Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2019-308-AC2, 2020

Atmospheric Measurement Techniques Discussions

Interactive comment on "Gradient Boosting Machine Learning to Improve Satellite-Derived Column Water Vapor Measurement Error" *by* Allan C. Just et al.

Allan C. Just et al.

allan.just@mssm.edu

Received and published: 12 March 2020

Response to RC2: We thank the reviewer for their comments and the opportunity for us to frame more clearly why our work will be a contribution to AMT. We have responded point-by-point to the comments below.

1. Comment: The manuscript deals with a machine learning concept to improve the MODIS/MAIAC column water vapour retrieval. Only machine learning aspects are discussed, and these aspects are just described. There is no chance for the reviewer to check the quality of the work. I have to believe what they write. This is rather unsatisfactory.

Response: We have worked with care and used rigorous scientific methods in our data analysis. We have added substantially to the detailed description of our methods (see also our response to Referee 1), and as we have listed in our manuscript, we have already placed full reproducible code and our datasets in the Open Access Zenodo repository (DOI 10.5281/zenodo.3266058) enabling anyone to rerun our full code and regenerate all of our results in the manuscript. This makes checking the quality and reproducibility of our work fully accessible. The contribution of machine-learning models to improving column water vapor retrievals is the main point of our manuscript and we present in our manuscript both a description of our approach and empirical results at AERONET stations and an independent validation dataset from SuomiNet.

2. Comment: MAIAC needs a large bunch of surface, atmospheric, and technical input parameters to successfully retrieve water vapour information. These input parameters have partly large uncertainties. The surface and atmospheric input data change from day to day, with time (morning vs afternoon), with season, with land use changes. Nevertheless, the MAIAC methodology seems to be very robust, the accuracy of the MAIAC products is very good (without any machine learning effort)! To my opinion, it is impossible to further improve the MAIAC column water vapour values!

Response: We wholeheartedly agree that the MAIAC suite of products are very good. However, in collaboration with the MAIAC PI (our co-author on this paper, Dr. Alexei Lyapustin), we have worked to identify opportunities to further understand and reduce retrieval error in the MAIAC column water vapor product. While the recently published global validation of the MAIAC column water vapor product (Martins et al. Atmos Res. 2019) has noted the temporal drift in Terra CWV records, ours is the first analysis that demonstrates an empirical correction. Furthermore, our machine-learning model accounts for the complex interactions of input parameters rather than considering each one separately. Our method clearly demonstrates an improvement in the MAIAC column water vapor values.

3. Comment: However, the authors of the manuscript want to convince the reader

that the machine learning concept overcomes this insurmountable wall of given and (unknown) uncertainties. It improves the results, and reduces the overall uncertainties although the given uncertainties are unknown! How is that possible? The paper gives no answer to this.

Response: Our empirical method to update the CWV values does not rely on propagation of estimated uncertainties for each of the inputs to the retrieval algorithm. Instead, we estimate the retrieval error versus AERONET stations and then build a statistical model to explain this retrieval error using a pre-defined set of input variables. Although we do not know the uncertainties in the retrieval parameters, our use of the new SHAP method for explainable machine-learning helps to quantify and rank which of the input variables we considered are the largest contributors to the retrieval error that we estimated.

4. Comment: The title is 'strange' , not logical! What does it mean: . . . to improve . . .. the error. . .? What does it mean: . . . satellite-derived . . . measurement??? The column water vapour is clearly a retrieval product. . .. There is no 'direct' measurement.

Response: In recognition of the referee's concerns we have modified our title to be: "Gradient Boosting Machine Learning to Reduce Satellite-Derived Column Water Vapor Retrieval Error"

5. Comment: Lines 85-90: In the introduction it is written: machine learning approaches such as XGBoost can model complex phenomena etc.. . . The resulting prediction model can provide an algorithm to reduce the retrieval errors. I conclude: yes, the model can do that provided the complex input parameter set is free of uncertainties. But many aspects (input data) are not well known in the case of the MAIAC retrieval, uncertainties in the input data are large and that is the reason for the uncertainties in the product.

Response: Our empirical results demonstrate that we are able to reduce retrieval error in the MAIAC product even without knowing the degree of uncertainty in the individual

input datasets. Our strategy of modeling the difference between MAIAC retrievals and a ground truth observation works because there are informative predictors that explain much of the retrieval error, whether these predictors are directly related to the source of uncertainty or are themselves proxies (such as time trend). We have a long-running collaboration with co-author and MAIAC PI Dr. Alexei Lyapustin and a track record of using this approach to quantify and reduce retrieval error.

6. Section 2: Line 118-120: Target modelling parameter is the difference between MAIAC and AERONET CWV. . . My question is: When the machine learning approach finds the best way for correction (e.g. based on all the 75 station of northeastern United States in Figure 1) can this approach then be applied to the rest of the world? I do not believe that this will work! Probably we have to find optimum ways for corrections again and again, region by region and all this for different seasons.

Response: We agree with the reviewer that the generalizability of our specific model to new regions is untested and we have acknowledged this in our limitations. However, our approach (and reproducible code) may be applied in other regions with ground AERONET stations in future applications. Our results include all seasons and all years through 2015 of the MODIS record and our validation at SuomiNet stations shows that our results hold across the Northeastern USA, including at ground stations that are hundreds of kilometers from the nearest AERONET station used in training.

7. Section 3 Some examples that explain my general feeling with the paper: Lines 147-149: The XGBoost package is used! Ok! But the reference for this is a conference contribution, grey literature!

Response: XGBoost is very widely used and has recently emerged as a leading tool often winning machine-learning competitions. We have added additional citations related to its performance but the convention in the rapidly evolving machine-learning discipline has included a greater use of competitive conference proceedings which are rigorously evaluated and empirically benchmarked through shared code (their field has

thus largely avoided the overhead of working with major commercial publishers). As an example, Google Scholar lists 4,856 citations since 2016 for this XGBoost paper that we have cited (as of February 27, 2020). We would not consider this publication to be grey literature. The XGBoost software implementation is also widely used and is developed by a sophisticated community of open source programmers - the software repository at https://github.com/dmlc/xgboost, which has 392 contributors as of February 28, 2020.

8. Lines 153 154: XGBoost is combined with DART (here the reference does not indicate any journal?). Can we believe, everything is ok with this procedure? Can we trust? Is all the material peer reviewed by machine learning experts?

Response: We thank both referees for drawing our attention to an incomplete citation. We have corrected our reference for the DART manuscript, which was developed in the Machine Learning department at Microsoft Research. Not only is this work peer-reviewed but also the implementation in XGBoost is open source software that is available for inspection and independent verification by anyone. For more information please see the documentation: https://xgboost.readthedocs.io/en/latest/tutorials/dart.html.

9. Lines 159-161: Bayesian optimization for hyperparameter tuning of XGBoost models was performed using the autoxgboost R package (Thomas et al, 2018). The reference points to arXiv. . .. This is a preprint archive (no peer review, nothing). So, what is this? Can we trust?

Response: While Bayesian hyperparameter tuning has some advantages, we have updated our code and manuscript in more recent revisions and no longer use the autoxgboost package. Instead, as we explain in our revised methods section, we use 50 sets of potential hyperparameter values that are evaluated for performance in a nested cross-validation (within the training data). An advantage of this approach is that it is considerably faster than the Bayesian optimization and also makes it easier

to evaluate the performance of varying combinations of hyperparameters. In our revised manuscript, we have included the following explanation of our updated approach to hyperparameter tuning: "XGBoost has several hyperparameters related to the desired size and complexity of the model that need to be set in training for each dataset. We had a priori selected to tune our XGBoost models with DART using six hyperparameters (Supplementary Table S2), while using default values for other potential hyperparameters based on previous modelling experience. Our tuning and evaluation approach used two-level (nested) cross-validation. Within each training fold for our outer cross-validation, we further randomly split the training data in half and performed a 2-fold cross-validation to compare the performance of XGBoost models using 50 random sets of potential hyperparameters selected with Latin hypercube sampling (Stein, 1987) to be well-spaced across the range of potential hyperparameter values. While this is more similar to a random search than a grid search, it is expected to more efficiently find well performing sets of hyperparameters than random search, because it decreases the likelihood of checking combinations that are trivially different or leaving unexplored regions in the six-dimensional space, which has too many combinations to effectively cover with a grid search. We selected the set of hyperparameters that minimized the RMSE within the withheld portion of the training data before refitting with all training data."

10. Lines 175-176. . . The contribution of each feature to cross-validated predictions was estimated by SHAP values (reference. . . arXiv). . . .again this preprint archive. . .

Response: Again, please see our previous responses on publication practices in the academic field of machine learning where preprints and conference proceedings are reviewed and widely accepted. For example, we cite the preprint as this is the most cited reference for this work but open reviews are also posted at: https://openreview.net/search?term=Consistent+feature+attribution+for+tree+ensembles&co

11. Lines 177-227: A lot of information and description is given by the authors, written

in a smart appearing way, but it does not help. The reader is lost! He/she just has to believe that everything is ok with this way. But he/she does not trust.

Response: We have made substantial edits within our introduction, methods, and discussion to add clarity and detail in our description of the machine-learning methods that we employ - particularly as their recent emergence in the field of machine learning means they have not had time yet to be widely adopted into atmospheric sciences. We particularly addressed the specific clarifications sought by referee #1 in their detailed comments (see response and revised manuscript). Again, as in our prior responses, we stress how we present rigorous scientific analyses including multiple approaches to cross-validation and comparison with an independent dataset (SuomiNet CWV). We again emphasize that all of our code and data are archived and are fully reproducible (see our Open Access Zenodo repository DOI: 10.5281/zenodo.3266058) enabling anyone to regenerate our results.

12. Section 4: results: I avoid to give my comments to the text. . . nobody can check what they state. . ., what is ok, what is not ok, what is trustworthy, what is not trustworthy. There is nothing to judge!

Response: Please see response #1 and response #11.

13. Figure 1: There is no hint where we are? no city name, e.g., Boston, New York, no name of any state. . . Maryland. . .. Figure 1 is a nice 'indicator' , ... of the feeling I have with the entire paper.

Response: While we note that our figure includes both latitude and longitude grid lines and labels as well as a descriptive figure legend that explains that the region shown is the Northeastern and Mid-Atlantic USA, we have now added labels for the major urban centers of Boston, New York, and Washington D.C. in our revised manuscript.

14. Figure 3 and the following figures tell me: MAIAC does a good job, seasonally dependent uncertainties are visible. This is ok, surface properties change and are

not perfectly considered in the retrieval. One should accept that. Machine learning procedures may purge the deviations in this specific 'learning region' of Northeast USA. But for any new region . . ..? We have to start again, I believe.

Response: As we make clear in our discussion of limitations, the generalizability of our findings to new regions was outside the scope of the detailed analyses that we present in this manuscript. Our demonstration of improved agreement of MAIAC CWV with an independent dataset of CWV from SuomiNet GWP stations at new locations in the Northeast USA is evidence that the MAIAC CWV retrieval error can be decreased through our empirical approach. As we discuss, our results are not perfect either although the similarity in the resulting RMSE for Aqua and Terra after applying our correction (both improved versus the use of raw MAIAC CWV values) also suggests that we have reached a plateau of what is possible within our approach. We demonstrate our results over a large region of the United States over 16 years including all seasons, and we make our code and data available for anyone who wants to apply our methods to new regions.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2019-308, 2019.