Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2019-308-RC1, 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.

Anonymous Referee #2

Received and published: 29 January 2020

The Authors propose a machine learning method to perform an a posteriori correction of MODIS water vapour retrievals produced by the MAIAC algorithm. Based on a training dataset of MODIS-AERONET coincidences, they identify a number of variables from which differences between MODIS and AERONET water vapour estimates can be predicted and they train a model to correct for such differences. MAIAC retrievals post-processed with the trained algorithm are then compared to SuomiNet water vapour data, and a reduction in the MAIAC-SuomiNet RMS difference is observed compared to the raw MAIAC retrievals.

During the access review I asked the Authors to somewhat expand their explanation

of the algorithm they adopt, and the Authors have accommodated my request to some extent. However, the description of the methodology still looks vague in some respects, and does not allow to easily understand which computational steps are performed in order to attain the presented results. The discussion still uses a number of terms without defining them and takes for granted concepts that do not look particularly obvious. Apart from that, the results look good, and the paper may be published after the Authors account for the following comments.

- L64-65. What do you mean by "weak predictor"? And "binary partitioning" of what?

- L113-122. The problem with this paragraph is similar to the one I highlighted during the access review, namely that it uses many specialized terms without defining them. Therefore, this paragraph is not very informative to a reader who does not have a background in this type of methods (a condition that is probably not uncommon among the readership of AMT), and probably is also not very informative to a reader who does. In particular, the following aspects are not clear to me.

1. Let us suppose that we have a number of predictors (e.g. solar zenith angle, viewing zenith angle, AOD, etc.). When you say that the model "specifies a few recursive binary splits of predictors etc.", do you mean that it defines a threshold for each predictor and returns a different output depending on whether the predictor is above or below the threshold? And does the next level of the tree apply similar operations to the result of this first thresholding, and so on? If so, make this point clearer in your discussion. I had to look inside the references to understand this, but such a basic level of detail should be already understandable from your paper, without forcing the reader to peruse the references.

2. Who decides which predictors should be split and whether they should be split independently or according to certain logical combinations (AND, OR, etc.)? Is it the user or is it the training algorithm that makes this decision? In addition, if this is up to the training algorithm, how is the system trained? How is the cost function defined and

how are the system parameters adjusted?

3. Again, what do you mean by "weak learner"? How are multiple learners combined? Who decides what weight should be given to each learner, and how?

4. What is the role of "gradient" in gradient boosting? Gradient of what with respect to what?

- L130. Please define the “several hyperparameters related to the desired size and complexity of the model”. Plus, why "hyperparameters" and not simply "parameters"?

- L132. What do you mean by “nested comparison”? In particular, in what sense “nested”?

- L133. Could you provide a reference for Latin hypercube sampling, and possibly summarize what it essentially does?

- In general, the fundamental question I have about Section 3 is: if I want to replicate your study or apply your method to another problem - e.g., by writing my own code - what do I actually need to do? What are the computational steps involved?

- L146. I think “Shapely” should actually read “Shapley”

- L442. Some details of the reference appear to be missing.

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