Atmos. Chem. Phys. Discuss., 12, C7891–C7904, 2012
www.atmos-chem-phys-discuss.net/12/C7891/2012/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Estimation of aerosol particle distributions with Kalman Filtering – Part 1: Theory, general aspects and statistical validity" *by* T. Viskari et al.

T. Viskari et al.

toni.viskari@fmi.fi

Received and published: 9 October 2012

Dear anonymous editor,

We would like to begin by thanking for your thoughtful comments. The points you raised are valid and were discussed before submitting the manuscript. Our responses to the questions are below.

1. "The validation compares raw measurements by the DMPS to simulated raw measurements derived from the particle size distributions. However, those particle size distributions were retrieved (by either Kalman filtering or inversion) using those same raw measurements. No independent observations are used for validation. It is not

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

obvious that the algorithms have found the appropriate solution (size distribution) and not one that is unphysical but nevertheless minimizes the differences between real and simulated raw measurements. The same issue must have confronted the developers of the inversion algorithm, so I suppose quite a bit of research has already been done. In any case, the authors should either supply a validation based on independent observations or present strong arguments why their current validation is sufficient."

Validation based on independent observations was not considered feasible because the true state is not known and all other measurements include uncertainties. Therefore the state estimate xEKF was compared directly to the DMPS raw observations y. The residual r=y-HxEKF has to meet two conditions: i) the bias and standard deviation of r has to be in the equal or better than the residual computed from the mathematical inversion, ii) large values of r are either due to measurement noise or special circumstances (e.g., precipitation, change of air mass). Because the true state is not known, this validation is admittedly subjective, but we feel that the set standards were reasonable as to establish the validity of the estimate.

2. "The tests for the two algorithms are not identical as observations are handled differently (in particular in the overlap region of the two DMPS). This begs the question whether any differences in results are due to the algorithms or the observations (I realize that only EKF can handle observations from both DMPS in the overlap). Have the authors conducted experiments where the EKF uses the same observations as the inversion?"

We have not tested a scenario where the observations are handled similarly to the observations. One of the key benefits of the EKF, and data assimilation in general, is that it is able to use information from multiple observations. In addition our primary motivation in the consideration was to establish that EKF is within the statistical limits considered acceptable for inversion methods. We did not consider there to be much added information from experimenting with a situation where the EKF uses the same observations as the inversion, as it would require us to hinder EKF.

3. “Related to both previous points: why are the differences between raw measurement and simulated measurement not zero for the inversion? I guess this is because the raw measurements in the overlap are averaged before inversion, but the validation is done with the original raw measurements?”

The overlap area is one of the reasons why the differences between the raw measurements and the inverse solution are not zero. Another is that the inversion solution here is calculated with a least-square non-negative pseudo-inverse method. As the method does mathematically determine a size distribution, it will not always be the same as the exact measurements in a particular size bin. There were, though, instances when the inverse solution momentarily diverged from the observations more than expected. In those cases it was determined that it was not due to the operator used for the comparison and thus, while the reasons for this difference were interested, it was decided not to focus on them as the work was about EKF, not numerical inversion method.

“Finally, it remains rather vague why the authors choose to use Kalman filtering instead of e.g. regularization techniques which seem more suited to their problem. Especially as the authors themselves point out that the time-information from the forward model calculations is detrimental to the solution when sudden changes occur (Sect 5.2.1). The main advantage that the Kalman filter has over the inversion is that it can handle observations from both instruments in the overlap, while the inversion requires averaging of those observations. The authors may want to stress this point.”

You are correct about our need to stress that point. Also, this work was meant to be an introductory to EKF and see how it functions with features common in aerosol physics. Thus by establishing it as a valid method, we could expand the method to different observations instruments in the next part. We will also stress this more. I would however point out that while the forward model calculations are detrimental when sudden changes occur, they are an advantage in most cases as they enforce a size distribution that is continuous over both time and particle diameter.

ACPD

12, C7891–C7904, 2012

Interactive Comment

“I think the Kalman filter (or other data assimilation tools) may become very useful in interpreting DMPS raw measurements in a configuration where model parameters are estimated. I’d like to hear the authors’ opinion on the feasibility of such an approach.”

It is a feasible approach and is indeed something we are currently considering as a possible next stage of the research.

Then to the specific comments.

“p. 18857, time evolution updating and observation updating might be replaced by the more common terms forecast and analysis. Observation updating is a confusing term anyway as no observations are updated.”

We follow the terminology set in Kaipio and Somersalo (2004). Thus we will, at least for now, continue using time evolution and observation updating.

“p. 18857, eq 1, I suggest you use M(x), instead of Mx to stress non-linearity. Also, why do you not include an error term here as you do in eq 2, that makes the text more logical.”

Good suggestion concerning M(x) and it has been changed in the manuscript. We do not include the error term in eq. 1, as eq. 2 represents the uncertainty in eq. 1. Thus it shouldn’t be included anymore in eq. 1.

“p. 18857, line 15: Q represents more unaccounted physics and chemistry and transport etc. it really represents structural model errors. Calling it system noise is not recommended because 1) your model may be biased (hence necessity of error term in eq 1); it gives the suggestion of numerical noise or some other implementation related issue. Also mention that Q cannot be specified. Finally, Bk is an approximation because the linear-tangent model is used.”

You are correct about the definition of Q and it has been corrected in the manuscript. We explained the problem with Q in section 4.1, basically that it cannot be currently specified, and as in section 2 we focus more on the theory of the method, we chose

ACPD

12, C7891–C7904, 2012

Interactive Comment

not to add it here. I added the Bk being an approximation to the manuscript.

“p. 18857 line 20: "Here H is a possibly non-linear observation operator, which produces the observation counterpart corresponding to the prior state". replace "observation counterpart" by simulated observation"”

Replaced.

“p. 18857 eq 3: H(x) instead of Hx”

Done.

“Section 2. The way I understand EKF, it is just KF but you use a non-linear model. Nothing really changes from KF, except the full model is used for forecast and the tangent linear for analysis. Please point this out explicitly. This also has consequences because the Kalman filter is only exact and optimal for a linear model. Please point this out as well.”

This has now been pointed out in the manuscript.

“section 3.1 it would be good if the authors detail measurement errors in yr (!, including temporal representation errors, see last paragraph of 3.1), dominant error sources in inversion, final error estimates for x and validation by independent observations. I’m guessing this has already been done by other groups so a few lines and some references will suffice.”

The individual articles referred to in the article discuss these matters. Work relating to the reliability of the inversion method has been done by other groups. A very detailed analysis of the performance and uncertainties of size distribution was presented by Wiedensohler et al. (2011). They compared several DMPS instruments side-by-side and found that under controlled laboratory conditions, the particle number size distributions from 20 to 200 nm determined by mobility particle size spectrometers of different design are within an uncertainty range of around $\pm10\%$ after correcting internal particle losses, while below and above this size range the discrepancies increased. For parti-

ACPD
12, C7891–C7904, 2012

Interactive
Comment

cles larger than 200 nm, the uncertainty range increased to 30 %, which they could not explain. As part of that work they also evaluated the contribution of the uncertainties of the inversion to the overall uncertainty. The same set of raw data, i.e., electrical mobility distributions was inverted by using various inversion codes. All agreed to within a few percent so they concluded that the main source of the uncertainties is not attributable to the inversion methods.

“p. 18859 line 25: is this transfer function the same as R? If not, what is their relation? "Transfer functions for both DMPSs are integrated separately for this diameter grid": why? Do you mean that within each size range, the transfer functions are integrated?

The transfer function is the same as R. And in the inversion algorihm we have, the transfer functions are integrated separately for both DMPSs. As far as we understand, it is mainly due to the DMPS I and II channels differing in width.

“section 3.2 How is the model initialized? What additional information is needed to run the model? For which time and spatial scales is it considered valid? Has the model been evaluated against observations (> results?)?””

We rewrote the manuscript so that it now better refers to the article in which the UHMA model was originally introduced in and in which those questions are mostly answered.

“For the EKF, did the authors develop a tangent linear version of the model, or are model results linearized ’on the fly’?”

Thank you for pointing out that we forgot to include this part. A tangent-linear version of the UHMA model has been previously been developed. A reference has been added to the manuscript for the article in which the tangent-linear model was introduced.

“section 3.3 p. 18861 line 20: "The observation operator was tested and validated by comparing raw observations to the values computed by H from a size distribution obtained with a mathematical inversion." Is this not a circular argument? The raw observations have to be inverted (RËĘ-1) and interpolated to the model grid (PËĘ-1)

ACPD
12, C7891–C7904, 2012

Interactive Comment

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

before you can use eq 7."

We don't consider the test to be a circular argument, as the inverse solution is solved with a different method. Instead we sought to establish that we could create a valid observation operator. As we did not alter R, this effectively means that we tested that interpolation scheme is good enough.

"sect 4.1 p. 18862 line 5: The authors state that the linear model is valid on time-scales of 30 min only, so carrying B forward repeatedly with the linear model should cause large errors and will be detrimental to the assimilation. Do they somehow condition the posterior B on the posterior x at each timestep?"

Good point on us being slightly unclear here. Both the forward operator and the tangent-linear model propagate the state by only 10 minutes, as that is the time between observations. After each 10 minute interval the uncertainty is determined and the model simulations are in essence re-initialized. Thus the tangent-linear hypothesis can be considered valid for the measurement interval. We have clarified this in the article.

"p. 18862 line 20: "As a consequence, the state error covariance may become gradually smaller and smaller, as was the case here". I know this effect exist in purely linear KF as well. In ensemble KF, inflation techniques are used to combat it, much like this paper does. Please mention this. To what extent is this problem amplified by the linear model for B?"

Based on our current understanding, the problem isn't greatly amplified by the linear model for B. Instead it appears to be much more due to the lack of understanding concerning the model errors and parameterizations. Also we have read on the inflation techniques used by the ensemble methods and it is something we are considering as a possibility to deal with the filter divergence. However, we did not want to include references to those works, because their inflation methods are much more sophisticated than ours and we did not want to give an erroneous image that we already used similar

ACPD
12, C7891–C7904, 2012

Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper

methods.

We added a line about Filter divergence being a common problem and a reference to Whitaker and Hamill (2002) concerning Filter divergence.

“p. 18863 line 20: "but representativeness and observation operator errors need statistical material, i.e. innovation sequences produced with EKF". I do not see how one can reliably determine those errors from data assimilation experiments. After all, various error sources all contribute to the statistics of the innovation. Either argue your case or remove this line.”

You are correct that the all the error sources contribute to the statistics of the innovation. However, there has been research on how to estimate unknown error covariances from those innovations. For instance Dee has published on the subject and we included a reference to his Dee and Da Silva (1999) to the article.

“p. 18863 line 25: The authors also assume that observation errors are uncorrelated across sizes. This seems unrealistic due to the size categorization use in DMPS. Is this assumption purely practical or is it also approximately valid (maybe other error sources are more important)? Please discuss.”

In our understanding you are correct, if you refer to the error correlations due to the transfer kernel R. Both particle charging probabilities and channel transfer functions do create correlated uncertainties. As R is a part of the observation operator for DMPS and observation operator uncertainties are included in observations errors, there should be error covariances also in the observation error.

The reason they are ignored is a mixture of practical reasons and valid approximations. It does ease matters to only consider background error covariances, but it is also our understanding for DMPS measurements, especially in Hyytiälä conditions, the observation error covariances should generally be so small that we can approximate the observation error to be uncorrelated. We added a few lines to the article to explain

ACPD
12, C7891–C7904, 2012

this.

“p. 18864 line 5: Do you actually mean that the errors in the smallest and largest size bins are correlated? Or do you mean that the errors in the smallest size bins are correlated and that the errors in the largest size bins are correlated? I guess you mean the latter. See also my comment before on observational error correlation.”

We actually meant the former, although not the largest size bins, but rather the size bins where the total surface is the largest. This due to the condensation and coagulation processes dynamically linking the uncertainties in those particle sizes together. The observation error correlation should be near zero in that situation.

“p. 18865 line 10: Again, how do you initialize the model? Why are model errors chosen to be 30%?”

The model was initialized with a particle size distribution interpolated from the inverse solution determined from measurements for 00:00 local winter time. The initial model errors were intentionally set as large so that the observations would have a large impact on the estimate at the very beginning. We tried to make this clearer in the article.

“p. 18865 line 15: EKF results are smoother because solutions are constrained by the time evolution of the model. Had the inversion used any kind of regularization, I guess it would also look more smooth?”

It is likely so.

“p. 18865 line 20: "The differences in the total number concentrations are partially due to the diameters for xEKF and xINV not being the same, which makes it difficult to limit xEKF to the same diameter range than xINV" Sorry, but I don’t understand what is meant here.”

What we meant was that the state estimate, xEKF, contains values from smaller particle sizes that the inverse solution, xINV. In order to compare the total number concentrations between xEKF and xINV, we need to limit xEKF to the same particle sizes than

ACPD

12, C7891–C7904, 2012

Interactive Comment

xINV. However, as xEKF and xINV have different diameters, which makes it really difficult to set the same lowest diameter xEKF and xINV. As the number concentrations are large in those particle sizes, this causes differences in the total number concentrations.

We tried to make this clearer in the article.

“p. 18866 line 5: I think this even suggests that there are (unsurprisingly) biases in the observations. If the observations for DMSP I and II were unbiased with properly assigned errors, results in the overlap should be better than outside (because one uses more observations), shouldn’t it (unless error magnitudes are much larger in the overlap)?”

You are correct in your assertion. It should be noted the relative measurement errors are usually larger at the edges of the instrument measurement areas, which in this case would include the overlapping area for both instruments. This, however, is more visible as an increased standard deviation in the overlapping measurement range.

“p. 18866 line 10: I am surprised that the inversion allows larger biases. You "validate" your results against the observations that you originally inverted. It seems the inversion allows less freedom in this (hence has larger errors) than EKF. But it is actually EKF that adds a lot of model information to the estimate of x. Please discuss. One thing that seems different is how you use the data in the overlap. For EKF you use all measurements by both instruments, but for the inversion, you use averaged data in the overlap, I believe.”

While technically EKF adds a lot of model information, that information is based on previous observations and that the model dynamics do not cause major changes in particle number concentrations in those particle diameters over a 10 minute time period. Thus the additional information from the background state is unlikely to increase the bias. Also it is important to remember that inverse solution is mathematically fitted from the observations, so errors are expected there.

ACPD

12, C7891–C7904, 2012

Interactive Comment

“Results in section 5 are not a proper validation as you do not use independent observations. Since the EKF introduces model information, it may actually yield worse results than the inversion when tested against independent observations.”

As mentioned in the beginning of this response, there are really no independent observations to compare to. We did, however, include the reasoning for our validation in to the manuscript. Also, we would again stress that the model information is primarily information from observations of prior states propagated over a short time period.

“p. 18867 line 10: so to what extent are the observations that you use for EKF and the inversion different? This should be specified in detail because it seems to have a significant impact. Why did the authors choose to use different observations? As a result it becomes impossible to separate data sampling effects from algorithm effects.”

EKF and the inverse solution method use the exactly same observations. The only difference is that for the inverse solution the observations from DMPS I and II are averaged in the overlapping measurement range. This, however, is how the inverse algorithm handles the overlapping observations and thus it is an algorithm effect.

“For a proper assimilation system, remaining biases should be smaller than the remaining standard deviations. It would appear that on this score, EKF fares worse than the inversion as it has smaller biases and smaller stddev. What is the authors’ take on this? Also, are stddev from raw measurements vs EKF similar to posterior stddevs? This is another important test for EKF. See also your point vi in Section 6.”

We are afraid that we do not understand your point here. The biases for EKF are generally smaller than the corresponding biases for inversion in this article, as can be seen in the figures. The few exceptions are in cases where the biases for EKF and inversion are approximately the same. The standard deviations for EKF are mostly larger than for the inversion. This was determined to be largely due to the measurement noise. The effects of time-dependent biases also increase the standard deviations, but that is actually a problem with the statistical calculation rather than the state estimates.

ACPD
12, C7891–C7904, 2012

These have all been discussed in the article.

The standard deviations shown here are for the posterior state, if we understand your question correctly. We also calculated the priori state standard deviations, which were naturally slightly larger than the posteriori state standard deviations.

“Section 5.2.1: The (incorrect) temporal information from the model when air masses or so change make me question the appropriateness of EKF for this problem. Wouldn’t an inversion with regularization be a better choice of algorithm? E.g. Philips-Thikonov regularization using assumed smoothness of the size distribution.”

The primary motivation in our article is to introduce a method that is able combine information from multiple simultaneous observations. Other benefits of the method are the decreased amount of observation noise and constrained physical continuity. As for the comparison with the regularized size distribution, we do not know. The focus of the article is not the inversion method.

The state estimate does appear to adjust to the large changes in the size distribution relatively fast, especially if the change is observed by several instruments. Additionally, this is something that can be improved in the future by either using the boundary information, e.g. temperature, moisture and wind, to indicate when the air mass changes or by properly statistically determining when the aerosol size distribution is too different between adjacent times. In both cases the state estimate can then be adjusted by increasing the weight of the observations.

We added in sections 6 to point I) a line about better including changes in air mass in the state estimate in the future.

“Section 6: I do not disagree with the points the authors make. They are, in a sense, general comments valid for a properly functioning KF. I feel the authors have not proven this is a properly working EKF as there are no independent observations to verify results. This impacts their statements i) and iii). In this particular paper, it seems the

ACPD
12, C7891–C7904, 2012

main advantage EKF holds over inversion is ii) better handling of multi-instrument retrievals. On the other hand v) and vi) seem rather unconvincing: due to the nature of a box-model it will be impossible to properly account for changing air masses etc.”

As we have already pointed out several times, we do not consider verification with independent results feasible in this case. We have also discussed the reasoning for our method of verification. We do not however understand why v) and vi) seem unconvincing, as they are important features in a Kalman Filter. And as we explained earlier, changing air masses can be accounted in box-models by either better incorporating information from the boundary conditions or by statistically determining when the aerosol particle size distribution experiences large, sudden changes.

“p 18861 line 15: "The interpolation matrix P is resolved" The authors mean: "The actual interpolation was performed". I have no idea what resolving a matrix implies, especially as there is no explicit equation.”

Changed.

“p 18862 line 15: "source term" Please use "error term" as source term often refers to emissions.”

Changed.

References:

Dee, D.P and Da Silva, A.: Maximum-Likelihood Estimation of Forecast and Observation Error Covariance Parameters. Part I: Methodology. Mon. Weather. Rev., 127, 1822-1834, 1999

Kaipio, J. and Somersalo, E.: Statistical and Computational Inverse Problems, Applied Mathematical Sciences, 160, Springer-Verlag. 339 pp. ISBN 0-387-22073-9, 2004.

Whitaker, J.S. and Hamill, T.H.: Ensemble data assimilation without perturbed observations. Mon. Weather Rev., 130, 1913-1924, 2002

ACPD

12, C7891–C7904, 2012

Interactive Comment

Wiedensohler, A., Birmili, W., Nowak., A., Sonntag, A., Weinhold, K., Merkel, M., Wehner, B., Tuch, T., Pfeifer, S., Fiebig, M., Fjäraa, A. M., Asmi, E., Sellegri, K., Venzac, H., Villani, P., Laj, P., Aalto, P., Ogren, J. A., Swietlicki, E., Roldin, P., Williams, P., Quincey, P., Hüglin, C., Schmidhauser, R., Gysel, M., Weingartner, E., Riccobono, F., Santos, S., Grüning, S., Fallon, K., Beddows, D., Monaha, C., Marioni, A., Horn, H.-R., Keck, L., Jiang, J., Scheckman, J., McMurry, P. H., Deng, Z. and Zhao, C. S.: Mobility particle size spectrometers: harmonization of technical standards and data structure to facilitate high quality long-term observations of atmospheric particle number size distributions, Atmos. Meas. Tech., 12, 657-685, 2012

Interactive comment on Atmos. Chem. Phys. Discuss., 12, 18853, 2012.

ACPD

12, C7891–C7904, 2012

Interactive Comment

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper