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.

Key points

- A machine learning based passive microwave (PMW) land PWV retrieval method is developed using the latest enhanced GPS PWV dataset
- With the addition of new features with clear physical meaning, the PWV retrieval accuracy improves by about 30%
- The proposed model performs well in areas that have been excluded in previous studies, such as open waters and permanently frozen areas

Abstract

Accurately retrieving precipitable water vapor (PWV) over wide-area land surface remains challenging. Unlike passive infrared remote sensing, passive microwave (PMW) remote sensing provides almost all-weather PWV retrievals. This study developed a PMW-based land PWV retrieval algorithm using the automated machine learning (AutoML). Data from the Advanced Microwave Scanning Radiometer 2 (AMSR-2) serves as the main predictor variables and high-quality Global Positioning System (GPS) PWV data as the target variable. Unprecedentedly large GPS training samples (over 50 million) from more than 12,000 stations worldwide are used to train the AutoML model. New predictors with clear physical mechanisms enable PWV retrieval over almost any land surface type, including snow cover and near open water. Validation shows good agreement between PWV retrievals and ground observations, with a root mean square error of 3.1 mm. This encouraging outcome suggests that the algorithm's potential for application with other PMW radiometers with similar wavelengths.

Plain Language Summary

Precipitable water vapor plays a critical role in the global hydrological cycle, but retrieving its value from remote-sensed data is challenging, especially for scientific purposes requiring high resolution and accuracy. This work proposes a new retrieval algorithm, which is attractive on three accounts. First is the use of information from the microwave radiometer onboard a solar-synchronous-orbit satellite, which has a high spatio-temporal resolution. The second attraction is the use of automated machine learning (AutoML), which could circumvent the complex model selection and tuning processes that are typically involved in machine-learning tasks. Thirdly, an unprecedentedly large ground-based dataset is gathered from GPS stations worldwide, which is to be used as target variables for AutoML training. The validation results reveal that the PWV retrieval is remarkably successful over all land surface types, which is previously rarely seen. The proposed algorithm can also be transferred and used with radiometers onboard other satellites.

1. Introduction

Albeit water vapor accounts for only a small fraction of the total amount of water in the atmosphere, the role it plays in many atmospheric processes such as atmospheric radiation or hydrological cycle must not be deemed at any rate unimportant. Water vapor is also Earth's most abundant greenhouse gas, in that, it contributes 70% of the total atmospheric radiation absorption, and thereby exerting significant positive feedbacks on climate warming (Bedka et al., 2010; Held and Soden, 2000; Wentz et al., 2007). Due to the high spatial and temporal variability, increasing the spatio-temporal resolution at which water vapor information can be acquired has attracted continuous attention of atmospheric scientists (Huntington, 2006; Lindstrot et al., 2014).

For precipitable water vapor (PWV), its ground-based measurements can be realized through the Global Positioning System (GPS). Although GPS measurements generally provide the most reliable and accurate results, their limited spatio-temporal coverage may not be sufficient for scientific studies. Considering the trade-off between accuracy and coverage, satellite remote sensing is almost always preferred. There are three classes of passive methods based on a satellite's solar reflectance (SR) channels, thermal-infrared (TIR) channels, and microwave (MW) channels, respectively. Underpinning all these three classes of methods is the fact that radiation is absorbed by water vapor as it transports in the atmosphere. Thus, the three distinct retrieval methods exploit different water vapor absorption bands (i.e., 0.9–1.0 μm for SR, 6.5–8.7 μm for TIR, and 1.64–13.5 mm for MW). Indeed, passive methods based on TIR satellite data have been widely used as a basis for PWV retrieving, with uncertainties ranging from 5 to 10% (Gao and Kaufman, 2003; Kaufman and Gao, 1992), but the retrieval is easily impacted by the presence of atmospheric aerosols and clouds (Du et al., 2015). Such dependence on clear-sky situations also affects methods based on SR satellite data. In contrast, microwave can penetrate clouds and even rain, thus enabling the retrieval of PWV under all-sky conditions, and making passive

microwave (PMW) an indispensable ingredient within the omni-source PWV observing system (Wentz, 1997; Wang et al., 2009; Gao et al., 2022; Ji et al., 2017).

PWV retrieval from PMW satellite observations is mature over the ocean, and operational products have been available for more than three decades (Deeter, 2007; Wentz, 1997). However, despite the attempts with varying levels of success, it remains a challenge over land originated from the low contrast between signals from the surface and atmosphere, as well as the strong heterogeneity of surface emissivity (Prakash et al., 2018). Among the most notable pioneer works on this matter is the one by Deeter (2007), who proposed a PWV retrieval method solely based on the polarization difference between brightness temperature (T_b) values, i.e., $\Delta T_b = T_b^V - T_b^H$, where the superscripts V and H annotate vertically and horizontally polarized T_b. The polarization-difference signal can be acquired from the space-borne instruments with dual-polarization scanning capabilities, such as the Advanced Microwave Scanning Radiometer 2. The retrieval mechanism is that ΔT_b can be concisely and precisely parameterized by PWV, liquid water path (LWP), surface temperature (T_s), as well as emissivity polarization difference ($\Delta\varepsilon = \varepsilon^V - \varepsilon^H$). By assuming $\Delta\varepsilon^{18.7} = \Delta\varepsilon^{23.8}$ (Ruston, 2004), T_s, and LWP are known, PWV could be analytically expressed as a function of the ratio of ΔT_b at 18.7 GHz to that at 23.8 GHz (termed as the microwave atmospheric water vapor index (MAWVI) by Jones et al. (2010)) and thus retrieved. The root mean square error (RMSE) of this method was about 6 mm over land with $\Delta\varepsilon > 0.03$, but the retrieval accuracy deteriorates substantially for scenarios with $\Delta\varepsilon < 0.03$. The is therefore defined as the ratio of the satellite-measured ΔT_b at 18.7 GHz to that at 23.8 GHz, as to derive PWV.

Inspired by this pioneer work, a rich literature on PWV retrieval seeks to express PWV as some analytic functions of predictor variables. For instance, Du et al. (2015) devised a multiple linear relationship between PWV and several parameters, such as altitude, surface temperature, or cloud liquid water (CLW), but not MAWVI. Although MAWVI was not utilized, a slight improvement in PWV retrieval accuracy (an RMSE of 4.7 mm) as compared to the value reported by Deeter (2007) was

achieved. Kazumori et al. (2018) developed a simple linear relationship between PWV and the logarithm of MAWVI, i.e., ln(MAWVI) = a×PWV + b; the RMSE was approximately 5.8 mm over scenes with large $\Delta\varepsilon$. It merits noting, however, that these aforementioned studies all make the assumption that the ratio of $\Delta\varepsilon$ at 18.7 GHz to that at 23.8 GHz being equal to 1, which is usually suitable only for the bare soil. In fact, the value of $\Delta\varepsilon^{18.7}/\Delta\varepsilon^{23.8}$ varies between 0.6 and 1.5 depending on the type of surface and the season (Ji et al., 2014). This assumption thus limits the accuracy and application range of PWV retrieval, especially in areas with dense vegetation cover and ice cover. Considering that $\Delta\varepsilon$ is not a constant and therefore its variability should be taken into account, Ji et al. (2014) developed a parameterization of $\Delta\varepsilon$ using other satellite data as well as surface elevation. Consequently, the modified algorithm is able to retrieve PWV with an RMSE of 4.85 mm when compared to the ground-based GPS PWV product.

Whereas using analytic relationships are conducive to interpreting the retrieval mechanism, such mathematical functions might lack flexibility. In PWV retrieval, land elevation and CLW have usually been simplified, if not ignored, in the process of deriving the physical model, leading to additional errors. On this point, Machine Learning (ML) is an emerging technology that opens new possibilities for satellite retrieval by using training data as much as possible to automatically learn a very complex function of the target variable on physically related predictors. For example, Gao et al. (2022) proposed a neural network method to retrieve PWV from PMW measurements with MAWVI, Ts, elevation, CLW, latitude, and longitude as input layers, which shows satisfactory results (an RMSE of 2.4 mm). Nevertheless, it is well known that the success of ML-based retrieval algorithms depends on the availability of high quality, complete and relevant training data. In this regard, previous studies in this area have often used very limited ground-based data, e.g., only 150 GPS stations are used by Gao et al. (2022), which may limit the eventual retrieval performance. In addition to the quality of training data, the choice of estimator, hyperparameters, sample size, and resampling strategy are also critical to the

construction of ML models, which directly affects the quality of prediction. Fortunately, with the automated machine learning (AutoML) framework, which balances the cost of data training and error evaluation, it is possible to find the "best" model faster and more accurately (Wang et al., 2021). In the atmospheric science community, this would be a very welcoming and useful tool. Indeed, Zheng et al. (2023) used the AutoML approach to estimate $PM_{2.5}$ over India, and the result demonstrated the bright prospects of AutoML in the atmosphere and environment.

Consolidating the limitations of previous works, the overarching aim of this paper is to develop a satellite-PMW-based PWV retrieval algorithm that is applicable to all types of land surfaces. With the most up-to-date enhanced GPS PWV product, a very large set of high-quality ground-based PWV data covering 16 surface types over land is used as the target variable in the AutoML-based PWV retrieval algorithm. To obtain comprehensive PWV retrieval over land, additional features which sensitive to the surface conditions are added to improve the accuracy of model prediction. The novelty of this work is threefold: (1) The latest enhanced GPS PWV dataset with high spatio-temporal resolution and accuracy is used as the target variable for AutoML. The dataset comes from 12,552 GPS sites worldwide, over the year 2020. To the best of our knowledge, this is the first time that such big training data is used in ML-based PWV retrieval algorithm development. (2) New predictors with a clear physical meaning are added to the AutoML-based retrieval algorithm, improving the generalizability and performance of the algorithm. (3) No external data other than satellite PMW measurements are used as predictors, making the proposal easily applicable to any other satellite PMW measurements.

2. Data

Four types of data were used to develop and validate the PWV retrieval algorithm: the T_b of AMSR-2, the land cover type from the Moderate Resolution

Imaging Spectroradiometer (MODIS), the enhanced GPS PWV product, and the Integrated Global Radiosonde Archive Version 2 (IGRA2) PWV data.

T_b, which is an essential parameter for PWV retrieval, is sourced from AMSR-2 onboard the Global Change Observation Mission-Water (GCOM-W1) solar synchronous orbit satellite launched in 2012 (Imaoka et al., 2012). AMSR-2 provides long-term and continuous data records to serve a better understanding on the global water cycle mechanism and the effects of climate change (Al-Yaari et al., 2014). AMSR-2 provides horizontal and vertical polarization T_b at 6 frequencies, i.e., 6.925, 10.65, 18.7, 23.8, 36.5 and 89.0 GHz and switches its descent and ascent orbits at 1:30 am and 1:30 pm, respectively. In this study, the AMSR-2 L1C product in 2020 is used.

The International Geosphere Biosphere Programme (IGBP) land cover type obtained from the MODIS product (MCD12C1) is employed as an additional predictor. The IGBP land cover type resides on a regular grid with a spatial resolution of 0.05° (Justice et al., 2002). There is a total of 17 IGBP land cover categories, the GPS sites included in this study cover 16 IGBP types (Figure S1 in support information S1); the only type that is not covered is the Deciduous Needleleaf Forests, because there are no GPS stations located on this surface type.

The GPS PWV product employed herein is an enhanced version of the operational GPS PWV dataset provided by the Nevada Geodetic Laboratory (NGL; Yuan et al., 2023), which serves as the target variable of the ML-based retrieval algorithm. It consists of high-quality global PWV measurements from 12,552 GPS stations (Figure S1 in Supporting Information S1). For the year 2020 alone, there are more than one billion data points. With the addition of the ERA-5, the spatiotemporal resolution of the product has been significantly improved. Compared to the operational version of GPS PWV, the mean absolute bias and standard deviation of the enhanced GPS PWV have been reduced by an average of 19.5% and 6.2%, respectively, using the situ measurements provided by radiosonde as a baseline (Yuan et al., 2023).

Aside from the GPS PWV, the radiosonde PWV measurements, which are widely used as the truth for validating other humidity measurements, are used as an independent calibration dataset. In this regard, the IGRA2 is the most comprehensive radiosonde dataset consisting of more than 770 stations worldwide in 2020 with regular daily observations at 00:00 and 12:00 UTC. PWV is calculated from the moisture profile when the profiles reach the surface and the pressure level at the top is at least 300 hPa and the pressure gaps should be less than 200 hPa.

3. Physical basis and ML algorithm development

3.1. Theoretical Basis

Ignoring the cosmic background radiation and atmospheric scattering, radiation received by satellite microwave radiometers can be characterized in a simple way (Merrikhpour and Rahimzadegan, 2017):

$$T_b(f,p,\theta) = T_s \times \varepsilon^p \times \Gamma_a(f,\theta) + T_a \times [1-\Gamma_a(f,\theta)], \qquad (1)$$

where f, p, θ denote frequency, polarization, and incident angle, respectively. T_b, as mentioned in the introduction, is the measured brightness temperature, which is a function of f, p, and θ. T_s is the surface temperature, ε^p is the land surface emissivity, T_a is the optical depth weighted effective atmospheric temperature, and Γ_a represents the atmospheric transmittance. The first term of the right hand of Eq. (1) represents the convolution effects of atmosphere and land surface on T_b, whereas the second term represents the upper emission of the atmosphere. The polarization difference in T_b, that is, $\Delta T_b = T_b^V - T_b^H$, can be approximated as follows (Jones et al., 2010, Du et al., 2015).

$$\Delta T_b = \Delta\varepsilon \times T_s \times \Gamma_a . \qquad (2)$$

Recall the definition of MAWVI, it is the ratio of the satellite-measured ΔT_b at 18.7 GHz to that at 23.8 GHz. Then following the approximation in Eq. (2), the following approximation obtains:

$$\mathrm{MAVWI}=\frac{\Delta T_b^{23.8}}{\Delta T_b^{18.7}}\approx\frac{\Delta\varepsilon^{23.8}}{\Delta\varepsilon^{18.7}}\times\frac{\Gamma_a^{23.8}}{\Gamma_a^{18.7}}. \quad (3)$$

The ratio of $\Delta\varepsilon^{23.8}$ to $\Delta\varepsilon^{18.7}$, relating to the land surface emissivity at different polarization and frequencies, is generally assumed to be a constant (close to 1). As such, MAWVI is extremely sensitive to the ratio of atmospheric transmittances at 23.8 and 18.7 GHz. It should be noted that the atmospheric transmittance is related to oxygen absorption, CLW and PWV. PWV can be directly derived directly from Eq. (1) if CLW and $\Delta\varepsilon$ are all known. Note that Eq. (1) is only applicable to scenes of bare soil. For the surface covered by vegetation, the emission and absorption of the plant canopy should be carefully considered (Mo et al., 1982). In areas where the land is mixed with open water, the satellite measured T_b is a weighted average of the radiation from land and water, therefore, the fraction of open water should also be considered (Jones et al., 2010).

From the literature review it can be summarized that the formerly published linear models and ML algorithms commonly use MAWVI, T_s, CLW, and altitude of the station as predictors for PWV retrieval, while ignoring the influence of vegetation, snow and open water. This is precisely the reason why these methods cannot perform very well in areas with small $\Delta\varepsilon$ values, e.g., over vegetation cover. If the ratio of $\Delta\varepsilon$ at 23.8 and 18.7 GHz was not accounted for carefully, large PWV retrieval errors ought to be expected (Ji et al., 2014). To increase the applicability of the retrieval model and improve the retrieval accuracy, we follow Jones et al. (2010) and introduce two additional input features to the ML algorithm, namely, F^H and P, to express vegetation transmissivity and open water fraction in terms of the simplified emission model. The derivation of F^H and P is given as follows:

$$F^H=\frac{T_b^{H,23.8}}{T_b^{H,18.7}} \quad \text{and} \quad P=\frac{T_b^{H,18.7}}{T_b^{V,18.7}} \quad (4)$$

The ratio of 23.8 GHz and 18.7 GHz in horizontal polarization is more responsive to vegetation canopy absorption, while the ratio of 18.7 GHz in horizontal

polarization and vertical polarization is sensitive to surface conditions. The advantage of using these parameters is that they can be directly derived from PMW measurements. The impacts of T_s and CLW on PWV retrieval are also accounted. For surface temperature T_s, the polarized brightness temperature $T_b^{V,36.5}$ may be used as a proxy (Jones et al., 2010). The 36.5 GHz and 89 GHz polarization difference ratio is very sensitive to CLW and is therefore used as a predictor to represent CLW effect (Jones et al., 2010). Given that PWV shows obvious seasonal and spatial variations, additional inputs to the ML algorithm include variables DOY (the sine of the ratio of day of year to 365), latitude, longitude and IGBP land type. Last but not least, PWV is also closely related to the altitude (h). Therefore, exp(–h) is also used as a predictor (Gao et al., 2022). Table S1 summarizes the input feature selection.

3.2. Collocation

As PWV exhibits strong spatio-temporal variation, collocation of data from various sources is thought to be important. In this work, the general criterion is that the maximum distance difference should not exceed 10-km and maximum time difference should not exceed 10-min when matching AMSR-2 T_b and GPS PWV. Samples from all sites are split into training and test sites according to the ratio of 4:1, where the number of test sites under different land cover types is directly proportional to the total number of sites of that land cover type in the world. The distribution of training and test sites is shown in Figure S1 in Supporting Information S1.

3.3. ML algorithm development

ML models extract relevant information from training data to make predictions. To achieve optimal performance, several critical considerations including model selection, hyperparameter tuning, feature selection, must be made. Numerous AutoML packages have, therefore, been developed to automate as far as possible (Wang et al., 2021). Among them, the Fast and Lightweight AutoML (FLAML) developed by Wang et al. (2021) is able to boost the rapidity of experimentation and facilitate efficient model optimization. FLAML focuses not only on the optimization of model

parameters, the model selection and the size of the dataset used, but also on the runtime of the model. It consists of two layers, an ML layer containing the candidate models and an AutoML layer, which includes a model proposer, a hyperparameter and sample size proposer, a validation strategy proposer and a controller.

In the predictor selection module, light gradient boosting machine (LGBM), extreme gradient boosting (XGBOOST), and random forest (RF) models are selected candidate models. The detailed model introduction is in Text S1 in Supporting Information S1. In the configuration of FLAML, we select the determination coefficient (R^2) as the optimizing metric and set the time budget to 3600s (note that FLAML also focuses on the runtime of the model). Two experiments with different input features are designed to illustrate the superiority of the new features introduced in section 3.1. The first case, which is taken as the control experiment, uses MAVWI, T_s, CLW, expH, Orbit, Lat and Lon as learning features to develop the tree-based ML models. In the second case, additional features including F^H, P, IGBP, and DOY are incorporated for comparison. The same training and test samples are used for both two cases, ensuring a consistent and fair comparison. Note that the matching data in areas with vegetation cover, open water, and permanent icing have always been excluded in previous studies, but they are retained here in the ML model development. This inclusion should improve the algorithm's performance in these specific areas.

4. Results

In the control experiment, we compare two cases of situations with and without additional features. The improvements in prediction accuracy due to new features are clearly evident. Quantitatively, the inclusion of the new features in RF significantly raises the accuracy of PWV retrieval by approximately 30% (the RMSE decreases from 5.43 mm to 3.76 mm in Figure S2 in Supporting Information S1). In addition, three tree-based models (RF, XGBOOST, and LGBM) included in FLAML were also evaluated and compared. When validated against the out-of-station GPS PWV observations, the R^2 values of XGBOOST and LGBM, as compared to that of RF,

both increase to 0.92 (Figure S3 in Supporting Information S1). 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 in Supporting Information S1 depicts the corresponding results, in that, the accuracy of PWV retrieval by all three ML models with IGRA2 as a reference. 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. Therefore, LGBM is selected as the best estimator for the following part of the work. More details are provided in Text S2 in Supporting Information S1.

4.1 Model performance over different surfaces

To test the applicability of the algorithm under a variety of surface conditions, validation results over all 16 included IGBP types are shown in Figure 1. For areas that are covered by ice (Persistent Snow and Ice) and heavily influenced by open water (such as Water Bodies and Permanent Wetlands), which are often ignored in previous studies, the RMSEs are 1.27 mm and 2.09 mm, respectively, and the present model can explain the variability of more than 90% (R^2), demonstrating excellent consistency with ground GPS PWV. In areas with bare soil or sparse vegetation (such as Barren and Closed Shrublands), the RMSEs are 2.27 mm and 1.85 mm, respectively, which presents as a significant improvement compared to previously reported values (4.7 mm in Du et al. 2015; 2.4 mm in Kazumori, 2018). In some forests densely covered with vegetation (such as Evergreen Needleleaf Forests and Deciduous Broadleaf Forests), the present algorithm can still maintain relatively high accuracy (RMSE is approximately 3.5 mm) thanks to the inclusion of new parameters. The results demonstrate that our algorithm is not only applicable to almost all land types, but also has excellent performance in all types.

The SHapley Additive exPlanations (SHAP) method (detailed introduction in Text S3 in Supporting Information S1) is used to calculate the marginal contribution

of each feature in the ML model. Figure 2 shows the SHAP values of all 11 features. It is evident that the two newly added features (F^H and P) hold the top two positions in terms of importance, which implies their high contribution to the retrieval process of PWV. Furthermore, their high values dominate the positive change in SHAP values, indicating that they are positively correlated with the PWV retrievals. Notably, T_s and Lat have also attained high ranks, which is consistent with our expectations. T_s exhibits a positive correlation with PWV, whereas Lat shows a negative correlation. This observation aligns well with the physical law of PWV spatial distribution. On the other hand, CLW ranks the lowest, showing a slight impact on PWV prediction, which is related to the exclusion of precipitation areas when training the ML model.

4.2 Global seasonal-averaged PWV distribution

Based on the AMSR-2 T_b and IGBP datasets in 2020, a 0.1°×0.1° resolution daily global PWV product is made using the trained LGBM model. Figure 3 (b) and (d) show the seasonal average PWV distributions in winter (December, January, and February) and summer (June, July, and August). The AIRS L3 product with a spatial resolution of 1°×1° in the same seasons is selected for comparison, as shown in Figure 3 (a) and (c).

In general, AMSR-2 PWV and AIRS PWV show similar spatial distribution patterns. The distribution of PWV decreases with increasing latitude. This phenomenon is consistent with the well-understood physical law (Seidel, 2002). In addition, the two products also show similar seasonal variations. In winter, as affected by temperature and solar radiation, the total PWV level in the northern hemisphere is low, about 5–15 mm. In central Africa and northern Oceania, the intensity of the AMSR-2 PWV is lower. In summer, the two products also show similar PWV distributions. Relatively extreme wet atmospheric conditions occur in Southeast Asia, South Asia, northern South America, and other regions north of the Equator. Similarly, the PWV of AMSR-2 is lower than that of AIRS in these regions, approximately 5 mm. In addition, it is noted that the spatial variability of PWV of AMSR-2 is more clearly visible in regions with low PWV values (such as the Qinghai-Tibet Plateau

and western North America). Although differences in estimating the highest PWV value, the AMSR-2 PWV product finely describes the PWV distribution of the global land.

5. Conclusion and discussions

In this work, a ML-based global land PWV retrieval algorithm is developed. Unlike previous studies, which only use limited samples for training, the most recent PWV data from more than 10,000 GPS sites are herein considered. Moreover, several new predictors with clear physical meaning are included as model inputs. As compared to PWV values retrieved using just traditional parameters, the newly added parameters (F^H, P, DOY, IGBP) improve the PWV retrieval accuracy by about 30%. At the same time, the SHAP analysis also confirms that the addition of new parameters makes significant contributions to the improvement of PWV retrieval accuracy.

When new parameters are added, the proposed ML model performs satisfactorily, with the RMSE being 3.13 mm and R^2 being 0.93. What is more is that our model also has a relatively stable performance across all 16 IGBP land cover types. The retrievals over Persistent Snow and Ice, Closed Shrublands land types exhibit the best performance with the overall RMSE less than 2 mm. Even in the worst performing areas (such as Evergreen Broadleaf Forest, Deciduous Broadleaf Forest), the RMSE remains around 3.5 mm, which is lower than the values reported in many former works. When using IGRA2 data for external verification, the results are also quite satisfactory (RMSE is 3.64 mm and R^2 is 0.87).

The proposed method in this work demonstrates the potential of using machine learning as an AMSR-2 PWV retrieval tool. It is thought that this method could be extended to other sensors with similar channels as AMSR-2, enabling the development of long-term continuous environmental datasets across multiple sensors.

Open Research

The enhanced GPS PWV product can be found at https://doi.org/10.5281/zenodo.6973528. The AMSR-2 L1C Tb data can be found at https://disc.gsfc.nasa.gov/datasets/GPM_1CGCOMW1AMSR2_07/summary?keywords=AMSR-2. The IGRA2 data is from https://www.ncei.noaa.gov/products/weather-balloon/integrated-global-radiosonde-archive. The MCD12C1 data is from https://ladsweb.modaps.eosdis.nasa.gov/archive/allData/6/MCD12C1. The retrieved global PWV dataset can be accessed online (at https://www.scidb.cn/s/UZbYzq).

Acknowledgments

The authors wish to thank the teams and members who provided the enhanced GPS PWV product, AMSR-2 L1C T_b data. This research was supported by National Natural Science of Foundation of China (42030608, 42075079, 42105128).

Reference

Al-Yaari, A., Wigneron, J.-P., Ducharne, A., Kerr, Y., De Rosnay, P., De Jeu, R., Govind, A., Al Bitar, A., Albergel, C., Muñoz-Sabater, J., Richaume, P., Mialon, A., 2014. Global-scale evaluation of two satellite-based passive microwave soil moisture datasets (SMOS and AMSR-E) with respect to Land Data Assimilation System estimates. Remote Sens. Environ. 149, 181–195. https://doi.org/10.1016/j.rse.2014.04.006

Bedka, S., Knuteson, R., Revercomb, H., Tobin, D., Turner, D., 2010. An assessment of the absolute accuracy of the Atmospheric Infrared Sounder v5 precipitable water vapor product at tropical, midlatitude, and arctic ground-truth sites: September 2002 through August 2008. J. Geophys. Res. 115, D17310. https://doi.org/10.1029/2009JD013139

Deeter, M.N., 2007. A new satellite retrieval method for precipitable water vapor over land and ocean. Geophys. Res. Lett. 34, L02815. https://doi.org/10.1029/2006GL028019

Du, J., Kimball, J.S., Jones, L.A., 2015. Satellite Microwave Retrieval of Total Precipitable Water Vapor and Surface Air Temperature Over Land From AMSR2. IEEE Trans. Geosci. Remote Sens. 53, 2520–2531. https://doi.org/10.1109/TGRS.2014.2361344

Gao, B.-C., Kaufman, Y.J., 2003. Water vapor retrievals using Moderate Resolution Imaging

Spectroradiometer (MODIS) near-infrared channels: WATER VAPOR RETRIEVALS USING MODIS. J. Geophys. Res. Atmospheres 108, n/a-n/a. https://doi.org/10.1029/2002JD003023

Gao, Z., Jiang, N., Xu, Y., Xu, T., Liu, Y., 2022. Precipitable Water Vapor Retrieval Over Land From GCOM-W/AMSR2 Based on a New Integrated Method. IEEE Trans. Geosci. Remote Sens. 60, 1–12. https://doi.org/10.1109/TGRS.2022.3151384

Held, I.M., Soden, B.J., 2000. Water Vapor Feedback and Global Warming. Annu. Rev. Energy Environ. 25, 441–475. https://doi.org/10.1146/annurev.energy.25.1.441

Huntington, T.G., 2006. Evidence for intensification of the global water cycle: Review and synthesis. J. Hydrol. 319, 83–95. https://doi.org/10.1016/j.jhydrol.2005.07.003

Imaoka, K., Maeda, T., Kachi, M., Kasahara, M., Ito, N., Nakagawa, K., 2012. Status of AMSR2 instrument on GCOM-W1, in: Shimoda, H., Xiong, X., Cao, C., Gu, X., Kim, C., Kiran Kumar, A.S. (Eds.), . Presented at the SPIE Asia-Pacific Remote Sensing, Kyoto, Japan, p. 852815. https://doi.org/10.1117/12.977774

Ji, D., Shi, J., 2014. Water Vapor Retrieval Over Cloud Cover Area on Land Using AMSR-E and MODIS. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 3105–3116. https://doi.org/10.1109/JSTARS.2014.2298979

Jones, L.A., Ferguson, C.R., Kimball, J.S., Zhang, K., Chan, S.T.K., McDonald, K.C., Njoku, E.G., Wood, E.F., 2010. Satellite Microwave Remote Sensing of Daily Land Surface Air Temperature Minima and Maxima From AMSR-E. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 3, 111–123. https://doi.org/10.1109/JSTARS.2010.2041530

Justice, C.O., Townshend, J.R.G., Vermote, E.F., Masuoka, E., Wolfe, R.E., Saleous, N., Roy, D.P., Morisette, J.T., 2002. An overview of MODIS Land data processing and product status. Remote Sens. Environ. 83, 3–15. https://doi.org/10.1016/S0034-4257(02)00084-6

Kaufman, Y.J., Gao, B.-C., 1992. Remote sensing of water vapor in the near IR from EOS/MODIS. IEEE Trans. Geosci. Remote Sens. 30, 871–884. https://doi.org/10.1109/36.175321

Lindstrot, R., Stengel, M., Schröder, M., Fischer, J., Preusker, R., Schneider, N., Steenbergen, T., Bojkov, B.R., 2014. A global climatology of total columnar water vapour from SSM/I and MERIS. Earth Syst. Sci. Data 6, 221–233. https://doi.org/10.5194/essd-6-221-2014

Merrikhpour, M.H., Rahimzadegan, M., 2017. An Introduction to an Algorithm for Extracting

Precipitable Water Vapor Over Land From AMSR2 Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10, 3975–3984. https://doi.org/10.1109/JSTARS.2017.2716403

Mo, T., Choudhury, B.J., Schmugge, T.J., Wang, J.R., Jackson, T.J., 1982. A model for microwave emission from vegetation-covered fields. J. Geophys. Res. 87, 11229. https://doi.org/10.1029/JC087iC13p11229

Prakash, S., Norouzi, H., Azarderakhsh, M., Blake, R., Prigent, C., Khanbilvardi, R., 2018. Estimation of Consistent Global Microwave Land Surface Emissivity from AMSR-E and AMSR2 Observations. J. Appl. Meteorol. Climatol. 57, 907–919. https://doi.org/10.1175/JAMC-D-17-0213.1

Ruston, B.C., 2004. Characterization of summertime microwave emissivities from the Special Sensor Microwave Imager over the conterminous United States. J. Geophys. Res. 109, D19103. https://doi.org/10.1029/2004JD004890

Wang, C., Wu, Q., Weimer, M., Zhu, E., 2021. FLAML: A Fast and Lightweight AutoML Library.

Wang, Y., Y. Fu, G. Liu, Q. Liu, and L. Sun (2009), A new water vapor algorithm for TRMM Microwave Imager (TMI) measurements based on a log linear relationship, J. Geophys. Res., 114, D21304, doi:10.1029/2008JD011057

Wentz, F.J., 1997. A well-calibrated ocean algorithm for special sensor microwave / imager. J. Geophys. Res. Oceans 102, 8703–8718. https://doi.org/10.1029/96JC01751

Wentz, F.J., Ricciardulli, L., Hilburn, K., Mears, C., 2007. How Much More Rain Will Global Warming Bring? Science 317, 233–235. https://doi.org/10.1126/science.1140746

Yuan, P., Blewitt, G., Kreemer, C., Hammond, W.C., Argus, D., Yin, X., Van Malderen, R., Mayer, M., Jiang, W., Awange, J., Kutterer, H., 2023. An enhanced integrated water vapour dataset from more than 10 000 global ground-based GPS stations in 2020. Earth Syst. Sci. Data 15, 723–743. https://doi.org/10.5194/essd-15-723-2023

Zheng, Z., Fiore, A.M., Westervelt, D.M., Milly, G.P., Goldsmith, J., Karambelas, A., Curci, G., Randles, C.A., Paiva, A.R., Wang, C., Wu, Q., Dey, S., 2023. Automated Machine Learning to Evaluate the Information Content of Tropospheric Trace Gas Columns for Fine Particle Estimates Over India: A Modeling Testbed. J. Adv. Model. Earth Syst. 15, e2022MS003099. https://doi.org/10.1029/2022MS003099

References From the Supporting Information

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

520

521

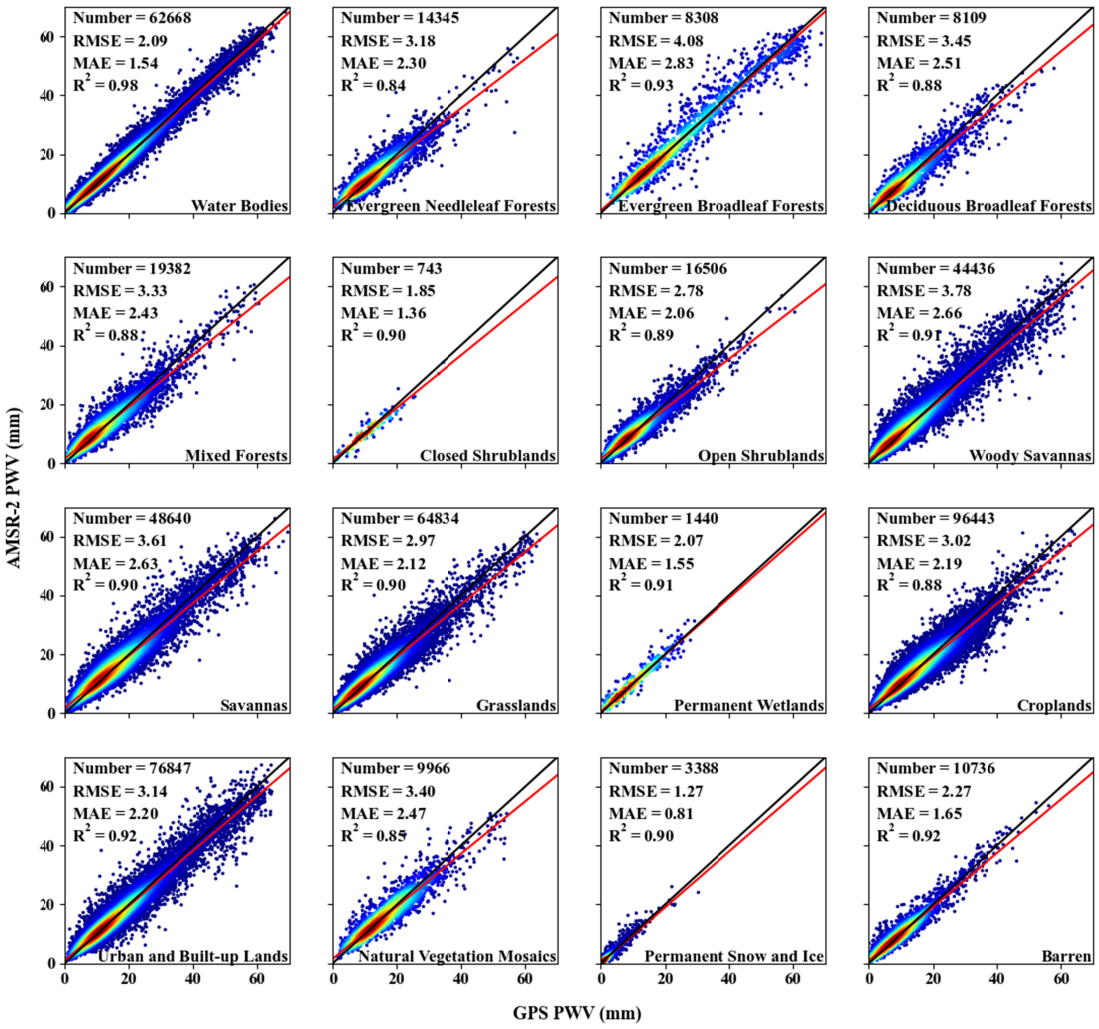
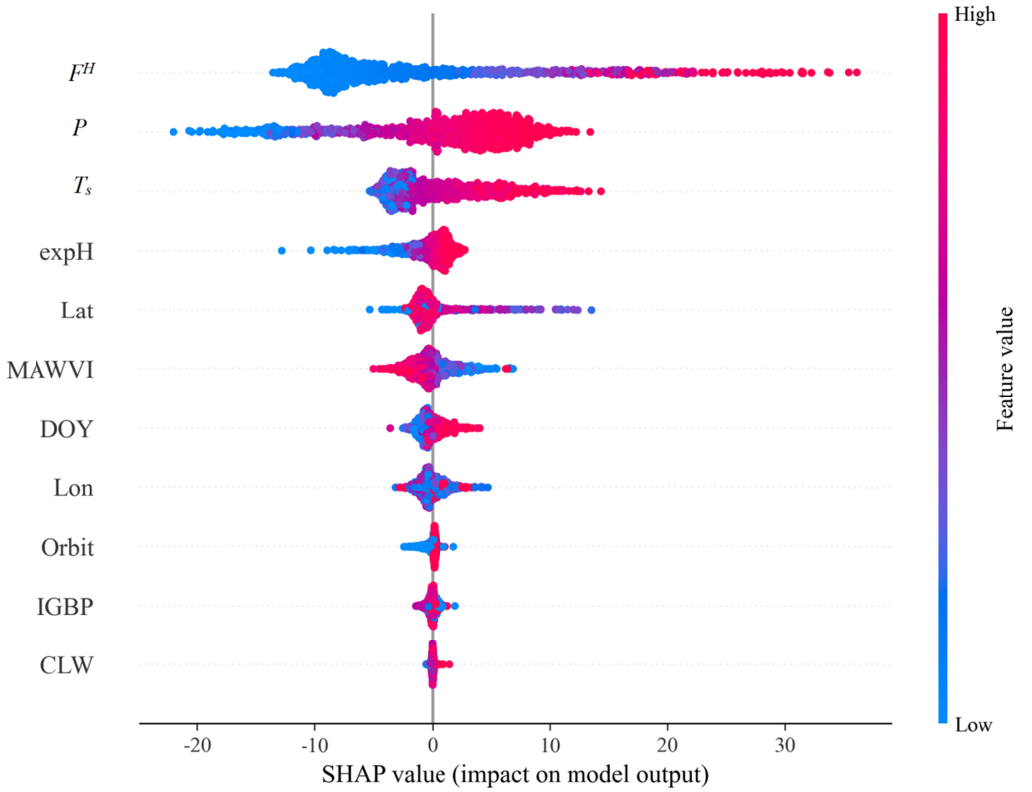


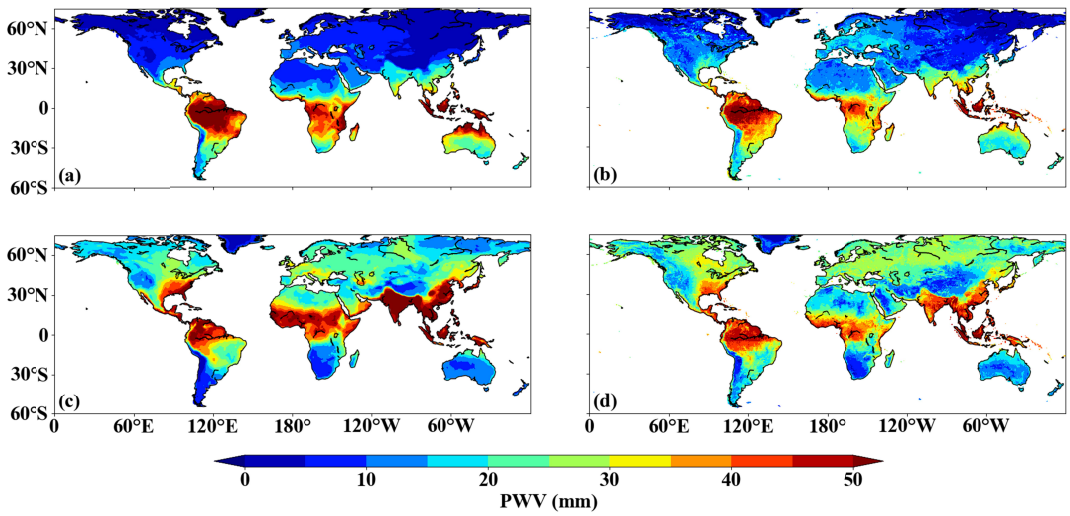
Figure 1. Accuracy comparison of AMSR-2 PWV over 16 MODIS IGBP types validated by GPS PWV (taking LGBM as estimator).



528

529 Figure 2. SHAP values of all input features (Contributions from high to low).

530



531

534 Figure 3. AMSR-2 global average PWV retrievals over land for winter (b) and

535 summer (d), compared to the AIRS L3 global product PWV over land for winter (a)

536 and summer (c).