Improving PM2.5 forecast over China by the joint adjustment of initial conditions and source emissions with an ensemble Kalman filter

Zhen Peng, Zhiquan Liu, Dan Chen, Junmei Ban

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).

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 analysing emissions may simply balance out some of these errors without improving the emissions. Note that we do no have observations to evaluate those emissions but this can not 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.

Abstract

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.

Introduction

P 2, L 40: The authors seem unaware of a lot of previous work on ensemble-based 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

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

Methodology

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?

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

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)

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

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.

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?

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

P 6, L 144: "β = 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?

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.

P 6, L 147: “concentration ratios ($\boldsymbol{\kappa}i,t$)inf served as the prior emission scaling factors $\boldsymbol{\lambda}i,t$” 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 parametrisations?

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.

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

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?

P 7, L 175: “the state variables of the analysis of the ICs were the 15 WRF-Chem/GOCART aerosol variables.” This should have been mentioned earlier, maybe line 101.

P 7, L 184: “($\boldsymbol{\lambda}$PM2.5, $\boldsymbol{\lambda}$SO2, $\boldsymbol{\lambda}$NO and $\boldsymbol{\lambda}$NH3)” This line and the following paragraph suggest that the authors keep the $\mathbf{E}$EC and $\mathbf{E}$ORG constant? They do not matter? I rather think they do. By the way, this paragraph might be rewritten to improve readability.

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.

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.

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

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

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

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

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

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.

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.

P 13, L 363: “the optimized PM2.5 scaling factor, $\boldsymbol{\lambda}$PM2.5a, showed an obvious variation with time, as did the optimized unspeciated primary sources of PM2.5, $\mathbf{E}$PM2.5a” From the authors explanation of how their system works, I do not understand why $\boldsymbol{\lambda}$PM2.5 and $\mathbf{E}$PM2.5 would have a different (if only slightly) time evolution. Is this because they are regional averages?

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.

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?

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?

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).

Figure 1: What is $\mathbf{F}$? How is it related to Eq 1?