Response to Reviewer #2's comments:

We thank Referee # 2 for their thoughtful comments and suggestions that have helped to improve this manuscript. Our responses to comments (in bold style) and the corresponding changes to the manuscript are detailed below. In particular, according to the reviewer's suggestions, we have added two more simulations; we also substantially rewrote the texts in both the major context and summary section to emphasize reviewer's questions.

General Comments:

The authors introduce a DA system based on an ensemble square root filter combined with WRF-Chem that assimilates surface observations of PM2.5 across China. The novelty is that they use both aerosol concentrations and emissions in their DA state vector (although it should be noted they did something very similar for CO2 in Peng et al ACP 2015).

The method in this work is very similar to that used by Peters et al. (2007) and Peng et al. (2015) for CO_2 emission inversion, but it is still of novelty for applications in aerosol anthropogenic emissions. In Peters et al. (2007), $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ were all 1. And only natural CO_2 emissions (i.e., biospheric and oceanic emissions) were assimilated at the ecological scale due to the 'signal-to-noise' problem. Thus, the uncertainty of anthropogenic and other CO_2 emissions were ignored. Besides, the framework is more advanced compared to our previous work. In Peng et al. (2015), in order to generate $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$, a set of ensemble forecasts were performed from time t to t+1 to produce the CO_2 concentration fields, forced by the prescribed net CO_2 surface fluxes using the previous assimilated concentration fields as initial conditions. That means that the ensemble forecast were performed twice in that DA system and it was time consuming. However, in order to save computing time, we used the chemical fields $\mathbf{C}_{i,t}^{\mathrm{f}}$ available in the previous assimilation cycle to calculate $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ in this work. Thus, WRF-Chem runs to forecast only once during a DA cycle.

We have added the above paragraph in Lines 187, Page 7 to Lines 200, Page 8.

While the main idea is interesting and the topic is certainly relevant to ACP, I recommend against publication for the following reasons: 1) no independent observations are used to evaluate results. While this is ok for the evaluation of forecasts, this is not good practice for the evaluation of analyses; 2) no proof is offered for the central contention that analyzing emissions *together with* concentrations improves results; 3) no proof is offered for the second central contention that this system improves emissions; 4) many assumptions are merely stated without due reference, deliberation or any kind of sensitivity study; 5) several conclusions are drawn based on irrelevant data (see my comments).

It should be noted that reviewer 1 mentions the first two points as well but is apparently more lenient.

Point 3 I find particularly important as this is a contention made by other authors as well (Tang et al, Miyazaki et al) with little in the form of proof. Models have errors, and analyzing emissions may simply balance out some of these errors without improving the emissions. Note that we do not have observations to evaluate those emissions but this cannot be used as an argument to forego proper scientific reasoning.

In addition I find the structure of the paper illogical, and missed important information on details of their DA system and several references to previous attempts at emission estimation. I hope the authors will continue this work but put more effort in stating their case convincingly, for this research topic is certainly worthwhile. Maybe my comments can be of some help towards improving this manuscript.

Thanks for those comments which did help improving this manuscript. Please see the point-to-point answers as below.

1) We have used the independent observations to evaluate both the analyses and the forecasts. Please see the details in the revised manuscript (Lines 354 to 355, Page 13; Lines 502 to 515, Page 19; Lines 632, Page 23 to Lines 691, Page25).

2) An experiment of pure assimilation chemical ICs and the corresponding 48-h forecasts experiment were also performed for comparisons in the revised manuscript. Please see the details in the revised manuscript (Lines 432 to 434, Page 16; Lines 448 to 452, Page 17; Lines 513 to 533, Page 19; Lines 620 to 622, Page 23; Lines 665 to 704, Page 25).

3) The analyzed emissions are only the results of a mathematical optimum by utilizing observations. They are influenced greatly by the model errors and the observation errors. In addition, only surface $PM_{2.5}$ observations were applied in this work, which may lack abundant constraint on the sources of the secondary aerosol precursors. Moreover, we do not have direct or exact emission information to evaluate the analyzed emissions, which was a challenging to many emission inversion research teams (e.g. Tang et al, 2011; Miyazaki et al., 2012; Ding et al., 2015; Mclinden et al., 2016; etc.). Different from the situations that standard national emission inventories are reported by government as in USA, European or other countries, the rapid economic development and complexity of emission sources in China lead to large uncertainties in the current public available emission inventories. Thus it's impossible for us to conduct the direct evaluation on emissions. For this reason, we weaken our judgment in the text.

Nevertheless, our system considering the emission assimilation provided better simulation results and the improvement of emissions can be verified in terms of two aspects, the diurnal variation and the location of increased emissions. The diurnal variation in the assimilated emissions can be used to verify our judgment to some extent. Especially in the PRD and YRD, $\mathbf{E}^{a}_{PM2.5}$ in the daytime were always larger than those in the night, which agreed well with Olivier et al. (2003), the WRAP (2006) and Wang et al. (2010). In addition, the locations of the larger values for the optimized $E^{a}_{PM2.5}$ in the JJJ region were in good agreement with the places of the crop residues burning traced by the environmental satellite of China. There were 10, 231, 37 and 3 crop residue burning spots in Hebei, Henan, Shandong and Shanxi province respectively from 5 to 11 October 2014 and the numbers are 7, 20, 5 and 21 respectively from 12 to 18 October 2014 (Weekly Crop Residue Burning Monitoring

Report traced by Environmental Satellite, 2015a, 2015b).

We have added the above paragraph in Lines 588, Page 21 to Line 613, Page 22.

4) and 5), we have revised the manuscript according to the reviewer's suggestions.

Abstract

1. P 1, L 13: "The forecast model of emission scaling factors was developed by associating the time smoothing operator with WRF-Chem forecast chemical concentrations". Please rephrase, this sentence is hard to understand without reading the paper first.

This sentence has been rephrased as: "The forecast model of emission scaling factors was developed by using the ensemble concentration ratios of the WRF-Chem forecast chemical concentrations and also the time smoothing operator".

We have rephrased these references in Lines 14 to 16, Page 1.

Introduction

2. P 2, L 40: The authors seem unaware of a lot of previous work on ensemblebased DA: Sekiyama et al ACP 2010, Schutgens et al. ACP 2010a, Schutgens et al ACP 2010b. , Dai et al, *Env. Pol*. 2014, Rubin et al. ACP 2016, , Yumimoto et al GRL 2016. Please include those references.

We have added these references in Lines 46 to 48, Page 2.

3. P 2, L 50: Again, several references seem to be missing i.c. emission estimation. For aerosol: Zhang et al JGR 2005, Sekiyama et al. ACP 2010, Huneeus et al ACP 2012, Schutgens et al. Rem Sens 2012, Huneeus et al ACP 2013

We have added these references in Lines 56 to 58, Page 3.

Methodology

4. **P 3, L 78: Please introduce the ENSRF in context of some other EnKF (EAKF, LEKF, LETKF). What is the reason for this choice of EnKF, what is it main strength/weakness?**

There are different versions of EnKF. The traditional EnKF with perturbed observations (Evensen 1994) introduces sampling errors by perturbing the observations. In contrast to the traditional EnKF, the EnSRF (Whitaker and Hamill, 2002) and the Ensemble Adjustment Kalman Filter (EAKF, developed by Anderson, 2001) obviate the need to perturb the observations. The local ensemble Kalman filtering (LEKF), a kind of EnSRF, was presented by Ott et al. (2002, 2004). It was computationally more efficient compared to the traditional EnKF, since it simultaneously assimilates the observations within a spatially local volume independently. The local Ensemble Transform Kalman Filter (LETKF, Hunt, 2007) integrates the advantages of the Ensemble Transform Kalman Filter (ETKF, developed by Bishop et al., 2001) and the LEKF. The computational cost of LETKF is much lower than that of the original LEKF because the former does not require an orthogonal basis. Though LETKF has more advantages, we still chose the same EnSRF as Schwartz et al. (2014) because we did not need to extend it to analyzing aerosol ICs, very similar to Schwartz et al. (2014).

We have added the above paragraph in Lines 205 to 219, Page 8.

5. P 54 L 94: Change "can be approximated" to "will be approximated". It is by no means certain that this is a good approximation. Part of the evaluation & tuning of an EnKF involves exactly the sampling errors introduced by Eq 5 & 6

We have changed this sentence in Line 235, Page 8.

6. P 3: Since the DA depends on the forecast model's details, I suggest to first discuss the forecast model (and introduce C and λ, and only then the ENSRF)

We have changed the orders of Section 2.1 and 2.2.

7. P 4, L 105: Please provide a bit more information on the base setup of the model: domain size, grid resolution, major aerosol species

We have added more information of the base set up of the model in Lines 101, Page 4 to Lines 114, Page 5.

8. P 4, L 106: "to forecast the emission scaling factors and the aerosol control variables". What are the control variables? I guess the authors mean aerosol concentrations, please change this. Note that both C and λ form the state vector.

We have revised this sentence in Line 88, Page 4.

9. P 5, L 123: "for the lowest eight vertical levels": so the emission inventory included heights at which the emissions were injected? These heights are all within the boundary layer? Why are only the lowest 8 layers considered?

In this work, the lowest 12 vertical levels were at ~ 12 m, 48 m, 98 m, 156 m, 232 m, 300 m, 400 m, 500 m, 600 m, 700 m, 850 m, and 1000 m respectively. So the lowest 12 layers were all within the boundary layer. And the lowest 8 layers were under 500 m.

The emission inventory did not include emission heights at which the emissions were injected, which may cause large uncertainties for model forecast. We prepared the prescribed emissions just following others research (Woo et al., 2003; de meij et al., 2006; Wang et al., 2010): the power generator emissions were interpolated for the lowest eight vertical levels. And other anthropogenic emissions were assigned totally to the 1^{st} level.

Emissions are very small above 500 m for all pollutants. So only the lowest 8 layers are considered.

We have added more discussions about the prescribed emissions in Lines 112 to 114, Page 5; in Lines 117 to 120, Page 5.

10. P 6, L 139: "$\boldsymbol{\kappa}_{i,t}$ are random". I wouldn't call them random. I realize they are distributed around the mean $\overline{\boldsymbol{\kappa}_t}$, , but they were calculated through a short-term forecast of WRF-Chem.

Yes. The ensemble concentration ratio ($\boldsymbol{\kappa}_{i,t}$) are distributed around the ensemble mean ($\overline{\boldsymbol{\kappa}_t}$). And $\overline{\boldsymbol{\kappa}_t} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\kappa}_{i,t} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{C}_{i,t}^{\mathrm{f}}/\overline{\mathbf{C}_t^{\mathrm{f}}} = \frac{1}{N*\overline{\mathbf{C}_t^{\mathrm{f}}}}\sum_{i=1}^{N}\mathbf{C}_{i,t}^{\mathrm{f}} = \overline{\mathbf{C}_t^{\mathrm{f}}}/\overline{\mathbf{C}_t^{\mathrm{f}}} = 1$. So they are actually distributed around 1.

We have removed random variables and changed this sentence as: 'so $\boldsymbol{\kappa}_{i,t}$ are numbers distributed 1 and with ensemble mean values of 1" in Line142, Page 6.

11. P 6, L 144: "$\beta = 1.5$ was chosen in this study": This sounds like an arbitrary choice? Normally β results from tuning a DA but no such exercise was done?

Peters et al (2007) first used the time smooth operator to evaluate the CO_2 fluxed scaling factors: $\boldsymbol{\lambda}_{i,t}^{\mathrm{f}} = (\boldsymbol{\lambda}_{i,t-2}^{\mathrm{a}} + \boldsymbol{\lambda}_{i,t-1}^{\mathrm{a}} + \boldsymbol{\lambda}_{i,t}^{\mathrm{p}})/3$ (P. 8, the last paragraph in Peters et al. 2007. Here, we use the same notation in our manuscript). In that work, $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ were all 1 (P. 11, below S3.3). The time smooth operator was very useful because $\boldsymbol{\lambda}_{i,t}^{\mathrm{f}}$ could gain useful information achieved by previous DA cycle through the using of $\boldsymbol{\lambda}_{i,t-2}^{\mathrm{a}}$ and $\boldsymbol{\lambda}_{i,t-1}^{\mathrm{a}}$. However, they had to assimilate natural CO2 emissions (i.e., biospheric and oceanic) at the ecological scale due to the 'signal-to-noise' problem. Thus, the uncertainty of anthropogenic and other CO2 emissions were ignored.

We used the time smooth operator following Peters et al. (2007). In order to optimize all CO_2 fluxes as a whole at grid scale, we first used the ensemble concentration ratio ($\boldsymbol{\kappa}_{i,t}$) to calculate the ensemble prior emission scaling factors $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ in Peng et al. (2015). $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ were artificial data to generate the ensemble emissions. It was difficult to give the ensemble members of $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ for the ensemble-based emission inversion system. Perhaps it was the simplest way to generate this data at every assimilation cycle by directly using the standard normal distribution function. But the

inversion system failed to optimize the prior fluxes at grid scale due to the 'signal-to-noise' problem (We have done the experiment for CO_2 inversion). From the other aspect, if following Peters et al. (2007) completely, the time smooth operator was applied and $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}} = 1$ was chosen. However, the scaling factors should be perturbed at the first assimilation cycle to generate the ensemble factors. Consequently, this inversion system failed to optimize the prior fluxes at grid scale due to the same 'signal-to-noise' problem (Peng et al., 2015). So other ways should be found to generate $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$. In Peng et al., $\boldsymbol{\kappa}_{i,t}$ was used to calculate $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$, and it seemed effective.

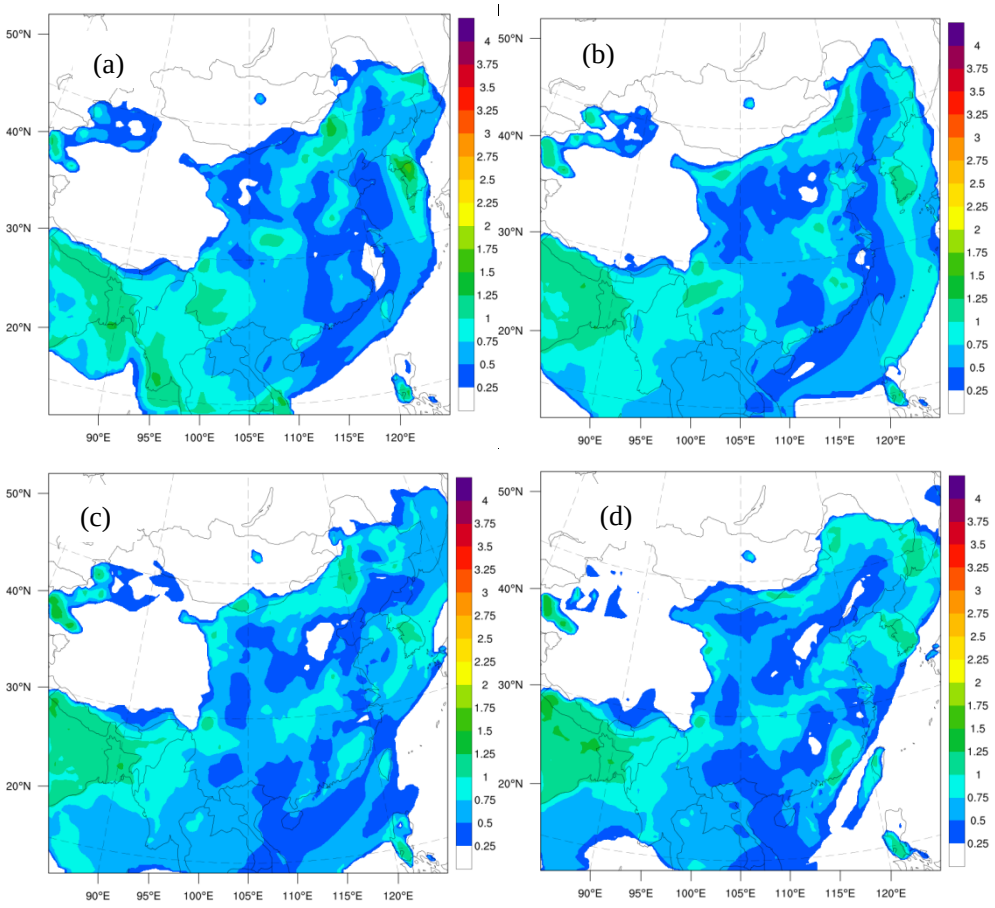
In Peng et al. (2015), the ensemble spread of $\boldsymbol{\kappa}_{i,t}$ was very small (ranging from 0 to 0.08 in most area at model-level 1), though the values of the ensemble spread of $\mathbf{C}_{i,t}^{\mathrm{f}}$ after inflation could reach 1 to 14 ppmv in most area at model-level 1. Therefore, covariance inflation was used to keep it at a certain level. After covariance inflation, the ensemble spread of $\lambda_{\mathrm{i,t}}^{\mathrm{a}}$ ranged from 0.1 to 0.8 in most model area for $\beta = 70$. Besides, several sensitive experiments were performed to investigate β (10, 50, 60, 70, 75, 80, 100). The ensemble spread of $\lambda_{\mathrm{i,t}}^{\mathrm{a}}$ ranged from 0.05 to 1.2 for β =60, 70, 75, 80. And the CO_2 DA system worked comparatively well for β =60, 70, 75, 80. Though CO_2 fluxes inversion was another topic, we mentioned it here because this experience was very helpful for us to develop the joint DA system for aerosol.

As for the $PM_{2.5}$ assimilation, we have done several sensitive experiments to determine the value of β (1.2, 1.5, 1.8, 2, 2.5) by using $PM_{2.5}$ measurements at the five U.S. Embassies stations in China (We did not gain the $PM_{2.5}$ observations from the Ministry of Environmental Protection of China at that time, in August 2015). It showed that the DA system worked comparatively well for $\beta = 1.2$, 1.5 and 1.8. For these cases, the ensemble spread of $\boldsymbol{\lambda}_{\mathrm{PM2.5}}^{\mathrm{f}}$ ranged from 0.1 to 1.25 in most model area. Thus, $\beta = 1.5$ was chosen for latter experiments.

The magnitudes of the ensemble spread of the emission scaling factors were very stable with time. For the joint DA experiment in this manuscript, the ensemble spread

of $\boldsymbol{\lambda}^{\mathrm{f}}_{\mathrm{PM2.5}}$ ranged from 0.25 to 1 in most model area except India where we were not interested in and no observations were available (see details in ReFig. 1). In the manuscript, hourly area-averaged time series of the ensemble spread for $\boldsymbol{\lambda}^{\mathrm{f}}_{\mathrm{PM2.5}}$ over JJJ, YRD, PRD were added in Figure 3d.

It is noted that there were very few negative values for $(\boldsymbol{\kappa}_{i,t})_{\mathrm{inf}}$ after inflation in some cases. A quality control procedure should be performed for $(\boldsymbol{\kappa}_{i,t})_{\mathrm{inf}}$ before further appliance: All these negative data were set as 0.001. Then $(\boldsymbol{\kappa}_{i,t})_{\mathrm{inf}}$ were re-centered to ensure the ensemble mean value of $(\boldsymbol{\kappa}_{i,t})_{\mathrm{inf}}$ were 1. We added this explanation in Lines 146 to 151, Page 6; Lines 489 to 496, Page 18.



ReFig. 1. Spatial distribution of the ensemble spread for $\boldsymbol{\lambda}^{\mathrm{f}}_{\mathrm{PM2.5}}$ at the lowest model level at (a) 0000 UTC 6 October 2014; (b) 0000 UTC 7 October 2014; (c) 0000 UTC 8 October 2014; (d) 0000 UTC 9 October 2014 for $\beta = 1.5$;

12. P 6, L 145: "As the concentrations were closely related to the emissions": if I assume this refers to emissions and concentrations in the same grid-box (given the mathematics of their DA system), this is a bold statement and needs some strong arguments. I can see that during the dust season, Beijing area will be heavily impacted by dust from Eastern China, invalidating your assumptions. Even for pollution emissions, transport may actually be very important.

It is true that transport is very important for aerosol or other air pollution. We corrected the text as "**As the concentrations were closely related to the emissions both locally and in the upwind regions.**"

As stated in Q11 in detail, the prior emission scaling factors $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ were artificial data to generate the ensemble emissions. We chose $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}} = (\boldsymbol{\kappa}_{i,t})_{\mathrm{inf}}$ (4) (original Eq. 9) only as a last resort. Though the concentrations are related to the emissions according to the mass conservation equation, Eq. (4) is not strongly supported. However, same as $(\boldsymbol{\kappa}_{i,t})_{\mathrm{inf}}$, $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ are numbers distributed around 1. From the perspective of generating the ensemble emissions, $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ can play the same role as other data, such as the random numbers created by using the standard normal distribution function. However, there are correlations among the grid-points of $(\boldsymbol{\kappa}_{i,t})_{\mathrm{inf}}$ because $(\boldsymbol{\kappa}_{i,t})_{\mathrm{inf}}$ are calculated through a short-term forecast of WRF-Chem. Thus, $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ have the same correlations as $(\boldsymbol{\kappa}_{i,t})_{\mathrm{inf}}$. While the random numbers are totally different. There are no correlations unless they are generated under certain correlations.

It is noted that the correlations among the grid-points of the prior emissions depend on $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$. Maybe these correlations deviate far from the truth. However, the correlations among the grid-points of the forecast emissions maybe come close to the truth due to the appliance of the smooth operator after multiple iterations.

We have revised the sentence "**As the concentrations were closely related to the emissions both locally and in the upwind regions**" in Lines 152 to 153, Page 6 and added the content of the above paragraph in Lines 157, Page 6 to Lines 165, Page 7.

13. P 6, L 147: "concentration ratios $(\boldsymbol{\kappa}_{i,t})_{\mathbf{inf}}$ served as the prior emission scaling factors $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$" So the concentrations themselves were not inflated, as is usually done in EnKF? What is the justification for this? Shouldn't the scaling factors be perturbed according to the uncertainty in emission inventories and parametrizations?

Posterior multiplicative inflation was applied for only the concentration analysis aiming to maintain ensemble spread.

As for the emission scaling factors, posterior multiplicative inflation was not used. Besides, they are not perturbed according to the uncertainty in emission inventories and parametrizations. Since $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ are calculated through a short-term forecast of WRF-Chem, $\boldsymbol{\lambda}_{i,t}^{\mathrm{f}}$ have deterministic values from the time smooth operator.

We have addressed the posterior multiplicative inflation, plus the covariance localization, in Lines 247 to 256, Page 10.

14. P 6, L 152: I suspect that Eq 10 is missing a factor 0.5. The prior and analysis scale factors are previous times are averaged.

It is right that the prior and analysis scale factors of previous times are averaged, but a factor 0.5 is not missed. In Equation (5) (original Eq. 10), j starts from t-M+1. Thus, M times of scale factors (the prior and M-1 analysis scale factors) are used to calculate $\boldsymbol{\lambda}_{i,t}^{\mathrm{f}}$. For example, in our manuscript, $\mathrm{M} = 4$. Thus, $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$, $\boldsymbol{\lambda}_{i,t-1}^{\mathrm{a}}$, $\boldsymbol{\lambda}_{i,t-2}^{\mathrm{a}}$, and $\boldsymbol{\lambda}_{i,t-3}^{\mathrm{a}}$ are used. Therefore, the denominator in the right hand of Equation (5) is 1/4.

15. P 6, L 153: Again, a rather arbitrary choice (M=4)? How does this relate to the DA cycle?

According to the smooth operator, the ensemble mean values of $\boldsymbol{\lambda}_{i,t}^{\mathrm{f}}$ depend on the ensemble mean of $\boldsymbol{\lambda}_{i,t-M+1}^{\mathrm{a}}$, ⋯, $\boldsymbol{\lambda}_{i,t-2}^{\mathrm{a}}$, $\boldsymbol{\lambda}_{i,t-1}^{\mathrm{a}}$, $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$, where the ensemble means of $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ are all 1. After multiple iterations, the smooth operator can give comparatively

good estimation for $\boldsymbol{\lambda}_{i,t}^{\mathrm{f}}$ since anthropogenic emissions are stable at a certain time scale (Mijling et al., 2012).

Peters et al (2007) chose M=3 ($\boldsymbol{\lambda}_{i,t-2}^{\mathrm{a}}, \boldsymbol{\lambda}_{i,t-1}^{\mathrm{a}}$ and $\boldsymbol{\lambda}_{i,t}^{\mathrm{p}}$ were used to calculate $\boldsymbol{\lambda}_{i,t}^{\mathrm{f}}$) for CO_2 fluxes inversion. They indicated that it was a compromise between prescribing prior CO_2 fluxes at each step and letting the system propagate all information from one step to the next without any guidance (in L 3, P 11). They also pointed out that the latter will work fine for the North American fluxes which were strongly constrained by observations. Similar to Peters et al. (2007), fewer states are used to calculate $\boldsymbol{\lambda}_{i,t}^{\mathrm{f}}$ for the joint DA system for aerosol in this manuscript.

In the revised manuscript, we have added some explanation in Lines 171 to 177, Page 7 and some results in Lines 539 to 541, Page 20.

16. P 6, L 159: "emission inventories". Except in the case of dust, sea-salt etc. Or are these not perturbed? If not, why are they not perturbed (surely they are uncertain as well)? Actually, the authors are rather sparse in their information. Is each species perturbed independently from the others? What is the level of perturbation? Are neighbouring grid-points perturbed independently or do you assume correlations?

In the assimilation part, we had applied 4 independent scaling factors: $\boldsymbol{\lambda}_{\mathrm{PM2.5}}$, $\boldsymbol{\lambda}_{\mathrm{SO2}}$, $\boldsymbol{\lambda}_{\mathrm{NO}}$ and $\boldsymbol{\lambda}_{\mathrm{NH3}}$. Both the forecast emissions (perturbed emissions) and the assimilated emissions were calculated according to EQ (6) : $\mathbf{E}_{i,t} = \boldsymbol{\lambda}_{i,t}\mathbf{E}_{t}^{\mathrm{p}}$ (original Eq. 11). $\boldsymbol{\lambda}_{\mathrm{PM2.5}}$ were used to calculated $\mathbf{E}_{\mathrm{PM2.5i}}$, $\mathbf{E}_{\mathrm{PM2.5j}}$, $\mathbf{E}_{\mathrm{SO4i}}$, $\mathbf{E}_{\mathrm{SO4j}}$, $\mathbf{E}_{\mathrm{NO3i}}$, and $\mathbf{E}_{\mathrm{NO3j}}$ (see details in 2.3.1). $\boldsymbol{\lambda}_{\mathrm{SO2}}$, $\boldsymbol{\lambda}_{\mathrm{NO}}$ and $\boldsymbol{\lambda}_{\mathrm{NH3}}$ were used to calculate $\mathbf{E}_{\mathrm{SO2}}$, $\mathbf{E}_{\mathrm{NO}}$ and $\mathbf{E}_{\mathrm{NH3}}$. In this study, only the species of the emission inventories mentioned above were perturbed (or updated according to the assimilated scaling factors).

Other inorganic species of the anthropogenic emission, such as $\mathbf{E}_{\mathrm{EC}}$ and $\mathbf{E}_{\mathrm{ORG}}$, are not perturbed for WRF-Chem, which is a limitation of this manuscript. However, other anthropogenic emissions, such as $\mathbf{E}_{\mathrm{PM2.5}}$, $\mathbf{E}_{\mathrm{SO4}}$ and $\mathbf{E}_{\mathrm{NO3}}$ are much larger

than $\mathbf{E}_{\mathrm{EC}}$ and $\mathbf{E}_{\mathrm{ORG}}$ in most area of China, and the ensemble spreads of the aerosol concentrations largely depend on the uncertainties of those anthropogenic emissions. Besides, model errors arisen from the meteorology, the emissions and the chemical model itself are compensated to some extent through the use of multiplicative inflation. In other words, the ensemble spread of the concentrations can be kept at a certain level though $\mathbf{E}_{\mathrm{EC}}$ and $\mathbf{E}_{\mathrm{ORG}}$, are not perturbed.

Natural emissions, such as dust and sea salt were not perturbed explicitly when the forecast emissions were generated. However, emissions of dust and sea salt were parameterized within the GOCART model (Chin et al., 2002). Within the DA system, varying meteorology across the members implicitly perturbed dust and sea salt emissions.

We have added the above two paragraphs in Lines 320, Page 12 to Lines 334, Page 13.

No other perturbations are added to the scaling factors. And no other correlations are assumed for the scaling factors. As stated above, both the forecast emissions (perturbed emissions) and the assimilated emissions were calculated according to EQ (6) : $\mathbf{E}_{i,t} = \boldsymbol{\lambda}_{i,t}\mathbf{E}_t^{\mathrm{p}}$ (original Eq. 11). The correlations among the grid-points of the forecast emissions depend on the correlations among the grid-points of $\boldsymbol{\lambda}_{i,t}^{\mathrm{f}}$. See some detail in Q.12 and in Line 182 to 186, Page 7.

17. P 7, L 175: "the state variables of the analysis of the ICs were the 15 WRFChem/ GOCART aerosol variables." This should have been mentioned earlier, maybe line 101.

We have moved this to lines 242 to 244, page 9.

18. P 7, L 184: "(λPM2.5, λSO2, λNO and λNH3)" This line and the following paragraph suggest that the authors keep the EEC and EORG constant? They do not matter? I rather think they do. By the way, this paragraph might be

rewritten to improve readability.

Yes, we keep the $\mathbf{E}_{\mathrm{EC}}$ and $\mathbf{E}_{\mathrm{ORG}}$ constant during the joint DA experiment, which is a limitation in this manuscript. It is true that these emissions are also important for the atmosphere aerosol. The reason we did not assimilate $\mathbf{E}_{\mathrm{EC}}$, $\mathbf{E}_{\mathrm{ORG}}$ is that only the $PM_{2.5}$ measurements are used in this DA experiment. However, the sources of the aerosols (especially organic aerosols) are so complex that our knowledge of their formation mechanisms is far from clear. Though it is technically possible to have all emissions assimilated, with such limited observations adding more control variables would cause much more uncertainties in the system which might lead to unreasonable analysis. This is our first attempt to simultaneously optimize the chemical ICs and emission input. In future work, when gas-phase observations of SO_2, NO_2 and O_3 are used and more aerosol species observations are available, perhaps more emissions are assimilated, similar to Tang et al. (2011).

We have added the above paragraph in Lines 300 to 308, Page 12.

We have also rewritten this paragraph in Lines 268 to 276, Page 10.

19. P 8, L 208: The authors never explain how the system is started up. Some initial perturbation in concentrations and/or emissions must be assumed.

We have rewritten some part of in Sec. 4.2 in Lines 424 to 431, Page 16.

20. P 9, L 247: "$\varepsilon_r = r\varepsilon_0\sqrt{\Delta x/\mathrm{L}}$," Can the authors provide a reference for this form of the representation error? Why do they choose L=3 km? How can it be that the representation error is a function of the measurement error? These are two independent error sources.

We calculated the representation errors completely following Schwartz et al. (2012), who followed Elbern et al. (2007) and Pagowski et al. (2010). Elbern et al. (2007) developed this scheme firstly based on the research of the European organizations. In Elbern et al. (2007), L= 20, 10, 4, 2, 1 and 3 km for Remote, Rural, Suburban, Urban, Traffic and Unknown station type (P 3758) respectively. We had

added some information of the scheme in Lines 366 to 367, Page 13.

21. P 9, L 252-255: Some statistics on how often this happened would be appreciated.

The numbers of the observations were about 17700. Among them 8 observations were discarded because they were larger than 800 μg m^{-3} and 243 (around 1.5%) were discarded due to the ensemble mean of the first guess departure exceeding 100 μg m^{-3}.

We added this statistics in Lines 373 to 375, Page 13.

22. P 10, L 261: "The horizontal grid spacing was 40.5 km and there were 262 57 vertical levels with the model top at 10 hPa." This sort of information should be in Sect 2.2.1

We have moved this sentence in Sect. 2.1.1.

23. P 1, L 265: "initialization and spin-up procedures" Please briefly state the spinup procedure. For how long was the ensemble run before the first DA happened?

We have done initialization experiments from 0000 UTC 1 October to 2300 UTC 4 October 2014. And we have rewritten the last paragraph in Sect. 4.1 in Lines 413 to 416, Page 15.

24. P11, L 279: "clean oceanic conditions." Does this mean that over land you assumed seasalt aerosol as LBC?

Actually the LBCs for chemistry/aerosol fields were idealized profiles embedded within the WRF/Chem model. It's not only for the clean oceanic conditions. We have corrected the text. The differences between the idealized profile and real boundary conditions may bring some errors for the boundary region but since our focus is centered in the JJJ, YRD and PRD regions that far from the boundary region. The impacts would be negligible.

We have corrected the text in Lines 398 to 399, Page 14.

25. P 11, L 280: "standard Gaussian random noise". Please briefly state what standard deviations you assumed, and how you dealt with negative emissions.

We perturbed the anthropogenic emissions following Schwartz et al. (2012).

For possible negative perturbed emissions, they were set as $\mathrm{E}_{ip}^{*}(\eta,t) = 0.001 * \mathrm{E}_{p}(\eta,t)$. This will increase the prescribed emissions more or less. However, only very few data were negative. So, this influence can be negligible.

It should be noted that the perturbed emissions were only used in the spin-up procedure and expC.

We have rewritten this part in Lines 403 to 412, Page 16.

26. P 13, L 336: "These statistics were calculated against observations over all the analyses" If I understand the authors, the same observations that were assimilated are here used to evaluate the results. This likely explains the high correlations. The authors should make it clear this is not an independent evaluation but merely a sanity check.

We have added the independent observations to evaluate the analysis in Lines 501 to 515, Page 19.

27. P 13, L 356: "These results indicate that DA greatly improved the ICs." This is rather bold as you have not used independent observations to evaluate the ICs. Obviously, if you nudge the model towards observations, the model will do better. Please remove this sentence.

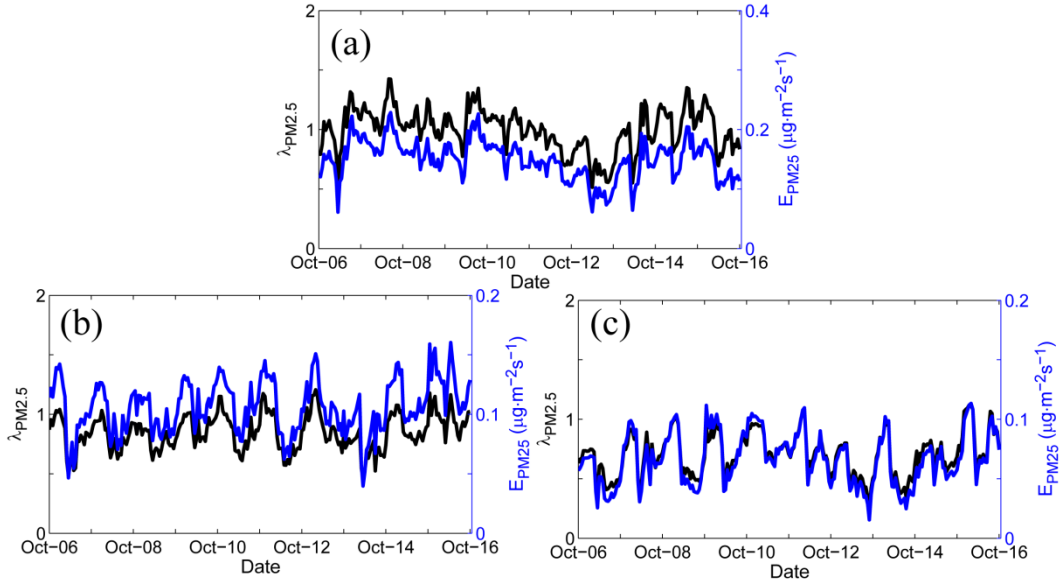
We have used the independent observations to evaluate the analysis. We also removed this sentence.

28. P 13, L 363: "the optimized PM2.5 scaling factor, λPM2.5a, showed an

obvious variation with time, as did the optimized unspeciated primary sources of PM2.5, EPM2.5a" From the authors explanation of how their system works, I do not understand why λPM2.5 and EPM2.5 would have a different (if only slightly) time evolution. Is this because they are regional averages?

Thanks for pointing out this error! The $\boldsymbol{\lambda}^{f}_{PM2.5}$, $\boldsymbol{\lambda}^{f}_{SO2}$, $\boldsymbol{\lambda}^{f}_{NO}$ and $\boldsymbol{\lambda}^{f}_{NH3}$ were 1 hour earlier than the $\mathbf{E}^{f}_{PM2.5}$, $\mathbf{E}^{f}_{SO2}$, $\mathbf{E}^{f}_{NO}$ and $\mathbf{E}^{f}_{NH3}$ in the original plot as I made a mistake when extracting those values.

ReFig. 2 (also updated in the manuscript) shows the right results. It shows that the $\mathbf{E}^{a}_{PM2.5}$ change along with $\boldsymbol{\lambda}^{a}_{PM2.5}$.



ReFig. 2. Hourly area-averaged time series of emission scaling factors (black) extracted from the ensemble mean of the analyzed $\boldsymbol{\lambda}^{a}_{PM2.5}$ and the corresponding analyzed unspeciated primary $PM_{2.5}$ emissions $\mathbf{E}^{a}_{PM2.5}$ (blue) over the three sub-regions: (a) Beijing–Tianjin–Hebei region; (b) Yangtze River delta; and (c) Pearl River delta.

29. P 13, L 379: "as the system is optimized based on ambient concentrations in which the transport and transformation processes are not directly taken into account" But surely transport is important? Maybe a Kalman smoother would have been a better system to solve this problem.

We think transport is as important as transformation. In our DA experiments, the

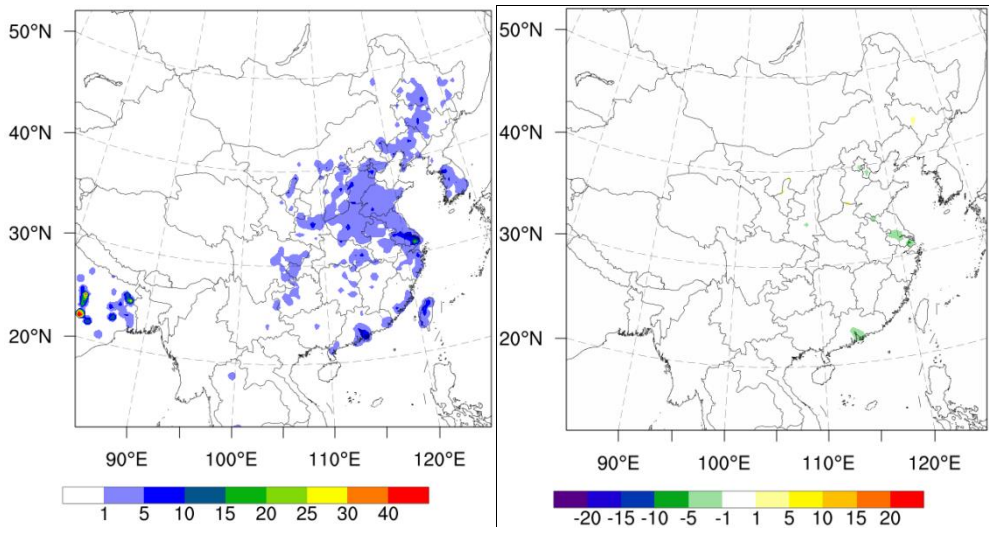
$PM_{2.5}$ measurements network was still spatially sparse and heterogeneous. Almost all measurements were located in the city and no data available in the rural region. However, the crop residues burning always occur in rural region. So the $PM_{2.5}$ measurements network can only capture the burning information a few hours later. It is right that a Kalman smoother would have been a better system to solve this problem.

We have added some explanation in Lines 557 to 565, Page 20.

30. P 14, L 388: "at the lowest model level" Why do you only discuss emissions at lowest level? Are they much larger than those at higher levels? Surely it is the vertically integrated emissions that is important for the amount of particulate matter entering the atmosphere?

Yes, the emissions at lowest level were much larger than those at higher levels. So the time-averaged differences between the ensemble mean analysis and the prior values of the unspeciated primary sources of $PM_{2.5}$ at higher levels were negligible (See ReFig. 3). Thus we only discussed emissions at lowest level.

We have added some explanation in Lines 572 to 574, Page 21.



ReFig. 3 Spatial distribution of (a) the prior unspeciated primary sources of $PM_{2.5}$ ($\mu g\ m^{-2}\ s^{-1}$) and (b) the time-averaged differences between the ensemble mean analysis and the prior values ($\mu g \cdot m^{-2}\ s^{-1}$) of the vertically integrated emissions from level 2 to level 8 averaged

over all hours from 6 to 16 October 2014.

31. P 15, L 406: "Our assimilated PM2.5 and NOx emissions were in good agreement with this trend". The DA experiments reported here cover a period of a few weeks, so how can you compare that to a trend over 15 years?

This conclusion was really arbitrary. We have removed related sentences.

32. P 17, L 470: "However, these results are still better than those obtained with the pure adjustment of ICs that lead to improvements in the first 12-h forecasts (Jiang et al., 2013; Schwartz et al., 2014)." This conclusion is baseless as Jiang et al use a different DA system (3D-VAR) with different observations (PM10) and Schwartz et al use a different domain (USA).

In the revised manuscript, the experiment of pure assimilation chemical ICs and the corresponding 48-h forecasts experiment were also performed for comparison. It seemed that the forecasts with the joint adjustment were always much better than the forecasts with only the optimized ICs for almost all the forecasts in the PRD and YRD. Please see the details in the manuscript (Lines 432 to 434, Page 16; Lines 448 to 452, Page 17; Lines 513 to 533, Page 19; Lines 620 to 622, Page 23; Lines 665 to 704, Page 25).

33. Figure 1: What is F? How is it related to Eq 1?

It was **E**. We have corrected it in Figure 1.