errors in the simulated concentrations are reduced, and

from that the a posteriori emissions are assumed to have lower uncertainty. However, this ignores the fact that there are unquantifiable model uncertainties included within the new emission estimates, and so I think it is dangerous to say the emission uncertainty itself is reduced without the support of an independent estimate.

Response: Thanks for this suggestion. Yes, we agree that the model error is an important uncertainty source for the estimates of emissions and is usually difficult to quantify. Generally, the reduction of uncertainties is used to test the effectiveness of an assimilation system to assimilate observational data (Chevallier et al., 2007; Takagi et al., 2011), it does not represent a real reduction in emissions uncertainty. We have revised that sentence as follows:

 "The results show that RAPAS has a good performance in assimilating ground in-situ observations, with the calculated emission uncertainties reduced by 44.4%, 45.0%, 34.3%, 51.8% and 56.1% for CO, SO_2, NO_x, $PPM_{2.5}$ and PMC, respectively. It can also significantly improve the simulations, the RMSEs of the simulated concentrations with posterior emissions decreased by 40.1-56.3%, and the CORRs increased from 0.26-0.66 to 0.69-0.87 for different species."

Chevallier, F., Breon, F.-M., and Rayner, P. J.: Contribution of the Orbiting Carbon Observatory to the estimation of CO2 sources and sinks: Theoretical study in a variational data assimilation framework, J. Geophys. Res.-Atmos., 112, d09307, https://doi.org/10.1029/2006JD007375, 2007.

Takagi, H., Saeki, T., Oda, T., Saito, M., Valsala, V., Belikov, D., Saito, R., Yoshida, Y., Morino, I., Uchino, O., Andres, R. J., Yokota, T., and Maksyutov, S.: On the Benefit of GOSAT Observations to the Estimation of Regional CO2 Fluxes, SOLA, 7, 161-164, https://doi.org/10.2151/sola.2011-041, 2011.

16. Line 872 onwards. This is probably also a symptom of the fact that the EnKF was not specifically developed to estimate state variables with significant bias errors. The incremental improvement described here are exactly symptomatic of this issue, and

examples like this are described in the literature on the topic cited above. This is an example of the text that will need to be revised in light of the required shift to speak of parameter estimation to solve for biases being treated by the EnKF.

Response: Thanks for this comment. Although the EnKF method has been widely used in the estimates of emissions and surface carbon fluxes (Brunner et al., 2012; Miyazaki et al., 2017; Peter et al., 2005, 2007; Wang et al., 2020; Zhang et al., 2015), we agree that there are problems with estimating emissions using the EnKF. The estimates of emissions are significantly related to the settings of the uncertainties of prior emissions, the model-data mismatch errors, the ensemble numbers, the localization scheme, and the choice of prior emissions. Generally, the uncertainty of prior emissions and the error of observations determine how closely the analysis is weighted toward the background and observation. Thus, it is very important to set reasonable values for above items. Therefore, we did many tests to choose the appropriate settings to ensure the reliability of the inversion results (section 4.3). The following words have been added in the revised manuscript. See lines 976-978, page 51.

"Therefore, when using an EnKF system for emission estimation, we have to be very careful about the setting of these errors."

Brunner, D., Henne, S., Keller, C. A., Reimann, S., Vollmer, M. K., O'Doherty, S., and Maione, M.: An extended Kalman-filter for regional scale inverse emission estimation, Atmos. Chem. Phys., 12, 3455-3478, 2012.

Miyazaki, K., Eskes, H., Sudo, K., Boersma, K. F., Bowman, K., and Kanaya, Y.: Decadal changes in global surface NOx emissions from multi-constituent satellite data assimilation, Atmospheric Chemistry and Physics, 17, 807-837, 2017.

Peters, W., Miller, J. B., Whitaker, J., Denning, A. S., Hirsch, A., Krol, M. C., Zupanski, D., Bruhwiler, L., and Tans, P. P.: An ensemble data assimilation system to estimate CO2 surface fluxes from atmospheric trace gas observations, J. Geophys. Res., 110, D24304.

Peters, W., Jacobson, A. R., Sweeney, C., Andrews, A. E., Conway, T. J., Masarie, K., Miller, J. B., Bruh-wiler, L. M. P., P´etron, G., Hirsch, A. I.,Worthy, D. E. J.,Werf, G. R. V. D., Randerson, J. T., Wennberg, P. O., Krol, M. C., and Tans, P. P.: An atmospheric perspective on North American carbon dioxide exchange:

CarbonTracker, P. Natl. Acad. Sci., 104, 18925–18930.

Wang, J., Feng, L., Palmer, P. I., Liu, Y., Fang, S., Bösch, H., O'Dell, C. W., Tang, X., Yang, D., Liu, L., and Xia, C.: Large Chinese land carbon sink estimated from atmospheric carbon dioxide data, Nature, 586, 720-723, 2020.

Zhang, S., Zheng, X., Chen, J. M., Chen, Z., Dan, B., Yi, X., Wang, L., and Wu, G.: A global carbon assimilation system using a modified ensemble Kalman filter, Geosci. Model Dev., 8, 805-816, 2015.

17. Line 878 to 883. There is an example of ozone observations being successfully used to estimate NOx and VOC emissions within the 4D-var framework (Hamer et al., 2015) that specifically deals with the problem of NOx-limited vs VOC limited conditions. I would suggest to the authors to consider including some text on the potential ways to address this problem. This might be one limitation of the EnKF method compared to the 4D-var?

Response: Thanks for this comment. Actually, there is also a similar limitation of the 4D-var method. Figure R11 shows one situation of limitation in adjusting NOx emissions using O3 observations with EnKF, which was adopted from Tang et al. (2016). Obviously, the relationship between O3 concentrations and NOx emissions is unreasonable due to highly nonlinearly chemical mechanism, because the system adjusts the emissions (blue dot) in the opposite direction to fit the observations. This relationship contained in the covariance matrix of EnKF is similar to the Jacobian matrix in 4D-var describing the forward-model response to perturbations to the emission parameters (Eq. 10 and Eq. 15 in Hamer et al. (2015)). Theoretically, one O3 observation may correspond to two NOx/VOC emission states. It is difficult to solve for VOC/NOx emissions using O3 observation only even with 4DVAR method unless other constraints are introduced, because O3 observations on their own would have problems identifying which NOx/VOC state is the correct one. We also confirmed this with Dr. Hamer. Although Hamer et al. (2015) conducted inversion analyses of NOx and VOC emissions under varying polluted photochemical regimes, their sensitivity analyses were performed within an idealised model. The photochemical regime (i.e., either NOx- or VOC-limited regimes) of each experiment was prescribed, and the

difference between prior and true emission scaling factors is small (their table 5). They also introduced constraints from observations of other species (e.g., NO2 in OCN experiment and HCHO in HCN experiment) on emissions. Therefore, in most cases, the posterior emission error was reduced (e.g., their table 8). Due to the complexity of pollution problem in China, the photochemical regime varies in different regions and at different times (Jin and Holloway, 2015; Li et al., 2021; Wang et al., 2019), thus it is difficult to adjust NO_x and VOC emissions using cross-species information within real observations in China. As suggested in Hamer et al. (2015), the analysis of the results was limited to identifying relationships between the observing scenario, the photochemical regime, the adjoint sensitivities and so on. The following words have been added in the revised manuscript. See lines 1133-1135, page 59.

"Although Hamer et al. (2015) successfully used O_3 observations to estimate NO_x and VOC emissions within the 4D-var framework within an idealised model, O_3 observations are not assimilated to improve NO_x and VOC emissions using cross-species information due to the strong nonlinear effects within the O_3-NO_x-VOC relationship (Wang et al., 2019), in which …"
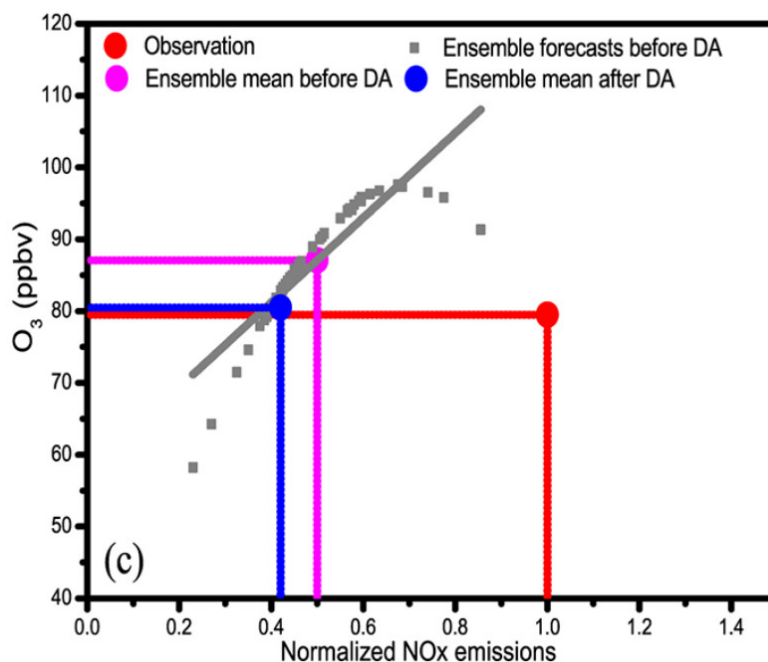

Figure R11. O3 concentrations (ppbv) and NOx emissions (no unit, normalized by the true NOx emission) before and after data assimilation and their ensemble samples. The

red dot represents the true state of the NOx emission and the observed O3 concentration (VOC-limited). The magenta dot represents the result of the ensemble mean of the grey squares (NOx-limited). The gray line represents a linear relationship calculated from the ensemble samples of O3 concentrations and NOx emissions. The analyzed O3 concentration and NOx emission are denoted by the blue dot. (Tang et al., 2016)

Hamer, P. D., Bowman, K. W., Henze, D. K., Attie, J. L., and Marecal, V.: The impact of observing characteristics on the ability to predict ozone under varying polluted photochemical regimes, Atmospheric Chemistry and Physics, 15, 10645-10667, 2015.

Jin, X. and Holloway, T.: Spatial and temporal variability of ozone sensitivity over China observed from the Ozone Monitoring Instrument, Journal of Geophysical Research-Atmospheres, 120, 7229-7246, 2015.

Li, R., Xu, M., Li, M., Chen, Z., Zhao, N., Gao, B., and Yao, Q.: Identifying the spatiotemporal variations in ozone formation regimes across China from 2005 to 2019 based on polynomial simulation and causality analysis, Atmospheric Chemistry and Physics, 21, 15631-15646, 2021.

Tang, X., Zhu, J., Wang, Z., Gbaguidi, A., Lin, C., Xin, J., Song, T., and Hu, B.: Limitations of ozone data assimilation with adjustment of NOx emissions: mixed effects on NO2 forecasts over Beijing and surrounding areas, Atmospheric Chemistry and Physics, 16, 6395-6405, 2016.

Wang, N., Lyu, X., Deng, X., Huang, X., Jiang, F., and Ding, A.: Aggravating O3 pollution due to NOx emission control in eastern China, Science of The Total Environment, 677, 732-744, 2019.

Technical changes:

Line 37. "It is capable of…"

Response: Thanks for this suggestion. We have changed "It is capable to …" to "It is capable of …". See line 37, page 2.

Line 38. "…assimilating spatially…"

Response: Thanks for this suggestion. We have changed "… assimilate spatially …" to "… assimilating spatially …". See line 38, page 2.

Line 44. "…subsystem in each data…"

Response: Thanks for this suggestion. We have changed "… subsystem in its each data …" to "… subsystem in each data …". See lines 44-45, page 2.

Line 103. "…and large amounts of …"

Response: Thanks for this suggestion. We have changed "… and large amount …" to "… and large amounts of …". See lines 104-105, page 4.

Line 126. "…because atmospheric…"

Response: Thanks for this suggestion. We have changed "… because atmopheric …" to "… because atmospheric …". See line 141, page 5.

Line 139. "…means that the assimilation windows are independent from each other, generally,…"

Response: Thanks for this suggestion. We have rephrased that sentence. See, lines 156-162, page 6.

Lines 176-177. "…and provides "perfect" chemical ICs…"

Response: Thanks for this suggestion. We have deleted that sentence and shorten the text of that paragraph. See line 207, page 8.

Line 199. "…can address the complex…"

Response: Thanks for this suggestion. We have changed "… could address the complex …" to "… can address the complex …". See line 237, page 9.

Line 204. "…and it covers the whole of mainland of China…"

Response: Thanks for this suggestion. We have changed "… and it covers the whole mainland of China …" to "… and it covers the whole of mainland of China …". See line 242, page 9.

Line 217. "…using optimized emission from the previous window…"

Response: Thanks for this suggestion. We have changed "… using optimized emission of previous window …" to "… using optimized emission from the previous window …". See line 258, page 10.

Line 242. "Additional work includes the…"

Response: Thanks for this suggestion. We have changed "Additional works include the …" to "Additional work includes the …". See line 284, page 13.

Lines 256-257. "Hourly mean surface pollution observations within a 1 hour window of the analysis…"

Response: Thanks for this suggestion. We have changed "Hourly surface pollution observations within 1 hour window of the analysis …" to "Hourly mean surface pollution observations within a 1 hour window of the analysis …". See line 298, page 14.

Line 259. "For gas concentrations that are directly…"

Response: Thanks for this suggestion. We have changed "For gas pollutions that are directly …" to "For gas concentrations that are directly …". See line 301, page 14.

Line 314-315. "The ground-level scale generally spreads 40-45 km…"Line 320. "In the EnKF…"

Response: Thanks for this suggestion. We have changed "The ground-level scale generally spread 40-45 km …" to "The ground-level scale generally spreads 40-45 km …". See line 358, page 17.

Line 401. "…with a lifetime more than 1 day."

Response: Thanks for this suggestion. We have changed "with lifetime more than 1 day …" to "with a lifetime more than 1 day …". See line 459, page 21.

Line 403. "In addition, NO2 is rather reactive…"

Response: Thanks for this suggestion. We have changed "In addition, NO2 is rather active …" to "In addition, NO2 is rather reactive …". See line 461, page 21.

Line 424. "During the inversion cycles, …"

Response:

Thanks for this suggestion. We have changed "During the cycling inversions …" to "During the inversion cycles …". See line 486, page 22.

Line 439. "…assimilation cycle…"

Response: Thanks for this suggestion. We have changed "… cycle assimilation …" to "… assimilation cycle …". See line 502, page 22.

Line 446. Missing last access date on link.

Response: Thanks for this suggestion. We have added last access date on link. See line 509, page 22.

Line 582. "…are over the rest of the areas…"

Response: Thanks for this suggestion. We have changed "… are over the rest areas …" to "… are over the rest of the areas …". See line 694, page 32.

Responses to the comments of Reviewer #2:

We are truly grateful to the positive comments and thoughtful suggestions. Those comments are all valuable and very helpful for revising and improving our paper, as well as the important guiding significance to our research. We have studied the comments/suggestions carefully and have made corresponding corrections. All changes made to the text are marked in red color. The point-by-point responses to the reviewers' comments are listed as follows:

General comments:

1. The study "A Regional multi-Air Pollutant Assimilation System (RAPAS v1.0) for emission estimates: system development and application by Shuzhuang Feng and colleagues describes a proposed method to estimate chemical and particulate matter emissions from observations by inversion. Emission rates of trace gases both of biogenic sources and anthropogenic chemical pollutants cannot be measured directly (apart from very special cases). Yet by evident reasons a precise knowledge is utterly important. While it is a long-discussed topic in atmospheric chemistry data assimilation and inversion, a solution of which is hampered by other factors, as initial and boundary data, deposition rates and many other parameters exert significant influence on the simulation result of chemistry transport models (CTM), which are the link between observations and emission estimates.

As their central objective the authors claim this study offer "a useful tool for accurately quantifying multi-species anthropogenic emissions at large scales and near-real time, which will serve better for monitoring emission changes and designing future emissions regulations and pollution control."

While the aspirations of the authors are remarkable, I could not find sound justifications for both a validated methodological approach and the practical proof of concept by the presented case study. I cannot recommend publication prior to a presentation of supporting evidence of the central claim.

Response: Thank you very much for this comment. We strongly agree that the conclusion that the system can reveal emissions changes well and provide effective assistance for pollution control can only be concluded after sufficient validation of the spatial and temporal changes in emissions inferred from the system. In this study, several methods were adopted to evaluate our system and the inversion results.

First, the uncertainty reduction rate (UR) is an important quantity to evaluate the performance of an assimilation system and the effectiveness of assimilating observations (Chevallier et al., 2007; Takagi et al., 2011). In this system, the overall UR of CO, SO2, NOx, PPM2.5 and PMC emissions were 44.4%, 45.0%, 34.3%, 51.8% and 56.1%, respectively, indicating that our system has good and comparable performance in optimizing the emissions of the 5 species.

Second, an Observing Systems Simulation Experiment (OSEE) was conducted to evaluate the performance of the RAPAS system. We used the MEIC 2016 inventory as a "true" emission, and reduced the "true" emission by 30% over the mainland of China as a prior emission. Overall, the errors of posterior emissions of CO, SO2, NOx, PPM2.5 and PMC over the mainland of China can be reduced by 78.4%, 86.1%, 78.8%, 77.6%, and 72.0%, respectively. Results of the OSSE show that the RAPAS can significantly reduce errors in prior emissions, and the method adopted in the RAPAS is reasonable and feasible.

Third, to diagnose the performance of the EnKF analysis, the chi-squared ($\chi2$) statistics was also calculated. Since we imposed a same uncertainty of prior emission at each DA window to partially compensate for the influence of model errors, $\chi2$ statistics showed small fluctuations, indicating that the system updates emissions consistently and stably.

Fourth, as only organized emissions from stacks have direct online monitoring in China, direct validation of the optimized emissions is impossible, and instead, we indirectly evaluate the posterior emissions by comparing the forward simulated atmospheric mixing ratios against measurement. The results showed that the RMSEs of the simulated concentrations with posterior emissions decreased by 40.1-56.3%, and the

CORRs increased from 0.26-0.66 to 0.69-0.87 for different species.

Fifth, according to the suggestion of another reviewer, the weekend effect of emissions was examined in the revised manuscript, which is a good way to test the reasonability of the day-to-day variation of the inverted emissions. But unfortunately, the weekend effect of China's emissions is not significant (Wang et al., 2014; Wang et al., 2015). We examined the day-to-day variations of the posterior NOx emissions, which is mainly emitted by transportation (Li et al., 2017), and found that there is no significant weekend effect, but the daily variation of the posterior emissions is in good agreement with the observations.

Finally, the emission changes of NOx during the 2020 COVID-19 Epidemic have been widely studied with satellite observations (Bauwens et al., 2020; Liu et al., 2020), and were relatively well known because of the lockdown policy. In our previous study (Feng et al., 2020), with the RAPAS system, we also estimated the spatial and temporal changes of the NOx emissions, and found that the inverted daily variations of NO2 emissions in different areas of China were consistent with their lockdown policies.

In this study, we further extended our system to synchronously assimilate the conventional ground observations of SO2, CO, NO2, PM2.5 and PM10 and infer the SO2, CO, NOx, primary PM2.5 and coarse PM10 emissions. Since the observations are released by the National Environmental Monitoring Station of China in real time, with this system, near-real time emissions could be estimated. Therefore, we believe that this system could serve as a useful tool for accurately quantifying the changes of anthropogenic emissions at near-real time, which will be helpful for the air pollution control in China, and the other regions around the world with ground observation networks.

Bauwens, M., Compernolle, S., Stavrakou, T., Müller, J.-F., van Gent, J., Eskes, H., Levelt, P. F., van der A, R., Veefkind, J. P., Vlietinck, J., Yu, H., and Zehner, C.: Impact of Coronavirus Outbreak on NO2 Pollution Assessed Using TROPOMI and OMI Observations, 47, e2020GL087978, 10.1029/2020gl087978, 2020.

Chevallier, F., Bréon, F.-M., and Rayner, P. J.: Contribution of the Orbiting Carbon Observatory to the estimation of CO2 sources and sinks: Theoretical study in a variational data assimilation framework, 112, 10.1029/2006JD007375, 2007.

Feng, S., Jiang, F., Wang, H., Wang, H., Ju, W., Shen, Y., Zheng, Y., Wu, Z., and Ding, A.: NOx Emission Changes Over China During the COVID-19 Epidemic Inferred From Surface NO2 Observations, Geophysical Research Letters, 47, 10.1029/2020gl090080, 2020.

Li, M., Zhang, Q., Kurokawa, J.-i., Woo, J.-H., He, K., Lu, Z., Ohara, T., Song, Y., Streets, D. G., Carmichael, G. R., Cheng, Y., Hong, C., Huo, H., Jiang, X., Kang, S., Liu, F., Su, H., and Zheng, B.: MIX: a mosaic Asian anthropogenic emission inventory under the international collaboration framework of the MICS-Asia and HTAP, Atmospheric Chemistry and Physics, 17, 935-963, 10.5194/acp-17-935-2017, 2017.

Liu, F., Page, A., Strode, S. A., Yoshida, Y., Choi, S., Zheng, B., Lamsal, L. N., Li, C., Krotkov, N. A., Eskes, H., van der A, R., Veefkind, P., Levelt, P. F., Hauser, O. P., and Joiner, J.: Abrupt decline in tropospheric nitrogen dioxide over China after the outbreak of COVID-19, Science Advances, 6, 2020.

Takagi, H., Saeki, T., Oda, T., Saito, M., Valsala, V., Belikov, D., Saito, R., Yoshida, Y., Morino, I., Uchino, O., Andres, R. J., Yokota, T., and Maksyutov, S.: On the Benefit of GOSAT Observations to the Estimation of Regional CO_2 Fluxes, SOLA, 7, 161-164, 10.2151/sola.2011-041, 2011.

Wang, Y. H., Hu, B., Ji, D. S., Liu, Z. R., Tang, G. Q., Xin, J. Y., Zhang, H. X., Song, T., Wang, L. L., Gao, W. K., Wang, X. K., and Wang, Y. S.: Ozone weekend effects in the Beijing-Tianjin-Hebei metropolitan area, China, Atmospheric Chemistry and Physics, 14, 2419-2429, 2014.

Wang, Z., Li, Y., Dong, X., Sun, R., Sun, N., and Pan, L.: Analysis on weekend effect of air pollutants in urban atmosphere of Beijing, Journal of University of Chinese Academy of Sciences, 32, 843-850, 2015.

Specific comments:

1. The introduction comprises related literature on emission inversion, however with a clear bias toward Kalman filtering. The variational approach, which plays a key role in radiatively active trace gases for greenhouse effects, is also used in reactive chemistry inversion. Yet this methodological branch is disposed of by remarking that this is "technically difficult and cumbersome for complex chemical transport models",

without addressing its features. Literature survey may be complemented to a appropriate level here.

Response: Thank you for this suggestion. We supplement some literature reviews on the characterization of the 4D-var method and related inversion studies in the revised manuscript. See lines 108-119, pages 4-5.

"4DVAR provides a global optimal analysis through minimizing a cost function. It shows implicit flow-dependent background error covariance and can reflect complex nonlinear constraint relationship (Lorenc, 2003). Additionally, the model error can be partly accounted for with a weak constraint 4DVAR method through the definition of a systematic error term in a cost function (Derber, 1989). For example, GEOS-Chem and TM5 4DVAR frameworks have been used to estimate CH_4 (Alexe et al., 2015; Schneising et al., 2009; Stanevich et al., 2021; Wecht et al., 2014) and CO_2 fluxes (Basu et al., 2013; Nassar et al., 2011; Wang et al., 2019) from different satellite retrieval products. Monteil et al. (2013) showed that the global patterns of CH_4 emissions derived from SCIAMACHY (with bias correction) and GOSAT retrievals are in remarkable agreement based on 15 months observations. Additionally, Jiang et al. (2017) used 4DVAR algorithm to estimate global CO emission trends from 2000–2015 using MOPITT retrievals. Kurokawa et al. (2009) and Stavrakou et al. (2008) …"

2. A central position next to emission optimization is initial value optimization as a prerequisite for unbiased emission rate estimates. A discussion on sensitivities of CTMs to additional parameter controls (see above), which demonstrate the authors' awareness of competing impacts on model simulations is lacking however. As it is presented, it is tacitly assumed that these sensitivities are minor. In this context, the authors' approach may well be valid. But it is a matter of scientific scrutiny to meticulously expose the underlying assumptions, which would not affect the value of the study.

Response: Thank you very much for this comment. We fully agree that the model-data mismatch error comes not only from the emissions, but also from the inherent model

errors arising from model structure, discretization and parameterizations, such as boundary condition error, deposition rates, lack of chemical reactions and processes of gas-particle transformation, etc. At present, there is still a lack of reasonable and effective algorithms to solve the model error in atmospheric data assimilation (Houtekamer and Zhang, 2016). Usually, it was assumed that the atmospheric chemical transport model is perfect. Neglecting model errors in the EnKF would attribute all uncertainties to emissions. We have added the following discussions about model error in the revised manuscript, which could also be found in lines 1096-1129, pages 58-59.

"The model-data mismatch error not only comes from the emissions, but also from the inherent model errors arising from model structure, discretization, parameterizations and the biases in the simulated meteorological fields. Neglecting model errors would attribute all uncertainties to emissions, and lead to considerable biases in the estimated emissions. In the version of CMAQ model used in this study, there is no heterogeneous reactions (Quan et al., 2015; Wang et al., 2017), the parameterization scheme for the formation of secondary organic aerosol (SOA) is imperfect (Carlton et al., 2008; Jiang et al., 2012; Yang et al., 2019), no feedback between chemistry and meteorology is considered, and we used an idea profile for chemical lateral boundary conditions. All of the above problems can lead to underestimated concentrations of pollutants, which in turn require more emissions to compensate, leading to overestimations in emissions. In addition, previous studies have shown that the emission of ammonia in the MEIC inventory was underestimated (Kong et al., 2019b; Paulot et al., 2014; Zhang et al., 2018). Due to lack of ammonia observations, our system does not include emission estimates of ammonia, which means that the concentration of ammonium aerosol was underestimated in this system, also resulting in an overestimation in the $PPM_{2.5}$ emission. Wind-blown dust was also not simulated here, thus the PMC emission inverted in this system do not only come from anthropogenic activities, but also from natural sources. Although some of these shortcomings could be solved in the future by updating the CTM model, there will still be errors in each parameterization and each process. Generally, parameter estimation method was used to reduce the model errors,

in which, some uncertain parameters were included in the augmented state vector and were optimized synchronously based on the available observations (Brandhorst et al., 2017; Evensen, 2009). However, it is still quite difficult to identify the key uncertain parameters of different species in different models, which generally comes not only from the complex atmospheric chemical model, but also from hundreds of model inputs (Tang et al., 2013). Another method is bias correction, which treats the model error as a bias term, and includes it in the augmented state vector (Brandhorst et al., 2017; De Lannoy et al., 2007; Keppenne et al., 2005). In addition, the weak-constraint 4D-Var method can also be used to reduce the model errors, which adds a correction term in the model integration to account for the different sources of model error (Sasaki, 1970). Although reliable diagnosis of model error is still a challenge at present (Laloyaux et al., 2020), it should be considered in an assimilation system. We will consider model errors in our system in the future to obtain better emission estimates."

Houtekamer, P. L. and Zhang, F.: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation, Monthly Weather Review, 144, 4489-4532, 2016.

Methodology comments:

3.a. Inversion and data assimilation methods perceived in the referenced literature address Best Linear Unbiased Estimators (BLUE). This property should also be granted by the two-step procedure combining three-dimensional variational (3DVAR) algorithm and the ensemble square root filter (EnSRF) algorithm, which is claimed to show "that the "two-step" scheme clearly outperformed the simultaneous assimilation of ICs and emissions ("one-step" scheme), …" (line 72). While splitting tasks into sequential steps is a well established strategy in numerics to reduce complexity and increase efficiency (e.g. ADI methods), essential care must be taken to ensure the convergence of both split (two-step) and combined (one-step) algorithms toward the

same result. The claimed robustness of the method is demonstrated here not in a sense of same result. Rather different background emissions are probed for convergence (e.g. line 153), which does not experimentally validate the two-step method. Moreover, it is a problem, that the two-step method is established by two different methods. Unless there is a sophisticated multivariate (that is multi-species crosscorrelating) background error covariance matrix (which is evidently not the case in this study) initial value estimation by 3DVAR cannot account for some realistic chemical consistency. Even monovariate cross-correlations are essential to be defined, both horizontally and vertically, to optimize unobserved adjacent areas/height levels. Any estimation errors made in step one are compensated by step two provided biased emission optimization, which is, by EnKF, also performed in time. (Say, too high/low estimated vertical concentrations aloft and mixed down to the observation site are adjusted by too low/high emissions). So the system is prone to propagate errors made in a single step, the methodologically weakest of which limits the overall success. So better analyse why (see L 71-72) it can be possible that "the "two-step" scheme clearly outperformed the simultaneous assimilation of ICs and emissions ("one-step" scheme)".

The authors should either proof the BLUE property of their two-step method with a unique solution by mathematical rigour, or adopt a one-step procedure based on their EnKF approach, or even better, an ensemble Kalman Smoother.

Response: Thank you for this comment. We want to prove the superiority of the system in the inversion strategy ("two-step") by comparing the performances of the "two-step" and "one-step" schemes. The latter has been widely used in previous studies (Miyazaki et al., 2017; Peng et al., 2018). However, in this study, for the "one-step" scheme, we use a combination assimilation method, namely 3DVAR for the optimizations of initial fields and EnKF for emission inversions in each DA window, which is similar as Jiang et al., (2017), but different from most previous studies. Because most previous "one-step" assimilation studies used only one method (i.e., EnKF). We agree that this combination method may cause the comparison less than perfect. However, It should be noted that, even using the same method (such as EnKF) to optimize the emission of

the current window and the initial field of the next window simultaneously (Peng et al., 2018), the initial field estimation errors will still be mixed in the simulated concentration field, resulting in unreasonable emission compensation in the next window. In "one-step" scheme, the essence is to build a good initial field through data assimilation for each DA window. The 3DVAR method used in this study has good performance in optimizing the initial field, which has been shown in section 4.1.2. Although the biases in the high levels were not evaluated, with only ground observations, we believe that the performance of the EnKF method in the high levels is similar. Schwartz et al. (2014) compared the performances of EnKF and 3DVAR in optimizing initial fields, and found that 3DVAR method can obtain a better initial field than EnKF method. Therefore, we believe that in this comparison, a combinatorial assimilation approach used in the "one-step" scheme is an acceptable approach.

During the comparison, three results were analyzed, including the evaluation of posterior emissions against observations, the daily variations of the posterior emissions, and the convergence performances with different prior emissions. The evaluation against observations shows that the RMSEs of the simulated concentrations with posterior emissions decreased by 40.1-56.3%, and the CORRs increased from 0.26-0.66 to 0.69-0.87 for different species; The daily variations show that the posterior emissions is in good agreement with the observations; and the convergence performances with different prior emissions show that the differences between the posterior emissions gradually decrease over time in the "two-step" scheme, but it is not found in the "one-step" scheme. Therefore, we believe that the "two-step" scheme has better performance than the "one-step" scheme in emission inversions.

Following descriptions are added in the revised manuscripts.

Lines 1046-1048, page 56.

"Overall, there is no significant difference between the two methods for NOx and SO2, but for CO, it can be clearly seen that the difference increases with the inversion (Figure S11)."

"On the contrary, this overestimation will be corrected quickly in the subsequent inversion using the "two-step" inversion scheme in this study (Figure S11), so as to ensure the stability of the system. Additionally, the other "one-step" experiment, taking MEIC 2012 as prior emissions, was conducted. However, the relative differences (Figure S12) in posterior emissions between this experiment and the EMS1 did not converge like that between EMDA and EMS2 with "two-step" scheme (Figure 13), which further demonstrates the advantages of the "two-step" scheme."

"When comparing the performances of the "two-step" and "one-step" schemes, for the "one-step" scheme, we use a combination assimilation method, namely 3DVAR for the optimizations of initial fields and EnKF for emission inversions in each DA window, which is similar as Jiang et al., (2017), but different from most previous studies (Miyazaki et al., 2017; Tang et al., 2013). Because most previous "one-step" assimilation studies used only one method (i.e., EnKF). This combination method may cause the comparison less than perfect. However, it should be noted that, even using the same method (such as EnKF) to optimize the emission of the current window and the initial field of the next window simultaneously (Peng et al., 2018), the initial field estimation errors will still be mixed in the simulated concentration field, resulting in unreasonable emission compensation in the next window. In "one-step" scheme, the essence is to build a good initial field in the high levels. Schwartz et al. (2014) compared the performances of EnKF and 3DVAR in optimizing initial fields, and found that 3DVAR method can obtain a better initial field than EnKF method. Therefore, we believe that in this comparison, a combinatorial assimilation approach used in the "one-step" scheme is an acceptable approach, and the conclusion is credible, that the "two-step" scheme has better performances than the "one-step" scheme in emission estimates."
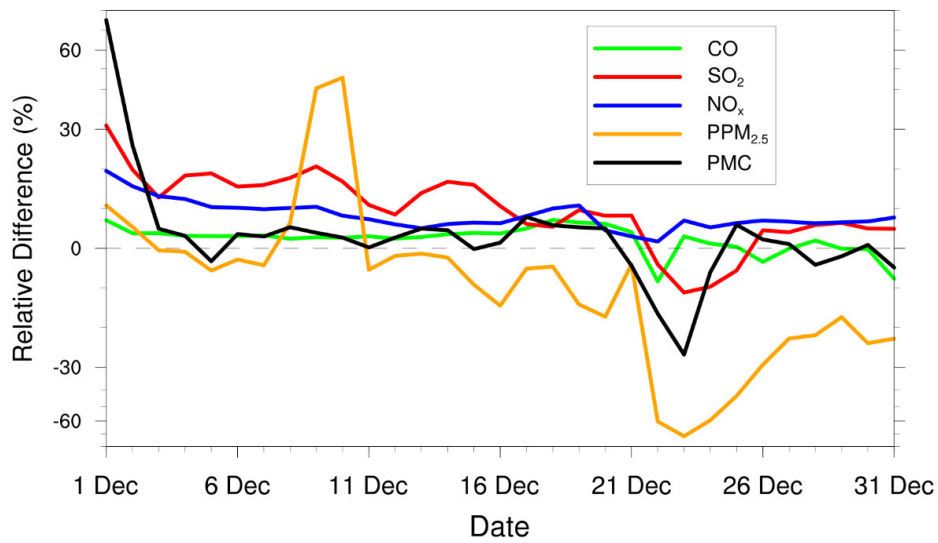
Figure R1. Relative differences in CO, SO2, NOx, PPM2.5 and PMC emissions (%) between the EMS1 and the other "one-step" experiment taking MEIC 2012 as prior emissions. Note that the X-axis scale is different from Figure 13. (Figure S12 in the revised manuscript)
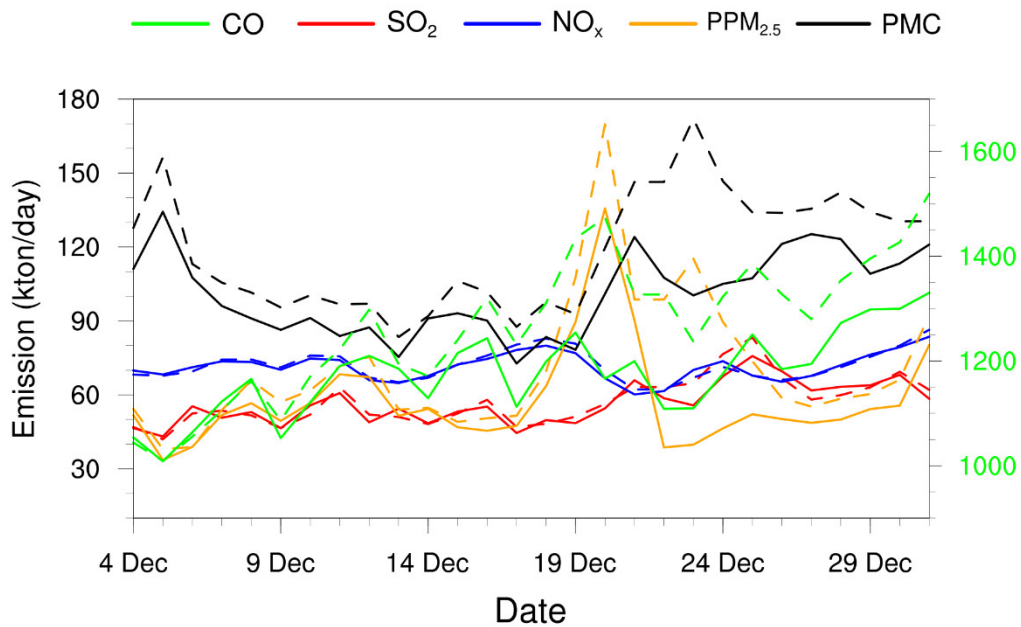

Figure R2 The temporal variability of the posterior emissions (kton/day) in EMSD (solid line) and EMS1 (dotted line) experiments over the whole mainland China. Labels on the right Y-axis refer to CO emissions. (Figure S11 in the revised manuscript)

Jiang, Z., Worden, J. R., Worden, H., Deeter, M., Jones, D. B. A., Arellano, A. F., and Henze, D. K.: A 15-year record of CO emissions constrained by MOPITT CO observations, Atmospheric Chemistry And Physics, 17, 4565-4583, 10.5194/acp-17-4565-2017, 2017.

Miyazaki, K., Eskes, H., Sudo, K., Boersma, K. F., Bowman, K., and Kanaya, Y.: Decadal changes in global surface NOx emissions from multi-constituent satellite data assimilation, Atmospheric Chemistry and Physics, 17, 807-837, 2017.

Peng, Z., Lei, L., Liu, Z., Su, J., Ding, A., Ban, J., Chen, D., Kou, X., and Chu, K.: The impact of multi-species surface chemical observation assimilation on air quality forecasts in China, Atmospheric Chemistry and Physics, 18, 10.5194/acp-18-17387-2018, 2018.

Schwartz, C. S., Liu, Z., Lin, H.-C., and Cetola, J. D.: Assimilating aerosol observations with a "hybrid" variational-ensemble data assimilation system, Journal Of Geophysical Research-Atmospheres, 119, 4043-4069, 10.1002/2013jd020937, 2014.

3.b. In addition, it is a well established validation procedure and common practice after implementation of a new data assimilation/inversion algorithm, to test the system by "identical twin" experiments and Observation System Experiments (OSEs), where a virtual reality is given by "nature runs", where "artificial observations" serve to estimate known chemical states and emission rates (see textbook by Daley, 1990 for details). There is no hint in the paper, that corresponding activities have been undertaken. These tests do not confirm the correctness of the theoretical approach. This being assumed, they seek a necessary, yet insufficient test being approved. The authors are strongly encouraged to validate and proof their approach by these tests.

Response: Thank you for this suggestion. According to this suggestion, we have added an Observing Systems Simulation Experiment (OSSE) to further validate our system in the revised manuscript. We also show the modifications as follows:

Lines 561-568, page 25.

"… coincides with the research stage. An Observing Systems Simulation Experiment (OSEE) was conducted to evaluate the performance of the RAPAS system, which has been widely used in previous assimilation systems development (Daley, 1997). In the

OSSE experiment, we used the MEIC 2016 inventory as a "true" emission, and reduced the "true" emission by 30% over the mainland of China as a prior emission. The simulations simulated using the "true" emission and sampled according to the locations and times of the real observations were used as artificial observations. The observation errors are the same as those in EMDA. To evaluate the IC improvements …"

Table 3

Exp. Type	Exp. Name	Period	IC of the first DA window	ICs of the subsequent DA window	Emission
Assimilation	OSSE	1-31 December	The same as EMDA	The same as EMDA	The same as EMDA, but with a decrease of 30% for CO, SO_2, NO_x, $PPM_{2.5}$, and PMC

Lines 819-828, pages 42-43.

4.1.6 Evaluation using OSSE

Figure 11 shows the spatial distribution of the error reduction in the posterior emissions of the five species. It can be found that after inversion, in most areas, the emission errors can be reduced by more than 80%, especially in the central and eastern regions with dense observation sites, while in remote areas far away from cities, due to the sparse observation sites, the emission errors are still not well adjusted. Overall, the error reduction rates of CO, SO_2, NO_x, PPM2.5, and PMC are 78.4%, 86.1%, 78.8%, 77.6%, and 72.0%, respectively, indicating that with the ground in-situ observations in China, RAPAS can significantly reduce emission errors, thus has good performance in emission estimates.
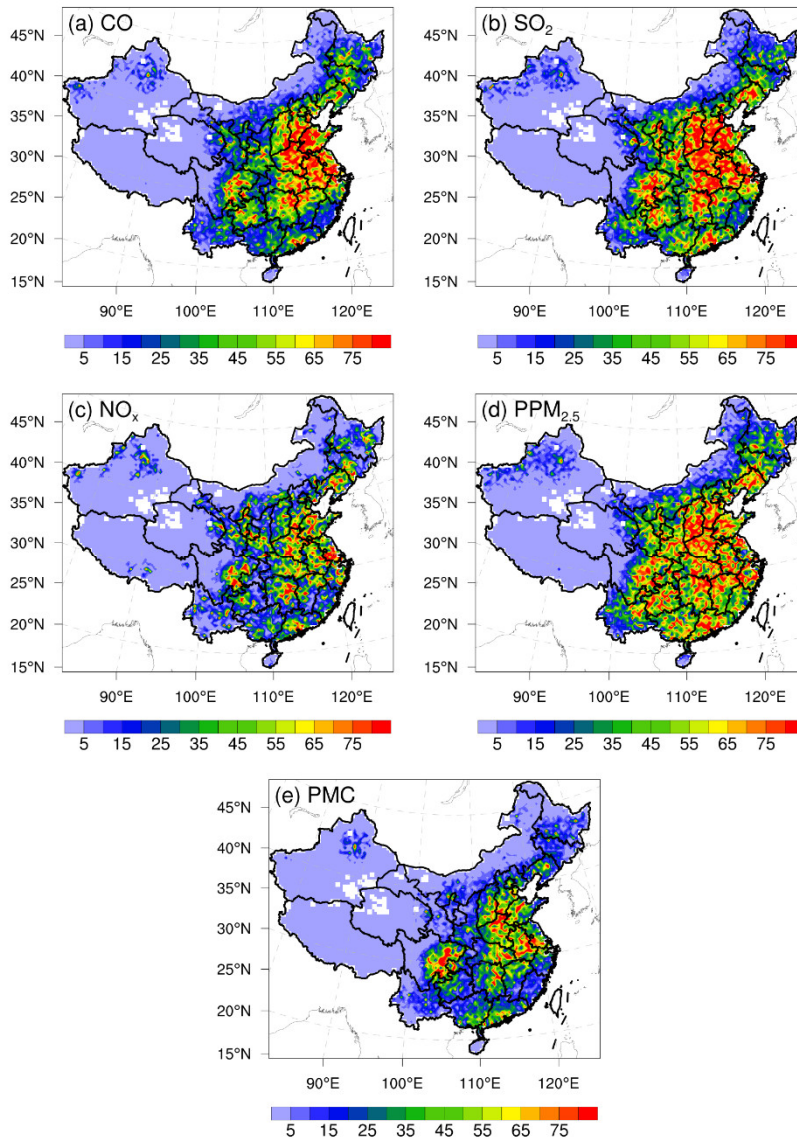
Figure R3 Spatial distribution of the error reduction (%) of posterior emissions in the OSSE. (Figure 11 in the revised manuscript)

4. The presentation of the EnKF is confusing (see equs. (7)-(12)): Below (7) we read "$\boldsymbol{\delta X_i^b}$ represents the randomly perturbed samples that are added to the prior emissions Xb0 to produce ensemble samples of the inputs $\boldsymbol{X_i^b}$ .....". Evidently these parameters are emissions. In equs (8) and (11), also others in between, these parameters are used as state parameters in traditional ensemble KF notation, revealing in (8) a comparison with observations ($\boldsymbol{y - H\overline{X^b}}$). I presume observation operator $\boldsymbol{H}$ does not link

emissions **X** with observations **y**. Rather concentrations are surely meant here (and not the extremely unlikely case of available eddy covariance tower flux measurements). A means to unravel this would be the approach of expanding the state vector of concentrations by emission rates following Wu at al. (2016). See their eq. (5.3). Take also note of their discussion of optimization and control on the finite assimilation window length. What is the assimilation frequency within a DA window of 1 day? Please clarify these points.

Response: Thank you for this comment. The observation operator H reflects the combined information of emissions, the physics and chemistry processes in simulations and the transformation of different species from model space to observation space (Peng et al., 2017). $H\overline{X^b}$ can be regarded as the simulations at the observation locations. Therefore, $y - H\overline{X^b}$ represents the deviations between the observations and the simulations. Wu at al. (2016) expanded the state variables to realize the joint optimization of concentrations and emissions, which is similar to "one-step" method. The relationship between emissions and observations was also expressed through the observation operator H (Wu et al., 2016, Section 3). The assimilation frequency in one assimilation window is once. Following descriptions are added in the revised manuscripts.

See lines 400-402, page 19.

"H is the observation operator that maps simulated concentrations from model space to observation space; $y - H\overline{X^b}$ reflects the differences between the simulated and observed concentrations;"

See lines 433-434, page 20.

"Due to the "super-observation" approach, only one assimilation is needed in one assimilation window."

Peng, Z., Liu, Z., Chen, D., and Ban, J.: Improving PM<sub>2. 5</sub> forecast over China by the joint adjustment of initial conditions and source emissions with an ensemble Kalman filter, Atmospheric Chemistry and Physics, 17, 4837-4855, 10.5194/acp-17-4837-2017, 2017.

5. Is there a test set demonstrating prognostic improvements? And if so, which forecast time is used. I was unable to identify such an experiment, which demonstrates a type of benefits from emission optimization.

Response: Thank you for this comment. In previous, many studied have been conducted to illustrate the benefits of emission optimization for air quality forecasting (ref.). In fact, improving air quality forecasts was one of our goals in developing this system, and during the development of this system, we indeed conducted experiments to test its ability to improve forecasts. We found that, compared the forecasts (72 hours) with original emission inventory, overall, the RMSEs of forecasts with the optimized emissions could decrease by 45.2%, 57.0%, 43.8%, 34.9% and 41.3% and the CORRs could increase by 82.0%, 140.0%, 19.9% 40.5% and 59.4% for CO, SO_2, NO_2, $PM_{2.5}$ and PM_{10}, respectively (Figure R4). Because the purpose of this paper is to introduce this system, we do not present in the paper the impact of emission optimization using this system on air quality forecasts..
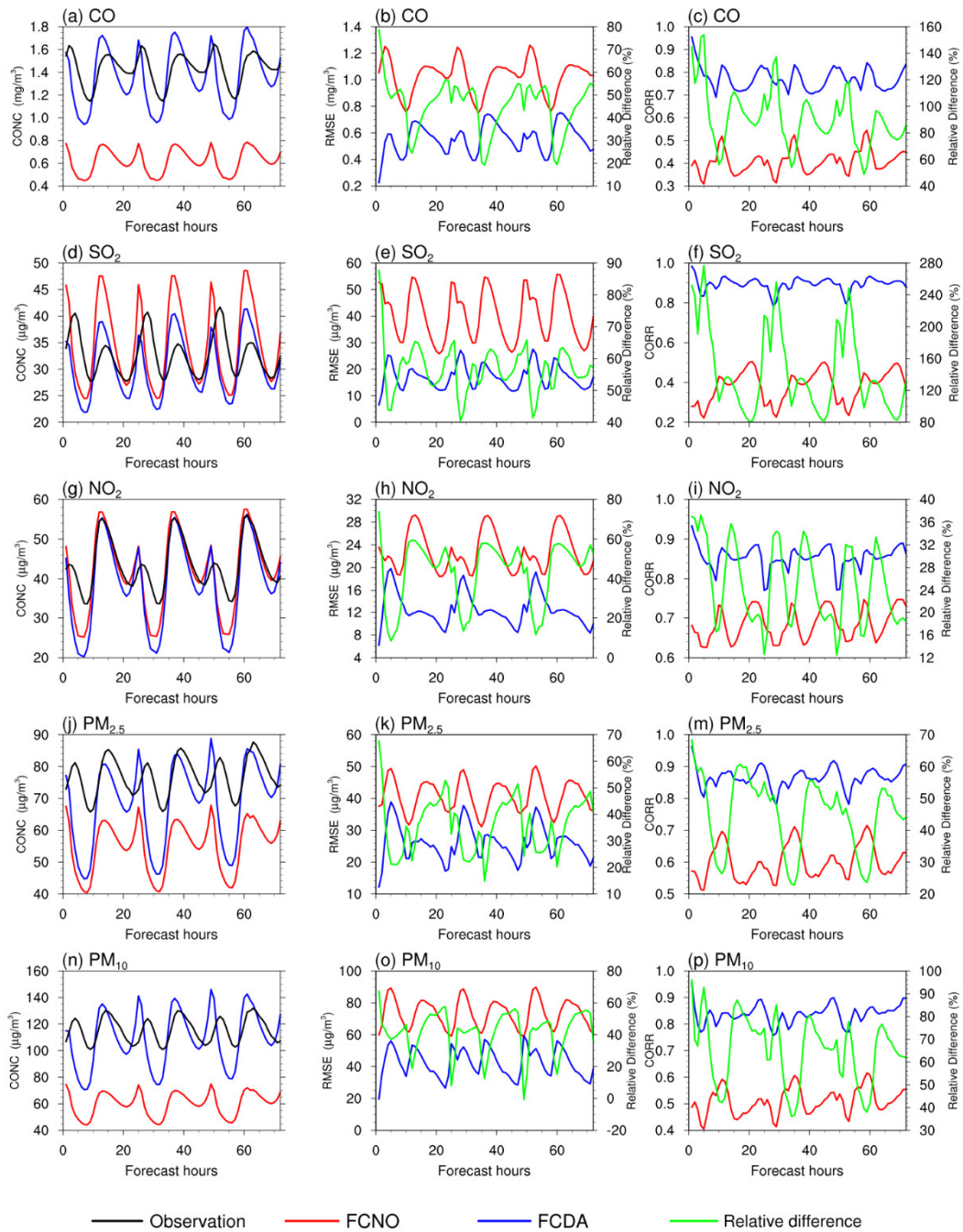
Figure R4 Mean forecasts and observations, RMSE (mg/m³ for CO and μg/m³ for other species) and CORR of the 72-hr forecasts, aggregated over 28 forecasts and verified against all surface monitoring data. The forecasts are performed at each 0000 UTC and the emissions of FCNO and FCDA are from the MEIC and a posteriori of the previous window, respectively.

6. L 203: The grid size of 36 km is very coarse. In fact to be judged in relation to your following statement:

L 878-883 "In additionally, O3 observations are not assimilated to improve NOx and VOC emissions using cross species information due to the strong nonlinear effects within the O3-NOx-VOC relationship, in which the O3 concentration and NOx (VOC) emissions are positively correlated in theNOx (VOC)-limited region and negatively correlated in the VOC (NOx)-limited region (Tang et al., 2011)." NOx chemistry is linked to VOCs and CO via O3. Due to the lack of VOC observations, the evolution of O3 (build up) is mandatory, as otherwise NO2 evolution is not properly analysable. This clearly indicates that the model set-up/coarseness is unable to simulate the core reactivity of tropos. chemistry properly and thus is unsuited to infer emission rates. It is suggested to rerun the experiment with asufficiently highly resolved nest on a densely observed area..

Response: Thanks for this suggestion. Indeed, the unoptimized VOC emissions due to the lack of observations will perturb the O3-NOx-VOC relationship, and then affect the optimization of NOx emissions. The purpose of those sentences is to show that due to the highly nonlinear relationship of NOx-O3-VOC, we cannot constrain NOx or VOC emissions by assimilating O_3 observations, not that the model cannot well simulate the tropospheric chemical reactions. CMAQ is a widely used and validated regional air quality model (Nolte et al., 2015), with the popular spatial resolution about 36 km at the regional scale (Chen et al., 2021; Holloway et al., 2012; Moniruzzaman et al., 2020; Sharma et al., 2016). Many studies used the global model (e.g., GEOS-Chem) to simulate O3 (Zhang et al., 2011; Jourdain et al., 2010; Parrington et al., 2008) or constrain NOx emissions (Jourdain et al., 2010; Vinken et al., 2014) with a resolution of tens to hundreds of kilometers at the regional scale, but they can still better capture the temporal and spatial changes of NOx and O3. There are reasons to believe that CMAQ has the ability to simulate O3 well at a higher resolution (36 km) at the regional scale. For example, Sharma et al. (2016) used CMAQ model with a resolution of 36 km to simulate ozone in India, and further investigated the effect of different precursor

species and sources on O3. The evaluations showed that the simulation was in good agreement with the observation. Using the same resolution, Mueller and Mallard (2011) used CMAQ model quantified the contribution of natural sources to ozone over the continental United States. Model evaluations showed that overall bias in daily maximum 8-h ozone was less than3 ppbv and the mean fractional bias was only 6%. In addition, we also evaluated the O3 simulation with posterior emissions inferred in this study and found that, compared with observation, the simulated O3 is still improved due to improved NOx estimation, with a decrease of RMSE from 35.8 to 33.3 $\mu g\ m^{-3}$ and an increase of correlation coefficient from 0.43 to 0.53. Gan et al. (2016) found that there was no significant difference for both simulated pollutant concentrations and radiation variables at 36 and 12 km resolution in CMAQ model. We also further conducted a nested emission inversion on a densely observed area (the Yangtze River Delta, China) with a grid spacing of 12 km (Feng et al., 2022). The study period is the same as this study. Results showed that the biases (-2.6 vs. -1.0 $\mu g\ m^{-3}$) , RMSEs (11.0 vs. 9.4 $\mu g\ m^{-3}$) and correlation coefficients (0.86 vs. 0.89) of simulated NO2 with two posterior emissions were comparable against the observation in the Yangtze River Delta. Moreover, the NOx emissions in the Yangtze River Delta retrieved at two resolutions are almost the same (14.7 kt/day vs. 13.4 kt/day), with a difference of 8.8%. These statistics indicate that the emissions can be adjusted effectively by RAPAS. We have added following discusses about the possible effect of coarse model settings on NOx emission inversion.

Lines 243-245, pages 9-10.

"This spatial resolution has been widely adopted in regional simulations and can provide good simulations of the spatiotemporal variations of air pollutants (Mueller and Mallard, 2011; Sharma et al. 2016)."

Lines 1141-1146, page 60.

"The optimization of the initial fields or emissions of NO_2 may also change the O_3-NO_x-VOC relationship. Assuming that NO_2 is underestimated, the NO_2 concentration

increases after assimilation, but the VOC concentration remains unchanged, then in the NOx (VOC)-limited region, the subsequent generation of O_3 will increase (decrease); Conversely, the ozone concentration errors caused by assimilating NO_2 will also affect the subsequent NOx emission inversion."

Lines 1150-1157, page 60.

"To evaluate the influence of O_3-NO_x-VOC relationship change and model resolution on inversion, we also further conducted a nested emission inversion on a densely observed area (the Yangtze River Delta, China) with a grid spacing of 12 km (Feng et al., 2022). The study period is the same as this study. Results showed that the NOx emissions in the Yangtze River Delta retrieved at two resolutions are almost the same (14.7 kt/day vs. 13.4 kt/day), with a difference of 8.8%, indicating that the emissions can be adjusted effectively by RAPAS."

Chen, Y., Shen, H., Kaiser, J., Hu, Y., Capps, S. L., Zhao, S., Hakami, A., Shih, J. S., Pavur, G. K., Turner, M. D., Henze, D. K., Resler, J., Nenes, A., Napelenok, S. L., Bash, J. O., Fahey, K. M., Carmichael, G. R., Chai, T., Clarisse, L., Coheur, P. F., Van Damme, M., and Russell, A. G.: High-resolution hybrid inversion of IASI ammonia columns to constrain US ammonia emissions using the CMAQ adjoint model, Atmos. Chem. Phys., 21, 2067-2082, 2021.

Feng, S., Jiang, F., Wang, H., Shen, Y., Zheng, Y., Zhang, L., Lou, C., and Ju, W.: Anthropogenic emissions estimated using surface observations and their impacts on PM2.5 source apportionment over the Yangtze River Delta, China, Science of The Total Environment, 828, 154522, 2022.

Gan, C.-M., Hogrefe, C., Mathur, R., Pleim, J., Xing, J., Wong, D., Gilliam, R., Pouliot, G., and Wei, C.: Assessment of the effects of horizontal grid resolution on long-term air quality trends using coupled WRF-CMAQ simulations, Atmospheric Environment, 132, 207-216, 2016.

Holloway, T., Voigt, C., Morton, J., Spak, S. N., Rutter, A. P., and Schauer, J. J.: An assessment of atmospheric mercury in the Community Multiscale Air Quality (CMAQ) model at an urban site and a rural site in the Great Lakes Region of North America, Atmos. Chem. Phys., 12, 7117-7133, 2012.

Jourdain, L., Kulawik, S. S., Worden, H. M., Pickering, K. E., Worden, J., and Thompson, A. M.: Lightning NO_x emissions over the USA constrained by TES ozone observations and the GEOS-Chem model, Atmos.

Chem. Phys., 10, 107-119, 2010.

Moniruzzaman, C. G., Bowden, J., and Arunachalam, S.: Aircraft landing and takeoff emission impacts on surface O3 and PM2.5 through aerosol direct feedback effects estimated by the coupled WRF-CMAQ model, Atmospheric Environment, 243, 117859, 2020.

Mueller, S. F. and Mallard, J. W.: Contributions of Natural Emissions to Ozone and PM2.5 as Simulated by the Community Multiscale Air Quality (CMAQ) Model, Environmental Science & Technology, 45, 4817-4823, 2011.

Nolte, C. G., Appel, K. W., Kelly, J. T., Bhave, P. V., Fahey, K. M., Collett Jr, J. L., Zhang, L., and Young, J. O.: Evaluation of the Community Multiscale Air Quality (CMAQ) model v5.0 against size-resolved measurements of inorganic particle composition across sites in North America, Geosci. Model Dev., 8, 2877-2892, 2015.

Parrington, M., Jones, D. B. A., Bowman, K. W., Horowitz, L. W., Thompson, A. M., Tarasick, D. W., and Witte, J. C.: Estimating the summertime tropospheric ozone distribution over North America through assimilation of observations from the Tropospheric Emission Spectrometer, Journal of Geophysical Research: Atmospheres, 113, 2008.

Sharma, S., Chatani, S., Mahtta, R., Goel, A., and Kumar, A.: Sensitivity analysis of ground level ozone in India using WRF-CMAQ models, Atmospheric Environment, 131, 29-40, 2016.

Vinken, G. C. M., Boersma, K. F., van Donkelaar, A., and Zhang, L.: Constraints on ship NO_x emissions in Europe using GEOS-Chem and OMI satellite NO2 observations, Atmos. Chem. Phys., 14, 1353-1369, 2014.

Zhang, L., Jacob, D. J., Downey, N. V., Wood, D. A., Blewitt, D., Carouge, C. C., van Donkelaar, A., Jones, D. B. A., Murray, L. T., and Wang, Y.: Improved estimate of the policy-relevant background ozone in the United States using the GEOS-Chem global model with 1/2° × 2/3° horizontal resolution over North America, Atmospheric Environment, 45, 6769-6776, 2011.

Technical corrections:

7. L. 59-63: What is meant in one case: Emission optimization with or without prior IC by 3D-var?

Response: Thanks. We have rephrased the sentence (See lines 61-69, page 3) as follows:

"Results showed that the simulated concentrations of CO, NO2, SO2, PM2.5 and PM10 with the prior inventory have large systematic biases, with relative biases in the range of -48.2-54.2%. In the IA subsystem, after 3DVAR, the root mean squared error (RMSE) of the simulated concentrations decreased by 50.0-73.2%, and the correlation coefficient (CORR) increased to 0.78-0.92 for the five species. In the EI subsystem, after emission inversions, the RMSE of the simulated concentrations decreased by 40.1-56.3%, and the CORR increased to 0.69-0.87."

8. L. 68-69: Selection of the Dec 2016 case study analysis: in view of typically episodic nature of mineral dust: does this high quantitative increment make sense for a validation of a novel algorithm? Why not another episode?

Response: Thanks for this comment. China has implemented strict pollution control measures in recent years to combat air pollution, especially in the cold season, and December is one of the most polluted months in China. Therefore, we believe that it is more meaningful to invert this month's emission changes for the pollution control. In addition, in China, sandstorms mainly occur in spring, and the month we choose is basically unaffected by sandstorms. However, because northern China is basically arid and semi-arid regions, the wind-blown dust exists throughout the year. According to the suggestion of another reviewer, we have also performed another inversion in July. The increase of PMC emissions (1178%) constrained by observations in July is comparable with that in December. The following discussions are added in the revised manuscript. See lines 549-551, page 24.

 "December is one of the months with most severe air pollution, while July is one of the least polluted months in China. Therefore, this study mainly tested the performance of the RAPAS system in these two months. For December, the IA …"

See lines 556-559, pages 24-25.

"For July, the system also operated in the same way as for December. It needs to be noted that due to the stronger atmospheric oxidation, the lifetime of NO2 in July is

significantly shorter than that in December, thus we adopted a smaller localization scale for NO2 (80 km)."

"The evaluation results for July show that the emission uncertainty can still be significantly reduced, and the performance of the system in July is comparable to that in December (Table S2). Additionally, the seasonal variation of emissions can be well reflected (Figures S4 and S5), which means that our system can perform well at different times of the year."
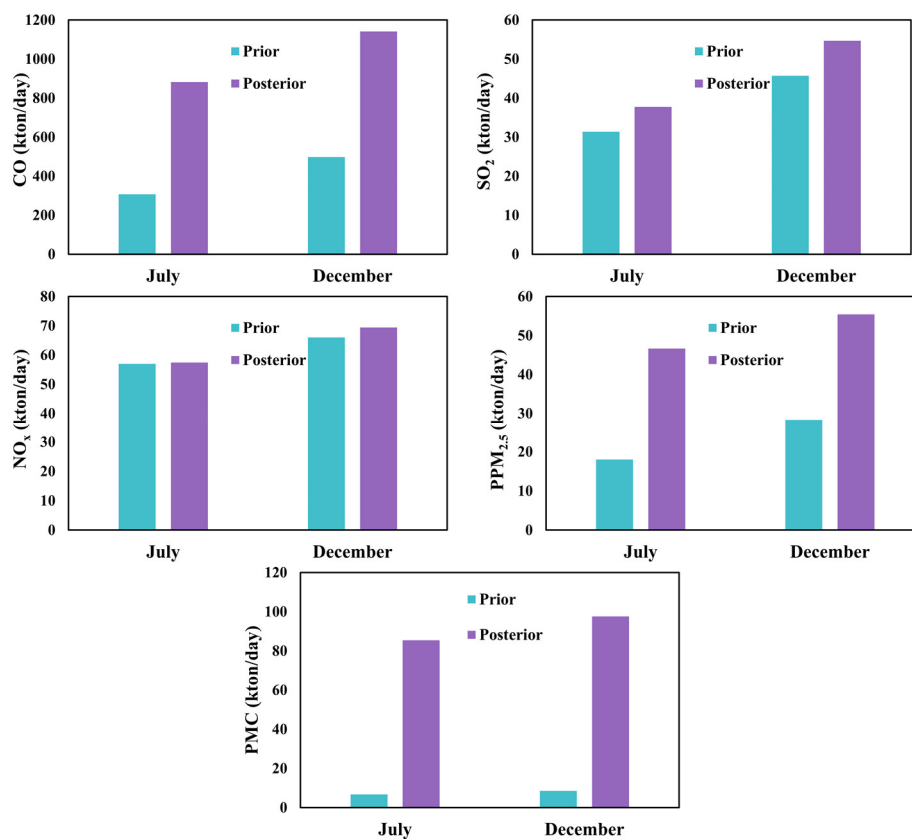

Figure R5 The prior and posterior emissions (kton/day) in July and December 2016 over the whole mainland China. (Figure S5 in the revised manuscript)
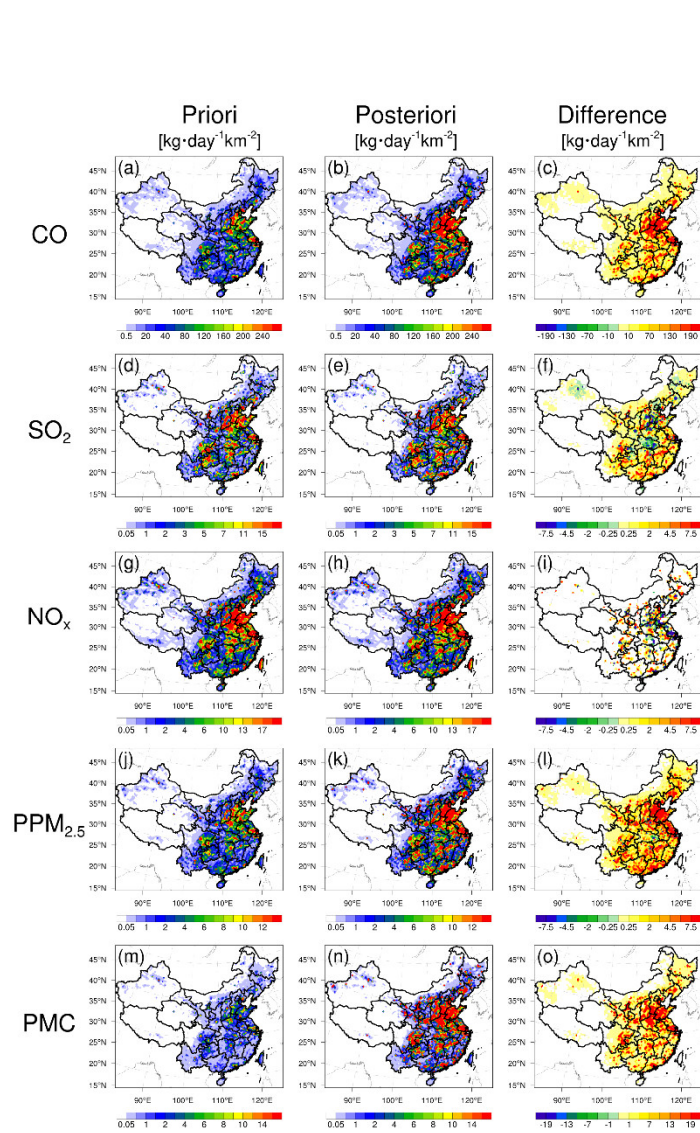
Figure R6 Spatial distribution of the time-averaged prior emissions (left column, MEIC 2016), posterior emissions (middle column), and differences (right column, posterior minus prior) in July 2016. (Figure S4 in the revised manuscript)

Table R1. Statistics comparing the pollution concentrations from the simulations with prior (CEP) and posterior (VEP) emissions against observations for July. CO unit: mg m^{-3}; others units: μg m^{-3}. (Table S2 in the revised manuscript)

Species	Mean Obs.	Mean Sim.		BIAS		RMSE		CORR	
		CEP8	VEP8	CEP8	VEP8	CEP8	VEP8	CEP8	VEP8
CO	0.79	0.33	0.63	-0.46	-0.16	0.58	0.35	0.25	0.65
SO$_2$	12.9	15.6	9.6	2.7	-3.3	19.7	7.2	0.12	0.71
NO$_2$	20.0	23.0	16.5	3.0	-3.5	22.0	6.8	0.47	0.81
PM$_{2.5}$	29.2	21.9	23.3	-7.3	-6.0	21.1	13.9	0.51	0.76
PMC	53.6	27.6	42.3	-26.0	-11.3	42.3	30.9	0.38	0.61

* BIAS, mean bias; RMSE, root mean square error; CORR, correlation coefficient

Beirle, S., Platt, U., Wenig, M., and Wagner, T.: Weekly cycle of NO2 by GOME measurements: a signature of anthropogenic sources, Atmospheric Chemistry and Physics, 3, 2225-2232, 2003.

9. L 122-124: Unclear: "Barbu et al. (2009) updated sulfur oxide (SOx) emissions with SO2 gas and sulfate aerosol observations and showed that forecasts were improved overall but degraded when derived only from SO2 or sulfate observations." Better formulate logics: ... do you mean alone?

Thanks for this comment. We have rephrased that sentence to make it clear. See lines 134-137, page 4.

 "Barbu et al. (2009) updated sulfur oxide (SO$_x$) emissions with SO2 and sulfate aerosol observations and found that simultaneous assimilation of both species had better performance than assimilating one of them alone."

10. L 176-177: "It runs only once and provides a "perfect" chemical ICs for the subsequent EI subsystem." Justify "perfect". Justify analysis of unobserved height

levels?

Response: Thanks for this comment. Even with data assimilation, it is still impossible to get a perfect initial field, so we put the quotes around the word of perfect. In the revised manuscript, we have changed "perfect" to "good" in that sentence (See line 42, page 2; line 530, page 23). In addition, for the concentrations at high levels, they could be constrained by the ground observations through the vertical correlation. Figure R7 shows the changes of vertical profiles between the assimilation and control experiments. Overall, it could be found that there are changes of concentrations almost throughout the whole column. However, due to lack of observations of vertical profiles, it is difficult for us to verify the constraints on the concentrations of the upper air. Anyway, the effect of only assimilating ground observations on the concentration in middle and upper layers of the boundary layer is necessarily limited, so there must be a bias. This also needs our attention, and may be partially resolved by further assimilating satellite observations in the future. We have added some discussions about this point.

See lines 666-670, page 30.

"These statistics indicate the initial fields of the ground level have been significantly improved. However, due to the lack of observations, we still do not know the simulation bias in the upper-middle boundary layer. Although concentrations at high altitudes can be constrained by ground-based observations through vertical correlations, the effect is limited, so the bias is still non-negligible."
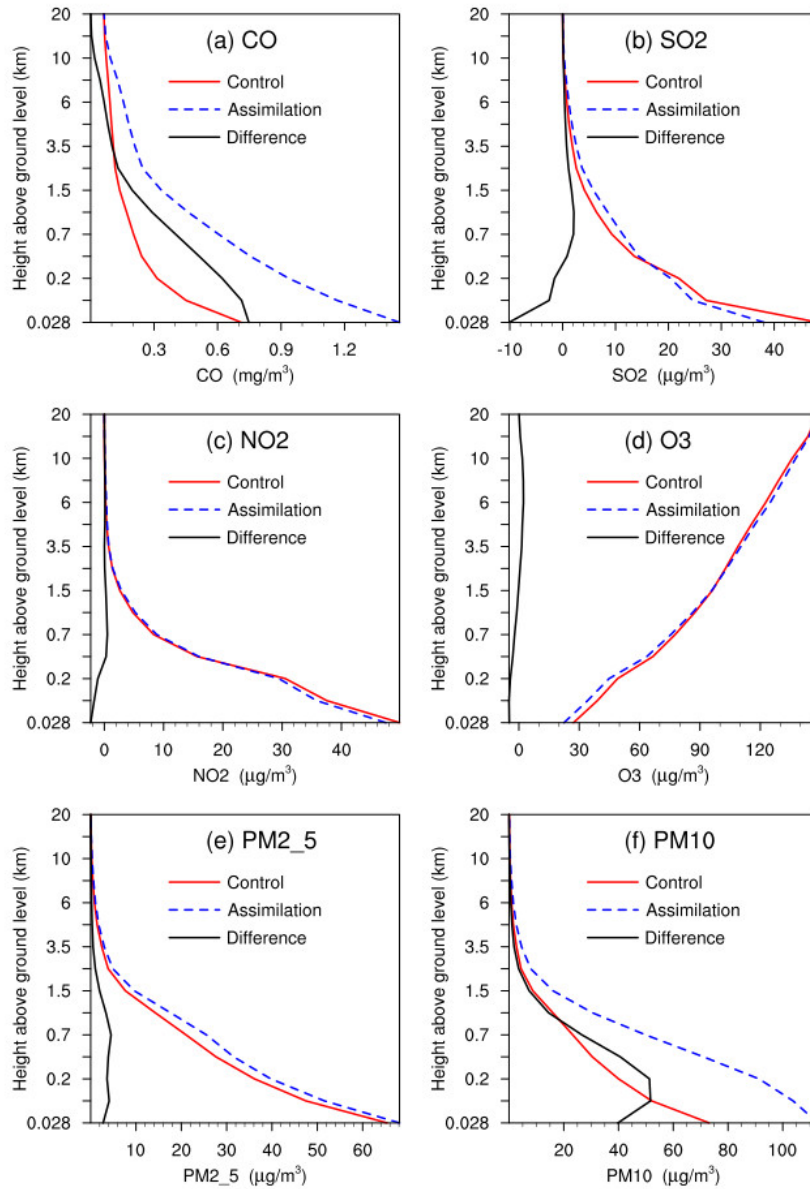
Figure R7 Mean vertical profiles of the background (red) and analysis (blue) fields, and their difference averaged over the China mainland.

11. L. 180-181: "which are then sampled according to the locations and times of the observations" Unclear, more technical details are needed here.

Response: Thanks for this comment. We have changed it to the following sentence. See lines 212-215, page 8.

"In the first step, the prior emissions (X^b) are perturbed and put into the CTM model to

simulate chemical concentration ensembles. The simulated concentrations of the lowest model level are then interpolated to the observation space according to the locations and times of the observations using the nearest neighbor interpolation method."

12. L. 184: Do you mean DA window length of 1 day?

Response: Yes, the DA window is set to 1 day. On the one hand, the model needs a longer time to integrate emission information into the concentration ensembles. On the other hand, due to the complexity of hourly emissions, it is very difficult to simulate hourly concentrations that can match the observations well. We have added descriptions about the length of DA window to Figure 1.

13. Figur1 a caption too short. More detail on precise times and data update frequency in the EnKF needed.

Response: Many thanks for this suggestion. The caption of Figure 1 has changed as follows:

"**Figure 1**. The composition and flow chart of RAPAS. The x_a and x_b represent the prior and posterior emissions. The 3DVAR assimilation stage lasts 5 days with data input frequency of 6 hours, and the DA window in the EI subsystem is set to 1 day."

14. L 225 "represent the measurement sites" Do you mean "chemistry"?

Response: Yes. We have changed "the measurement sites" to "the air pollution monitoring sites". See line 268, page 12.

15. L 242: "… of surface air pollutant observation operators,…" Explain please.

Response: Thanks for this comment. The surface air pollutant observation operator means the interpolation of simulated concentrations from model space to observation

space according to the locations of observation data. In this study, we need to construct observation operators for different pollutants. It has been explained in lines 297-319, page 14.

16. L 312 "the horizontal length scales decrease with increasing heights," Please explain why, and not the opposite.

Response: Thanks. We made a mistake here. As shown in Figure 3, for gaseous pollutants and most individual aerosol components, the horizontal length scales increase with height. We have corrected it in the revised manuscript. See lines 354-355, page 17.

17. L 314 " The ground-level scale generally spread 40-45 km for all control variables on average." This is little more than a grid cell. How defined?

Response: Thanks for this comment. Although the ground-level scale calculated in Formular 6 generally spread 40-45 km on average, in GSI namelist, two variables, namely scale factor and weight factor, were used for further tuning horizontal scale (e.g., 40-45 km) specified for horizontal smoothing and weight setting. The final background error covariance matrix used in analysis are the content multiplied by these factors. Therefore, the actual increment usually spreads multiple grids.

17. L 315-318: More detailed explanation required: Do you mean stack overshooting vs. surface emissions? A result of NMC? Is boundary height and related mixing considered?

Response: Yes, this result was calculated using the "NMC" method. The boundary height and related mixing were considered because these processes are included in the CMAQ model. The vertical length scale of most species increases first and then decreases with height, which is related to the vertical mixing and stack emissions at

about 200 m height. Ma et al (2018) also found high correlation occurred in the middle boundary layer other than the upper or lower levels.

The sentence (See lines 358-360, page 17) has been rephrased as follows:

"The vertical length scale of most species increases first and then decreases with height, which may be related to the vertical mixing (Kahnert, 2008) and stack emissions at about 200 m height."

Ma, C., Wang, T., Zang, Z., and Li, Z.: Comparisons of Three-Dimensional Variational Data Assimilation and Model Output Statistics in Improving Atmospheric Chemistry Forecasts, Advances in Atmospheric Sciences, 35, 813-825, 2018.

18. L 323-324: Whitaker and Hamil is a meteorological application without emission optimisation. For square root filtering technique fundamentals please refer to much earlier textbook literature, eg. Bierman: 1977, or Maybeck 1979

Response: Thanks! We have updated the references. See lines 368-369, page 17.

19. L 336 variable localization: meaning?

Response: The variable localization was defined in Miyazaki et al. (2012) and Ma et al. (2019). It means that the emission of one species is only constrained with its corresponding air pollutant observation. We have rephrased that sentence (See lines 383-388, page 18) as follows:

"We used variable localization to update the analysis, which means that the covariance among different state variables was not considered, and the emission of one species was only constrained with its corresponding air pollutant observation. This method has been widely used in chemical data assimilation systems to avoid spurious correlations among species. (Ma et al., 2019; Miyazaki et al., 2012b)."

Ma, C., Wang, T., Mizzi, A. P., Anderson, J. L., Zhuang, B., Xie, M., and Wu, R.:

Multiconstituent Data Assimilation With WRF-Chem/DART: Potential for Adjusting Anthropogenic Emissions and Improving Air Quality Forecasts Over Eastern China, 124, 7393-7412, 2019.

Miyazaki, K., Eskes, H. J., Sudo, K., Takigawa, M., van Weele, M., and Boersma, K. F.: Simultaneous assimilation of satellite NO2, O-3, CO, and HNO3 data for the analysis of tropospheric chemical composition and emissions, Atmospheric Chemistry and Physics, 12, 9545-9579, 2012.

20. L 336-337: Do you mean also at the same location?

Response: We reinterpreted that sentence. Please see Response 19.

21. Eq (12) Matrix in a denominator? Please explain

Response: Yes, the denominator is a matrix. We have revised Eq (12) to make it clear.

$$\widetilde{K} = \left(1 + \sqrt{R \Big/ (HP^bH^T + R)}\right)^{-1} K \qquad (12)$$

22. L. 384-386: This is too much a try-and-error like discussion. Emissions are not constant or diurnially constant over these times: weather changes, working days, holidays, ... Please be more precise about your reasoning

Response: Thanks for your comment. That sentence (see lines 438-442, page 20) has been changed as follows:

"Kang et al. (2012) conducted OSSEs and demonstrated that due to the errors of transport and increase the spurious correlation, a longer DA window (e.g., 3 weeks) would cause the analysis system to blur out the essential emission information far away from the observation."

23. L425 arithmetically

Response: Thank you. We have corrected it. See lines 488, page 22.

"… and the ensemble-estimated error covariance matrix is arithmetically likely to be underestimated."

24. L 446 are distributed

Response: We have changed "distribute" to "are distributed". See line 510, page 22.

25. L 449 unrepresentative

Response: We have changed "nonspatially representative" to "unrepresentative". See line 513, page 21.

26. L 452-453 Explain relations, formulae better

Response: We have rewritten that formulae and added some explains in the revised manuscript. See lines 516-519, page 23.

"… a function of $max(|O(t) - O(t \pm 1)|) \leq f(t)$, where $O(t)$ and $O(t \pm 1)$ represent observations at time t and $t \pm 1$, respectively, and $f(t) = T_a + T_b \times O_t$. That means that both concentration differences between time *t* and time *t+1* and *t-1* should be less than $f(t)$."

27. L461 25/336 is a fairly low fraction. Please justify.

Response: For better constraints on emissions, we need to select relatively more observations for assimilation. The 25 independent observations are relatively evenly distributed in the central and eastern regions where emissions are relatively strong. We believe that it is enough for the evaluation of the posterior emissions.

28. L 534- 540 For emission inversion the boundary layer height and its prevalent stability is more decisive than any other parameter. Can you please make efforts to validate this by radiosonde data, if available?

Response: Thank you for this suggestion. We have obtained 92 sites (http://weather.uwyo.edu/upperair/sounding.html, last access: 10 March 2022) of sounding data (including virtual potential temperature, geopotential height, wind speed, wind direction, etc) from the University of Wyoming (http://weather.uwyo.edu/upperair/sounding.html) for the validations of the simulated planetary boundary layer height (PBLH), which are in 12 hours interval (00 and 12 UTC). The observed PBLH was calculated with these sounding data using the bulk Richardson number method (Grachev et al., 2013; Richardson et al., 2013). Overall, the WRF simulations show an overall systematic underestimation of PBLH, with bias of -41.1 m (-15.4%). Similar underestimation was also found in Banks et al. (2015) when evaluated eight different schemes in WRF. Lower PBLH corresponds to worse diffusion and transport of air pollutants, which results in higher concentrations in the atmosphere, and accordingly, less emissions were inferred to compensate this underestimation. It should be noted that the low vertical resolution of most sounding data in China may cause a certain bias in the calculated PBLH. We have added evaluations about PBLH in the revised paper.

lines 622-639, page 28.

"To quantitatively evaluate the performance of the WRF simulations, the mean bias (BIAS), root mean square error (RMSE), and correlation coefficient (CORR) were calculated against the surface meteorological observations measured at 400 stations and the planetary boundary layer height (PBLH) calculated using the sounding data at 92 sites. The surface observations were obtained from the National Climate Data Center (NCDC) integrated surface database (http://www.ncdc.noaa.gov/oa/ncdc.html, last access: 25 October 2021), and the sounding data were obtained from the website of the University of Wyoming (http://weather.uwyo.edu/upperair/sounding.html, last access: 10 March 2022). The sounding data are in 12 hours interval. The observed PBLH were

calculated using the sound data through the bulk Richardson number method (Richardson et al., 2013). The spatial distribution of the meteorological stations is shown in Figure 2. The simulated temperature at 2 m (T2), relative humidity at 2 m (RH2), wind speed at 10 m (WS10), and PBLH from 26 November to 31 December 2016 are evaluated against the observations. Table 4 summarizes the statistical results of the evaluations of the simulated meteorological parameters. Overall, the T2, RH2 and PBLH are slightly underestimated, with biases of -0.1 °C, -3.8% and -41.1 m, respectively. The CORRs are approximately 0.98 for T2, 0.94 for RH2 and 0.90 for PBLH, showing good consistency…"

Banks, R. F., Tiana-Alsina, J., Rocadenbosch, F., and Baldasano, J. M.: Performance Evaluation of the Boundary-Layer Height from Lidar and the Weather Research and Forecasting Model at an Urban Coastal Site in the North-East Iberian Peninsula, Boundary-Layer Meteorology, 157, 265-292, 2015.

Grachev, A. A., Andreas, E. L., Fairall, C. W., Guest, P. S., and Persson, P. O. G.: The Critical Richardson Number and Limits of Applicability of Local Similarity Theory in the Stable Boundary Layer, Boundary-Layer Meteorology, 147, 51-82, 2013.

Richardson, H., Basu, S., and Holtslag, A. A. M.: Improving Stable Boundary-Layer Height Estimation Using a Stability-Dependent Critical Bulk Richardson Number, Boundary-Layer Meteorology, 148, 93-109, 2013.

29. "These statistics indicate that the initial fields can be adjusted effectively by our IA subsystem." I think this claim is not sustained, despite the fact that emissions are not yet corrected: What about height levels above? How are they validated? How can be made sure that CB5 chemsitry fields are in the proper chemical balance of the system (notably NO_x, O3, CO, VOCs)?

Response: Yes, this conclusion is indeed not rigorous enough. The evaluations only show that the initial fields of the ground level have been significantly improved. Due to lack of observations, we do not evaluate the concentrations at high levels. Although the concentrations at high levels could be constrained by the ground observations

through the vertical correlation, the effect is limited, so there must be a bias. We modified that sentence in the revised paper, see lines 665-670, page 30. The changes are also listed as follows.

"These statistics indicate the initial fields of the ground level have been significantly improved. However, due to the lack of observations, we still do not know the simulation bias in the upper-middle boundary layer. Although concentrations at high altitudes can be constrained by ground-based observations through vertical correlations, the effect is limited, so the bias is still non-negligible."

30. "L 619: This applies more to the NOx-O3-CO/VOC chemistry rather than PM2.5.

Response: Thank you for this comment. Yes, O_3 has complex precursors, and strong nonlinear relationship with its precursors. Although there are nonlinear processes during the formation of secondary particulate matter, overall, the more PM2.5 precursors are emitted, the more secondary particulate matters are formed, resulting in higher $PM_{2.5}$ concentrations. We have revised that sentences as follows:.

"… It may also be related to the complex precursors and complex homogeneous and heterogeneous chemical reactions and transformation processes of secondary $PM_{2.5}$, and the fact that …"

31. L 662: Where did you get the \sigma from? For KF an analysis error covariance matrix A should be available for this.

Response: Thank you for this comment. In the EnKF-based inverse estimation scheme, the uncertainty is represented by the spread of the ensemble samples (Tang et al., 2013). The posterior and prior uncertainties are the standard deviations of the prior and posterior perturbations of X_i^b and X_i^a. X_i^b was perturbed from the prior emissions X_0^b by adding a randomly perturbed item of δX_i^b, which was drawn from Gaussian distributions with a mean of zero and the standard deviation of the prior emission

uncertainty in each grid. After constrained using observations, the perturbed emissions of X_i^b is changed to X_i^a according to Eq. 2 ~ 5.

$$X_i^b = X_0^b + \delta X_i^b \quad , i = 1, 2, \ldots, N \tag{1}$$

$$X_i^a = X^a + (X_i^b - X^b) - \tilde{K}H(X_i^b - X^b) \tag{2}$$

$$\tilde{K} = \left(1 + \sqrt{R \Big/ (HP^bH^T + R)}\right)^{-1} K \tag{3}$$

$$K = P^bH^T(HP^bH^T + R)^{-1} \tag{4}$$

$$P^b = \frac{1}{N-1}\sum_{i=1}^N (X_i^b - \overline{X}^b)(X_i^b - \overline{X}^b)^T \tag{5}$$

We have added this calculation method of the posterior and prior uncertainties in the supplemental material.

"… where $\sigma_{posterior}$ and σ_{prior} are the posterior and prior uncertainties, respectively, which were calculated using the standard deviations of the prior and posterior perturbations (Text S3)."

Tang, X., Zhu, J., Wang, Z. F., Wang, M., Gbaguidi, A., Li, J., Shao, M., Tang, G. Q., and Ji, D. S.: Inversion of CO emissions over Beijing and its surrounding areas with ensemble Kalman filter, Atmospheric Environment, 81, 676-686, 2013.

32. L 701-702: With 1000% excess is linearisation still acceptable? A scrutinized analysis is appropriate.

Response: Thank you for this comment. If we assume that the 1000% increment of PMC is all from natural dust, that means the contribution of natural dust accounts for 90% of total PMC emissions, which is reasonable in China. Qiu et al. (2016) conducted source apportionment of PM10 and PM2.5 in different functional areas of Lanzhou, China, and the results showed that the contribution of wind-blown dust accounts for more than 90% of total PMC emissions. Another study in Changsha in central China showed that the contribution of wind-blown dust account for about 75% of the primary

PM10 concentrations (Li et al., 2010).

We also realize that this problem in the inversion results. Therefore, we used a larger prior uncertainty in the first three windows, a "two-step" inversion scheme and cyclic iteration to quickly converge emissions. Figure 10 shows the time series of the relative differences between the prior and posterior emissions in each window. In addition to the relatively large adjustment range of emissions in the first three windows, the adjustment range of the PMC is basically within the uncertainty range (e.g., 40%), indicating that with this scheme, EnKF still has a good performance in emission inversion. For PMC, because the posterior emission includes the contribution of natural dust, which is not considered in the original emission, there is a large increment in the first three windows, which is consistent with Ma et al (2019).

We have added related discussions in the revised paper, see lines 861-865, page 44. The changes are also listed as follows:

"Overall, PM10 emissions (PPM2.5+PMC) increased by 318%. If we assume that all the increment in PM_{10} emissions is all from natural dust, that means the contribution of natural dust accounts for 75% of total PM_{10} emissions, which is consistent with the source apportionment of PM_{10} of 75% in Changsha in Central China (Li et al., 2010)."
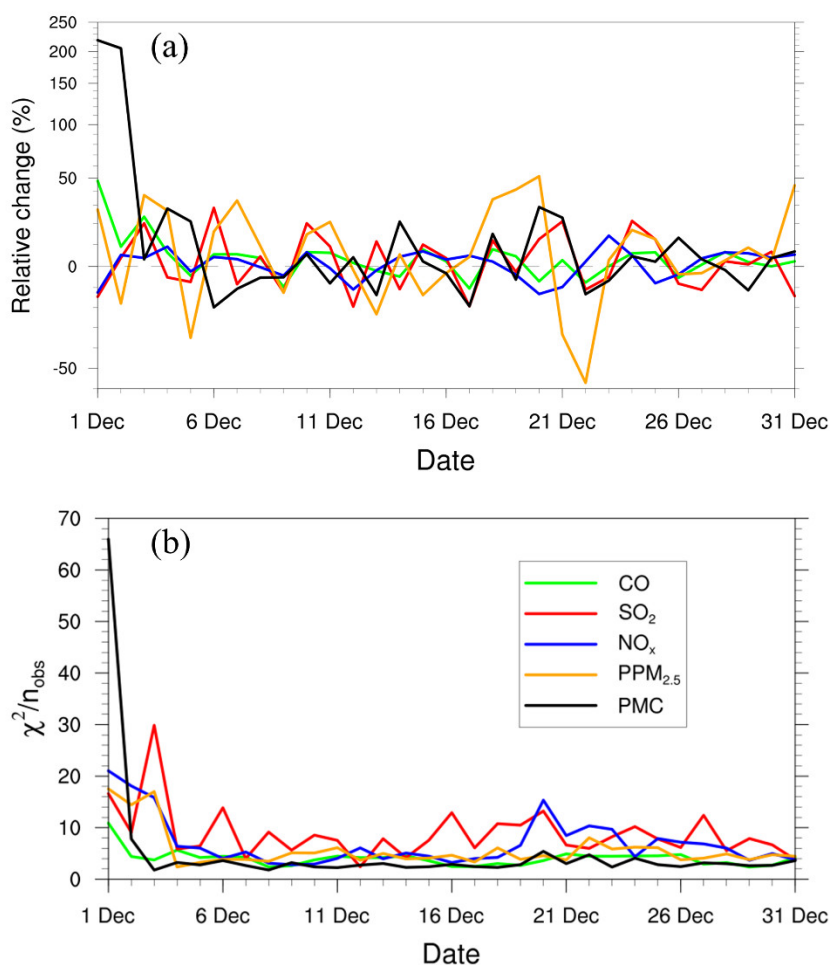
Figure R8 Relative changes (a) in a posteriori emission estimates of CO, SO_2, NO_x, $PPM_{2.5}$ and PMC, and χ^2 statistics (b) of these state vectors in each window. (Figure 10 in the revised manuscript)

Li, J.-d., Deng, Q.-h., Lu, C., and Huang, B.-l.: Chemical compositions and source apportionment of atmospheric PM10 in suburban area of Changsha, China, Journal of Central South University of Technology, 17, 509-515, 2010.

Ma, C., Wang, T., Mizzi, A. P., Anderson, J. L., Zhuang, B., Xie, M., and Wu, R.: Multiconstituent Data Assimilation With WRF-Chem/DART: Potential for Adjusting Anthropogenic Emissions and Improving Air Quality Forecasts Over Eastern China, 124, 7393-7412, 10.1029/2019jd030421, 2019.

Qiu, X., Duan, L., Gao, J., Wang, S., Chai, F., Hu, J., Zhang, J., and Yun, Y.: Chemical composition and source apportionment of PM10 and PM2.5 in different functional areas of Lanzhou, China, Journal of Environmental Sciences, 40, 75-83, 2016

33. L 705 "In addition, without dust may be another reason, since no wind blowing dust scheme was applied in this study as mentioned above." Check English meaning.

Response: Thanks! We have revised that sentence. . See lines 858-860, page 44.

"In addition, the absence of natural dust is another reason, as the wind-blown dust scheme was not applied in this study."

34. L727-728 : Please demonstrate this statement.

Response: Thanks for this suggestion. In the second step of the "two-step" scheme, the optimized emissions are entered again into the CTM model to generate the initial fields of the next DA window. In other words, the unresolved posterior emission error will be fed back to the initial field of the next window through simulation in the current assimilation window of one day, and then fed back to the emission again in the next window for timely optimization. Previous studies (Byun, 2002; de Almeida Albuquerque et al., 2018) have demonstrated that a spin-up time of 1-2 days can provide realistic initial conditions for the CMAQ runs, even if the simulation starts from the clean atmospheric conditions, which shows that the emissions can be well mixed in the atmosphere within the window time of one day in this study. Additionally, with this method, the system can maintain mass conservation (Zhang et al., 2015), thus it can update emissions more consistently and stably.

The sentence (See lines 1048-1053, page 456) has been rephrased as follows:

"As mentioned previously, in the "two-step" scheme, the unresolved posterior emission error will be fed back to the initial field of the next window through sufficient mixed simulation within one day for timely optimization. Meanwhile, the system always maintains the mass balance of pollutants. In this way, the system updates emissions more consistently and stably."

Byun, D. W.: A study of photochemical processes of the Houston-Galveston

metropolitan airshed with EPA CMAQ.

de Almeida Albuquerque, T. T., Andrade, M. d. F., Ynoue, R. Y., Moreira, D. M., Andreao, W. L., dos Santos, F. S., and Sperandio Nascimento, E. G.: WRF-SMOKE-CMAQ modeling system for air quality evaluation in SAo Paulo megacity with a 2008 experimental campaign data, Environmental Science and Pollution Research, 25, 36555-36569, 2018.

Zhang, S., Zheng, X., Chen, J. M., Chen, Z., Dan, B., Yi, X., Wang, L., and Wu, G.: A global carbon assimilation system using a modified ensemble Kalman filter, Geosci. Model Dev., 8, 805-816, 2015.

35. L 746-747: This is well known, but demonstrate that this flaw cannot happen. Please my mathematical means.

Response: Thanks for this comment. During the heavy pollution period, the severe pollution enhances the feedback between the boundary layer and air pollution (Huang et al., 2020). However, WRF-CMAQ is an off-line model, which can not consider the concentration rise caused by the feedback process. Therefore, during the inversion, more emissions will be inferred to compensate this underestimation. This process is a flaw of the off-line atmospheric chemical transport model. In the "two-step" inversion, when such serious feedback leads to overestimation of emissions, the unresolved posterior emission error will be fed back to the initial field. In the next window, the system will quickly adjust the overestimated emissions to compensate for overestimated simulations. The system corrects the emissions by means of error transfer, and thus avoid the impact of error accumulation. The purpose of the "two-step" inversion strategy is not to avoid this flaw, but to make the unsolved emission inversion error be corrected quickly and stably.

Huang, X., Ding, A., Wang, Z., Ding, K., Gao, J., Chai, F., and Fu, C.: Amplified transboundary transport of haze by aerosol-boundary layer interaction in China, Nature Geoscience, 13, 428-434, 2020.

36. Fig 13 emission changes

Response: Thanks for this suggestion. We have changed "emissions changes" to "emission changes". See line 985, page 53.