Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2017-64-RC2, 2017

Atmospheric Measurement Techniques
Discussions

Interactive comment on “Cirrus cloud retrieval with MSG/SEVIRI using artificial neural networks” *by* Johan Strandgren et al.

Anonymous Referee #3

Received and published: 12 July 2017

The manuscript presents new inversions of MSG/SEVIRI data, providing information on cirrus clouds. The retrieval products are cloud top height, optical thickness and ice water path. As the authors describe, there exists a lack of measurements of cirrus properties. The best cirrus data are today provided by CALIOP and CloudSat, that are both active instruments flying together in a sun synchronous orbit and have both swath widths of about 2 km. This gives poor spatial and temporal coverage, and complementing retrievals by passive instruments are required. Making use of a geostationary instrument, such as MSG/SEVIRI, limits the geographical coverage but excellent diurnal coverage can obtained. The authors also selected to just use infrared observations to obtain 24 h coverage. In addition, SEVIRI provides 15 min resolution. Accordingly, the manuscript has a good justification and the topic fits well with AMT.

The authors selected to apply artificial neural networks (ANNs), and a large fraction of the manuscript describes the procedure for selecting net topology and training approach. I would say that ANNs today are used quite broadly and this part is too detailed (more below). On the other hand, the core element in the training dataset is CALIOP retrievals and there is basically no discussion of the limitations and accuracy of those retrievals. This information is required as all limitations in the training dataset are inherited by the ANN retrievals. In addition, there is no motivation to why CALIOP-only retrievals were selected for the training.

Further, as the other referee, I note the lack of a case specific error characterisation. In fact, there is not even a proper general error characterisation as the errors inherited from CALIOP are not considered. There should also be errors caused by the collocation procedure, as discussed below. In my opinion, this is not satisfactory. However, this is a general issue for ANN retrievals, and it is probably easy to find similar examples published recently. I happen to notice that the authors have submitted a new manuscript, with a title indicating that an extended error analysis now is at hand. I leave it to the editor to judge the overall situation, and potentially consider if these two manuscripts should be joined into a single manuscript.

General comments:

As indicated above, Section 3 could be shortened considerable. In fact, I think the section would become much more clear if the final net topology and training are simply presented "as given". If there is any general experience to draw from the tests performed to reach the final configuration, summarise these separately. The present detailed description of the tests just obscures the final outcome, and it is very hard to extract if there is any experience of general interest.

Some comments on terminology used around the ANN training. In "machine learning" one is usually supposed to work with three datasets: training, validation and test set. The training set should be used to train one or several methods, validation set should be used to select the best one, and finally the test set should be used to evaluate the final system. In the manuscript the authors seem to mix up these sets. They use the validation set to monitor the training of the ANNs and then use the test set to select the best one and evaluate the performance. This is probably not critical for large datasets, nevertheless from a conceptual point of view this is not very nice. The authors refer to those datasets as training data, internal validation data, and (final) validation data, respectively.

Since this is also relevant for the follow-up paper, I suggest the authors create an additional test set that is used exclusively for evaluation.

The authors should reflect upon if using CALIOP alone is the best choice for training data. Why not use some combined CloudSat and CALIOP retrievals, such as DARDAR? As far as I understand, the CALIOP lidar signal is quickly attenuated and I assume that the SEVIRI IR channels have a deeper penetration into the clouds. That is, CALIOP alone does not span a sufficient range of IWP. This problem should vanish if using e.g. DARDAR for training. This comment is a hint, no demand for redoing the work.

However, this touches upon the comment made above, that the errors of the CALIOP retrievals must be reported. These errors propagate directly into errors in the ANN product. As the test dataset is taken from the same CALIOP retrievals, the inherent CALIOP errors are not revealed. Conservative, quantitative, values on the dynamic range (i.e. coverage of optical thickness and IWP) and accuracy of the CALIOP product used for training shall be given.

Further, there are also errors originating in the collocation procedure. Probably most important is the fact that CALIOP has a swath smaller than the resolution of SEVIRI. This results in that CALIOP covers only a part of the SEVIRI footprint, and this gives an additional uncertainty in the empirical relationships between CALIOP retrievals and SEVIRI measurements that the ANN is trained to represent. In any case, there are no

comments at all of possible errors caused by imperfections in the collocation procedure.

The points raised in the last two paragraphs, are they considered in the new manuscript targeting errors?

Specific Comments:

p 5, l 23: Also the bias neurons need to be assigned the correct values.

Sec 3.3: The section fails to clearly report what spatial resolution that is applied. For me this became clear first when reaching p 12, l 31. As 5 km anyhow is used, is the main discussion in Sec 3.3 actually needed? It seems to refer to an older version. That is, this section could be shortened.

p 13, l 24: This is the only place where the authors mention the activation functions of the networks. This information should not be given below "Training data", but as indicated below. Further, why don't the authors use identity activation functions on the output layers of the regression ANNs instead of re-scaling, which would be the more common approach? This could also have an effect on the learning of extreme values since the gradient vanishes at both limits of the output range.

p 15, l 1-10: When introducing the structure of the networks the authors should mention which activation functions are used in the hidden layers and the output layer.

p 15, l 28-29: 1 CPU@3.4 GHz does not describe the computer sufficiently especially since ANN inference is highly parallelizable. The authors should at least give number of cores and processor model.

p 16, l 25: Figure 3 does not seem compare the performance after the final training to the performance before, so the reference here seems pointless.

p 17, l 5-13: The detection threshold for the CCF is a parameter of the classification ANN and is thus prone to overfitting. Its value should not be determined based on the performance on the test set. This touches upon the general comment above. Also, a plot of the POD against FAR for different thresholds would be good as it gives an additional perspective on the performance of the classifier.

p 18, l 4: ’the values corresponds’ should be ’the values correspond’

p 18, l 9: Also here the authors claim the test data was excluded from the training but it seems that it has been used for the selection of the network structure and meta parameters.

p 19, l 3 - 6: See comment p 17, l 5 - 13

p 21, l 5: ’might seems high’ should be ’seem’

p 21, l 7 - 14: The authors should explain in more detail what they mean with uncertainty and solution and/or provide a reference for their claims on the behaviour of ANNs.

p21, l 15 - 21: Isn’t that the reason for the ’uncertainty’ mentioned in the paragraph (see comment p21, l 7 - 14) in the training data? (I.e. low signal to noise ratio for thin clouds?)

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2017-64, 2017.