Atmos. Meas. Tech. Discuss., 6, C4795–C4818, 2014
www.atmos-meas-tech-discuss.net/6/C4795/2014/


Atmospheric Measurement Techniques Discussions

# *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.

**M. Taylor et al.**

patternizer@gmail.com

Received and published: 25 April 2014

We are grateful to Dr Noia for taking the time to detail many insightful comments which we believe have helped improve the presentation of our new methodology and the interpretation of our findings in the revised manuscript.

"Dear Authors,

As I am also currently investigating the application of neural networks to aerosol re-



trievals, 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 concentrate 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."

Below we answer Dr Di Noia's comments point by point:

"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."

We would like to start by thanking Dr Noia for his support of the underlying rationale and associated bibliography presented in the introduction of the manuscript.

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.

We are grateful to Dr Di Noia for this observation which is correct. The rationale behind the choice of CASES 1-4 was in part, to demonstrate the effect of different input information on the NN retrieval. Regarding the CASES 1 NN, the reviewer is right that a single PC contributed over 98% to the variance of the 3 AOD inputs and 7 output PCs contributed over 98% to the variance of the 38-variable AVSD, SSA, CRI and ASYM outputs. As Dr Di Noia points out, only 2 back-propagation iterations (epochs) were



performed by the optimal CASE 1 NN that comprised 10 hidden Tanh neurons and which used 90% of the AERONET training data (3427 data records). Despite achieving a high NN output versus AERONET output correlation (R=0.998), of the 4 CASES considered, the CASE 1 NN performed the worst in terms of its ability to retrieve the AERONET values at the test site Dakar. As the reviewer points out, this happens because the input information is limited (only AODs). We include CASE 1 in our paper to highlight the improvement when using more information in the input than only the AODs (CASES 2, 3 and 4).

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."

Dr Di Noia raises a very good point here. In a follow-up paper in preparation, we performed cluster analysis on chemical emission data to partition the globe into 10 distinct (in terms of composition) aerosol type regions. In each regional cluster, we extracted AERONET inversion data for the AVSD alone and trained a NN using co-located and synchronous MODIS AOD and H2O inputs (i.e. with OMI inputs). In the context of the submitted manuscript, this corresponds to a CASE 2 NN with satellite inputs and with a reduced set of outputs (only the AVSD). As an example, in Figs. A, B, C and D below we show some results that suggest that this approach appears to be plausible:

Fig. A: The "marine sulphate" cluster comprises 656 data records of AERONET AVSD in 5 pixels (1x1 degrees).

Fig. B: The daily AVSD retrieved from the trained NN compared with co-located AERONET AVSDs.

Fig. C: The mean AVSD retrieved from the trained NN and from AERONET for such a CASE 2 NN trained on MODIS AOD + H2O inputs. The variability of the sample is shown via the 5%, 25%, 50%, 75% and 95% quantiles.

Fig. D: The correlation coefficient in each radial bin.

In Fig. D the maximum correlation coefficient is ≈0.55 and in the submitted manuscript it is ≈0.47. These findings suggest that, even here, there is no clear indication of an improvement in the performance with less output variables. The focus of the submitted manuscript is on the outlining of a methodology that could potentially provide quantifiable and verifiable estimates of as large and physically useful as possible set of aerosol microphysical and optical parameters from NNs. As such, we decided against using a reduced subset of target parameters.

"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 Âń TE2 but VE1 Âż 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?

We thank Dr Di Noia for this insight. We understand and agree with your line of thought here. In the Levenberg-Marquardt scheme used here to update the weights and biases of the NN, the training data is used to adjust neuron connection weights, to back-propagate the error through the NN and then to calculate the accuracy of the NN output. The validation data (also part of the NN learning process) is used to ascertain whether or not the NN accuracy meets a threshold accuracy condition (or "goal"). For each NN in the grid of NNs, we set a rather stringent goal at 1/100th of the mean variance of the target PCs. This brings us to a discussion of our rationale for using a combination of the training accuracy and the validation accuracy to select the optimal NN from the grid. As Dr Di Noia correctly points out, it is customary to optimize a NN on the validation data rather than the training data. This was also our initial approach when we began analyzing the data. However, we found that the generalization performance of the NN at the test site (Dakar) was maximized when we coupled the training accuracy assessment (via the MSE) with the validation accuracy assessment – even though we are aware that this does not mean that the performance will be maximized for all cases. It is possible that this is problem-specific (i.e. works only for the dataset used here). Equally, one may ask why we chose to use the sum of the MSEs as a selection criterion and not, for example, a weighted sum (i.e. the training % times the MSE of the training data + validation % times the MSE of the validation data)? Our (rather naïve) answer

to this is that the simple sum provided the best results in terms of the performance of the resultant NN on unseen data. The detailed reason as to why exactly this works best is not yet understood by us and we remain open to suggestions on this point. Our short answer to Dr Di Noia is that by using the sum of validation + training error provided the best results in terms of the performance of the resultant NN on unseen data of Dakar. We recognize that the NN is not built to work in the general case (i.e. to retrieve dust properties worldwide), but it works well for the Northern Africa region where we performed our study. We hope to address the generalization problem in a future publication.

"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."

Again, we thank Dr Di Noia for his detailed consideration of this point. We wish to emphasize that the methodology presented here is a first attempt at objectivizing the choice of NN architecture and is not ideal. It would be interesting to investigate this and also Dr Di Noia's previous comment in a targeted study using surrogate data. In



the revised manuscript we have added the following paragraph on Page 10967, line 14 and added the relevant citation to the references:

"We wish to emphasize that the methodology presented here is a first attempt at objectivizing the choice of NN architecture and is not ideal. For example, the discrete steps in (neuron, training%)-space could be made finer (i.e. instead of steps of 5% in the training% we could have used a 1% step size). In addition, a "bootstrapping" approach could be adopted that would allow several different instances at the same training%/validation% ratio to be evaluated. It should also be noted that it is customary to optimize a NN on the validation data rather than the training data (Bishop, 1995). This was also our initial approach. However we found that the performance of the resultant NN on unseen data at the test site (see Section 4) was maximized when we coupled the training and validation MSE. We recognize that the NN is not built to work in the general case (i.e. to retrieve dust properties worldwide), but it works well for the Northern Africa region where we performed our study. We hope to address the generalization problem in a future publication. For a thorough description of data handling in the context of constructing and testing function approximating NNs, we refer the reader to Bishop (1995)."

"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."

Thank you. Yes, the training error is significantly larger than the validation error in CASES 1-4. The percentage fractional error (= 100% * (training MSE – validation MSE) / training MSE) is +15.8%, +17.2%, +27.8% and +12.1% for each case respectively. While we attempted to implement a strict quality filter via aerosol typing and the exclusion of outliers at the 68% level of confidence with Grubb's Test, it does appear that Dr Di Noia is correct here - that the input vector does not contain information for recovering the target vector fully. On the one hand, the percentage fractional error does not appear to depend on the size of the sample (the CASE 1-4 NNs have N=3808, 3808, 353 and 213 training data records respectively). However, we cannot exclude the possibility that there may be data vectors which are associated with input-output values that occur less frequently and which are therefore not learned well by the NN. To investigate this, in Figs. E, F and G below we show histograms of the input and output parameters in the filtered training dataset used to train the CASE 2 NN:

Fig. E: A histogram of the inputs used in training the optimal CASE 2 NN.

Fig. F: A histogram of the optical outputs used in training the optimal CASE 2 NN.

Fig. G: A surface plot of the daily AVSD output used in training the optimal CASE 2 NN.

Our first comment is that the distributions of the AOD, SSA and CRI-I all exhibit a single long tail and are non-symmetrical (i.e. skewed). The distributions of the CRI-R and ASYM are more symmetrical (a test for normality could be used to establish whether or not they are Gaussian or not here). The distribution of H2O is clearly bi-modal and while the AVSD surface shows both variability in magnitude and location in the coarse region, it is as expected for dust in this geographical region. It is clearly likely



that a small sample of these (filtered) training data vectors contain input and output values (particularly for the SSA and CRI-I) which are in the long tail of their respective distributions. This is suggested by the training results presented in Table 3 where for example the value of the correlation coefficient for CRI-I is lower than CRI-R and that for the SSA is lower than ASYM for the CASE 2 NN. In the revised manuscript on Page 10971, line 10, we have added the following paragraph:

"One thing to note from Table 2 is that the training error in CASES 1-4 is substantially larger than the validation error (having percentage fractional errors of +15.8%, +17.2%, +27.8% and +12.1% respectively). This can be due to outliers in the dataset, although we attempted to implement a strict quality filter via aerosol typing and the exclusion of outliers at the 68% level of confidence with Grubb's Test. While the percentage fractional error does not appear to depend on the size of the sample (the CASE 1-4 NNs have N=3808, 3808, 353 and 213 training data records respectively), we cannot exclude the possibility (even in CASES 2-4) that there may be data vectors which are associated with input-output values that occur less frequently and which are therefore not learned well by the NN. The second thing to note is that for the CASE 1 NN, convergence was achieved very rapidly (2 epochs), suggesting that the input vector is clearly not containing the information needed to recover the target vector."

"- 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?"

We would like to thank the reviewer for this observation and question. Although Figure 4 plots the number of neurons out to 24 neurons (and not 40 neurons), Dr Di Noia's observation is correct. Figure 4a shows that as the number of neurons increases, a positive gradient emerges in the training MSE with training % (most clearly visible in the lower panel of Figure 4a when the number of neurons is greater than about 12

neurons) – i.e. for a fixed number of neurons the training MSE is increasing with training %. While this may be somewhat counter-intuitive, it is possible that by increasing the training data sample, increases the likelihood of also including a couple of records in the long tail of the parameter distributions (referred to in our reply to the previous comment). As Dr Di Noia pointed out above, the MSE is sensitive to deviations due to the squaring of differences. It is plausible to us that this is what is being manifested here. In the revised manuscript we have re-written the paragraph on Page 10971, lines 20-23 to emphasize this:

"Figure 4a shows (as expected) that the training MSE tends to decrease as the number of hidden neurons is increased. Furthermore, it shows that as the number of neurons increases, a positive gradient emerges in the training MSE with training% (most clearly visible in the lower panel of Figure 4a when the number of neurons is greater than about 12 neurons) – i.e. for a fixed number of neurons the training MSE is increasing with training%. While this may be somewhat counter-intuitive, it is possible that by increasing the training data sample, we increase the likelihood of including a couple of records from the long tail of the parameter distributions which are not easily retrieved, resulting in larger MSEs. Figure 4b shows that the validation MSE increases slowly with the number of hidden Tanh neurons."

"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 in Paper 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."

We would like to thank Dr Di Noia once more for his careful consideration of these important technical points. Our choice of exploratory statistics (the MAE, MARE, correlation coefficient and also the mean values of both the NN-retrieved parameters and AERONET parameters over both the training and simulation datasets) was motivated

by the need to assess how well the NN was able to retrieve the variability of the aerosol parameters and not simply just their mean value. It is not clear that the NN is able to capture the variability pattern at any timescale - since significant variations appear not to be retrieved. We agree that the NN retrieves around the climatological mean of the training dataset and the acceptable answer from Dakar testing is most probably because Dakar data has similar climatological mean values with the training dataset. It is hoped that further validation studies using a cohort of larger datasets will be able to provide a more clear assessment of NN performance. In the revised manuscript, on Page 10983, line 10 we have added the paragraph:

"Finally in this section, we wish to emphasize that, while the MAE and MARE show that the CASE 4 NN appears to be successfully retrieving the mean value of the variability of optical parameters, it is apparently not able to capture the variability itself. For example, Fig. 11 shows the CASE 4 NN retrieval of the coarse mode volume concentration at the daily and seasonal (3-monthly) time scales compared with AERONET data at Dakar. While the mean values are almost indistinguishable, the standard deviation of the NN retrieval is approximately 50% of the standard deviation of the AERONET data at both time scales. This adds further support to the idea that the input information used to train the NN is not sufficient to fully retrieve the variability in the target data. Moreover, since the NN retrieves around the climatological mean of the training dataset, the acceptable performance at Dakar testing is most probably because Dakar data have similar climatological mean values with the training dataset. It is hoped that further validation studies using a cohort of larger datasets will be able to provide a more clear assessment of the NN performance."

The following (new) Figure 11 (with accompanying caption) has been added in the revised manuscript:

"Fig. 11. Test results at Dakar for the volume concentration of the coarse mode V(c) at the seasonal (3-monthly) time scale (upper panels) and the daily time scale (lower panels). Note that while the mean value of the NN retrieval and the AERONET target



data are almost equal, the standard deviation of the NN retrieval is approximately half of that associated with the AERONET target data."

"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."

Thank you. As a cross-check with our findings, we performed Test (A) on the CASE 1 NN (which performed worst) as a check on the ability of the NN to retrieve more than just the average value over the training dataset. The results are presented in the table below:

Parameter MAE (NN-AERONET) MAE (AERONET-mean) PFE [%] V(c) 0.0365 0.154 -321.9 CRI-R(440) 0.041283 0.040864 -9.2 CRI-R(675) 0.036997 0.038041 -15.9 CRI-R(870 0.035801 0.037886 -16.5 CRI-R(1020) 0.036539 0.039494 -17.0 CRI-I(440) 0.002091 0.002264 1.0 CRI-I(675) 0.001735 0.001903 -2.8 CRI-I(870) 0.001671 0.001864 -5.8 CRI-I(1020) 0.001665 0.001884 -8.1 SSA(440) 0.018866 0.020601 -8.3 SSA(675) 0.017837 0.020680 -9.7 SSA(870) 0.017744

0.020675 -11.6 SSA(1020) 0.017599 0.020582 -13.2 ASYM(440) 0.011873 0.013333 -12.3 ASYM(675) 0.014654 0.016723 -14.1 ASYM(870) 0.014869 0.016467 -10.7 ASYM(1020) 0.014753 0.016281 -10.4

The PFE is the percentage fractional error between the MAE error (NN-AERONET) and the MAE difference (AERONET-mean). We can see that the MAE errors for all of these parameters are lower than the MAE difference of the training dataset. The PFE for the coarse mode volume appears to be surprisingly high. This may be due to the 0/0 effect associated with calculating fractional errors for distributions (the AVSD in this case) containing values close to zero. While it is hoped this brief analysis helps answer Dr Di Noia's concerns, we believe it is a lot of detail and may distract from the central message of the paper. We would be happy to present this analysis in the form of supplementary material if requested. Regarding Test (B), we hope to have satisfied Dr Di Noia that the paper has analyzed the performance of the optimal NNs at various timescales (including the seasonal timescale). As mentioned in the manuscript, we selected all available AERONET sites contributing inversion data situated on the peak of the dust AOD. We are extremely grateful to Dr Di Noia for this helpful advice and while, for sake of brevity, we have not applied the NN to test sites outside the Sahara dust peak, we hope to have demonstrated the feasibility of the general approach described in this paper.

"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?"

We are grateful to Dr Di Noia for pointing out this source of confusion. He is correct – it is not clear. Table 3 presents the statistics computed on the training dataset (having filtered the inputs with Grubbs Test). The calculations (mean values and correlation



coefficients) are given for the AERONET "Target" data and for the NN output data (which was confusingly labelled as "Validation". In the revised manuscript we have replaced "Validation" with "NN" in the table columns and also in the caption.

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

We have added this important reference which is cited in the new paragraph added on Page 10967, line 14 (please see our answer to Dr Di Noia's earlier comment). We would like to express our gratitude to Dr Di Noia for offering his expertise and advice which we hope have helped to clarify the methodological concerns he has raised.

---

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

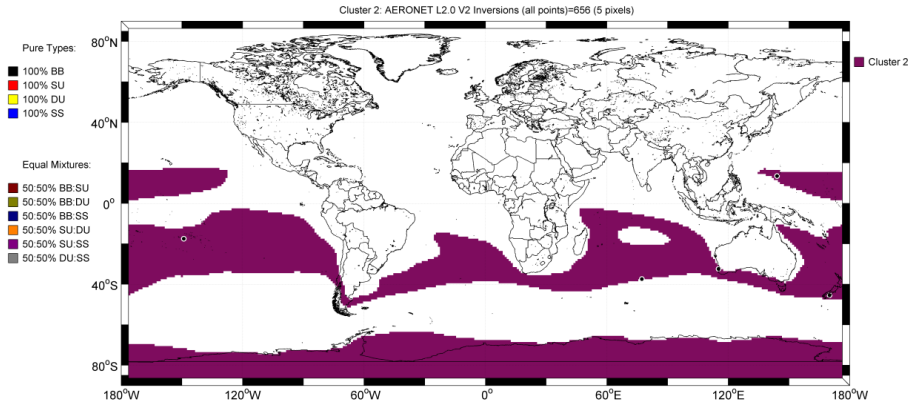


**Fig. 1.** Fig. A: The "marine sulphate" cluster comprises 656 data records of AERONET AVSD in 5 pixels (1x1 degrees).



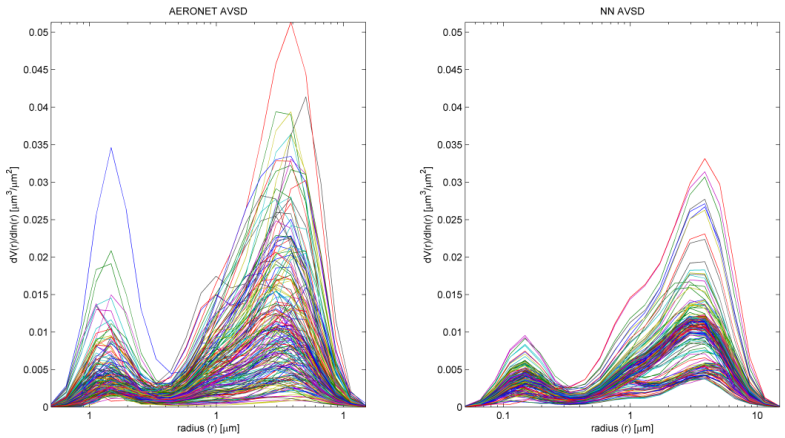


**Fig. 2.** Fig. B: The daily AVSD retrieved from the trained NN compared with co-located AERONET AVSDs.

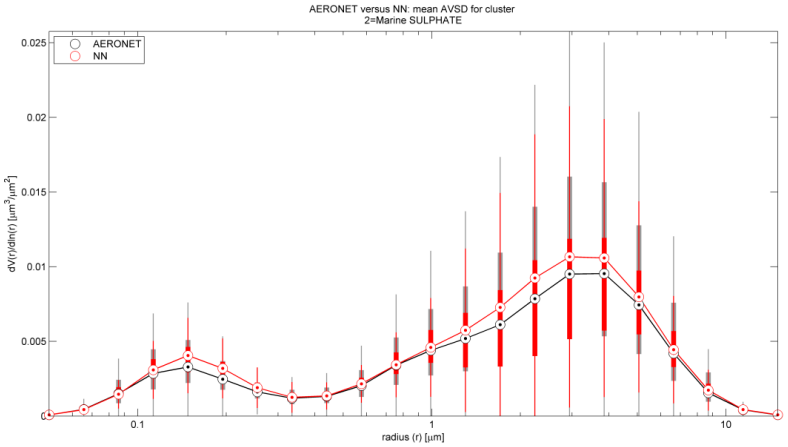


**Fig. 3.** Fig. C: The mean AVSD retrieved from the trained NN and from AERONET for such a CASE 2 NN trained on MODIS AOD + H2O inputs. The variability of the sample is shown via the 5%, 25%, 50%, 75% and 95% qu



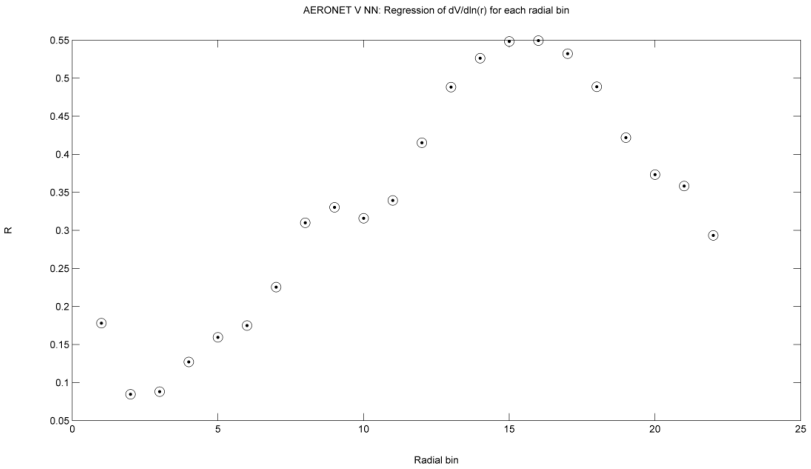


**Fig. 4.** Fig. D: The correlation coefficient in each radial bin.

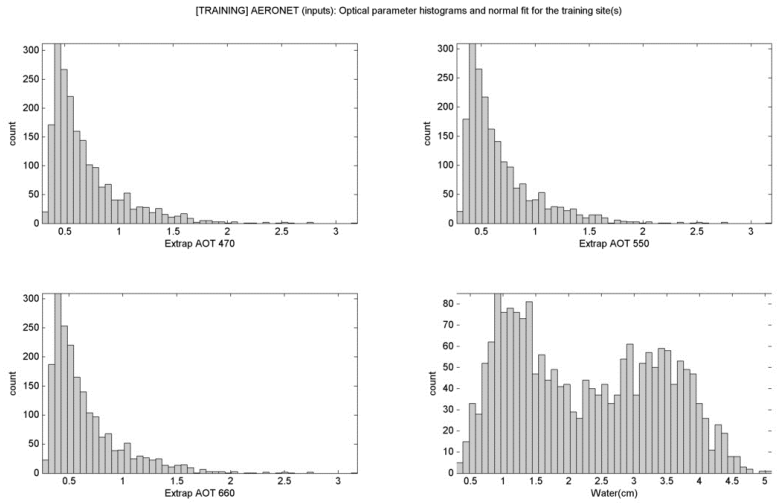


**Fig. 5.** Fig. E: A histogram of the inputs used in training the optimal CASE 2 NN.



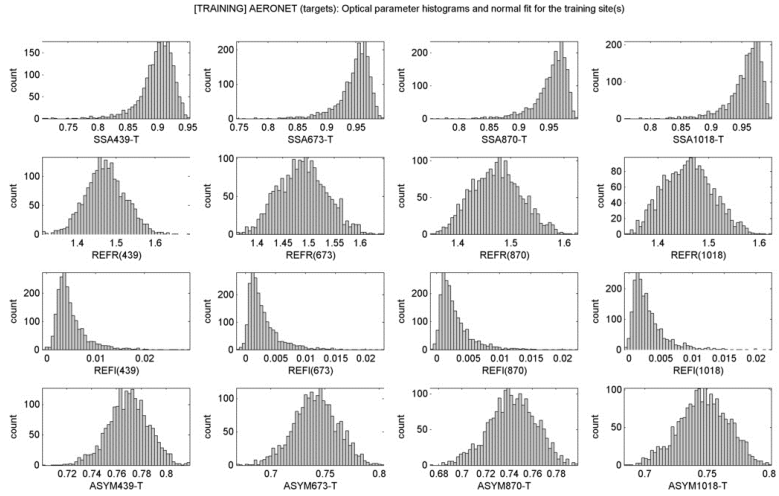


**Fig. 6.** Fig. F: A histogram of the optical outputs used in training the optimal CASE 2 NN.

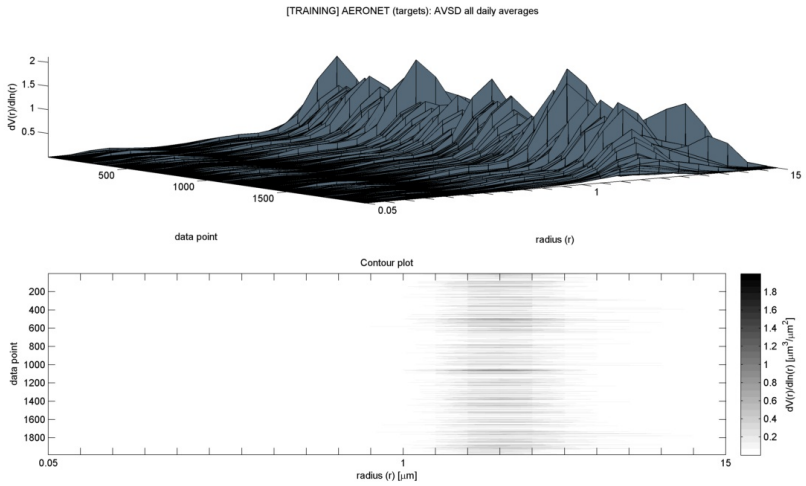


**Fig. 7.** Fig. G: A surface plot of the daily AVSD output used in training the optimal CASE 2 NN.



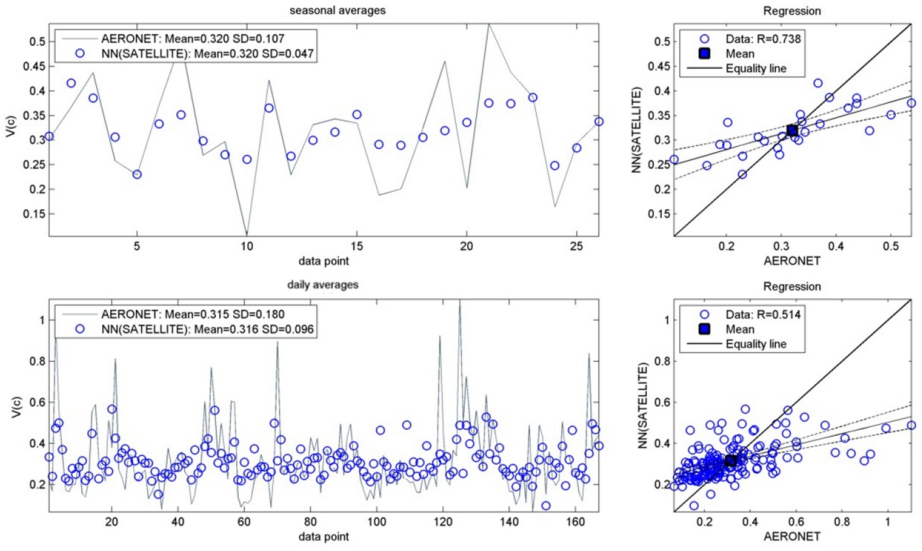


**Fig. 8.** Fig. 11. Test results at Dakar for the volume concentration of the coarse mode V(c) at the seasonal (3-monthly) time scale (upper panels) and the daily time scale (lower panels).