*GEOPHYSICAL RESEARCH LETTERS*

Supporting Information for

# Retrieving precipitable water vapor over land from satellite passive microwave radiometer measurements using automated machine learning

Xinran Xia[1,2,†], Disong Fu[2,3,†], Wei Shao[1], Rubin Jiang[2,4], Shengli Wu[5], Peng Zhang[5], Dazhi Yang[6], Xiangao Xia[2,3,4,*]

[1] School of Hydrology and Water Resources, Nanjing University of Information Science and Technology, Nanjing 210044, China

[2] LAGEO, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

[3] University of Chinese Academy of Sciences, Beijing, China

[4] Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters (CIC-FEMD), Nanjing University of Information Science & Technology, Nanjing 210044, China

[5] National Satellite Meteorological Center, Chinese Meteorological Administration, Beijing 100089, China

[6] School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin 150001, China

[†] Xinran Xia and Disong Fu contributed equally to this work.

Corresponding author: Xiangao Xia, xxa@mail.iap.ac.cn.

**Contents of this file**

Figures S1 to S4

Tables S1

**Text S1.**

Amongst the "shallow" network structure, tree-based methods have been repeatedly shown to be superior in terms of performance (Belgiu and Drăguţ, 2016; Ma et al., 2022). To that end, three tree-based ML methods, namely, random forest (RF), extreme gradient boosting (XGBOOST), and light gradient boosting machine (LGBM), are selected as the candidate methods. In short, RF is an ensemble learning method that combines the outputs of multiple base decision trees to make final predictions. Each decision tree is built by recursively partitioning the data based on the value ranges of different features. RF models have advantages in handling high-dimensional data, outliers and missing data (Liu et al., 2022; Lundberg et al., 2020). XGBOOST is also an ensemble learning framework, which seeks to build an ensemble of weak decision trees and combine them using the gradient boosting technique. Each subsequent tree corrects the discrepancies between the prediction of the previous tree and the target value. It incorporates regularization techniques to prevent overfitting and has gained popularity for its high performance (Chen and Guestrin, 2016; Fu et al., 2023). LGBM is another gradient boosting framework that aims to offer faster training speed and lower memory usage compared to other implementations. It incorporates a technique called "gradient-based one-side sampling" to select the most informative samples during the tree building process. Moreover, the histogram-based gradient estimation, which leverages the advantage of binning to compute efficiently, is applied in LGBM (Choi et al., 2023).

**Text S2.**

PWV retrievals by the RF model with two sets of learning features are compared to the out-of-sample GPS PWV observations in Figure S2, in the form of scatter plots. In these scatter plots, the number of samples within a neighborhood is represented by the color

of that neighborhood; the identity line and the least-squares linear fit line are also presented; lastly, several popular summary statistics that gauge different aspects of prediction quality are listed.

Based on the spread reduction of scatter points in Figure S2, the improvements in prediction accuracy due to new features are clearly evident. Quantitatively, the RMSE decreases from 5.43 mm to 3.76 mm and the mean absolute error (MAE) decreases from 4.00 mm to 2.76 mm in two cases of controlled experiments, respectively, and $R^2$ increases from 0.75 to 0.89. Collectively, these indicators confirm that the inclusion of the new features significantly is able to raise the accuracy of PWV retrieval by approximately 30%. Readers are noted that similar comparison experiments have been performed on the other two models (XGBOOST and LGBM), though the results are not shown here for brevity, the improved performance can also be observed.

In comparison with RF, LGBM and XGBOOST are updated tree-based models with significant improvement in accuracy and computational efficiency. Figure S3 shows that the results of XGBOOST and LGBM models validated against the out-of-sample GPS PWV observations. In these two models, the $R^2$ values, as compared to that of RF, both increase to 0.92. The RMSE and MAE values lower to 3.1 mm and 2.2 mm, respectively. In this regard, one may conclude that the utilization of these two models leads to a further enhancement in PWV retrieval accuracy.

As another important way to obtain PWV worldwide, IGRA2 data could be used for independent validation. Figure S4 depicts the corresponding results, in that, the accuracy of PWV retrieval by all three ML models with IGRA2 as a baseline. Both the XGBOOST and LGBM models show an improvement of 0.03 in $R^2$ and a reduction of approximately 0.5 mm in RMSE as compared to RF model. Among them, the PWV generated by LGBM model exhibits the best consistency with IGRA2 PWV data, with an RMSE of 3.64 mm, an $R^2$ of 0.87 and an MAE of 2.71 mm.

**Text S3.**

ML models are usually perceived as "black boxes," because of the low interpretability of the regression mechanisms in contrast to conventional statistical counterparts. To understand how the ML model works and gets the predicted values, some methods that seek to explain the feature importance have been devised. In this work, the SHapley Additive exPlanations (SHAP) method is used to calculate the marginal contribution of each feature in the model (Lundberg and Lee, 2017). SHAP method assumes that the original ML model can be approximately interpreted by s a linear combination of variables:

$$f(x) = \theta_0 + \sum_{i=1}^{N} \theta_i x_i \qquad (1)$$

where N is the number of input features, $\theta_i$ is the contribution of feature $i$ and $x$ is an input sample. In this way, traditional statistical frameworks for model interpretability could be of use.

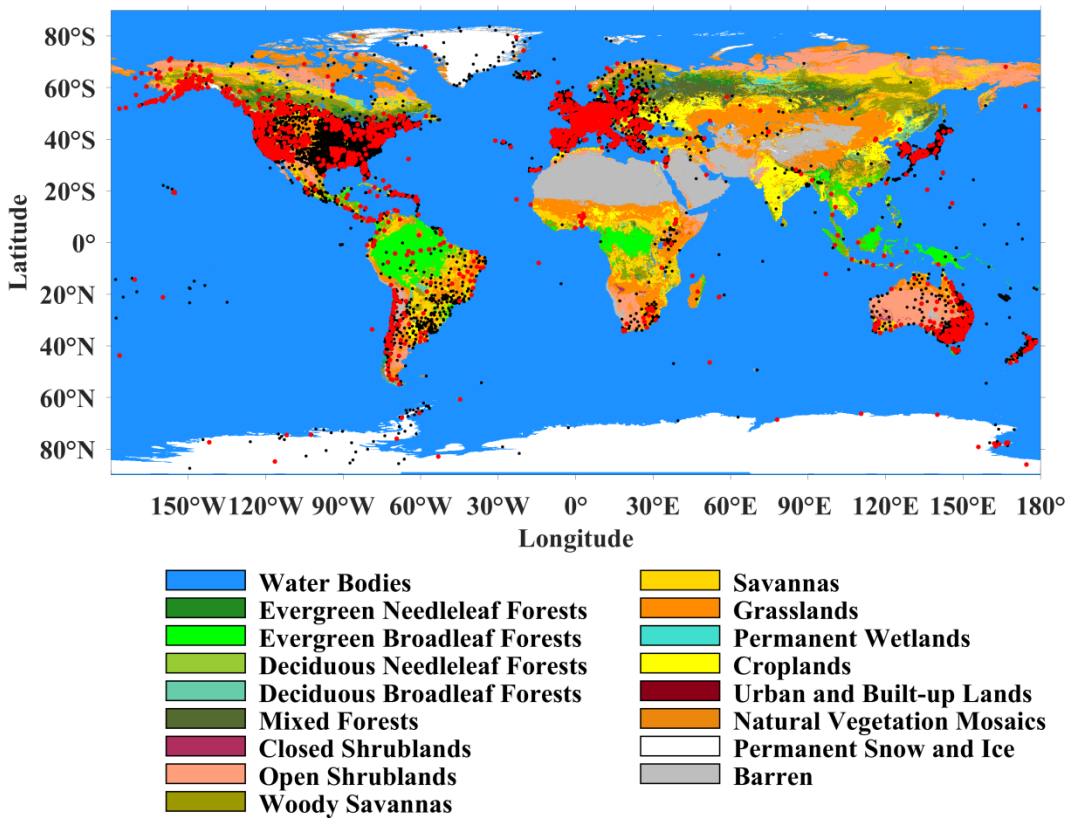


**Figure S1.** Spatial distribution of 12552 Enhanced GPS PWV sites over the MODIS IGBP global land cover map, contain training sites (black dots) and test sites (red dots), used for developing and testing the AMSR2 PWV retrieval algorithm.

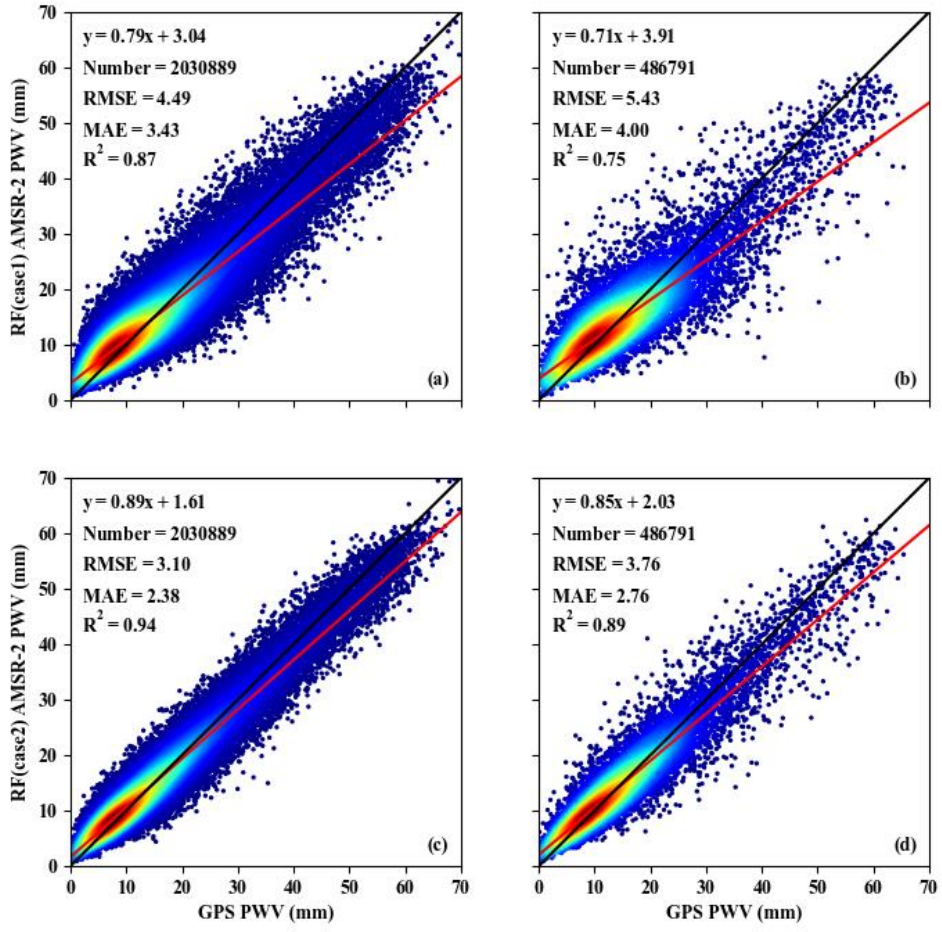


**Figure S2.** Comparisons between AMSR2 PWV retrievals based on 2 RF models validated by GPS PWV, from top (case1: RF model without additional features in training data (a) and test data (b)) to bottom (case2: RF with additional $F^H$, $P$, IGBP, and DOY as new features in training data (c) and test data(d)).

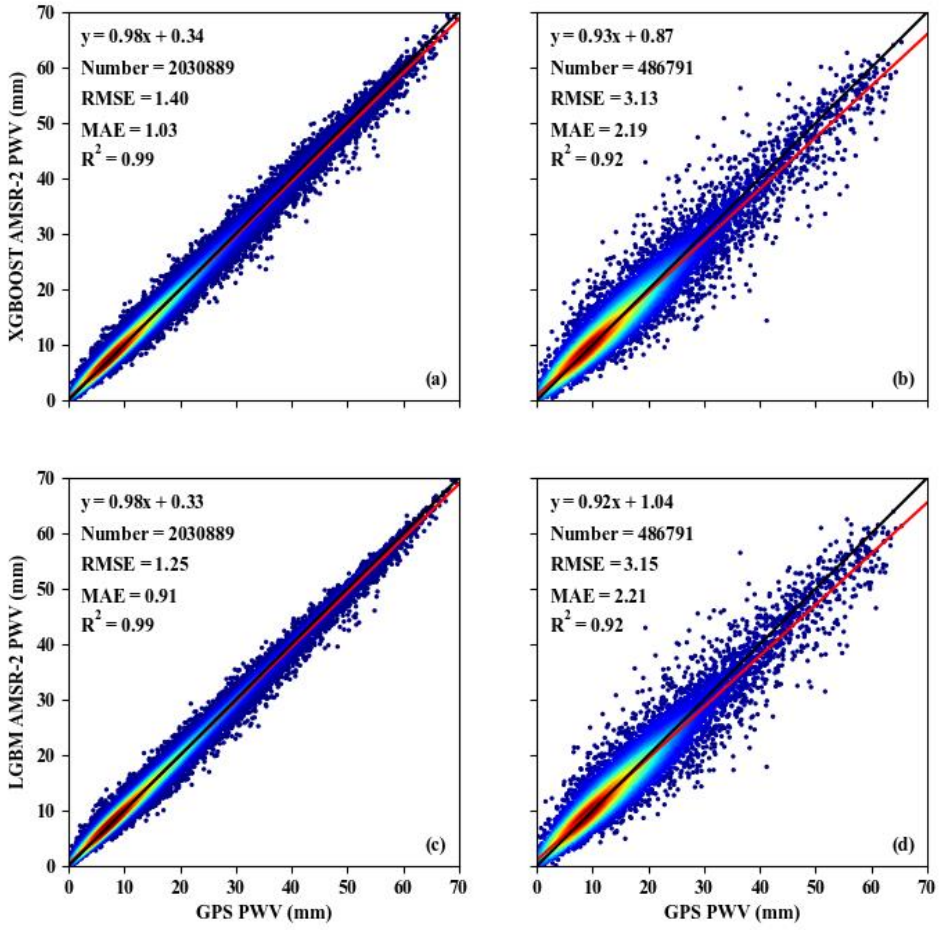


**Figure S3.** Comparisons between AMSR2 PWV retrievals based on XGBOOST and LGBM validated by GPS PWV, from top (XGBOOST in training data (a) and test data (b)) to bottom (LGBM in training data (c) and test data (d)).

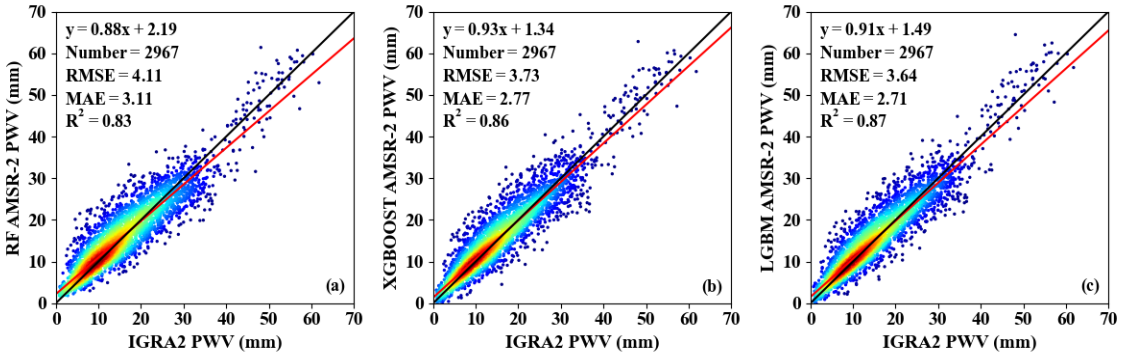


**Figure S4.** Independent verification based on IGRA2 PWV of AMSR-2 PWV generated by 3 ML models (from left to right: RF (a), XGBOOST (b) and LGBM (c)).

| **Feature name** | **Formula** | **Supplement** |
|---|---|---|
| Lon | Longitude | Longitude of each location |
| Lat | Latitude | Latitude of each location |
| MAVWI | $\frac{\Delta T_b^{23.8}}{\Delta T_b^{18.7}}$ | Microwave Atmospheric Water Vapor Index (Du et al., 2015) |
| $T_s$ | $T_b^{V,36.5}$ | Sensitive to land surface temperature (Duan et al., 2020) |
| CLW | $\log(\frac{\Delta T_b^{89}}{\Delta T_b^{36.5}})$ | Sensitive to cloud liquid water (Du et al., 2015) |
| DEM | $\exp(-h)$ | $h$ represent the altitude of each location (Du et al., 2015) |
| Orbit | Ascending or descending | Orbit marker, ascending is 0, descending is 1 |
| $F^H$ | $\frac{T_b^{H,23.8}}{T_b^{H,18.7}}$ | Determine vegetation transmissivity and open water fraction (Jones et al., 2010) |

| | | |
|---|---|---|
| $P$ | $\frac{T_b^{H,18.7}}{T_b^{V,18.7}}$ | Determine vegetation transmissivity and open water fraction (Jones et al., 2010) |
| DOY | $\sin\left(\frac{doy}{366}\right)$ | *doy* represents the day of the year for each location. |
| IGBP | IGBP type | IGBP classification from 0 to 16 representing different types |

.

**Table S1.** Features name and representation.

## Reference


Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Presented at the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Choi, H., Park, S., Kang, Y., Im, J., Song, S., 2023. Retrieval of hourly PM2.5 using top-of-atmosphere reflectance from geostationary ocean color imagers I and II. Environ. Pollut. 323, 121169. https://doi.org/10.1016/j.envpol.2023.121169

Fu, D., Gueymard, C.A., Yang, D., Zheng, Y., Xia, X., Bian, J., 2023. Improving aerosol optical depth retrievals from Himawari-8 with ensemble learning enhancement: Validation over Asia. Atmospheric Res. 284, 106624. https://doi.org/10.1016/j.atmosres.2023.106624

Liu, C., Yang, S., Di, D., Yang, Y., Zhou, C., Hu, X., Sohn, B.-J., 2022. A Machine Learning-based Cloud Detection Algorithm for the Himawari-8 Spectral Image. Adv. Atmospheric Sci. 39, 1994–2007. https://doi.org/10.1007/s00376-021-0366-x

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions.

Ma, X., Yao, Y., Zhang, B., He, C., 2022. Retrieval of high spatial resolution precipitable water vapor maps using heterogeneous earth observation data. Remote Sens. Environ. 278, 113100. https://doi.org/10.1016/j.rse.2022.113100