

Review for "Assimilation of IASI partial tropospheric columns with an Ensemble Kalman Filter over Europe"

Remus Hanea

The presented paper deals with the problem of state estimation in air quality modeling by assimilating satellite data in an Ensemble Kalman filter (EnKF) approach to solve an inverse problem. The ozone concentrations are estimated combining the predictions of the CHIMERE atmospheric model and the measurements coming from the satellite data, both sources of information being uncertain. The EnKF was the algorithm employed to solve the inverse problem. The setup of the data assimilation experiment is well constructed and after a sensitivity analysis was performed, two parameters were appointed as most influential: dimension of the ensemble and the local patch size for the localization.

Results of the experiments are presented using real IASI data and the performance of the whole approach is depicted well throughout the paper. The paper is well written and well explained. The novelty of the approach is clear and well positioned in the context of other work in the same area of research. I believe that the article should be accepted with minor comments.

Consequently, there are few issues that need clarification and those are discussed in the following comments:

1 Introduction

1. Page 3, paragraph 35. The authors are describing the principal objective of the data assimilation in case of weather prediction as the improvement of the initial condition estimation. I think it is good to specify in the same place the different target for data assimilation in atmospheric chemistry : state and parameter estimation.

There is a big difference in the performance of certain methods/algorithms for data assimilation when applied to state estimation or to parameter estimation (especially for the EnKF). There is a general believe in the data assimilation community that for the state estimation the sequential methods (ensemble methods) are giving great results and for the parameter estimation the variational methods are more suitable. BUT, that does not mean that the sequential methods can not be applied successfully to parameter estimation problems. I think it is worthy to be mentioned and noted that the state and parameter estimation are two different problems with different issues to be taken care.

2. Page 3, paragraph 55. "...together with a bias/rmse reduction ...". The abbreviation for rmse was not defined prior in the text.
3. Page 4, paragraph 95. At the end of the Introduction the authors are presenting the whole summary of their papers and the structure of the paper. I do believe that the paper shows clearly the added value of the IASI data in the assimilation process and the impact on the quality of the estimation. Therefore, the authors should be more bold in presenting their apport to the research in hand and replace verbs like "... the aim of the present study was **to examine** ..." with "... was **to show** ..." and drop the "**eventual** gain" word out the text, because the paper shows clearly the gain of assimilating the IASI data. It might send the wrong signal from the beginning of the paper, meaning that the results of the paper can be considered as a trial and not as a certainty.

2 The Ensemble Kalman filter method

The EnKF is explained in Chapter 4, section4.1 presenting the main equations and introducing the issues that usually show when we work with a complex model and use real data in our assimilation processes.

The use of an ensemble with small number of members leads to spurious correlations(nonphysical). Also , the use of a small ensemble size makes the assimilation of large amount of independent data very difficult. This is the case when using satellite data. One well known solution for these kind of problems is to use localization. But, even then there are issues in choosing the right type of localization (distance based, adaptive localization, etc.).

1. Page 11, paragraph 280. "Moreover, an ensemble with a limited number of members becomes unable to estimate the background error Thus "covariance localization" has become a very widely used technique . . . of the background covariance matrix."
The terms background errors and background covariance matrix were not defined prior. The authors discussed about forecast and analyzed covariance matrix when explaining the EnKF. It is a bit confusing taking because the use of the term "background" for errors and covariances comes from the variational approaches and the terms forecast and analyzed are used usually in case of sequential methods.
2. Page 11, paragraph 280 "Thus "covariance localization" has become . . ." and Page 12, paragraph 320 " Another option made here was to apply a local analysis (as mentioned above) in order to avoid spurious correlations ". The terms covariance localization and local analysis are two different ways to tackle localization (see Sakov, P, Bertino, L . 2011 "Relation between two common localization methods for the EnKF"). It is not obvious from the text which of the two methods were applied. Taking into account that the local patch is an important parameter for the assimilation in the presented paper, I think a more detailed presentation of the localization is needed to cover the basic issues, to present the two possible ways of handling localization and to explain the choices made in this paper.

3 Set-up of the assimilation experiments

1. Page 11, paragraph 290. "We allow the model to contain unknown errors and use the information both from the data and the model to improve the actual model state, which contains all the concentrations for all the species in each cell grid"
So there is model error present in the assimilation process. It is clear that is not coming from the unknown parameters. But is not clear what really is the cause for the model error and how that is taking into account in the set-up of the assimilation.
It is due to un-modeled physical phenomenons? or due to errors in the model when describing a certain chemical reaction?
How the model error is taking into account? It is an additive white/colored

noise applied to the forecast equation? I think that a more detailed explanation for the set-up of the assimilation process is needed. EnKF is an extension of the classical Kalman filter approach for linear models. Therefore, the basic state space representation of the system and the observations plays an important role and can give a better understanding of all the elements involved.

$$\begin{aligned} x_{k+1} &= \mathcal{F}(x_k) + w_k \\ y_k &= \mathcal{H}(x_k) + v_k \end{aligned} \tag{1}$$

The definitions for x as the the model state, \mathcal{F} as the CHIMERE atmospheric chemistry model, w_k as the model error, \mathcal{H} as the observation operator and v_k as the measurement error should be given and explained in more details.

Is the observation operator from eq. 5 page 12 a non linear function? If yes, how do we deal with that?

2. As we have already discussed the reason for spurious correlations is a small number of members and that can be solved with localization. What if we increase the dimension of the ensemble to 200? The improvements in the RMSE for the estimated concentrations will improve enough and the spurious correlation will fade away? Is this practical? Even then, the huge amount of data to be assimilated will lead to the use of localization due to the difficulties of the EnKF to handle large sets of data. I think this discussion should be made more clear in the text.

4 Results

The evaluation of the whole data assimilation setup and the improvements of different species estimation is well written and explained and the added value of this exercise is clear. I have only one question regarding the possibility of forecasting with the model after the model state was optimally updated by the EnKF.

There were any tests or experiments made where we move towards an ensemble prediction system and hope that the results obtained with the improved/optimal model (after data assimilation) can and will predict certain

dangerous peaks of ozone concentration or strong smog episodes? Consequently, what will be the quality of these predictions related with the quality of our model?