Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 13, No. 3, 2024

Machine Learning-Based Calibration Study of MERSI-II Atmospheric Precipitable Water Vapor Product

Mengnan Li

School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454003, China

Abstract: Existing studies have found that there is a systematic underestimation of the terrestrial atmospheric precipitable water products of the FY3D satellite. In order to correct the bias and improve the accuracy of the products, this paper takes the data from the global AERONET ground observatory and the data from the FY3D atmospheric precipitable water products as the data source and conducts the modeling according to the ground observatory as the center of the circle, 0.05° as the spatial radius, the time of the satellite transit, and half an hour before and after the time scale. Spatio-temporal matching is used to obtain the modeling data, and the Random Forest Model (RF) is used to model the data and correct the FY3D atmospheric precipitable products. The results show that the application of the RF model can correct the product bias and improve the quality of the products.

Keywords: FY3D; AERONET; MERSI-II; PWV; Random Forest.

1. Introduction

Precipitable Water Vapor (PWV), or atmospheric water vapor, is the total atmospheric water vapor in a vertical column of a cross-sectional area extending from the Earth's surface to the top of the atmosphere, plays an irreplaceable role on Earth despite its relatively low proportion of the atmosphere, 0.1% ~ 4% of the total [1]. Atmospheric precipitable water is a key factor in the Earth's circulatory system and plays an important role in multiple Earth climate change processes [2]. Atmospheric precipitable water is very active on the spatial and temporal scales of the atmosphere, and it is also a fundamental cause of weather changes [3]. At the same time, water vapor is one of the most important input parameters for the inversion of atmospheric correction data, and water vapor can also be considered as complementary information to other geophysical parameters (e.g., temperature at the Earth's surface, etc.) [4]. Therefore, the water vapor content is of great practical importance for the study of remote sensing quantitative inversion. Due to the influence of spatial scale, the atmospheric precipitable water content in different regions fluctuates very much, so it is very difficult to obtain atmospheric precipitable water content with high accuracy, and the absence of high-precision and real-time atmospheric precipitable water information data will inevitably cause obstacles to the further development of the weather changes and the climate forecasting system [5]. Existing ground station (AERONET) data can provide high-precision real-time atmospheric precipitable water information in the region, while space satellites can provide weather observation information on a global scale [6]. If the advantages of these two can be combined, it is believed that a high-precision atmospheric precipitable water product with a wider spatial and temporal range and in real time can be obtained. This is also the result that this paper wants to realize.

2. Data and Methodology

2.1. FY3D MERSI-II PWV

The data used in this paper are land-based atmospheric precipitable water (PWV) product data from FY3D with MERSI-II on board [7]. The product is divided into four categories: segment, day, month, and ten days. The resolutions are 1 km, 5 km, 5 km, and 5 km, respectively, with the segment products in particular having the highest resolution and the largest amount of data. Therefore, the MERSI-II near-infrared atmospheric precipitable water segment product is selected as the experimental data in this paper. The total amount of atmospheric precipitable water over the land area is obtained by inverting the water vapor absorption channels of 0.905 mm, 0.936 mm, and 0.94 mm of the MERSI-II NIR channel data with the other two atmospheric window channels of 0.865 mm and 1.03 mm [8]. The segment product has an image acquisition interval of 5 minutes and a resolution of 1 km at the point below the star, and the product can be used to acquire the number of images for each day from April 30, 2019, to the present on a global scale. In addition, since the MERSI-II near-infrared atmospheric precipitable segment product data does not have related geolocation information data, this paper needs to download the simultaneous, same-resolution L1 GEO file of the FY3D satellite, which also carries the MERSI-II sensor. The geolocation data product is called FY-3D Moderate Resolution Spectral Imager L1 data (1KM_GEO). The physical significance of this product is that it houses Earth observation 1 km resolution MERSI geolocation data after geolocation preprocessing. In addition to being used to provide geolocation reference information for this study, the product can be widely used to assist in the generation of 1 km spatially resolved atmospheric, oceanic, and terrestrial remote sensing products for the same area at the same time. The above product data can be downloaded free of charge from the Fengyun Satellite Remote Sensing Data Service Network: http://satellite.nsmc.org.cn/PortalSite/Data/Satellite.aspx.

2.2. AERONET PWV

As shown in Figure 1, AERONET, the global automated observing network, is a ground-based aerosol remote sensing observing network jointly established by NASA and CNRS [9]. Now there are about 300 solar photometers in the network, covering more than 500 ground-based observation sites around the world. Therefore, AERONET can provide high-

precision ground-based PWV data. In the official platform of AERONET, there are three levels of data that can be downloaded for free: Level 1.0, Level 1.5, and Level 2.0. In the official statement of the platform, Level 1.0 means that the data are unfiltered and may not have the final calibration applied, and Level 1.5 means that the data are those that have had the cloud cleanup and the quality control applied, but these data may not have the final calibration applied, and these data may be subject to change; Level 2.0 refers to data to which automated cloud clearing has been applied and quality is assured by applying pre- and post-field calibration. After comparative analysis, Level 1.0 data are unscreened and have a high level of data complexity, while Level 2.0 data suffer from problems such as missing data. In this paper, Level 1.5 data is selected as the data source for extracting the PWV and other information from ground-based observatories. The product data can be downloaded for free from the AERONET website: https://aeronet.gsfc.nasa.gov/.

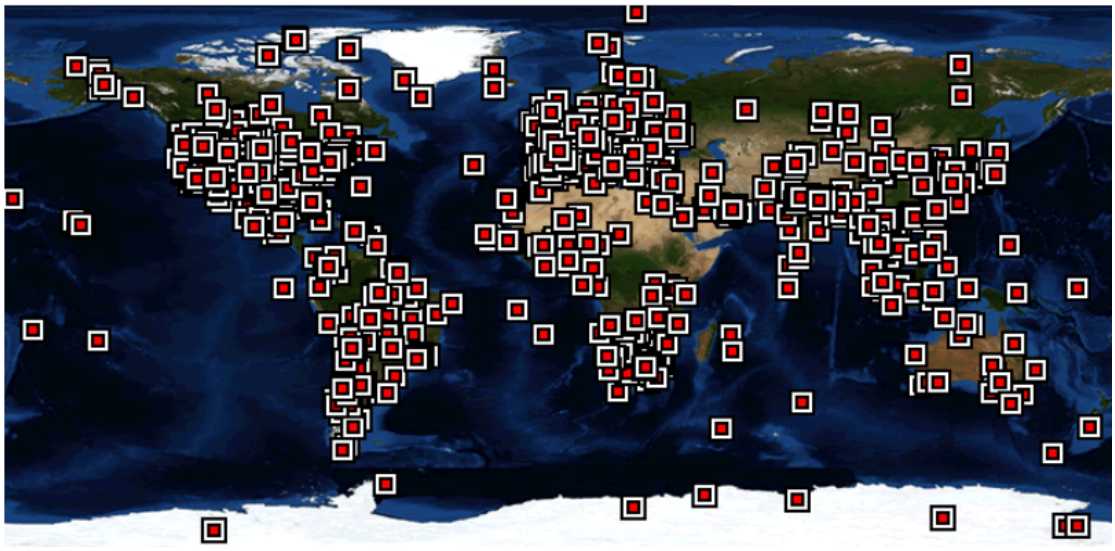

Figure 1. 2019-2020 AERONET global terrestrial site distribution

2.3. Spatio-temporal matching method

In accordance with the scientific and rigorous research attitude, this study carries out the corresponding spatio-temporal matching for the data that have been explicitly extracted [10]: (1) Spatial matching: Take the PWV dataset of the satellite water vapor product as an example (MERSI PWV); take the latitude and longitude of the ground-based observatory station as the center of the circle, with a radius of 0.05° as the radius; determine a range of the data; and then discard the filler values and invalid values within the range, and then take the average value. Then take the average value; we can get the spatial matching value of the corresponding satellite water vapor product of the ground-based observatory, i.e., the satellite PWV data; (2) time matching, AERONET ground-based observatory to obtain the data of the time is not fixed, and sometimes it can be recorded in a day within 24 hours of the detailed observational data, but there are sometimes a day of very few or even a continuous number of days of the missing observational data, so from the point of view of the experimental rigor, it is necessary to match the satellite data with the data from the ground-based observatory and then take the average value. Therefore, from the point of view of experimental rigor, it is necessary to time-match satellite data and ground-based observatory data. Taking the ground-based observatory station observation data as the data source and the satellite transit time as the fundamental reference, and the half-hour before and after as the time range, a data range is determined, and the filled and invalid values within the range are discarded and then averaged, then the time-matched value corresponding to the satellite transit time of the ground-based observatory station can be derived, i.e., the ground-based PWV data.

2.4. Statistical Indicators

A total of four statistical parameters were used in this study to evaluate the quality of MERSI-II PWV products before and after correction. These include the Pearson correlation coefficient (R, equation (1)), root mean square error (RMSE, equation (2)), mean bias (MB, equation (3)), and mean absolute error (MAE, equation (4)).

$$R = \frac{\sum_{i=1}^{n}\left(X_{i,satellite} - \overline{X}_{satellite}\right)\left(Y_{i,AERONET} - \overline{Y}_{AERONET}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_{i,satellite} - \overline{X}_{satellite}\right)^2 \sum_{i=1}^{n}\left(Y_{i,AERONET} - \overline{Y}_{AERONET}\right)^2}} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{i,satellite} - Y_{i,AERONET}\right)^2} \quad (2)$$

$$MB = \frac{1}{n}\sum_{i=1}^{n}\left(X_{i,satellite} - Y_{i,AERONET}\right) \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|X_{i,satellite} - Y_{i,AERONET}\right| \quad (4)$$

Where n is the number of matching points between the satellite and the ground-based observations. Xsatellite and YAERONET represent satellite retrievals and AERONET data, respectively. $\overline{X}$satellite and $\overline{Y}$AERONET represent the average values of satellite retrievals and AERONET data, respectively.

3. Results

3.1. Comparison of results before and after calibration of fusion PWV products

The research principle used in this paper is the post-processing correction satellite inversion method of the enhanced model [11], and the research process is as follows: Download the satellite water vapor products using the Wind and Cloud Remote Sensing Data Service (WRSS) network, and download the ground station water vapor products from the official website of AERONET, and then, after spatio-temporal matching of these data, build the model in accordance with the Random Forest (RF), and divide the training set, the validation set, and the test set in accordance with the way of 7:2:1, and the main hyper-parameters The main hyperparameters are configured as follows: n_estimators=100, max_depth=None, min_samples_split=2, min_samples_leaf=1, and max_features=n_features. Finally, the corresponding satellite inversion results of the model are obtained. The auxiliary data are solar zenith angle (SZA), satellite zenith angle (VZA), solar azimuth angle (SAA), satellite azimuth angle (VAA), ground elevation information (DEM), longitude, latitude, and date. The essence of the principle is a post-processing correction of the satellite water vapor product based on the ground station water vapor product and the necessary ancillary data, with the fundamental aim of obtaining a satellite data reprocessing product with a higher degree of accuracy (using the ground station data product as a reference).

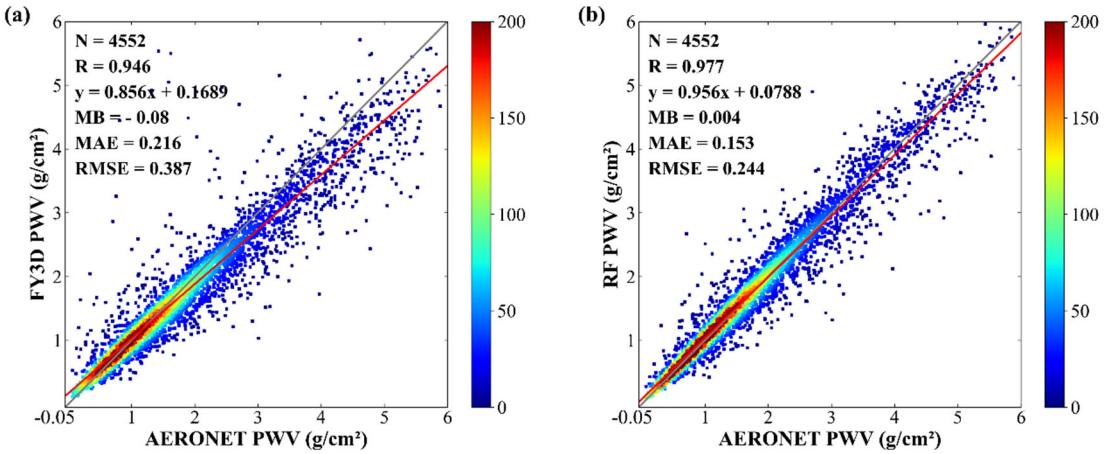

Figure 2. Scatterplot of validation results for the MERSI-II PWV fusion dataset with AERONET PWV data: (a) validation results before correction; (b) validation results after correction. The gray line is the y=x line. The red line is the linear fitting curve. The color of each point indicates the density of the location of the point.

Figure 2 shows the density scatterplot of the MERSI-II fusion PWV product relative to the AERONET PWV before and after correction by the RF model during 2020. Among them, the matching results in the density scatterplot are concentrated in the range of 0-3 g/cm2.Fig. (a) shows that the fusion PWV product exhibits good correlation with the AERONET PWV product, with a correlation coefficient of R of 0.946.However, most of the matches in its density scatterplot are centered below the straight line of y = x, which suggests that there may be a systematic underestimation of the fusion PWV product. Compared with Fig. (a), Fig. (b) after the RF model correction shows better results: overall, the matching points in the density scatter plot are more evenly distributed on both sides of the y=x straight line, and the linear fitting curve is very close to the y=x straight line; in terms of the statistical parameters, the correlation coefficient R improves from 0.946 to 0.977, and the mean bias (MB) decreases from -0.08 g/cm2 to 0.004 g/cm2, mean absolute error (MAE) from 0.216 g/cm2 to 0.153 g/cm2, and root mean square error (RMSE) from 0.387 g/cm2 to 0.244 g/cm2.These results indicate that the application of the RF model can improve the quality of the fused PWV products.

4. Conclusion

In the results obtained in this study, the following conclusions can be obtained by comparing and analyzing the validation results before and after model correction. Compared with the observation data of AERONET ground station, the MERSI-II PWV product before the Random Forest model correction has higher accuracy, but there is an obvious underestimation problem in the product; the accuracy of the MERSI-II PWV product after the model correction has been greatly improved in all indexes. The average deviation tends to be close to 0. The correlation coefficient is improved from 0.946 to 0.977. The average absolute error accuracy is improved by about 29.2%. The precision of root mean square error improves about 37.0%.

In summary, the application of the Random Forest model can correct the systematic bias present in the MERSI-II PWV product and improve the PWV product accuracy. This suggests that well-trained machine learning models can be used to improve atmospheric precipitable water product accuracy on a global scale. This processing can be extended to correct other satellite atmospheric products as well. This will provide more accurate data for weather forecasting, climate change monitoring and other atmospheric studies.

References

[1] Aumann, H. H., and Coauthors, 2003: AIRS/AMSU/HSB on the aqua mission: design, science objectives, data products, and processing systems. IEEE Transactions on Geoscience and Remote Sensing, 41, 253-264, https://doi.org/10.1109/tgrs.2002.808356.

[2] Held, I., and B. Soden, 2000: Water vapor feedback and global warming. Annual review of energy and the environment, 25, 441-475, https://doi.org/10.1146/annurev.energy.25.1.441.

[3] Xu, J., and Z. Liu, 2023c: Long-Term Calibration of Satellite-Based All-Weather Precipitable Water Vapor Product From

FengYun-3A MERSI Near-Infrared Bands From 2010 to 2017 in China. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-14, https://doi.org/10.1109/tgrs.2023.3300880.

[4] Xie, Y., and Coauthors, 2021: Validation of FY-3D MERSI-2 Precipitable Water Vapor (PWV) Datasets Using Ground-Based PWV Data from AERONET. Remote Sensing, 13, https://doi.org/10.3390/rs13163246.

[5] Campmany, E., J. Bech, J. Rodríguez-Marcos, Y. Sola, and J. Lorente, 2010: A comparison of total precipitable water measurements from radiosonde and sunphotometers. Atmospheric Research, 97, 385-392, https://doi.org/10.1016/j.atmosres.2010.04.016.

[6] Zhang, W., L. Wang, Y. Yu, G. Xu, X. Hu, Z. Fu, and C. Cui, 2021: Global evaluation of the precipitable-water-vapor product from MERSI-II (Medium Resolution Spectral Imager) on board the Fengyun-3D satellite. Atmospheric Measurement Techniques, 14, 7821-7834, https://doi.org/10.5194/amt-14-7821-2021.

[7] Zhang, X., and Coauthors, 2020: The development and application of satellite remote sensing for atmospheric compositions in China. Atmospheric Research, 245, https://doi.org/10.1016/j.atmosres.2020.105056.

[8] Wang, L., X. Hu, N. Xu, and L. Chen, 2020: Water Vapor Retrievals from Near-infrared Channels of the Advanced Medium Resolution Spectral Imager Instrument onboard the Fengyun-3D Satellite. Advances in Atmospheric Sciences, 38, 1351-1366, https://doi.org/10.1007/s00376-020-0174-8.

[9] Holben, B. N., and Coauthors, 1998: AERONET—A Federated Instrument Network and Data Archive for Aerosol Characterization. Remote Sensing of Environment, 66, 1-16, https://doi.org/10.1016/S0034-4257(98)00031-5.

[10] Pang, X., Y. Mu, X. Lee, Y. Zhang, and Z. Xu, 2009: Influences of characteristic meteorological conditions on atmospheric carbonyls in Beijing, China. Atmospheric Research, 93, 913-919, https://doi.org/10.1016/j.atmosres.2009.05.001.

[11] Lipponen, A., and Coauthors, 2021: Model-enforced post-process correction of satellite aerosol retrievals. Atmospheric Measurement Techniques, 14, 2981-2992, https://doi.org/10.5194/amt-14-2981-2021.