Atmos. Meas. Tech. Discuss., 6, C4756–C4763, 2014
www.atmos-meas-tech-discuss.net/6/C4756/2014/

Atmospheric Measurement Techniques
Discussions
Open Access

Interactive comment on "Satellite retrieval of aerosol microphysical and optical parameters using neural networks: a new methodology applied to the Sahara desert dust peak" *by* M. Taylor et al.

A. Di Noia

a.di.noia@sron.nl

Received and published: 5 April 2014

Dear Authors,

As I am also currently investigating the application of neural networks to aerosol retrievals, I have read your paper with interest. Nevertheless, I can't help expressing some concerns about the proposed methodology. I would like to bring such concerns to your attention in case you judge them helpful for improving the manuscript during the revision phase. I will not go through a full review of the manuscript but will only con-

centrate on a number of key points that captured my attention. The concerns I have, that are explained in detail below, involve the following aspects:

1. The attempt to estimate a large number of aerosol parameters from a limited amount of input information.

2. The attempt to optimize too many neural network features using the same validation set, and the use of a performance metric that depends on the results on the training set.

3. The methodology and the statistics used to evaluate the quality of the aerosol retrievals produced by the neural network.

I send you my best regards and wish you a successful completion of the review process.

Antonio Di Noia

——————SPECIFIC COMMENTS ——————

1. As correctly noted in the introduction, accurate global datasets of the aerosol microphysical properties are needed for a good understanding of the aerosol impacts on climate, and satellite remote sensing is, in principle, the most suitable tool to produce such datasets. As correctly noted as well, MODIS products deliver global daily estimates of the AOT, but do not currently deliver reliable estimates of other microphysical parameters. Some complementary aerosol information is provided by instruments such as OMI , that deliver information on UV absorbing aerosols and an absorption AOT at 550 nm. The aerosol optical thickness is physically related to the aerosol size distribution and to other aerosol microphysical parameters. In fact, the AOT is essentially the integral of the aerosol extinction coefficient over the vertical coordinate, and the extinction coefficient depends on the aerosol size distribution and on the aerosol refractive index through the extinction cross section. As a result, several authors have developed retrieval techniques aimed at inferring the aerosol size distribution from multispectral

AOT measurements. Examples of such techniques are correctly cited in the paper. To the best of my knowledge, such techniques usually rely on some assumptions on the aerosol refractive index and single scattering albedo. Simultaneously retrieving the aerosol size distribution, refractive index and possibly single scattering albedo as a function of the wavelength is a much more complex task. In order to perform these simultaneous retrievals, it is important to complement the multispectral capability with multiangle observations and polarization measurements. The recognition of this is basically what has inspired the development and launch of instruments such as MISR (multiangular) and POLDER (multiangular and polarimetric), and what continues to drive much of the research in this field. Neural networks are a valuable tool for satellite retrievals, but they cannot make miracles if the input information that is provided to them is not sufficient. It seems to me that some of the neural networks presented in this paper suffer from this problem quite severely, especially the case 1 NN, where it seems that an attempt is made to estimate 7 uncorrelated pieces of information on the aerosol parameters from only 1 piece of information in the input data (I am referring to the results of the PCA). It is not surprising to see that the best case 1 NN was deemed as “trained” after as few as 2 epochs (if I correctly interpret Table 2). According to my experience, this typically happens when an attempt is made to train a NN with data that contain very limited information on the target quantity.

In view of these considerations, and also based on the results of your analysis (e.g. Table 7), I was wondering if it would not be worth to try to reduce the set of estimated aerosol parameters. This would, in my opinion, ensure a more “transparent” use of the input information in the retrievals, avoiding the delivery of parameters that are probably not really “retrieved” by the neural network scheme because of the lack of information in the input data.

2. The NN architecture and the fraction of data to be used in the training dataset are selected by means of a heuristic procedure. The whole dataset is split in a training and a validation subset with different ratios between number of training data and number

of validation data (which from now on I will call T/V ratio for brevity). Several neural networks, each characterized by a certain number of hidden neurons and trained with a certain T/V ratio, are compared, and the metric used for this comparison is the sum between the training and the validation MSE. To my awareness, this metric is quite unusual, as the most common procedure for heuristic NN model selection is to compare different neural models based on their generalization performance, thus evaluating them only on a set of unseen cases and not including the training MSE in the metric for model optimization. While I think that such an approach is quite natural, even from a theoretical standpoint, I honestly do not see a particular merit in using the training MSE as part of a criterion for NN model selection. I am also concerned that the metric you use has, in some cases, the potential of biasing the model selection towards heavily overfitting models rather than on models that generalize well. For example, let us consider two networks NN1 and NN2, let TE1 and VE1 be the training and validation MSE of NN1 respectively, and let TE2 and VE2 be the analogous quantities for NN2 on the same sets. If TE1 « TE2 but VE1 » VE2, then this might be an indication that NN1 has merely memorized the training set without building any generalization capability. A performance metric only based on the validation MSE would select NN2 as the best model. A metric based on the training + validation MSE, instead, might lead to the selection of NN1 if TE1 + VE1 < TE2 + VE2, despite NN1 is, by hypothesis, much less capable than NN2 of producing correct results on unseen cases. In my opinion, the most appropriate performance metric would be the one that selects NN2, because after all what we are interested in is to use a NN in unseen situations, rather than just use it to memorize a dataset. May I ask you to justify the choice of using the training + validation MSE metric for the selection of the best NN?

The method used to select the best T/V ratio raises similar concerns. It is true that summing the statistics on the training and on the validation subsets is probably a way for you to ensure that the statistics for the comparison are calculated over the same sample, but still the training + validation MSE metric might favor nets with higher T/V ratio if the MSE on the training set decreases faster than the MSE on the validation set with an increase in the T/V ratio. Furthermore, I am not sure that the difference between using 85% or 90% of the training set (in two cases 85% turns out to be better) is really significant, or is just related to the intrinsic randomness in the data splitting and in the NN initialization. In any case, if you think it is important to optimize the T/V ratio, the only way I can see to do it in a meaningful and conclusive way involves: (i) using a third set of unseen cases as a benchmark set for all the T/V ratios between the training and the validation subset; and (ii) repeating the process for several realizations of the data splitting and several initializations of the neural networks, in order to see if the results are really an indication that a certain T/V ratio really leads to a better NN than another one.

Other minor considerations:

- From Table 2 I also notice two facts: (i) the selected NNs in for all the 4 cases were stopped after a very limited number of epochs (see comment 1); (ii) for all the selected NNs, the training error seems to be significantly larger than the validation error. One one hand, you may argue that this might remove (but only for these particular cases and not on a methodological note) the concern I have raised above about the used performance metric possibly favoring overfitting networks. However, on the other hand the only way I can imagine to explain this fact is that in the training set there are some some cases that are not learned well by the NN. These "outliers" might be bad training data (but it seems strange, since you seem to have carried out a strict quality filtering), or be simply cases in which the input vector does not contain information for recovering the target vector. The MSE is a statistic that is quite sensitive to outliers (because there is a square power involved, which enhances the weight of high values), and since you use it for performing comparisons in support of a model selection, you might really want to check what is going on there.

- A thing that looks a bit surprising from Figure 4 is that both the training and the validation MSE for the CASE 4 NNs with about 40 neurons seem to increase for an increasing fraction of training data. I find it surprising especially for the training MSE. Is

it really like that or am I misinterpreting the figure? Do you have at least a preliminary explanation for that?

3. Validation. In Table 5, the percentage of retrievals whose absolute deviation from AERONET at Dakar falls within the threshold set by Mishchenko et al. (2007) is reported for each retrieved parameter. Based on these results, it is concluded that the quality of most of the NN retrievals is within the thresholds. However, I am a bit afraid that this kind of analysis might be slightly misleading if the (seasonal?) cycle, or at least the mean value, is not removed from the AERONET data prior to the comparison. Let's look, for example, at Figure 10, where the daily averages of the AERONET and the retrieved SSA(440) are shown, together with a scatterplot and a histogram of the residuals. What I see from the scatterplot is that the correlation between AERONET and the NN is quite weak, and the NN always returns something that is similar to the average of those data, which is in turn close to the average of SSA(440) on the training dataset. So it seems to me that, as long as the AERONET data stay close to their mean value, the NN retrievals are “good” and probably within the Mishchenko's threshold, but when there are anomalous values of the SSA (i.e. values that are well above or below the mean) the NN does not seem to get them right. But significant deviations from the means are, in a sense, the most interesting part of the retrieval. For SSA(440), Table 5 says that 73.37% of the retrievals were “certain” (i.e. error within Mishchenko's threshold). But which percentage of those AERONET data deviate from their averages more than the same threshold? This comment applies to all the aerosol parameters. From neural network theory it is known (e.g. Bishop, 1995, Chapter 6) that a neural network trained with a variable X as input and a variable Y as target returns an approximation of the conditional mean E[Y|X]. If Y is not sensitive (weakly sensitive) to X, then E[Y|X] is equal (close) to E[Y], and I would expect more or less this as the value returned by the NN in such a case. This is just to say that in order to be reasonably sure that a NN is actually “retrieving” something, it is really necessary to make sure that the target quantity is estimated with a reasonable accuracy also when it deviates significantly from its average value over the training set. For instance, the

average value of SSA(440) at Dakar is (Table 5) 0.901, which is very close to that of the same parameter over the training set (0.900, if I correctly interpret Table 3). I think it would be interesting to see what happens if you test your NN on an AERONET site where this parameter has a different climatological behaviour. Would the NN be able to get at least close to the actual climatological mean of that site or would it still return values close to 0.9? This would tell you if the NN is really using information from the input data to estimate SSA(440) or it is just returning something close to E[SSA(440)] whatever value the input vector takes. As a general methodology, I would suggest you to perform the following sequence of tests on the results of your NN:

A. For each aerosol parameter estimated by the NN, check if the RMS (or the mean absolute) error of the NN on that parameter is significantly lower than the RMS (or the mean absolute) difference between the target values of that parameter and its average value over the training set. If this is not the case, then it is highly likely that the NN is not using any input information to estimate that parameter, but has only learned to return its average value over the training set.

B. The parameters that do not pass the previous test are actually not retrieved by the NN. For the parameters that pass the previous test, perform an analysis over a number of locations, and for each location (and season) check which features of the time series of those parameters are well reproduced by the NN, e.g. only its “climatological” value (which would be already something) or also shorter term variations. I have the impression that this procedure would lead to a more reliable assessment of the actual retrieval capabilities of your algorithm.

A further minor comment:

- I have had some difficulties in interpreting Table 3. In the text and in the caption of the table, you talk about “training results”. But the column containing the mean values of the input parameter is named “Validation”. Are these statistics then computed on the training set or on the validation set?

References

Bishop, C. M. (1995), “Neural Networks for Pattern Recognition”, Oxford University Press, New York, NY, USA.

Interactive comment on Atmos. Meas. Tech. Discuss., 6, 10955, 2013.