Review of:

„A random forest algorithm for the prediction of cloud liquid water content from combined CloudSat/CALIPSO observations“

by Richard Schulte et al.

General comment

In the manuscript named „A random forest algorithm for the prediction of cloud liquid water content from combined CloudSat/CALIPSO observations“, the authors use machine learning (random forests) to predict cloud optical depth and cloud top effective radius with CloudSat and CALIPSO observables, independent of the typically used radar reflectivity profile. The idea is to a) fill gaps in existing CloudSat radar-based products during daytime and b) estimating cloud water profiles during nighttime where this information is missing completely.

The manuscript touches on an important topic for climate/cloud research, in particular the potential of the described method to derive nighttime cloud microphysics is large, and of interest to the readers of AMT. The manuscript is well written and the figures support the conclusions drawn by the authors. I recommend the manuscript to be published after my minor concerns and comments are adequately addressed.

Minor points that need revision

- I was a bit surprised by the authors choice regarding the architecture of the random forests: with about 25 million data points during training, the authors train only 100 trees, but balance this by allowing to grow the trees very deep (max depth of 50). This is somewhat opposing the original idea of random forests: to have many weak (shallow) decorrelated learners. I would assume that with the current model choice the model setup it would overfit, but this is not reported.
- The authors evaluate the random forest predictions with MODIS observations, and aim at using the data for filling the gaps of the CloudSat 2b-CWC-RVOD LWC data. In my opinion, it would be good to provide an additional evaluation for these specific situations (during daytime), as a) the situations where gaps exist tend to feature particularly low LWCs (thus insufficient reflectivity) and b) the random forests tend to overestimate cloud optical thickness and effective radius when their values are low (i.e. in these situations, as shown in figure 2). This also somewhat affects the use case of filling the gaps, as these gaps are likely filled with this bias.
- The authors analyze the importance of the random forests input features. I appreciate this, but it does not really help to provide an answer for the speculation that the improved skill of the random forests (compared to the linear regression) is due to the nonlinear capabilities (on page 15), e.g. capturing the saturation of the TB94 signal. This could be easily analyzed using partial dependencies.
- Random forests are incapable of extrapolating, and I am wondering if 1 year of data is enough to capture all variability of the input feature distributions (the authors argue that there is a „slightly different climate“ between 2008 (training) and 2009 (test data)). I am interested to learn if the authors have compared the distributions of the input data between the training and test data. If there are differences in the distributions, a different machine learning technique would likely be more appropriate.