Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2019-327-RC1, 2019

Atmospheric Measurement Techniques Discussions

Interactive comment on "Low-level liquid cloud properties during ORACLES retrieved using airborne polarimetric measurements and a neural network algorithm" *by* Daniel J. Miller et al.

Anonymous Referee #1

Received and published: 14 October 2019

The Authors present a neural network algorithm to retrieve the microphysical properties of liquid water clouds from multi-spectral multi-angle polarimetric measurements such as those carried out by the RSP instrument. They trained a number of NN algorithms from synthetic data and applied them to RSP measurements acquired over the south-eastern Atlantic Ocean during the ORACLES campaigns held in 2016 and 2017. On these campaign data, they compared their NN retrievals to two other retrieval schemes, one using polarization and one based on a typical Nakajima-King radiance look-up table.

The results look partly encouraging and partly in need for further investigation. As I

Printer-friendly version

Discussion paper

will detail below, a number of design choices raise some questions (keeping the cloud height fixed in the generation of the training dataset, going for a neural network with an enormous number of hidden neurons, training on a relatively coarse grid for some parameters, an input scaling mechanism that is supposed to force the NN to give less weight to more uncertain inputs but looks somehow dubious to me). Therefore, I would invite the Authors to elaborate more on the rationale behind such choices. Furthermore, I am a bit surprised by the fact that the NN does not seem capable of retrieving effective variance, as work published earlier this year suggested that this is possible with multi-angle polarimetry, even with instruments that observe at fewer angles and wavelengths than RSP (i.e. POLDER).

Below are my detailed comments.

- P1, L1-2, I would suggest to replace "relate ... microphysical properties" with "retrieve cloud microphysical properties from multi-angular and multi-spectral polarimetric remote sensing observations"

- P3, L33-34, sentence "in a manner unbiased by prior understanding of that relationship". This is not entirely true, or at least needs rephrasing. If you use a forward model to generate the training dataset for your NN, the dataset will exactly capture "your prior understanding of that relationship". If, instead, you mean that by using a training dataset you don't impose a specific analytical parameterization on that relationship, then your statement is correct, but it needs to be expressed better.

- P7, L29. If you exclude SWIR channels, what is the use of radiance in estimating effective radius? Wouldn't it be better to train a specific NN only using polarized radiance, as done in Di Noia et al. (2019)?

- P8, L9. I guess "This screening criteria was" should read "These screening criteria were" (by the way, here I am assuming that you mean the two criteria involving HSRL-2 among those you listed above)

- P12, L10-11. The reason for the choice of fixing the cloud top height at 1 km in the training dataset looks a bit puzzling. Why fix it and not change it just like the other parameters? After all, I guess cloud height is definitely going to have an impact at least on polarization at shorter wavelengths.

- P13, Table 1. What did you do with the parameters affecting the reflectance of the ocean surface, such as wind speed or ocean color? I guess this may well be important over optically thin or broken clouds.

- P13, L7, “principle” should read “principal”

- P14, L7-9. Standardization is just one of the typical pre-processing steps. Another one, just as typical, is to linearly scale the NN inputs to a specified range, such as [-1,1]. Together with a good initialization of the NN weights (which are typically initialized within the same interval), this choice makes the cost function for training easier to optimize.

- P14, L17-20. Usually the goal of standardization (or linear scaling) as a pre-processing step for NN training is to bring all the input and output variables to the same variability range, as this is usually beneficial to any gradient-based optimization problem (Nocedal and Wright, 1999, Chapters 2 and 4). If your method doesn’t do that, then I would say it’s questionable, as it shapes the cost function in a more unfavourable way for convergence. I am not even sure that your standardization method will help the NN to give less weight to reflectance, as the training process may adjust the neural network weights in a way that compensates for that. You should verify if this is the case by looking at the final values for the weights, and possibly at the derivatives of the NN output with respect to the input variables. In general, if one wants to force the NN to give less weight to certain inputs, I guess that one would have to act on the derivatives of the training cost function with respect to the weights (via regularization), rather than on how the input variability range is scaled.

- P14, Section 3.3. It would be interesting to see how PP, NJK and RFT perform on the

synthetic dataset you used to evaluate the NN retrieval.

- P14, L25. With 4 hidden layers with 1,024 nodes each, your NN architecture will have millions of weights, a number that – if I am correctly interpreting the captions of Tables 1 and 2 – is about 1-2 orders of magnitude larger than the amount of data you used for training. This is often not considered good design practice in terms of overfitting avoidance. Are there any compelling reasons why you have chosen this architecture? Have you compared it to other architectural choices? Later on you say that this was done to handle the increase in the input layer dimensionality, but I do not see any obvious reason why an increase in the input layer dimension should require a similar increase in the hidden layers, especially because the main difference between your NN and the previous NN you are referring to is that you are no longer doing PCA, so your input vector has about the same information content as before.

- P14, L26. By "trained" do you mean "presented to the NN"? It doesn't make much sense to say that "a batch of samples is trained".

- P15, L12. Does the added noise variance reflect what is expected for RSP? Was the noise regenerated at each training cycle?

- P15, L22-23. The results for v_e look a bit concerning. Di Noia et al. (2019) report good results for synthetic v_e retrievals from POLDER (RMSE $\sim$0.04), and RSP should be even better suited than POLDER for retrieving effective variance, as it can sample the rainbow region at more viewing angles and more wavelengths. Are you sure your underperformance is not caused by some design choice? Is the rainbow always sampled in the evaluation dataset?

- P16, L3-4. In what sense "the behavior of the initial output layer"? Do you mean the NN retrieval or what?

- P16, L20, sentence "in the framework of NN it is difficult to diagnose the source of this error". Please express more precisely what you mean. Which methods would you use

to diagnose it if you were using a different technique and why can't the same methods be used in the framework of NNs? Is the forward model used in the PP retrievals the same as the one you used to train the NN? If you pass your NN retrievals back to the forward model you used to train the NN, how does the synthetic measurement compare with the observation? And if you pass the PP retrievals to the same forward model? Furthermore, what happens if you pass the PP retrievals to the forward model and try to invert the output with the NN? If, instead, the forward model is different from the model used to create the PP database, and you feed your NN retrievals back to both, how do the simulated measurements compare?

- P18, L13-15, sentence "An evident feature ... below each PDF". I also see some spikes in the NJK histograms. Do they also correspond to LUT points? Going back to the NN histogram, could it be that your NN model is too flexible and does not generalize well outside the discrete grid points you used for training? After all, your NN is remarkably large (4 hidden nodes with 1024 neurons each sounds really huge to me). This again relates to my question how did you come to the conclusion that this NN architecture was suitable for your task. I would also recommend trying to sample the training data from a continuous distribution, but I see you already mention something like that in the conclusions.

- P20, L15. I guess "the behavior of" should not be there.

- P21, L4-5. You attribute the poor correlation in r_e retrievals to the absence of SWIR data, but doesn't polarization give you sensitivity to r_e even if you don't have SWIR, at least in the rainbow angles? Doesn't that mean that for your NN it is still difficult to find a good balance between using radiance and polarization in the r_e retrieval? This goes back to the question of whether it would be better to try to infer r_e with a NN that only uses polarization.

- P22, L1, about the flattening of the tau scatter plot for retrievals using SWIR. How does SWIR reflectance behave as a function of tau? Does it saturate for a smaller COT

than at shorter wavelengths? If so, this may explain why using SWIR is detrimental in your case. Later in the paper you say that SWIR channels are also more affected by 3D radiative effects, so is this maybe the reason?

- P25, L12, I guess there should be an “of” between “impact” and “atmospheric absorption”

- P28, L3. I guess “While” should be removed.

- P28, L17, statement “it lacks a clearly traceable relationship between observations and retrievals”. I agree with you if you are establishing a comparison between NNs and simple retrievals such as Nakajima-King, where the relationship between observations and retrievals is so simple that it can be even represented visually. However, unless you can give a precise definition of “physical traceability”, I do not totally agree if you are also comparing NNs to more complicated fitting schemes such as optimal estimation and Phillips-Tikhonov (as you seem to do two lines below), as the convergence of such schemes is also affected by a number of subjective choices (e.g., choice of covariance of regularization matrix, choice of the regularization parameter, choice of the first guess, choice of boundaries for the parameter values, and I am probably forgetting something), in an often unpredictable way. If you are referring to the fact that OE or PT retrievals optimize a clearly defined cost function involving the observation and the state vector, you should note that with an increasing number of training data the NN retrieval should asymptotically converge to the conditional expectation of the state vector given the measurements (Bishop, 1995) which, for example, for a linear Gaussian problem would coincide with the “traditional” OE solution (which, instead of the conditional expectation gives you the "conditional mode"). Thus also NNs have a "traceable" rationale behind them.

REFERENCES

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

Di Noia, A., et al. (2019), Retrieval of liquid water cloud properties from POLDER-3 measurements using a neural network ensemble approach, Atmos. Meas. Tech, 12, 1697-1716, doi: 10.5194/amt-12-1697-2019.

Nocedal, J., and Wright, S. J. (1999), Numerical Optimization, Springer.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2019-327, 2019.

Printer-friendly version

Discussion paper