

Geosci. Model Dev. Discuss., referee comment RC2
<https://doi.org/10.5194/gmd-2022-207-RC2>, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on gmd-2022-207

Anonymous Referee #2

Referee comment on "Implementation and application of ensemble optimal interpolation on an operational chemistry weather model for improving PM_{2.5} and visibility predictions" by Siting Li et al., Geosci. Model Dev. Discuss., <https://doi.org/10.5194/gmd-2022-207-RC2>, 2022

The authors present the implementation of an optimal interpolation framework for PM_{2.5} based on the regional model GRAPES_Meso5.1/CUACE, which computes on-line meteorology as well as chemistry, and PM_{2.5} atmospheric observations from surface sites in China. The aim of the data assimilation scheme implemented here is to provide initial fields of PM_{2.5} concentrations in order to improve the forecasted PM_{2.5} fields and visibility. After sensitivity tests to adjust some parameters of the ensemble optimal interpolation scheme, the impact of the analysed initial PM_{2.5} field on the forecasted PM_{2.5} and visibility for the day are assessed.

General comments

The data assimilation scheme described in this paper is well-known and the interesting points are its adaptation to the application to PM_{2.5} and visibility in a large domain, China, with very contrasted concentrations both in time and in space. Therefore, such a topic is relevant for publication in GMD. Nevertheless, I am not convinced that the work described here actually brings anything new or state-of-the-art regarding data assimilation for air quality forecast.

The introduction to the paper does not give all the required information e.g. definitions of "haze period" (l.37), "the forecast accuracy of air quality forecasting" (l.46), "advantageous" (l.68) are not defined, nor are the terms directly referring to the method such as "spatial length scale of the covariance localization" (l.70); it is not always clear if the references are relevant: they must not be cited as lists of works but each of them must be linked to an aspect of the topic of the paper; conversely, references seem to be missing e.g. "4D-Var requires coding the adjoint model, which is difficult to perform for complex systems": true, but it has been done nevertheless, even for PMs.

In Section 2, the presentation of the EnOI is not very clear regarding which parts are the well-known mathematical method and which are the contributions of this work. For example, l.118 "To avoid all observations affecting the same model grid,...": why should it be an issue? It generally is not, it may be quite the reverse. Other examples include

values which are assigned to parameters without any explanation such as "alpha is taken as 0.9 in this study" (l.115) or "the time step is 100s" (l.164).

In Section 3, the sensitivity experiments are not designed in a very convincing way. The case with only two sites is more a pedagogical illustration than a result or proof. Some possibilities are not explored or even discarded with an explanation: for example, why not adapt the localization length-scale to the representativeness of each station? Why don't we obtain the expected conclusion on the size of the ensemble i.e the larger the ensemble, the better the results of the DA in Section 3.2? The aim of Section 3.3 is not very clear: the first paragraph seems to be dedicated to evaluating the impact of DA against non-assimilated sites but the following ones seem to use all sites for DA (l.261). Therefore, the results showing the impact of DA are trivial: the assimilation is designed to obtain better statistics at the assimilated sites; they just show that there is not an obvious bug in the implementation. The comments on the 40-km length scale and the "discrete" aspect of the posterior field are a bit strange. Obtaining such a result suggests that the spatial resolution of the model (what is it?) is not relevant compared to the representativity of the measurements. This can be linked to the issue of the "overlapping" of the influence areas of sites. If the corrections for each site are independent from the others, then the reasoning in term of "field" (i.e. continuous mapping) is not really valid. The impact on forecasted concentration fields and visibility are not presented in a clear enough way to be convincing. This is linked to the above remarks, also to the ones in the Specific comments, but mainly to the lack of information of what is aimed at for the forecast: is the aim to get better scores about peak detection / visibility loss forecasting? To avoid false positive? False negative? Moreover, the possibilities of a simpler correction of the initial field or improvements to the model are not discussed.

Several important remarks on the way the text is written may explain why it is not convincing to the reader:

- please avoid using "etc": either provide the full information or indicate in the sentence that it's only the most recent/most relevant examples which are given.
- please don't use any phrase such as "it is obvious that" (l.258), "is evident" (l.283). Nothing is obvious or evident that is not trivial in a scientific paper.
- qualitative statements do not bring actual information: what is an "unreasonable result" (e.g. l. 234)? What does "significantly" means here (e.g. l.240, l.258)? What is "light" or "heavy" pollution (l.265-266)?
- please justify why so many significant numbers are provided, in particular for statistical indicators such as MB, ME, RMSE - or stick to a meaningful rounding.

The legends of the Figures are not detailed enough: several do not contain all the information required to understand what is shown without looking up in the text which color is which variable or what an acronym means.

Specific comments

Introduction

- l.37: what is "the haze period"?
- l.39: please detail the "etc": health?
- l.44-45: "a deviation of air quality forecast results from observed comparisons, which can reach 30-50%": which results compared to which measurements can lead to a difference of 30 to 50% (and not -30 to -50%)? Do you compare PM2.5 concentrations? Visibility? Other variables?
- l.46: " the forecast accuracy": what is intended here: to forecast concentrations of PM2.5 and/or visibility or to forecast air quality in terms of peaks and other indicators? The "accuracy" is not the same depending on what is intended.
- l.54: what is in "etc"?
- l.50 to 61: Avoid lists of methods and references if there is no message relevant to the topic of the paper to be derived from them.
- l.62-63: "which is difficult to perform for complex systems": it has nevertheless be done once or twice, please look up the references.
- l.68: "which is an advantageous data assimilation method": advantageous with regards to what? Practical implementation? Computing time and/or power? Why is it not always used (eg what about non-linear problems, such as is the case here)?
- l.69-75: too many terms which are not defined before, not adapted to an introduction.
- l.79: "the CTMs are strongly non-linear systems and the assumption of Gaussian variables and non-biased do not apply": 1) the sentence is not clear, at least one word is missing after "non-biased" and "variables" is not clear (shouldn't it be errors??) 2) it looks like the linearity and the Gaussian assumption are linked, which is not the case.
- l.85-86: "which makes the computation greatly reduced": at least one work is missing: the computation time?
- l.90: "still relatively rare": not very precise, it's been done in the past; also, explain why this is the case: are there any drawbacks? Which are they? Or maybe it's not efficient?
- l.92: same remark as before about the "forecast accuracy" but is it all the more crucial here that the assessment of the results depend on this definition.

Methods and Data

Generally, in this section, the difference between the well-known mathematics of the method and the choices made for this particular study must be made clearer.

- Eq.1: the notation used for H indicate that it is considered linear. But l.79, it is stated that the case studied here is strongly non-linear - which is what is expected with PMs. An explanation of why assuming H to be linear is justified is required.
- please ensure that all notations and acronyms are defined at the first use. Example: AF l.113.
- l.115: "(0,1]": why (? Shouldn't it be])?
- l.115: "alpha is taken as 0.9 in this study": why?
- l.118: the localization scheme is not part of the general method. A explanation of why it is needed, and stating its formulation and link with the previous equations is required, probably in a new subsection.
- generally, in this
- l.119-120: "limit the influence of a single observation by the Kalman update equation[...]": this sentence is not clear for me.
- l.124: "the observations can be reused": for what?
- l.135-137: long list of references but what is the message?
- l.143: "Etc": which processes are included in this etc?
- l.155: which sectors are included in the "etc"?
- l.164: "the time step is 100s": why this choice?
- l.179: "the missing localization": not clear, do you mean "not using a localization

scheme"?

- I.180: "localization was performed in selecting the optimal ensemble size": not clear what the link between localization and the ensemble size may be. Please explain.

Results and discussion

- I.191-205: this two-site case is interesting as a pedagogical illustration but does not seem very relevant for the case study.

- I.206: "the initial fields from 15 to 23 December 2016 were performed": not clear, DA is performed on these fields?

- I.208: "means that more the simulation is closer to the observation": broken English.

- I.210: "[...] of the DA experiment are smaller than those of the CR": it is not clear if the statistical indicators are computed against the validation sites or against assimilated sites. If it is the later, then the results are trivial.

- I.219: what are "spurious increments"? How are they detected?

- I.221: what is "an unreasonable initial field"?

- I.222: concluding on the localization length-scale as the one allowing "the best assimilation" is too strong based on the previous results. What is the "best assimilation"? The one for which initial fields match the validation data? The one which provides forecast concentrations closer to validation data?

- I.227: what is "the r field"?

- I.229-230: "the range of positive CORR at sites A and B gradually increases with the range of CORR greater than 0.7" I don't understand this sentence, please rephrase.

- I.230 : can be considered small": relative to what?

- I.232: "which exaggerates the correlation of each area": what does "exaggerates" mean here? Overestimates? The correlation of each area with what?

- I.234: what are "unreasonable results"?

- I.240: "changed significantly": 1) define "significant" for each statistical indicator here; 2) if the statistics are computed against sites used in the DA, the result is trivial.

- I.240-245: put numbers in a Table.

- I.247-248: the expected result would be that the larger the ensemble, the better the results. Why this is not the case must be explained.

- I.258: "it is obvious that": please avoid this kind of phrase, everything must be demonstrated.

- I.258: what is "significant" for a correction of the initial field?

- I.260: "affects other areas": that is the role of the localization length-scale, isn't it?

- I.265: "BF, AF < AFI": please ensure all abbreviations are defined at the first occurrence.

- I.265-266: please define "light" and "heavy" pollution. Are there official thresholds involved?

- I.270: "after assimilating the BFs": strangely put. Usually, it's more relevant to explain which observations are assimilated.

- I.270-271: "the AFs PM2.5 concentration distribution changes from sheet-like to discrete, which is due to the update of the model data in a length-scale of 40 km range": as stated in the General Comments, this looks like a discrepancy between the use of a large modelled domain (with what horizontal resolution?) and a comparatively short localization length. Please justify fully.

- I.274-276: this sentence is not complete, please re-write.

- I.277-282: please summarize the results in a Table so that the text is easier to read.

Beware also of the number of significant numbers!

- I.282-283: "The results show that the correction effect of DA on the initial fields is evident": 1) nothing is evident or 2) if it is evaluated against assimilated sites, it is trivial.

- I.286-289: the sentence is too long.

- I.289-291: please put the information about which color is what in the legend of the Figures. The same applies below in the section.

- I.291: "relatively consistent in (20)": I don't understand, please rephrase.

- I.293: "gradually overlaps": not very precise or clear. Do you mean the effect of DA is at a very short term (a few hours)?
- I.296: "assimilating the initial field improves the PM2.5 forecast field throughout the assimilation time window": this is not very clear. The assimilation time window is 0 since DA is only performed for the initial field. This field is then used for simulating forecast for a given simulation length but no DA is performed during this period.
- I.297: "strongest": please quantify.
- I.297-303: what is the message?
- I.311-313: "assimilating the initial field [...] a significant impact": all this is not clear for me. Please rephrase.
- I.313-316: this concluding remark is very general and is not specific to the case study. It would be more logical to have it in the method section.
- I.322-323: same remark about the "discrete" DA fields as for the PM2.5 concentrations.
- I.324-329: description of the Figure, not required in the text.
- I.329-330: "It proves that [...]": what is the link with the DA discussed here?
- I.334seq: put elements about Figure 11 in the legend of said Figure.
- I.342: "It is obvious that": no, everything that is worth mentioning must be explained/demonstrated.
- I.346: "the inaccuracy of the humidity simulation here and inaccurate visibility parameterization scheme for the model": wouldn't it be more relevant to improve the model than to perform DA? With a poorly adapted model, the impact of improving initial fields cannot be very large.
- I.348-349: "other objects of assimilation": the priority really seems to be the improvement of the model.

Conclusion

- I.364: "the DA can significantly improve the model initial field": see above 1) define what is significant 2) trivial if not evaluated against validation sites.
- I.369-370: "was most pronounced in the first 12 hours and gradually decreased": very vague statement for the actual result expected from the implementation of DA.
- I.372-373: "efficiency is highest with the largest distance between the model simulation and observation": here again, if the comparison is made against assimilated sites, it is trivial. The poorer the prior compared to the assimilated observations, the more spectacular the shift towards the observations after assimilation - provided there are no bugs in the implementation.
- I.379-380: "but this positive correlation is not particularly obvious": a correlation is so much, it is not obvious or not.
- I.383-384: considering the improvements which could be done in the model and the difficulties of assimilating such data streams as satellite and surface together, this perspective seems very ambitious.

Tables and Figures

- Tables 2 and 3: please check the number of significant numbers.
- Figure 1: "ensemble generate": explain how; "calculate B": this is not shown in the equations; "C-B": what is C? It is not defined in the text when the reader is referred to the Fig; "verify assimilation result": how?
- Figure 4: "The num":? ; also make sure that all abbreviations are defined in the legend (same remark for all Figures)
- Figure 7: what is a "most serious" PM2.5 pollution?

Technical corrections

- l.158: "Fig. 1" -> should be Fig.2