Dear Editor:

As requested, we have responded to the reviewers' comments point-by-point. We have also made many responsive revisions to our manuscript which is substantially improved and includes greater clarity and detail on the machine-learning methods that we demonstrate in our empirical results. We would also like to point out that two previous referees stated in the access review that they felt the manuscript was an appropriate fit for inclusion in AMT. We have responded point-by-point and presented evidence of our scientific rigor in our data analysis. Finally, the ultimate test of the replicability of our empirical results lies in our use of the Open Access archival repository Zenodo where we have placed an irrevocable copy of all of our data and R code to regenerate all of our results. We thank the editor for their consideration of our work and for the opportunity to improve our manuscript through the AMT discussion process.

Response to RC1: We thank the reviewer for their helpful and specific comments. We completely agree that revisions contributing to more accessible methods and discussion sections will increase the impact of our work. We have added plain language descriptions that explain the meaning behind specialized terms that are commonly used in machine learning. More importantly, our more accessible language is to be taken in conjunction with the reproducible data + open source R code we have archived in the Zenodo open-access digital repository (`(https://doi.org/10.5281/zenodo.3568449)`, which will ease the integration of the methods and ideas that we employ into other atmospheric measurement applications. We have substantively revised the language we use to describe methods within the introduction, methods, and discussion sections. We add additional details, define terms coming from the field of machine learning, and seek to clarify the points of confusion raised by the reviewer. We have also included references to two strong introductory articles to guide readers that are seeking additional information on the inner workings of the machine-learning methods that we describe. We hope that the referee also sees the improved clarity and accessibility of the revised manuscript which we consider substantially improved. We include a point-by-point response to comments below:

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

   Response: We have added additional information to clarify this sentence in our introduction which previously read, "Gradient boosting involves fitting a large number of tree-based models where each subsequent tree is made a weak predictor of the error from the previous trees." The revised section now reads, "XGBoost involves fitting a large number of tree-based models. Each subsequent tree is fit to the error from the previous trees and the predictions of all the trees are added together. Each tree's prediction is multiplied by a shrinkage factor (or "learning rate") η, a number between 0 and 1. By adding successive trees, XGBoost descends the gradient of the loss function. The component trees use a recursive binary partitioning of the predictors that accommodates varying types and scales of predictor variables and is

robust to outliers (Elith et al., 2008).” To further clarify, both the shrinkage factor η and the use of random dropout (we explain the DART method elsewhere), are used to decrease overfitting that occurs when a model is directly applied to residual error. Our revisions have clarified that the binary partitioning used in regression trees is a splitting of the predictor variables.

2. - 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.

Response: We have added additional detail to explain specialized terms and have added a reference to an accessible introduction to regression trees. Below we include specific examples of how we have clarified points that were not clear to the reviewer.

3. In particular, the following aspects are not clear to me. 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.

Response: Following the reviewer’s suggestion, we have revised this section to explain in greater detail how tree-based regression models operate. While avoiding jargon, we also see our paper as an opportunity to inform a new readership on the terminology of machine learning and how it can be applied in earth sciences. It now reads: “For an introduction to regression trees, see Strobl et al. (2009). A regression tree is a model that specifies recursive binary splits of predictors and assigns a constant value to all cases that end up in the same terminal node (namely, their mean on the dependent variable). The algorithm chooses the splits across all predictors that minimize the variance of the residuals. The maximum number of splits within each tree (also known as the maximum depth) can be set as a hyperparameter. A set of multiple trees can be used for prediction by combining the outputs of the individual trees for each case. Such a set of trees can accommodate complex relationships including non-linearities and interactions while being robust to outliers. Boosting is a method of fitting a series of models iteratively, with each model fit on the residuals of the previous models. While each tree may individually perform relatively poorly at predicting the outcome (and thus is known as a “weak learner”), the combination of many trees can collectively describe complex relationships and account for the impact of many predictors. Further, because boosting includes

sequentially learning by combining many iteratively fit trees that address the error in previous trees, this technique performs well, achieving low testing error. The XGBoost package is a scalable gradient boosting implementation with additional features including penalties to avoid overfitting and optimized computational speed (Chen and Guestrin, 2016)." We believe our revisions achieve a reasonable balance of including sufficient descriptive information and useful references without swamping the reader with excessive detail.

4. 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?

Response: As clarified in our revised Statistical Methods section, the regression tree algorithm selects at each step the split across all predictors that minimizes the variance of the residuals. The selection of splits is done recursively and the number of splits (depth of the tree) is a hyperparameter that is tuned by the analyst (we discuss our hyperparameter tuning approach below). While a split upon another split of a tree constructs a logical AND statement, the addition of sufficient splits can approximate an OR statement. The user does not select split points.

5. 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?

Response: the "weak learner" term was explained in the revised manuscript. It now reads: "While each tree may individually perform relatively poorly at predicting the outcome (and thus is known as a "weak learner"), the combination of many trees can collectively describe complex relationships and account for the impact of many predictors. "

Instead of fitting a single large decision tree to all the data with many splits, which will perform poorly in prediction (having low bias but very high variance), the boosting approach learns slowly by fitting a smaller decision tree to the residuals of the model and slowly improving the model in areas where it does not perform well. In general, statistical learning approaches that learn slowly tend to perform well by producing both low bias and low variance.

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

Response: The gradient in question is the residuals (the observed values minus the predicted values), which are the gradient of squared-error loss with respect to the residuals. A gradient-boosting model adds trees in order to minimize this gradient.

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

Response: In machine learning, a hyperparameter is a configuration set before the learning process begins and that takes values that cannot be directly estimated from the data. Parameters, such as regression coefficients or split points in tree-based models, are estimated from the data. Simple algorithms like linear regression don't require hyperparameters while more complex algorithms may have several hyperparameters. For these more complex machine-learning algorithms, hyperparameters need to be predefined by researchers within a certain range of values. Often a set of appropriate hyperparameter values that result in improved performance are selected through a cross-validation process. We have added the following sentences to clarify, “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.”

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

Response: To avoid overfitting, it is important that all learning, including the selection of hyperparameter values, occurs within the training dataset. However, the evaluation of performance should still be done on data that was not used in algorithm training and thus requires cross-validation within the training dataset. This leads to nested cross-validation, where the training data is being further split in half in order to evaluate the performance of different hyperparameter values. Our revised section now reads, “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.”

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

Response: We use Latin Hypercube Sampling (LHS) to generate random combinations of parameter values. It is based on the Latin square design, which has a single sample in each row and column. One-dimensional LHS involves dividing your cumulative density function into n equal partitions; and then choosing a random data point in each partition. We are using a multi-dimensional LHS because we have six hyperparameters to tune across simultaneously. LHS helps to ensure that samples are representative of the real variability in the distribution. We have added additional explanation of the purpose of the Latin hypercube sampling and our section now reads, “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. - 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?

Response: We have taken several steps to assist our readers with replication. We have added substantially to the detailed description of our methods, and as we have listed in our manuscript, we have already placed full reproducible R code, including all computational steps, and our datasets in an 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 and will assist readers in replicating this study in new datasets or in applying these methods to other problems in atmospheric measurement science.

11. - L146. I think "Shapely" should actually read "Shapley"

Response: we have corrected this in the revised manuscript.

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

Response: we have corrected the reference which appeared to be missing journal information.

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&content=all&group=ICML.cc/2017/WHI&source=all

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.

Gradient Boosting Machine Learning to Improve Satellite-Derived Column Water Vapor Measurement Error

Allan C. Just[1*], Yang Liu[1], Meytar Sorek-Hamer[2,3], Johnathan Rush[1], Michael Dorman[4], Robert Chatfield[3], Yujie Wang[5,6], Alexei Lyapustin[6], Itai Kloog[4]

[1]Icahn School of Medicine at Mount Sinai, New York, USA
[2]Universities Space Research Association (USRA), California, USA
[3]NASA Ames Research Center, California, USA
[4]Ben-Gurion University of the Negev, Beer Sheva, Israel
[5]University of Maryland Baltimore County, Maryland, USA
[6]NASA Goddard Space Flight Center, Maryland, USA

*Correspondence to: Allan C. Just (*allan.just@mssm.edu)

Abstract. The atmospheric products of the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm include column water vapor (CWV) at 1 km resolution, derived from daily overpasses of NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) instruments aboard the Aqua and Terra satellites. We have recently shown that machine learning using extreme gradient boosting (XGBoost) can improve the estimation of MAIAC aerosol optical depth (AOD). Although MAIAC CWV is generally well validated (Pearson's R >0.97 versus CWV from AERONET sun photometers), it has not yet been assessed whether machine-learning approaches can further improve CWV. Using a novel spatiotemporal cross-validation approach to avoid overfitting, our XGBoost model with nine features derived from land use terms, date, and ancillary variables from the MAIAC retrieval, quantifies and can correct a substantial portion of measurement error relative to collocated measures at AERONET sites (26.9% and 16.5% decrease in Root Mean Square Error (RMSE) for Terra and Aqua datasets, respectively) in the Northeastern USA, 2000-2015. We use machine-learning interpretation tools to illustrate complex patterns of measurement error and describe a positive bias in MAIAC Terra CWV worsening in recent summertime conditions. We validate our predictive model on MAIAC CWV estimates at independent stations from the SuomiNet GPS network where our corrections decrease the RMSE by 19.7% and 9.5% for Terra and Aqua MAIAC CWV. Empirically correcting for measurement error with machine-learning algorithms is a post-processing opportunity to improve satellite-derived CWV data for Earth science and remote sensing applications.

1 Introduction

Water vapor represents a small but environmentally significant constituent of the atmosphere. The integrated water vapor from ground to space is defined as the column water vapor (CWV), in units of centimeters (i.e. precipitable water vapor) (Gao and Goetz, 1990). CWV has important applications in many fields, such as atmospheric correction of remote sensing images, Earth energy balance and global climate change, land surface temperature retrieval in thermal remote sensing, and astronomy. Thus, high resolution CWV values with global coverage have multiple uses in Earth science and remote sensing. CWV has been measured by multiple technologies and monitoring networks, including sun photometers, GPS sensors (e.g. SuomiNet), Aerosol Robotic Network (AERONET) sun photometers, and satellite remote-sensing. The AERONET sun photometer network measures CWV in approximately 400 stations worldwide, in channels centered at 940nm and provided to the user in Level 2, which is the highest data-quality level provided by AERONET (Pérez-Ramírez et al., 2014). The AERONET CWV data have been well validated with the U.S. Department of Energy Atmospheric Radiation Measurement Program (ARM) radiosonde observations and other ground-based retrieval techniques such as microwave radiometry (MWR) and SuomiNet GPS receivers, and do not observe any dependence of biases with the zenith angle (Pérez-Ramírez et al., 2014). AERONET CWV has been used in studies that examine aerosol optical, microphysical, and radiative properties in Africa (Adesina et al., 2014; Boiyo et al., 2019; Kumar et al., 2013), in the Brazilian tropics (Schafer et al., 2008) , and in Beijing and Kanpur (Wang et al., 2011). Global satellite-borne CWV is available at high resolution (1 km), from the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm derived from daily overpasses of NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) instruments aboard the Aqua and Terra satellites. The MAIAC CWV is computed using MODIS near-infrared (NIR) channels centered at 940nm. This method applies two ratios of channels to compute the water vapor transmittance, and then compute the amount of water vapor using look-up-tables (Lyapustin et al., 2014). The MAIAC CWV algorithm was validated against ground measurements of CWV from 265 AERONET stations worldwide, with a relatively strong association (Pearson's $R > 0.95$; root mean squared error [RMSE] < 0.25 cm; average accuracy of ±15%) (Martins et al., 2019). These datasets were collocated by averaging MAIAC values within 9 x 9 pixels and AERONET values ±30 minutes of the satellite overpass in cloud-free conditions. A significant upward trend ($p < .05$) for MAIAC TERRA was found over most regions, although this was not significant over the Northeastern USA. Globally, the highest average correlation between MAIAC CWV retrievals from both Aqua and Terra with AERONET CWV have been shown in Asia and both Northern and Southern regions of the USA.

In spite of the strong performance of MAIAC CWV in multiple locations, comparing it with collocated AERONET CWV, there may be opportunities to characterize and correct complex interactions and challenging conditions that increase satellite retrieval error. However, it has not yet been assessed whether machine-learning approaches can improve the estimation of satellite-borne CWV. We have recently demonstrated that machine learning using eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) can improve the estimation of MAIAC aerosol optical depth (AOD) parameters over AERONET stations (43% decrease in cross-validated RMSE) (Just et al., 2018). For an introduction to gradient boosted regression trees, please see the work of Elith et al. (2008). XGBoost involves fitting a large number of tree-based models. Each subsequent tree is fit to the error from the previous trees and the predictions of all the trees are added together. Each tree's prediction is multiplied by a shrinkage factor (or "learning rate") η, a number between 0 and 1. By adding successive trees, XGBoost descends the gradient of the loss function. The component trees use a recursive binary partitioning of the predictors that accommodates varying types and scales of predictor variables and is robust to outliers (Elith et al., 2008). An advantage of flexible algorithmic machine-learning approaches such as XGBoost is that they can model complex phenomena (Chen and He, 2015), including interactions of multiple features (e.g. retrieval angles, seasonality, and surface characteristics). The resulting prediction model can be used as an algorithm to reduce the retrieval error. Machine-learning tools for model interpretation can also help explain the contributions of these features to retrieval error and guide feature selection to build parsimonious models.

While the satellite data record continues to grow, the ground monitoring networks that can be used for validation and algorithmic measurement error correction of satellite retrieval products are still sparse. Collocated ground-satellite datasets may thus have important non-independent spatiotemporal structure if they rely on observations that occur in only a few locations. Flexible machine-learning models would overfit to the characteristics of these particular stations or the days when AERONET data are available if cross-validation assumed independence of observations. While machine-learning applications in aerosol research have begun to adopt group-K-fold cross validation for assessing model fit across fixed monitoring networks (Di et al., 2016), we propose a novel cross-validation approach taking into consideration data structure due to both fixed sites and correlation of observations from the same day.

The goals of this work are to 1) evaluate whether machine-learning gradient boosting models can improve satellite-based CWV retrievals, and 2) understand the contributions of different features as well as spatial and temporal structures of the ground station measures to

Author
Deleted: Elith

Author
Deleted: Gradient boosting

Author
Deleted: where

Author
Deleted: e

Author
Deleted: is made a weak predictor of

Author
Deleted: then averaged into the ensemble of previous trees after multiplication

Author
Deleted: less than

Author
Deleted: (this hyperparameter is also known as the learning rate and named *eta* in XGBoost) to avoid overfitting

Author
Deleted: the boosting algorithm

Author
Deleted: machine

Author
Deleted: provide

predictions of error in the estimated CWV. The data and machine-learning methods are described in section 2, followed by a discussion of the results in sections 3 and 4.

2 Description of the Data

In order to assess the agreement of the MAIAC estimates of CWV with those from AERONET, collocated datasets were built using MAIAC data where AOD was available (representing clear sky conditions) from both Terra and Aqua (separately) collocated to the nearest 1 km * 1 km grid centroid and the closest observation in time (no more than 60 minutes) with cloud-screened (version 2, level 2.0) (Smirnov et al., 2000) measures of CWV from the AERONET network of sun photometers over the Northeastern USA (including 13 states and the District of Columbia from Maine to Virginia). The study period included 10,247 observations (from 75 AERONET stations) for Terra (2000-2015) and 8,536 observations (from 71 stations) for Aqua (2002-2015). All analyses were performed for Terra and Aqua datasets separately. AERONET stations in the Northeast are largely urban and coastal (Fig. 1). We defined our target modeling parameter as the difference between MAIAC and AERONET CWV (ΔCWV = MAIAC CWV - AERONET CWV) such that any variation from zero indicated a component of measurement error that we sought to explain.

After exploratory scatterplots of ΔCWV versus time showed a temporal cluster of large outliers coming from a single AERONET station (City College of New York), observations from this site between 2007-06-17 and 2009-01-01 were dropped from further analysis, including 99 observations collocated for Terra and 95 observations for Aqua datasets. This particular period, which was flanked on both sides with months without values at that station, showed a clear deviation from the monitor's typical trend across the remainder of the study period.

The date range for the collocated Terra dataset was from 2000-02-25 to 2015-12-27, including observations from 3,024 unique days (52% of days during this interval). The date range for the collocated Aqua dataset was from 2002-07-04 to 2015-12-28, including observations from 2,627 unique days (53% of days during this interval).

3 Statistical Methods

We examined the use of XGBoost (Chen and Guestrin, 2016) for improving satellite-based MAIAC CWV retrievals and decreasing estimation error, as this method had previously outperformed two related supervised learning approaches using regression trees, namely gradient boosting and random forests, in a similar application (Just et al., 2018). The XGBoost algorithm is a popular implementation of boosted regression trees (Friedman, 2001). For an introduction to regression trees, see Strobl et al. (Strobl et al., 2009). A regression tree is a model that specifies recursive binary splits of predictors and assigns a constant value to all cases that end up in the same terminal node (namely, their mean on the dependent variable). The algorithm chooses the splits across all predictors that minimize the variance of the residuals. The maximum number of splits within each tree (also known as the maximum depth) can be set as a hyperparameter. A set of multiple trees can be used for prediction by combining the outputs of the individual trees for each case. Such a set of trees can accommodate complex relationships including non-linearities and interactions while being robust to outliers. Boosting is a method of fitting a series of models iteratively, with each model fit on the residuals of the previous models. While each tree may individually perform relatively poorly at predicting the outcome (and thus is known as a "weak learner"), the combination of many trees can collectively describe complex relationships and account for the impact of many predictors. Further, because boosting includes sequentially learning by combining many iteratively fit trees that address the error in previous trees, this technique performs well, achieving low testing error. The XGBoost package is a scalable gradient boosting implementation with additional features including penalties to avoid overfitting and optimized computational speed (Chen and Guestrin, 2016).

We end up with more parsimonious XGBoost models, i.e. fewer trees, by adopting the concept of 'dropout' from deep learning, in which individual learners are randomly dropped during training. Specifically, we used Dropout meets Additive Regression Trees (DART) (Rashmi and Gilad-Bachrach, 2015). Dropping trees helps to avoid the diminishing contributions and over-specialization of later trees in XGBoost. This is particularly important in our application given the low number of AERONET stations and relatively small size of the collocated datasets for machine-learning algorithms. 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-

Author
Deleted: ,

Author
Deleted: ,

Author
Deleted: a few

Author
Deleted: for

Author
Deleted: and may vary between applications

Author
Deleted: simple

Author
Deleted: is

Author
Deleted: a weak learner

Author
Deleted: is process of

Author
Deleted: XGBoost models were fit with tree dropout using

Author
Deleted: : Dropout meets Additive Regression Trees

Author
Deleted: The tree dropout implementation in DART

Author
Deleted: s

Author
Deleted: gradient boosted regression trees

Author
Deleted: ,

Author
Deleted: ,

Allan Just 3/10/2020 10:45 AM
Deleted: eight

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.

Prior to feature selection, initial analyses included 25 candidates features such as: MAIAC variables including an uncertainty parameter related to blue band surface reflectance, relative azimuth angle, and AOD; time trend (integer date); elevation; several land use terms from the National Land Cover Database 2011 aggregated to proportions within 1 km * 1 km grid cells as well as proportion of water within 5-15 km buffers; and distance to major water bodies (Great Lakes and the Atlantic Ocean). Feature engineering calculated candidate features based on spatial patterns in non-missing MAIAC data including the number of contiguous non-missing grid cells (clump size) and the number of non-missing observations in focal windows of side lengths from 30-510 km. Details on the data sources and feature engineering for all candidate features are included in Supplementary Materials. No external meteorology or assimilated data were included.

The contributions of each feature to cross-validated predictions were estimated from Shapley Additive Explanations (SHAP) values (Lundberg et al., 2018). These SHAP values form an additive feature attribution measure to interpret complex machine-learning models. SHAP values estimate the contributions of each feature to each individual prediction (for ΔCWV, this is in units of cm). Specifically, the SHAP value for a given predictor and a given observation is the difference in the output, i.e. a predicted ΔCWV, if the model is fit with or without the predictor. For each observation, the sum of all SHAP values, plus the bias term (the overall mean of ΔCWV), equals the prediction from the XGBoost model. The resulting matrix of SHAP values can be summarized to understand how a predictor contributes to the predictions. The mean absolute SHAP value across all observations summarizes the global feature importance and more local model interpretation is possible through exploratory data visualizations such as scatterplots of individual predictors versus their SHAP values.

Allan Just 3/9/2020 4:01 PM
Deleted: Each time we trained a model

Author
Deleted: nested

Allan Just 3/9/2020 4:03 PM
Deleted: ison, nested within the training data, of

Allan Just 3/9/2020 4:32 PM
Deleted: generated to be well-spaced across the range of potential hyperparameter values using

Allan Just 3/9/2020 4:33 PM
Deleted: this

Allan Just 3/9/2020 4:34 PM
Deleted: increase the efficiency of

Allan Just 3/9/2020 4:34 PM
Deleted: ing

Allan Just 3/9/2020 4:34 PM
Deleted: highly performant

Allan Just 3/9/2020 4:34 PM
Deleted: versus

Allan Just 3/10/2020 11:19 AM
Deleted: eight

Allan Just 3/9/2020 4:35 PM
Deleted:

Allan Just 3/9/2020 4:35 PM
Deleted: (

Allan Just 3/9/2020 4:35 PM
Deleted:)

Author
Deleted: e

Because a more parsimonious set of features can ease future efforts to build large spatiotemporal datasets for algorithmic correction, an initial feature selection approach was performed prior to evaluating overall model performance. Feature selection was performed in a randomly selected 20% subset of the data to avoid overfitting prior to later model evaluation steps. Within this subset, we evaluated both the mean absolute SHAP values as a measure of global feature importance within a full model with all 25 candidate features, as well as a recursive stepwise procedure. We adopted 5-fold cross-validation split by MAIAC stations to alleviate overfitting to spatial features of the relatively low number of unique stations. In each round of cross-validation, backwards feature selection was applied to rank and remove the features by increasing importance. Starting with the XGBoost model containing all 25 candidate features, the overall RMSE was calculated from the out-of-sample predictions after cross-validation. Then the feature importance was ranked by mean absolute SHAP values for all the features in the model from low to high. This step was repeated removing the least important feature at each step. After plotting the overall RMSE from the cross-validated predictions against the number of features, we selected the model with the lowest RMSE for Aqua and Terra separately. We then pooled the set of top-ranked features from both satellites to facilitate comparisons between the Aqua and Terra models examined in the full dataset.

Using the selected features, grouped ten-by-ten-fold cross-validation randomly splitting the data by both station and day was performed on the whole dataset. In each training iteration, all observations from one fold of stations and from one fold of days were withheld with the remaining dataset containing roughly 0.9*0.9 = 81% of the training data (a similar share of training data as in 5-fold cross-validation). However, for each combination of withheld data, predictions for evaluating model performance and the corresponding SHAP values were only made in the intersection of withheld days and monitors (~1% of the data). Thus predictions for each observation were made on a model trained without any observations from the same day or station (see Fig. 2). For comparison, we also evaluated model performance using grouped-5-fold cross-validation separately splitting the data by station or by day. Hyperparameter tuning of the XGBoost model was performed separately in each round of cross-validation.

While we used an aggregated measure of the mean absolute SHAP value for each feature as a measure of feature importance in our variable selection, we also plotted the out-of-sample SHAP in order to aid model interpretability. In particular, we plotted frequencies of SHAP values by variable and in bivariate scatterplots versus observed values.

Finally, we conducted an additional external validation of our final model by comparing both the original MAIAC CWV and our corrected CWV with an independent dataset of CWV measured by GPS-based stations in the SuomiNet dataset (Ware et al., 2000), within our Northeastern USA study region - many of which are quite distant from the AERONET sites. All SuomiNet stations use precision survey-quality dual-frequency GPS receivers and antennas. The water-lag derived CWV measures from GPS-based stations are generally considered to have excellent precision (5-10%), exceeding that from sun photometers (Pérez-Ramírez et al., 2014).

4 Results

4.1 Descriptive analysis of CWV and ΔCWV

The overall agreement of the original MAIAC CWV and AERONET CWV was quite good with a Pearson's correlation of 0.976 and 0.984 for Terra and Aqua, respectively, in agreement with the global MAIAC CWV validation (Martins et al., 2019). However, outlying values and a positive bias in Terra-derived MAIAC CWV particularly indicate a potential for improvement in MAIAC CWV relative to AERONET. The target parameter of ΔCWV (based on the difference between MAIAC and AERONET) was approximately symmetrically distributed and had a mean of 0.043 cm and -0.054 cm, and a standard deviation of 0.25 cm and 0.18 cm for the collocated Terra and Aqua datasets, respectively (Table 1). Descriptive scatterplots of the ΔCWV versus individual predictors showed some clear patterns prior to modeling (Fig. 3 and Supplementary Fig. S1). For example, there is a clear seasonal pattern with a larger SD of ΔCWV in the summer (0.32 cm and 0.25 cm for Terra and Aqua) when the SD of AERONET CWV is also highest (0.90 cm). This seasonal pattern and the positive bias for Terra (MAIAC CWV overestimates AERONET CWV) is seen to grow larger in more recent years (e.g. 2010-2015). This trend is related to the trend in MODIS Terra calibration, as previously reported (Martins et al., 2017).

4.2 Feature Selection and Model Performance

Variable selection using feature importance from SHAP was run in a 20% subset for both Terra and Aqua datasets. Using both global feature importance from a full model and a stepwise backward selection calculating RMSE at each step after ranking variable importance by mean absolute SHAP, we selected 6 features for the Terra model and selected 7 features for the Aqua model. The 4 features shared by both models were time trend (date represented as an integer), MAIAC CWV, blue band uncertainty, and MAIAC AOD. The other variables selected for Terra

were elevation and distance to major water body, and for Aqua were the proportion of forest in a 1x1 km square, relative azimuth angle, and the proportion of developed area in a 1x1 km square. Pooling these features from both satellites brought the original set of 25 features down to a more parsimonious set of 9 with little loss of model performance (results not shown).

Using the reduced feature set, we implemented the cross-validation in the full dataset to evaluate model performance. In the collocated Terra dataset, the predicted ΔCWV evaluated with the grouped monitor-by-day cross-validation (10*10 fold) explained 45.0% (R^2) of the variance in ΔCWV and reduced the RMSE from 0.252 cm (the root mean squared difference between MAIAC and AERONET CWV) to 0.184 cm, a 26.9% decrease in RMSE. In the collocated Aqua dataset, the predicted ΔCWV explained 24.1% of the variance in ΔCWV (R^2) and reduced the RMSE from 0.189 cm to 0.158 cm, a 16.5% decrease in RMSE.

The evaluation of model performance was substantively different depending on how the cross-validation strategy reflected the data structure. Ignoring the non-independence of the training data by site and withholding unique days for grouped 5-fold cross-validation (training on 80% of the data), RMSE for Terra was 0.146 and for Aqua was 0.145 (Table 2), a much better performance (smaller RMSE) that indicates over-fitting to the particular sites in the training dataset. Similarly, the RMSE from cross-validation split by station (and not by day) was also slightly lower than the RMSE from station-by-day cross-validation suggesting a much smaller degree of overfitting also to the specific dates in the training set.

After applying the XGBoost model, the measurement error of ΔCWV was corrected to be closer to zero, particularly for the largest magnitude ΔCWV values. For Terra and Aqua respectively, 87% and 93% of the ΔCWV observations beyond one standard deviation (outside of the dotted lines in Fig. 4, making up 24% of the collocated observations in Terra and 19% in Aqua) had lower measurement error (|ΔCWV|) by an average magnitude of 41% smaller in Terra and 53% smaller in Aqua after XGBoost correction.

We describe the variation in hyperparameters from XGBoost models across the 100 runs of the site-by-day 10*10 fold cross-validation. Greater variation in the selected hyperparameter

values across folds with very similar training datasets may indicate a lower impact on model performance (Supplementary Table S2).

4.3 Variable Importance Assessment

Although the final model had already been restricted to include only the top variables from our variable selection approach, we further interpreted variable importance and the contribution of these variables with SHAP values estimated in the grouped-cross-validation (at monitors and on days not included in the training data for each fold). SHAP values describe the additive contribution to the prediction from every variable for each observation.

The SHAP overview plot illustrated different patterns of feature importance in Terra and Aqua (Fig. 5). The rank of the mean absolute SHAP values suggested that the top key contributing variables to predicting the magnitude of ΔCWV in the Terra dataset were time trend (even though all of the data in the testing set were from days not included in the training data, there was still clear seasonality when plotting the SHAP estimates), the magnitude of the MAIAC CWV itself, the blue band uncertainty estimate from MAIAC, the MAIAC AOD, the distance to the nearest major water body, and the elevation. For the Aqua dataset, the blue band uncertainty ranked at the top, followed by the MAIAC AOD, the MAIAC CWV, the proportion of developed area in a 1x1 km square, time trend, the proportion of forest coverage in a 1x1 km square, and the relative azimuth angle. The SHAP values ranged from -0.52 to 0.82 cm for Aqua, and from -0.55 to 0.30 cm for Terra, aligning with the higher overall error in the Terra dataset.

For Terra, predicted ΔCWV values became larger in more recent years (Fig. 6.a). This suggests the observed positive bias has been getting stronger since ~2010. This trend was not observed in Aqua for which the time trend was a much weaker predictor. Similarly, a higher MAIAC CWV was also more likely to generate higher ΔCWV in Terra (a positive bias), and the influence was getting stronger along the time trend (Fig. 6.b). In contrast, in Aqua the model suggested that MAIAC CWV conservatively underestimated extreme values in both seasons, although the overall impact was weaker (SHAP values closer to zero) and more stationary across time compared to Terra.

The impact of the rest of the features was similar for both Terra and Aqua (Fig. 7). Some outlying large AOD values had negative effects on the ΔCWV. Larger blue band uncertainty, higher elevation, or relative azimuth angle around 45 and 145 degrees increased the error. The SHAP estimates of global feature and individual datum contributions clearly diagnoses two main factors: 1) changing calibration of MODIS Terra NIR bands at 940nm over time, resulting in a trend of CWV bias from Terra, and 2) growing underestimation of CWV with increase in AOD. MAIAC CWV retrieval neglects the effect of aerosol scattering, which increases the measured radiances and the band ratios, resulting in underestimation of CWV.

4.4 Prediction with new data

To predict into a new dataset, we refit our XGBoost models by again running our nested random hyperparameter tuning using DART tree dropout, this time on the entire training dataset. For both models fit to the Aqua and Terra datasets, the optimal set of hyperparameters (selected from the same set of 50 candidates) was the same, including both L1 and L2 regularization (alpha and lambda), the deepest trees we permitted (maximum depth of 9), and no more dropout (rate drop of 0) than the minimal random selection of one tree per model that had been fixed a priori (with the one drop option) (Table 3).

The resulting trained algorithm can generate ~3 million MAIAC CWV measurement error estimates per minute (on 4 cores) in new locations using the XGBoost predict function and these can be subtracted from the MAIAC CWV value to generate a corrected CWV estimate for downstream use.

4.5 Validation with SuomiNet GPS CWV

As an external validation, we applied our XGBoost models to MAIAC data in 1 km * 1 km grid cells containing SuomiNet GPS stations. We removed about 20 observations (0.1% of the merged datasets) with outlying CWV values above 9 cm which were almost all from SuomiNet site P776 in year 2011. The resulting validation dataset with collocated Terra or Aqua MAIAC CWV and SuomiNet CWV included 17,469 and 16,466 day-observations respectively from 57 SuomiNet stations (from years 2005 to 2015). SuomiNet CWV in the Terra collocated dataset had a mean of 1.57 ± 1.04 cm, while the Aqua collocated dataset had a mean of 1.50 ± 1.01 cm. The SuomiNet CWV had a more right-skewed distribution than MAIAC CWV.

After applying our correction, the MAIAC CWV had lower RMSE versus SuomiNet CWV compared with the raw MAIAC CWV in 53/57 sites for Terra and 56/57 sites for Aqua. The RMSE for agreement with SuomiNet CWV in the full validation dataset improved by 19.7% from 0.28 to 0.22 cm for Terra, and by 9.5% from 0.25 to 0.23 cm for Aqua. The Pearson's correlations of MAIAC CWV with SuomiNet CWV were improved by 1 percentage point from 0.969 to 0.978 for the Terra collocated dataset, and by 0.4 percentage points from 0.974 to 0.978 for the Aqua dataset. Plotting the RMSE after correction at SuomiNet locations (grid cells where we make predictions), we observed higher RMSE (worse performance) near Lake Ontario and on the Atlantic coastline (Fig. 8.a). Most sites show improved RMSE after correction except 4 sites for Terra and 1 site for Aqua (Fig 8.b).

Another goal of addressing measurement error in satellite retrievals is to improve the comparability of different instruments. Given changing atmospheric conditions within the same day between overpass times (Terra in late morning and Aqua in early afternoon), we use the corresponding two SuomiNet CWV measures to estimate the expected agreement. When restricting to days with both Terra and Aqua CWV observations collocated with SuomiNet observations, we had 9,940 station-days with all four measures. Raw MAIAC CWV had a Pearson correlation of 0.975 between Terra and Aqua, and after applying our correction this increased to 0.977, although this was still slightly below that of the two corresponding within-day SuomiNet CWV measures with a correlation of 0.982. We demonstrate that our algorithmic correction slightly improves on the already excellent agreement of MAIAC CWV from Terra versus Aqua, but is still not quite as close as comparing pairs of within-day measures from the same ground instruments.

5 Discussion

The Northeastern USA exhibits large seasonal variation in CWV. While satellite retrievals using the MAIAC algorithm are overall excellent at estimating CWV, they also have seasonality in their measurement error versus ground measurements from AERONET sun photometers. We show this measurement error has notable heteroscedasticity (larger errors with greater CWV) and has been worsening, with time, for data derived from Terra. Satellite retrievals using MODIS and similar platforms have considerable strengths for measurement of CWV based on their global daily coverage and reconstruction of longer-term records during the satellite era. Our analysis demonstrates that gradient boosting with XGBoost and features including satellite retrieval quality assurance, aerosol optical depth estimates, land use terms, and time trends can substantially refine satellite-derived retrievals of 1 km * 1 km resolution CWV compared

with sun photometer measures of CWV on test days and at sites that were withheld from training data. Even with this rigorous cross-validation, our model explains 45.0% of the measurement error from Terra CWV (R^2), and 24.1% of the measurement error in Aqua CWV. This is an impressive proportion of the difference of MAIAC and AERONET CWV to explain given that the MAIAC CWV is already quite accurate with a ΔCWV standard deviation of only 0.252 cm and 0.189 cm for Terra and Aqua respectively, in spite of comparing a 1 km * 1 km satellite retrieval with point measurements from the AERONET sun photometers.

Strategies for model training and cross-validation of powerful algorithmic predictive methods need to reflect the structure of the underlying data and the intended use of prediction models - otherwise overfitting may lead to an inaccurate assessment of model performance. Given the sparsity of the collocated AERONET data, we decided to assess performance in cross-validation that mimicked the desire to predict to new places (without AERONET stations) and on dates without AERONET data (e.g. when sun photometers are out of service for recalibration).

While our XGBoost models are complex ensembles of one hundred boosted regression trees, we use the powerful new SHAP method for interpretation of the importance of each variable and their contributions to individual predictions. Contextualizing the magnitude of the SHAP value (for each variable) and examining the SHAP-based contribution in visualizations along with the feature value distribution can also hint where retrieval algorithms can be modified for better results. For example, although the measurement error was lower for Aqua, scatterplots for the top two variables by SHAP suggest that MAIAC may underestimate CWV when the blue band uncertainty is very low and may underestimate CWV at higher AOD values. For Terra, the date as an integer is the most important feature, even though our cross-validation approach meant that all SHAP values were estimated for predictions made on dates that did not occur in the training data. Based on the SHAP plots, the date predictor describes seasonal and long-term trends related to an emerging positive bias for Terra that is worse in the summertime.

Demonstrating that there is an improvement in the agreement of corrected MAIAC CWV with the SuomiNet measures is a strong validation for several reasons. First, the SuomiNet stations offer a well-validated measure of CWV that relies on a different principle (tropospheric delay) from the sun photometry of the AERONET and the MODIS satellite retrieval of the MAIAC algorithm. The second strength of this validation is that the SuomiNet validation occurred at

locations that are unique (not included in the training data from AERONET sites), including many that are far away from the largely coastal AERONET stations in the Northeastern USA. Although Terra CWV also had a larger measurement error versus SuomiNet CWV measures than Aqua CWV, after our correction using XGBoost, the updated MAIAC CWV for Terra and Aqua both had lower RMSE values of 0.221 cm and 0.226 cm versus SuomiNet stations - suggesting that we may have achieved parity and perhaps reached the limits of this approach to correct for the sources of measurement error we considered in comparing this satellite retrieval product with point measures from ground stations.

Strengths of our empirical machine-learning approach include a fast algorithm that uses only a few variables, primarily already included in the MAIAC retrieval suite and derived land use terms, to correct measurement error in CWV. Limitations of using MODIS-derived CWV from MAIAC include the availability of few measurements per day (versus geostationary satellites) and restriction to cloud-free and daytime values. Our measurement error model has not yet been evaluated for how well it would have worked in a region with substantially fewer AERONET stations or very different climate conditions.

6 Conclusions

Empirically correcting for measurement error with machine-learning algorithms is a relatively easy post-processing opportunity to improve satellite-derived CWV data quality for Earth science and remote sensing applications. Furthermore, the use of machine-learning interpretation tools points to potential sources of measurement error (e.g. a positive bias in CWV retrievals from Terra that is worse in more recent years) that can help when refining satellite retrieval strategies. We demonstrate that a parsimonious nine-predictor XGBoost model for updating satellite-based column water vapor from the MAIAC retrieval based on AERONET values can decrease measurement error as validated at an independent network of ground sensors across the North Eastern USA.

Author contribution

AJ designed the study and supervised analysis. YL carried out the analysis including developing code and running the models. YW and AL provided MAIAC data and valuable guidance. JR, MD and IK created the analytic datasets. RC and MSH provided guidance on remote sensing principal approaches. AJ and MSH prepared the manuscript with contributions from all co-authors.

The authors declare that they have no conflict of interest.

Acknowledgments

We thank Kodi B. Arfer for assistance with the XGBoost hyperparameter tuning implementation. We thank the AERONET federation PIs and their staff for establishing and maintaining the 75 sun photometer sites used in this investigation available from https://aeronet.gsfc.nasa.gov/. We thank the UCAR and the SuomiNet program for the university-based GPS network data available at https://www.suominet.ucar.edu/. MAIAC data were downloaded from ftp://dataportal.nccs.nasa.gov/DataRelease on 2016-10-16. Elevation data was obtained from the National Elevation Dataset at https://catalog.data.gov/dataset/usgs-national-elevation-dataset-ned. Land use data were obtained from the National Land Cover Database 2011 at https://www.mrlc.gov/.

All data and code to reproduce the analyses in this study are available and archived at https://doi.org/10.5281/zenodo.3568449

This research was funded by NIH grants UH3 OD023337 and P30 ES023515 and grant 2017277 from the Binational Science Foundation. A.C.J. was supported by NIH grant R00 ES023450.

References

Adesina, A. J., Kumar, K. R., Sivakumar, V. and Griffith, D.: Direct radiative forcing of urban aerosols over Pretoria (25.75°S, 28.28°E) using AERONET Sunphotometer data: first scientific results and environmental impact., J. Environ. Sci. (China), 26(12), 2459–2474, doi:10.1016/j.jes.2014.04.006, 2014.

Boiyo, R., Kumar, K. R., Zhao, T. and Guo, J.: A 10-Year Record of Aerosol Optical Properties and Radiative Forcing Over Three Environmentally Distinct AERONET Sites in Kenya, East Africa, J. Geophys. Res. Atmos., 124(3), 1596–1617, doi:10.1029/2018JD029461, 2019.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD' ' ’16, pp. 785–794, ACM Press, New York, New York, USA., 2016.

Chen, T. and He, T.: Higgs boson discovery with boosted trees., NIPS 2014 workshop on high-energy physics and machine learning, 69–80, 2015.

Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y. and Schwartz, J.: Assessing PM2.5 Exposures with High Spatiotemporal Resolution across the Continental United States., Environ. Sci. Technol., 50(9), 4712–4721, doi:10.1021/acs.est.5b06121, 2016.

Elith, J., Leathwick, J. R. and Hastie, T.: A working guide to boosted regression trees., J. Anim. Ecol., 77(4), 802–813, doi:10.1111/j.1365-2656.2008.01390.x, 2008.

Friedman, J. H.: Greedy function approximation: a gradient boosting machine., Ann. Statist., 29(5), 1189–1232, doi:10.1214/aos/1013203451, 2001.

Gao, B.-C. and Goetz, A. F. H.: Column atmospheric water vapor and vegetation liquid water retrievals from Airborne Imaging Spectrometer data, J. Geophys. Res., 95(D4), 3549, doi:10.1029/JD095iD04p03549, 1990.

Just, A. C., De Carli, M. M., Shtein, A., Dorman, M., Lyapustin, A. and Kloog, I.: Correcting Measurement Error in Satellite Aerosol Optical Depth with Machine Learning for Modeling PM2.5 in the Northeastern USA., Remote Sens (Basel), 10(5), doi:10.3390/rs10050803, 2018.

Kumar, K. R., Sivakumar, V., Reddy, R. R., Gopal, K. R. and Adesina, A. J.: Inferring wavelength dependence of AOD and Ångström exponent over a sub-tropical station in South Africa using AERONET data: influence of meteorology, long-range transport and curvature effect., Sci. Total Environ., 461-462, 397–408, doi:10.1016/j.scitotenv.2013.04.095, 2013.

Lundberg, S. M., Erion, G. G. and Lee, S.-I.: Consistent Individualized Feature Attribution for Tree Ensembles, arXiv, arXiv:1802.03888, 2018.

Lyapustin, A., Alexander, M. J., Ott, L., Molod, A., Holben, B., Susskind, J. and Wang, Y.: Observation of mountain lee waves with MODIS NIR column water vapor, Geophys. Res. Lett., 41(2), 710–716, doi:10.1002/2013GL058770, 2014.

Martins, V. S., Lyapustin, A., de Carvalho, L. A. S., Barbosa, C. C. F. and Novo, E. M. L. M.: Validation of high-resolution MAIAC aerosol product over South America, J. Geophys. Res. Atmos., 122(14), 7537–7559, doi:10.1002/2016JD026301, 2017.

Martins, V. S., Lyapustin, A., Wang, Y., Giles, D. M., Smirnov, A., Slutsker, I. and Korkin, S.: Global validation of columnar water vapor derived from EOS MODIS-MAIAC algorithm against the ground-based AERONET observations, Atmos. Res., 225, 181–192, doi:10.1016/j.atmosres.2019.04.005, 2019.

Pérez-Ramírez, D., Whiteman, D. N., Smirnov, A., Lyamani, H., Holben, B. N., Pinker, R., Andrade, M. and Alados-Arboledas, L.: Evaluation of AERONET precipitable water vapor versus microwave radiometry, GPS, and radiosondes at ARM sites, J. Geophys. Res. Atmos., 119(15), 9596–9613, doi:10.1002/2014JD021730, 2014.

Rashmi, K. V. and Gilad-Bachrach, R.: DART: Dropouts meet Multiple Additive Regression Trees., PMLR, 38, 489–497, 2015.

Schafer, J. S., Eck, T. F., Holben, B. N., Artaxo, P. and Duarte, A. F.: Characterization of the optical properties of atmospheric aerosols in Amazônia from long-term AERONET monitoring (1993–1995 and 1999–2006), J. Geophys. Res., 113(D4), doi:10.1029/2007JD009319, 2008.

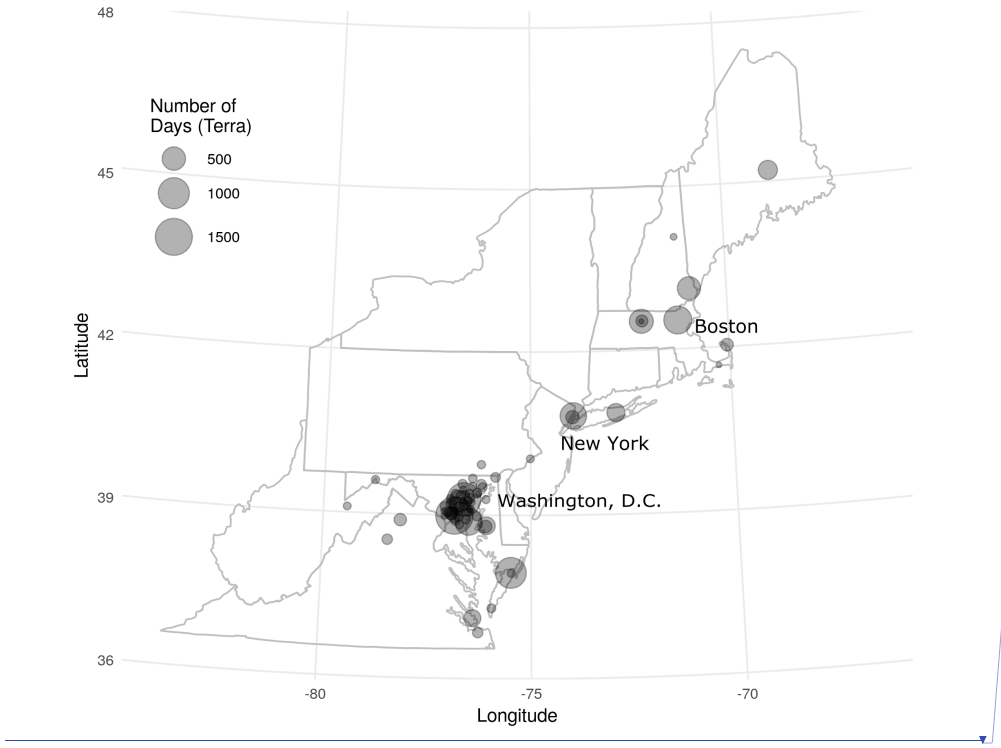
Smirnov, A., Holben, B. N., Eck, T. F., Dubovik, O. and Slutsker, I.: Cloud-Screening and Quality Control Algorithms for the AERONET Database, Remote Sens. Environ., 73(3), 337–349, doi:10.1016/S0034-4257(00)00109-7, 2000.

Stein, M.: Large Sample Properties of Simulations Using Latin Hypercube Sampling, Technometrics, 29(2), 143–151, doi:10.1080/00401706.1987.10488205, 1987.

Strobl, C., Malley, J. and Tutz, G.: An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests., Psychol. Methods, 14(4), 323–348, doi:10.1037/a0016973, 2009.

Wang, S., Fang, L., Gu, X., Yu, T. and Gao, J.: Comparison of aerosol optical properties from Beijing and Kanpur, Atmos. Environ., 45(39), 7406–7414, doi:10.1016/j.atmosenv.2011.06.055, 2011.

Ware, R. H., Fulker, D. W., Stein, S. A., Anderson, D. N., Avery, S. K., Clark, R. D., Droegemeier, K. K., Kuettner, J. P., Minster, J. B. and Sorooshian, S.: Suominet: A real–time national GPS network for atmospheric research and education, Bull. Amer. Meteor. Soc., 81(4), 677–694, doi:10.1175/1520-0477(2000)081<0677:SARNGN>2.3.CO;2, 2000.



Allan Just 4/16/2020 1:22 AM

Deleted:

Figure 1. Study region in Northeastern and Mid-Atlantic USA with 75 unique AERONET stations showing the number of days with observations from the collocated Terra dataset.

Fold	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
Station 1	Training				Training					
Station 2										
Station 3										
Station 4				Testing						
Station 5	Training				Training					
Station 6										
Station 7										
Station 8										
Station 9										
Station 10										

Figure 2. Example of training (blue) and testing (brown) datasets of one fold in ten-by-ten-fold cross-validation. Prediction models are only evaluated on days and at stations that were not used in model training to avoid overfitting.

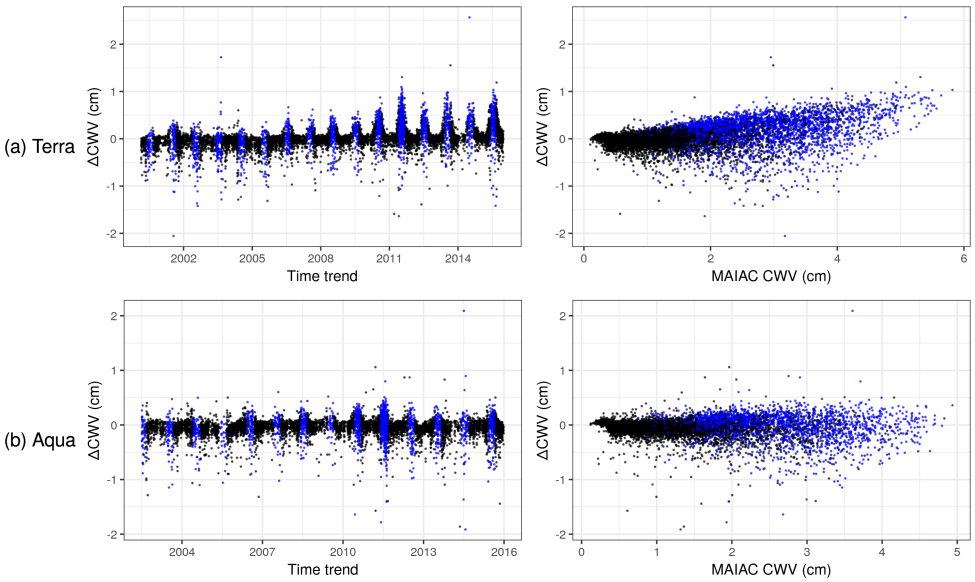


Figure 3. Scatterplots of ΔCWV versus time trend and MAIAC CWV in Terra (a) and Aqua (b). Observations in the summer months (June - August) are colored in blue.

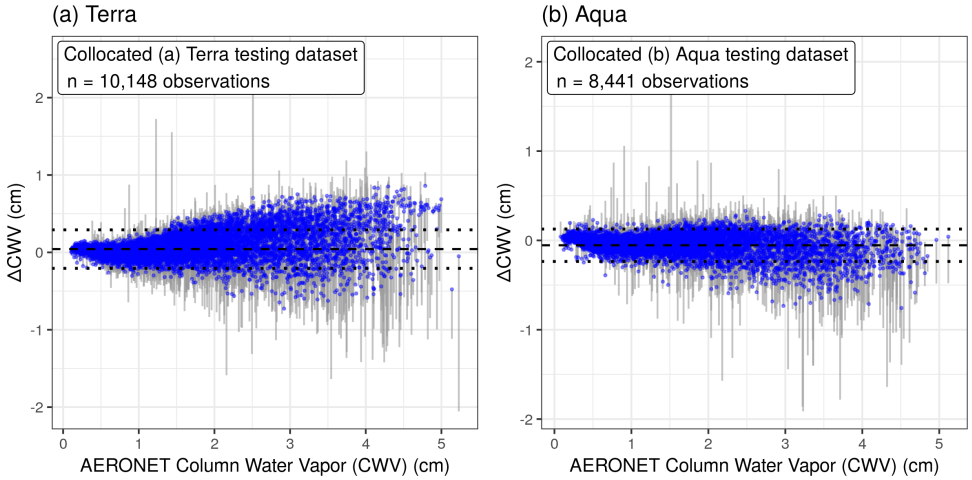


Figure 4. The difference between MAIAC and AERONET CWV values (ΔCWV) was reduced in cross-validation of collocated (a) Terra and (b) Aqua data. The corrected values of ΔCWV are shown with blue points, segments connect back to the measurement error from the raw ΔCWV. The dotted lines show one standard deviation from the mean (the dashed line near zero).

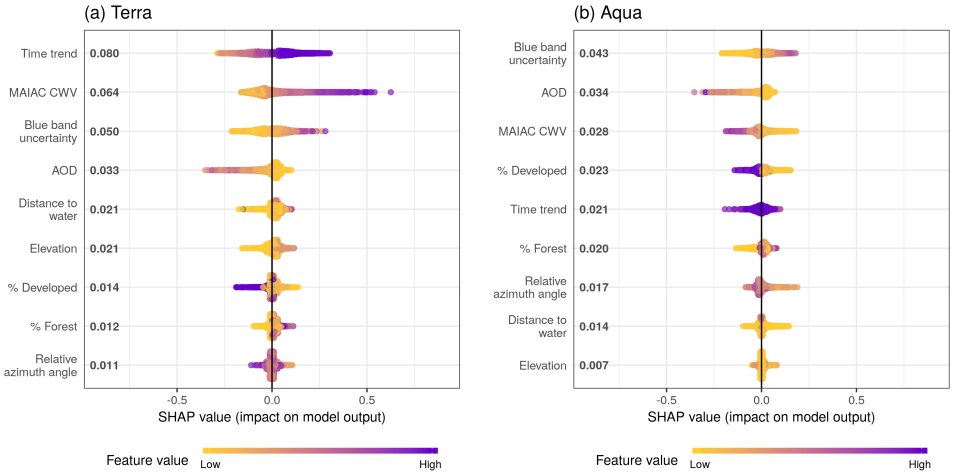


Figure 5. Sina plots show the distribution of feature contributions to predictions of CWV measurement error using SHAP values of each feature for every observation. The x-axis is set between -1 and 1 to facilitate comparison across subpanels

showing models for Terra and Aqua datasets. Features were ordered on the y-axis by their mean absolute SHAP values over all observations (bold on the right of the variable names, units are the same as ΔCWV predictions in cm). The color is scaled to the feature value (purple high, yellow low).

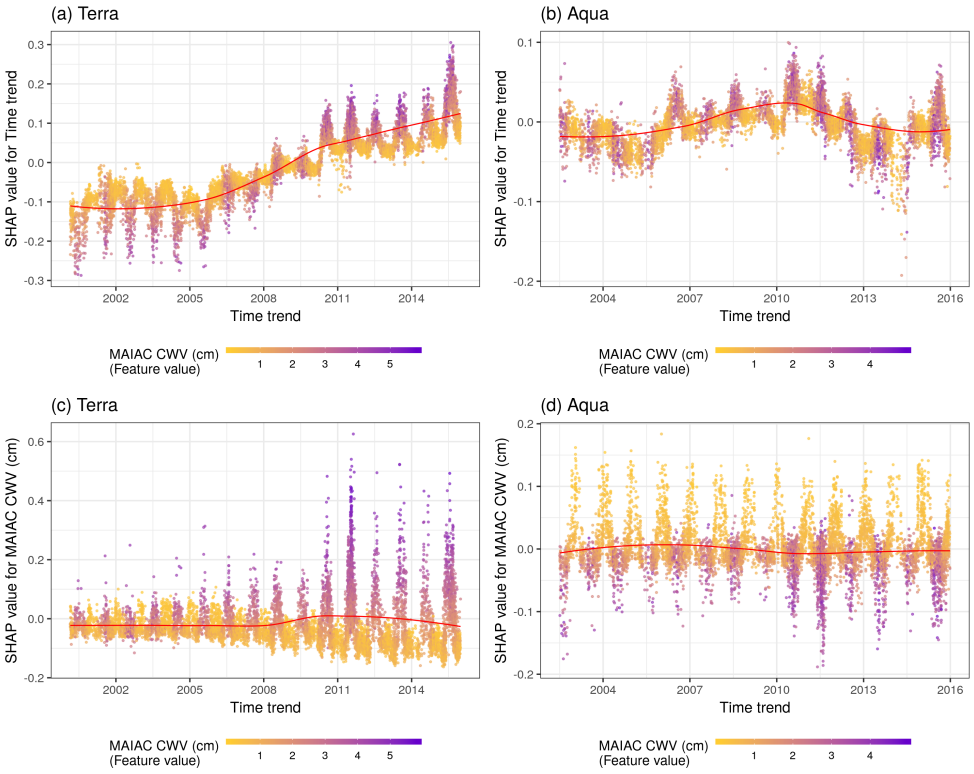


Figure 6. SHAP values showing the contribution of the time trend to predictions for Terra (a) and Aqua (b). The color represents the MAIAC CWV for each observation (purple high, yellow low). The LOESS (locally estimated scatterplot smoothing) curve is overlaid in red. Terra (c) and Aqua (d) SHAP values showing the contribution of the MAIAC CWV to predictions of CWV measurement error shown across the time period of the study. Note distinct y-axis scales for Terra and Aqua datasets. The color represents the MAIAC CWV for each observation (purple high, yellow low).

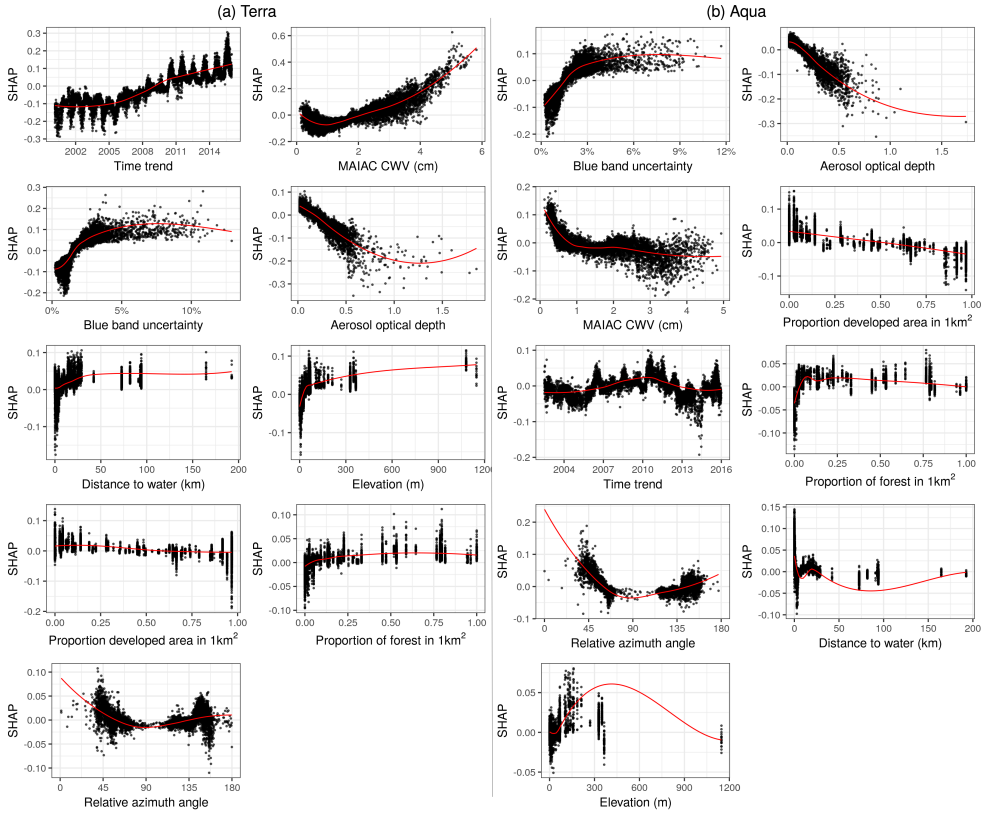


Figure 7. Descriptive scatterplots of the features versus their SHAP scores approximating their contribution to the predictions for ΔCWV (cm) on the y-axis. Subplots are ordered by overall variable importance (mean absolute SHAP score, see Fig. 5) by satellite.

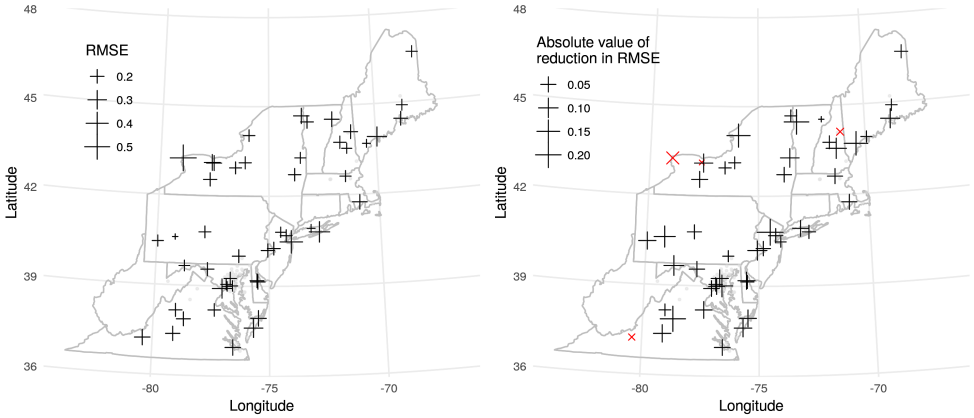


Figure 8. (a) RMSE between algorithmically-corrected Terra MAIAC CWV and GPS-based CWV for each SuomiNet station shown as crosses. AERONET sites used to train the model are shown as points. (b) the difference in RMSE versus GPS CWV using the corrected MAIAC CWV relative to using the original MAIAC CWV by SuomiNet station. The four sites (out of 57 total) having higher (worse) RMSE after correction are shown with red X symbols.

Table 1. Descriptive Statistics of MAIAC, AERONET CWV and the ΔCWV for Terra and Aqua by Season.

Terra

Mean (SD) (cm)	Spring (N = 2,265)	Summer (N = 3,150)	Fall (N = 3,101)	Winter (N = 1,632)	Total (N = 10,148)
MAIAC CWV	1.18 ± 0.82	2.71 ± 0.94	1.39 ± 0.82	0.56 ± 0.30	1.62 ± 1.12
Aeronet CWV	1.20 ± 0.80	2.54 ± 0.90	1.40 ± 0.79	0.60 ± 0.34	1.58 ± 1.04
ΔCWV	-0.014 ± 0.192	0.172 ± 0.322	-0.009 ± 0.202	-0.032 ± 0.095	0.043 ± 0.249

Aqua

Mean (SD) (cm)	Spring (N = 1,921)	Summer (N = 2,276)	Fall (N = 2,715)	Winter (N = 1,529)	Total (N = 8,441)
MAIAC CWV	1.12 ± 0.74	2.49 ± 0.84	1.33 ± 0.73	0.55 ± 0.27	1.45 ± 0.98
Aeronet CWV	1.16 ± 0.78	2.52 ± 0.90	1.42 ± 0.78	0.60 ± 0.33	1.51 ± 1.01
ΔCWV	-0.049 ± 0.150	-0.027 ± 0.253	-0.086 ± 0.159	-0.047 ± 0.101	-0.054 ± 0.181

Note. Means and standard deviations (units of cm) are shown for three-month seasons (Spring: MAM; Summer: JJA; Fall: SON; Winter: DJF) across all the years and the total.

Table 2. Predictive Performance in the Testing Dataset Comparing Three Cross-validation Strategies

	Terra dataset	Aqua dataset
Overall variation	SD 0.25 cm	SD 0.19 cm
Split by day (5 fold)	RMSE 0.15 (57.8%)	RMSE 0.14 (76.3%)
	R^2 = 65.6%	R^2 = 36.5%
Split by station (5 fold)	RMSE 0.18 (71.4%)	RMSE 0.16 (83.8%)
	R^2 = 47.5%	R^2 = 23.5%
Split by station and day (10*10 fold)	RMSE 0.18 (73.1%)	RMSE 0.16 (83.5%)
	R^2 = 45.0%	R^2 = 24.1%

Note. The relative percentage of RMSE compared to overall variation (SD) is listed beside RMSE.

Table 3. Hyperparameters for the Fully Trained XGBoost Model

	Selected values
eta	0.44
max_depth	9
gamma	0.099
lambda	38
alpha	0.0023
rate_drop	0
one_drop (fixed a priori)	True
nrounds (fixed a priori)	100