Assimilation of ground based lidar and ceilometer observations of aerosols from the European E-Profile network into ECMWF's Integrated Forecasting System (IFS-COMPO, CY49R1)

Michael Kahnert, Melanie Ades, Mickaël Bacles, Johannes Flemming, Vincent Guidard, Alexander Haefele, Robin J. Hogan, Samuel Rémy, and Eric Sauvageat

Response to reviewer comments 1 by Jeff Reid

(Reviewer comment are given in blue, our response is rendered in black.)

1. This paper presents numerical experiments on the assimilation of E-Profile data into CAMS. E-Profile is a well-populated network of ceilometers with what appears to be a robust and well-thought-out data distribution system, making it eligible for the Benedetti et al. (2018) critieria of being "fast, accessible, and well-documented" for use in operational aerosol assimilation. And the good news is that, if taken from a purely statistical point of view, the authors show that the assimilation of E-Profile data is beneficial to the analysis of other variables such as PM2.5. This is in itself important: assimilating one set of parameters (here ceilometer-derived attenuated backscatter) helps the analysis of a fully independent set—in this case, PM2.5. If this is the exit criteria, done. The authors and study participants have done their job and have a result that is quite worthy of publishing in GMT.

This is, indeed, the main criterium, since the model setup we investigate is the global component of the CAMS operational system, and the primary goal is to improve the operational analysis and forecast.

2. But (and in data assimilation there is always a "but"), while the authors show a positive impact for region-wide error statistics, it is unclear to the reader what is going on—and in particular, where the impact is relative to significant events. I very much appreciated the appendices, but even here specifics seem to be unclear. I can understand any possible reticence of the authors to add more details, as data assimilation investigations easily lead to endless rabbit warrens that can go in unexpected directions (think Watership Down). But I think there are many things that can be made clear through more direct language, and also some additional discussion, in particular of their example cases. I don't expect the authors to perform any major reruns of their

analysis, but what I suggest can take a little bit of time without being particularly onerous or lengthy.

The very first (pre-submission) version of our manuscript did not have an appendix. The entire analysis of all 20 experiments was contained in the main body of the paper. However, there was concern among the co-authors that only a small number of readers would be valiant enough to plough through such a large number of details, which would have severely limited the impact of the paper. Therefore, we decided to move all the details into an appendix and only retain the results for the optimum system setup in the main body of the paper. This way even more casual readers will easily get the main message of the paper and not find it buried in details. We are aware that this compromise is less ideal for the most enthusiastic readers, such as the two reviewers, who appreciate the large amount of information that can be obtained from such a comprehensive number of experiments. But for the benefit of the majority of readers and the potential impact of the paper, we are, indeed, reluctant to significantly expand the analysis or revert to the original structuring of the paper without an appendix.

Having said this, we do find many comments by Jeff Reid highly insightful and relevant, and we will make an effort to incorporate most of them, while keeping the size of the paper manageable. Also, we appreaciate the suggestion to accomplish much of this by a more effective use of language.

3. The problem at hand is that data assimilation is not unlike a half-filled water balloon. You constrain it somewhere, and based on the workings of the model, things can pop out in other places. For CAMS, as the authors point out, the baseline is AOD assimilation. E-Profile data modified this baseline through the instruments' and the model's attenuated backscatter, which is an underdetermined parameter. There are lots of ways to distribute an increment in attenuated backscatter in the model, and yet there is no discussion whatsoever of the 4D-Var adjoint. Based on my previous conversation with Angela Benedetti, the adjoint for aerosol is quite simple (as it should be). Nevertheless, for the case studies, I would strongly recommend adding direct details of what is changing in the model. I am guessing the aerosol speciation fraction is not changing. If so, it would help if it was stated clearly. If it is, how so? Details of how the AOT error changes would also be helpful. There are many AERONET sites in Europe, and I found some nearby for the two test cases shown.

Yes, the aerosol speciation is not changing; the control variable is the total mass mixing ratio. This *was* clearly stated in the discussion paper in lines 67-69. We can try to reformulate this sentence in a way that makes it even clearer. It is a good idea to show a few AERONET sites in the revised paper.

4. It is pretty trivial to make your model look like your observation, but the real question is to what extent you break something else in the process. Returning the the half filled water balloon analogy, they have optimized for pan-European bias and RMSE. But we don't know what is happening at sites and more significant events. We then have a long throw from two incomplete case studies to domain-averaged bias and RMSE. This is equivalent to a forecast contest when you guess the next day's high temperature ahead of a cold front. Get the timing of the front wrong and you have lost. But statistically, if you pick the mean temperature climatologically, you have a good shot at winning—not in spirit, mind you, but by the metrics at hand. So I, as a reader, would appreciate a few more statistics—at the very least for picking more significant versus background events, event skill scores, and/or a PDF of sites that see improvement versus sites that may be diminished.

   A good way to accommodate this comment without inflating the paper too much will be to show and discuss a few AERONET sites for the two events. We can also add more map plots showing not only Bias and RMSE but also temporal correlations at different sites. To really do statistics on event skill scores etc will possibly require a longer time series of monitored data encompassing a sufficiently large number of events. We are presently working on ingesting E-profile observations through the standard BUFR pathway (instead of the pathway of experimental observations), and we will routinely monitor the data as part of the CAMS operational suite. This will, eventually, allow us to assess the event skill over an extended period of time.

5. Some of what is going on is somewhat obfuscated in the appendices. There is a great deal of discussion on "updates to the optical model," but in the context of this work, all they need to say is, "We essentially changed the lidar ratio for the near-infrared from X to Y." Such a change could easily result in a big change/improvement in attenuated backscatter and vertical distribution without changing AOD. But it would help if they showed that. Likewise, adding a VIIRS image and an AOT plot would be greatly beneficial. Maybe even time-height cross-sections for Figures 2 and 3 would make the authors' case more effectively.

   The second reviewer asked us to explicitly show the lidar ratios we used, not just the change in lidar ratio due to using spheroids instead of spheres. Doing this will help to clarify the effect of the optics model. Also, we agree with the suggestion of showing time-height cross sections for each of the two events. In fact, one of our main concerns before starting this study was that assimilating data that are geographically highly limited into a global model may produce artifacts at the boundaries. The time-height profiles help to disperse such concerns. We will add and discuss these extra images in the revised manuscript.

6. Anyways, I did enjoy reading the paper, and it made me think. These comments are offered out of admiration for the work. I leave it to the authors to decide how far they want to take this.

   Thanks, we are very grateful for this encouraging review.

7. Specific comments.

   Ln 6: AERONET, not Aeronet.

   Agreed.

8. Figure 2&3: Figure text is way way too small. Just a suggestion, can you please add as visible dots the model midpoints to the profiles. Also, maybe add below a timecodes of the baseline ceilometer data? Lastly, please revise before the typesetters ask you to. Also, please add the change in AOD for these cases.

   Yes, the text needs to be enlarged. We will probably restructure the figures with the requested addition of vertical curtains and AERONET stations. We are not sure what the reviewer means by “timecodes of the baseline ceilometer data”.

9. Section 2.1: I don’t mean to extend the paper, but I would appreciate a few more details on the assimilation process up front in Section 2, instead of having to rummage through the supplemental materials. The authors note, on one hand, that the model is at full resolution, but the minimization is at a lower resolution. I would appreciate the authors being a bit more specific on these points in the main body.

   We need to reconcile this with some of the comments made by the second reviewer, which seem to point in the opposite direction. The different resolutions are defined in terms of the spectral truncation. The second reviewer commented, when we mention the spectral truncation T511 of the model, that he did not know what this means. Indeed, this paper will, hopefully, be interesting to a broad community of experimentalists and data users, and not be limited to modellers and data assimilation experts. Thus, we agree with the second reviewer’s criticism of our improvident use of technical terms. Listing the resolutions in different parts of the minimisation process would not be much of an effort, but it would require us to take a lengthy detour into explaining different model grids and spectral representation of fields, none of which is part of the new developments presented in this paper. Therefore, we refer those readers who are interested in a complete technical description to the relevant chapters of the IFS documentation, which are cited in the paper.

10. Where this is going is that their assimilation experiment is using high-density ceilometers, but what is the nature of the data assimilation resolution versus, say, meso-alpha features that may be important in Europe, where terrain variability can be significant? Again, I am not asking for a full analysis, just a little context for the experiment and what it captures and what it doesn't. For example, the use of wavelets seems to be important, but the main body and the referenced appendices are quite thin on this method.

    Concerning the use of high-density ceilometers: We agree that the use of dense stationary observations in a global model with a horizontal resolution of 35º by 35º presents challenges, especially related to the representativeness error of stations with very different locations (e.g. coastal and mountain stations vs continental stations in populated areas). We do not address these questions in our pilot study, but we agree that this needs to be mentioned in the revised version at a relevant place. Concerning the wavelets: see our response to comment #17.

11. Section 2.2.1 / Figure 1: Comparing the left and right panels, it looks like most (but not all) of the stations for validation are also collocated with stations for assimilation—or are so close as to make no difference. How are you holding back data then? Temporal removal? A bit more description here would be appreciated.

    This problem was rather unavoidable. The alternative would have been to withold data in data-sparse regions, such as the high Arctic, Greece, or Eastern Europe. This would have compromised the assimilation at the European boundaries, and, in the absence of any nearby stations, one would probably not have seen much of an effect in the validation stations. Also, please recall that E-profile is not a centrally administrated network that someone has designed on the drawing board. Rather, it is an ad hoc collection of independent data providers. This is reflected in the geographically inhomogeneous distribution of stations and highly inhomogeneous distribution of different types of instruments. This makes it very challenging to group the stations into those for assimilation and for validation, respectively.

12. Section 3.1/Figure 2 &3: I very much appreciate case studies—they are where you can learn what works and what doesn't. I recommend significant

expansion here, not only in the figures (e.g., speciated AODs would really help), but also by adding a time element. The case you showed for September 6, 2023, is on the backside of a big event, it seems, based on my own analysis. So, that site was hit with a big event on the 4th and 5th and cleared out on the 6th. But CAMS did not clear it out. Hence, E-Profile made the beneficial correction, but you never really know that unless you look at a bit of a timeseries. This kind of context helps the reader really understand what is happening four-dimensionally in the model. Oh, and please increase font size before the typesetter asks you.

We agree. We will add timeseries for a selection of relevant AERONET stations. (See our response to comments 3, 4, and 8).

13. Figure 4 and 5. With the colorbar it is sometimes hard to see what is going adding on. I suggest adding line plots for say 3 levels (surface, 1 km, 3 km?). You could also add individual stations for a good middling and low performers.

As stated before, we will add a few illustrative station plots (see our response to comments 3, 4, 8, and 12).

Concerning the colourbar: Our original plots were using a colourbar that was much clearer, but it was not accepted by the journal. Although it was suitable for allmost all conditions of colour vision deficiencies (CVD), the colour scale was not adequate for Blue Cone Monochromacy (which affects 0.001% of the population). We are keenly aware that the colour scale we are using now is substantially less ideal for those not affected by Blue Cone Monochromacy. But we tested different colour scales, using the online tool Coblis (as advised by the journal), and we found that all other colour scales that comply with all conceivable CVD conditions were even worse. So, the use of this colour scale is due to journal policy constraints.

The problems with the journal-imposed colour bars were raised by both reviewers (see also comment #8 of the second reviewer). Following the reviewer's suggestion, we now tried out line plots at different altitudes (using CVD-compliant line colours). These plots also have their issus, and we are not sure they will yield an improvement that justifies inflating the number of figures. But we will take this suggestion into consideration in our revisions.

At any rate, replacing colour plots by line graphs may not be a sustainable long-term solution. Presently, it feels like the journal is imposing requirements which are causing issues that the authors are expected to resolve. We would like to ask the editor to communicate the reaction of the two reviewers to the board of editors and find a viable reconciliation between the legitimate interests of CVD-affected and non-affected readers.

14. Section 3.2: Maybe a bit more discussion on what regions benefitted the most/least and a discussion why you think that is? You mention you think it is because some regions have higher/lower quality instruments (e.g., line 200), but does any of this also correlated with meteorological or terrain features? If you think it is instrument related, shouldn't that be one of the core conclusions on what kind of instruments are most appropriate? Also, the discussion of PM2.5 bias (Lines 210) again leads to a rabbit's warren on optical properties. So again, what ultimately is the lidar ratio you are deriving here? Does it make sense?

The lidar ratio was also requested by the second reviewer. We will give and discuss those figures in the revised manuscript.

Concerning geographic regions and instrument types: It is a very interesting idea to investigate in more detail if we can find a correlation between the benefit of the assimilation in a region and the dominant type of instrument in the vicinity. But we are not sure if this sort of investigation will be trivial, given the highly inhomogeneous geographic distribution of instruments, which will make it next to impossible to disentangle the effects of instrument quality from the effects of geographic location. However, we will investigate during the revisions if this can lead anywhere. If yes, we will include it in the revised version. If not, we will reword our conjectures about the impact of instrument quality more cautiously.

15. Figure A1. Maybe make (e)-(h) on a log scale?

We presume the reviewer wants to be able to distinguish transmittances very close to 1, but a log scale would not help with this as that is more for distinguishing values very close to zero. The optical depth in the top row is already on a log scale so we can distinguish the magnitudes of the weakly absorbing gases there. The purpose of the lower row is really to show that some gases have a lot of structure when plotted on a log optical depth scale, but their effect on transmittance (which is what matters for modelling lidar attenuated backscatter) is nonetheless negligible so can be safely either ignored or treated as a weak grey absorber.

16. Figure A3. Can you also please provide the baseline lidar ratio you are using for spherical or nonspherical (your preference), But there are factors of 2 floating of around and I am curious where this system sits relative to field observations.

    Agreed (see our response to comment 14).

17. D4: Perhaps a bit more discussion in the impact of AOD? Given the changes in Figure D3, I would imagine the AOD change would be huge. Can you give specific numbers for this? The mean bias plot (D4) would necessarily be of low magnitude because the mean AODs are low. How about for significant events? Maybe break it up between, low, middle, high AODs? The wavelet analysis finding is interesting. Care to elaborate?

    Concerning the impact on specific events, e.g. skill scores, see our response to comment 4.
    Concerning the wavelet analysis: We can only speculate here, because the way the wavelet correlations work is complex and the link between the correlations and the impact on the analysis increment is not necessarily straightforward. The vertical correlations vary geographically and spectrally. They are defined for grid points that have a coarse resolution for small wavenumbers (large geographical affect) and high resolution for high wavenumbers (local effects). The overall impact takes both the geographical location and the spectral information into account when calculating the analysis increment. The figure we show in the paper is therefore an average over the vertical correlations for all the wavenumbers at a specific location. The majority of locations that we have considered for aerosols show a similar form, that the correlation between model levels towards the surface is higher and extends more through the boundary layer. Taken with the higher standard deviations at the surface, we hypothesise that the system leans towards giving a larger increment at the surface. Removing any correlations between vertical model levels just gives the system more freedom to update vertically based on the amount of aerosol present and only the standard deviation as a constraint. Really understanding this link though, would probably require detailed work with single observations under different scenarios.
    To address the reviewer's question, we will state in the revised paper that removing vertical correlations from the wavelet allows the minimisation process more freedom in deciding on the vertical profile of the analysis increment.