Atmos. Meas. Tech. Discuss., 7, C4082–C4087, 2014
www.atmos-meas-tech-discuss.net/7/C4082/2014/

Atmospheric Measurement Techniques Discussions

Interactive comment on "Use of neural networks in ground-based aerosol retrievals from multi-angle spectropolarimetric observations" *by* A. Di Noia et al.

A. Di Noia et al.

a.di.noia@sron.nl

Received and published: 4 December 2014

We thank Reviewer 3 for his/her constructive comments, which help us highlight some aspects of the methodology. Our detailed replies to the Referee's comments follow. Reviewer's comments are in bold, replies are in plain text.

2. Introduction, P. 9050, line 19. The following paper might also be worth citing here: Radosavljevic, V., S. Vucetic, and Z. Obradovic, 2010. A Data-Mining Technique for Aerosol Retrieval Across Multiple Accuracy Measures. IEEE Geosci. Remt. Sens. Lett. 7, pp. 411-415.

We thank the Reviewer for the suggestion, but the paper is already cited in the specified line. The Reviewer may have overlooked it.

3. P 9054, Table 1. The LUT values given here are not necessarily optimal. For example, there can be some advantage in creating tighter grid spacing in regions of scattering-angle space where particle single-scattering phase functions tend to vary most rapidly. Similarly, an adaptive grid in AOT space can also improve LUT performance. This comment does not detract from the neural net approach favored here, but does suggest that might be ways to obtain better results from a LUT as well. Some indication of the interpolation error tolerance for the LUT values chosen would be helpful.

We agree with the Reviewer that the LUT entries are not chosen in an optimal way, and better results may indeed be achieved with the LUT approach either by increasing the number of entries (as we also mention in the introduction) or perhaps by defining the grid in a more clever way. However, it is difficult to devise in advance a way to do this with some guarantee of success, and this is also one of the reasons why we decided to directly propose a neural network approach rather than trying to optimize the LUT. Also an idea of the interpolation error tolerance for the LUT values is difficult to give without setting up a dedicated experiment that would take quite some time.

4. P 9056, lines 17-20. Approximately 90% of the simulated data was used to train the NN, and only 10% to test the result. Do the 10% adequately cover the range of conditions in a statistically meaningful way?

Actually the partitioning of the dataset was 70% training, 15% validation and 15% test (training set with slightly less than 8×10^5 examples, validation and test sets with 1.65×10^5 examples each, ratios are about $8/11.3 \sim 0.7$, $1.65/11.3 \sim 0.15$). By looking at the histograms and at the correlation matrices of the test data and comparing them with those of the overall dataset we did not find evidence of sampling biases. The absolute dimension of the test dataset (1.65×10^5 data) is also fairly large, giving us reasonable

confidence that the range of conditions is adequately sampled by the test dataset.

5. P 9057, line 17. I'm wondering why the error is assessed against the generic, noise-free (y) rather than the original measurements. This seems to imply a very high confidence in identifying noise in the original data. (I see now that you get to this to some extent later in the paper.)

The personal experience of the first author is that, if a minimum error reconstruction metric is used (as we have done, instead of the percentage of explained variance) and the reconstructed noisy data are compared to the original noisy data themselves, the procedure tends to get biased towards a high number of retained principal components, probably because optimally reconstructing noisy data also implies reconstructing the noise part, which usually is better done if many principal components are used, but which we would like to suppress as much as possible. Of course we assume that our instrument noise model is good enough that the conclusions we draw on simulated data are not too unrealistic, but if we look at how the algorithm performed on real data it seems that this assumptions did not create major problems.

6. P 9064, line 14. I'm not surprised that the NN provides a better initial guess than the LUT, so convergence is faster, as expected. But why would the PT systematically not reach as good a solution when initialized by the LUT, if convergence is achieved? (According to Figure 3, convergence is achieved in essentially all cases before the 20-iteration cutoff.) Is it that the PT finds local minima when initialized by the LUT, whereas the NN finds a global minimum, and if so, why might the LUT guess wrong so consistently?

In general it can never be said that the NN finds a global minimum of the cost function. Since the cost function is usually nonconvex, the only way a global minimum can be identified with certainty would consist of scanning the whole parameter space. Our retrieval approach is a Gauss-Newton type iterative scheme, and as such it can never escape a local minimum if it reaches one. Probably the reason why the PT performs

worse when initialized by the LUT is that the NN tends to drive the iterative scheme towards better local minima than does the LUT. This makes sense, because the LUT has a very partial knowledge of the correspondence between the parameter and the measurement space, whereas the NN, if well trained, tends to return an approximation of the conditional expectation of the state vector given the measurement vector (Bishop, 1995), that is probably likely to be around a better local minimum of the cost function than a best matching value taken from a somewhat arbitrary list of alternatives. Once again, it is possible that a denser or differently defined LUT exists that may compete with the NN, but a denser LUT is very demanding in terms of memory, and it is not trivial to find a different LUT with a similar amount of entries that satisfies a predefined accuracy requirement.

7. P 9065, lines 19-20. There might be a reason the AERONET Level 2.0 (quality assured) particle property data are not available. See Note 10 below.

Please see response to note 10.

8. P 9066, lines 11-14. Do the six points in Figure 4 having values <1 for the NN and values >6 for the LUT have some underlying characteristics in common? For example, are they all outside the range of applicability of the parameter space defined for the LUT? Similar question for the points that failed to converge altogether for the LUT but not the NN approach.

We have not noticed relevant characteristics common to all the points with $\chi^2 < 1$ for the NN and $\chi^2 > 6$ for the LUT. If we look at Table 1 and assume that AERONET microphysical retrievals, even if not extremely accurate, are not extremely far from the truth, it does not seem that there were any situations that were outside the parameter space defined by the LUT.

9. P 9066, Figures 6 and 7. It is difficult to see what is going on here in any detail. Perhaps you could plot the difference between the AERONET validation data and the LUT+PT or NN+PT values.

Since AERONET microphysical retrievals are not extremely accurate, we think that plotting differences with respect to an unstable reference could be misleading and lend itself to overinterpretation. After all, the main message of Figures 6 and 7 is that there is qualitatively a good agreement between groundSPEX and AERONET retrievals, and it seems to us that this message emerges more clearly if absolute values are plotted instead of differences.

10. P 9067, lines 10-13. AERONET sky scan retrievals are not considered to be of good quality unless AOT_440 > 0.4 [e.g., Dubovik et al. JGR 2000]. Except perhaps for the AOT peaks on 07 and 09 July, this appears not to be true. This raises a question about the results of Figure 6 and especially 7, specifically for AERONET, but perhaps also for the other retrievals.

We have checked the AOT at 440 nm and the Referee is right: the AOT at 440 nm is larger than 0.4 only during the AOT peaks on July 7 and 9 (and close to this value during the peak on July 8). However, we prefer to leave the results of Figure 6 and 7, because after all the level 1.5 data are still the only comparison we can exhibit for these retrievals, which we believe is still better than showing no comparisons at all. Also the consistency between the AERONET and the groundSPEX retrieval approaches, that are fairly different (azimuthal scans and no polarization versus principal plane scans with polarization) gives us some confidence that these comparisons are not useless.

11. Maybe it would be worth comparing Angstrom exponents, as these are reported from AERONET direct sun measurements, which are Level 2.0, and although they are less specific than fine-mode AOT, etc., about particle size, they are also less dependent on the definitions of the modes.

We have followed the Reviewer's suggestion and compared the 440-675 nm Ångström exponents. A new figure has been introduced into the revised paper accordingly. We observed a good agreement between the Angstrom exponents during July 2013, whereas on September 5 groundSPEX retrieved systematically lower Ångström expo-

nents than AERONET. This complicates our interpretation of the sensitivity of groundSPEX retrievals to large particles, because it seems to contradict the results found for the coarse mode effective radius and the coarse mode AOT. However, at the moment we are not able to offer a systematic explanation for this effect.

12. P 9067, line 27 ff. What happens if an actual atmospheric column contains an aerosol mixture not consistent with the assumed bi-modal distribution, either because the individual aerosol components are not represented in the particle microphysical property parameter space, or because there are more than two modes present?

It is quite difficult to predict what happens if the actual aerosol size distribution has a shape that differs from the bimodal log-normal distribution. The retrieval would still try to fit the parameters of the log-normal distribution in order to reproduce the measurements as good as possible. A response to this interesting question would also require dedicated analyses that are far outside the scope of this work.

REFERENCES

Bishop (1995), "Neural Networks for Pattern Recognition", Oxford University Press.

Interactive comment on Atmos. Meas. Tech. Discuss., 7, 9047, 2014.