Reply to the reviewer #2

The authors developed a global terrestrial precipitable water volume (PWV) dataset from 2012 to 2020 by applying a machine learning model using Microwave Radiation Imager (MWRI) observations on board the Fengyun satellites series. The accuracy of dataset is evaluated by comparing with the products of SuomiNet GPS and Integrated Global Radiosonde Archive Version 2 (IGRA2) PWV. This work contributes to representing spatial and temporal PWV variations and providing valuable resource for atmospheric research. The manuscript may be considered for publication after being major revised in accordance with the following comments:

R: Thank you for your patient review. Your insights and suggestions are extremely helpful in refining our manuscript. A detailed response to each of your points is outlined below, with the issues raised presented in black and our responses highlighted in blue.

General:

1. The introduction could benefit from a more comprehensive discussion of the significance of PWV dataset in the context. This could include a brief overview of existing challenges and gaps in PWV dataset construction, and how this dataset addresses them. Additionally, the literature review should be expanded to include more recent studies on PWV retrieval employing machine learning techniques. This may help establish the novelty and

contribution of approach proposed in this study.

R: Thank you for your advice. We revised the manuscript according to your invaluable suggestions. 1) To demonstrate the importance of the PWV dataset, in line 60, we added: “With all-weather global PWV records, researchers are expected to use them to study the role of PWV in weather patterns, refine precipitation forecasts, and validate climate simulations”. In line 414 we added: “It will be instrumental in detecting atmospheric rivers, understanding moisture distribution, and assessing its effects on weather systems and climate. Moreover, the dataset is invaluable for hydrological models that simulate the water cycle, aiding in water resource management, drought assessment, and flood risk evaluation. Additionally, it provides a key reference for validating and improving other satellite-based precipitable water vapor products, thereby enhancing the overall accuracy of satellite observations”. 2) For the recent advances in PWV retrieval based on machine learning techniques. in line 84 we added: “Jiang et al. (2022) developed a back-propagation neural network (BPNN) to retrieve PWV over land with the RMSE of 3.87 mm”. We also cite more related articles up to date as follow:

Zhou, S., Cheng, J.: A physics-based atmospheric precipitable water vapor retrieval algorithm by synchronizing MODIS near-infrared and thermal infrared measurements. Remote Sens. Environ. 317, 114523. https://doi.org/10.1016/j.rse.2024.114523, 2025.

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

Jiang, N., Xu, Y., Xu, T., Li, S., Gao, Z.: Land Water Vapor Retrieval for AMSR2 Using a Deep Learning Method. IEEE Trans. Geosci. Remote Sensing 60, 1–11. https://doi.org/10.1109/TGRS.2022.3162222, 2022.

Jiang, N., Wu, Y., Li, S., Xu, Y., Wang, Y., Xu, T.: First PWV Retrieval Using MERSI-LL Onboard FY-3E and Cross Validation With Co-Platform Occultation and Ground GNSS. Geophysical Research Letters 51, e2024GL108681. https://doi.org/10.1029/2024GL108681, 2024.

He, W., Chen, H., Xia, X., Wu, S., Zhang, P.: Evaluation of the Long-term Performance of Microwave Radiation Imager Onboard Chinese Fengyun Satellites. Adv. Atmos. Sci. 40, 1257–1268. https://doi.org/10.1007/s00376-023-2199-2, 2023.

Li, R., Hu, J., Wu, S., Zhang, P., Letu, H., Wang, Y., Wang, X., Fu, Y., Zhou, R., Sun, L.: Spatiotemporal Variations of Microwave Land Surface Emissivity (MLSE) over China Derived from Four-Year Recalibrated Fengyun 3B MWRI Data. Adv. Atmospheric Sci. 39, 1536–1560. https://doi.org/10.1007/s00376-022-1314-0, 2022.

Wang, Y., Jiang, N., Wu, Y., Xu, Y., Kaufmann, H., Xu, T.: An Improved Model for the Retrieval of Precipitable Water Vapor in All-Weather

Conditions (RCMNT) Based on NIR and TIR Recordings of MODIS. IEEE Trans. Geosci. Remote Sens. 62, 1–12. https://doi.org/10.1109/TGRS.2024.3381750, 2024.

Su, H., Yang, T., Wang, K., Sun, B., Yang, X.: Evaluation of Precipitable Water Vapor Retrieval from Homogeneously Reprocessed Long-Term GNSS Tropospheric Zenith Wet Delay, and Multi-Technique. Remote Sensing 13, 4490. https://doi.org/10.3390/rs13214490, 2021.

He, J., Liu, Z.: Water Vapor Retrieval from MODIS NIR Channels Using Ground-Based GPS Data. IEEE Trans. Geosci. Remote Sensing 58 (5), 3726–37. https://doi.org/10.1109/TGRS.2019.2962057, 2020.

He, J., Liu, Z.: Refining MODIS NIR Atmospheric Water Vapor Retrieval Algorithm Using GPS-Derived Water Vapor Data. IEEE Trans. Geosci. Remote Sensing 59, 3682–3694. https://doi.org/10.1109/TGRS.2020.3016655, 2021.

2. In the method section, the authors choose Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost) and Random Forest to train the model. The reasons for selecting these models should be supplemented.

R: Thank you for your advice. We added the reason why we chose the three ML models in Line 198 with: “Among the shallow network structures, tree-based models have been consistently shown superior performance. Briefly, RF is an ensemble learning method that combines the outputs of multiple basic decision trees to make final predictions. Each decision tree is built by recursively partitioning the data based on the value ranges of various features. RF models have advantages in dealing with high-dimensional data, outliers and missing data (Belgiu and Drăguţ, 2016; Lundberg et al., 2020). XGBoost is an ensemble learning framework designed to construct an ensemble of weak decision trees that are combined using the gradient boosting technique. Each successive tree corrects the discrepancies between the prediction of the previous tree and the target value. By incorporating regularization techniques to prevent overfitting, XGBoost has gained popularity for to its high performance and reliability (Chen and Guestrin, 2016). LGBM is another gradient boosting framework that aims to provide faster training speed and lower memory consumption compared to other frameworks. It incorporates a technique called gradient-based one-sided sampling to select the most informative samples during the tree-building process. In addition, histogram-based gradient estimation, which takes advantage of binning for efficient computation, is used (Ke et al., 2017)".

3. In the conclusion, it is essential to articulate not only the strengths of the dataset but also to elucidate its constraints and limitations.

R: Thank you for your comments. To summarize the shortcomings of our product and what can be improved in the future, we have added line 454 with: “The MWRI PWV retrievals are still improved under extreme precipitation events, which may be resolved to some extent by combing MWHS measurements with much more channels”.

4. Please check the grammar in the manuscript to improve the text quality. For example, the subject of the sentence that “With the development of computer science, and in particular the proliferation of machine learning (ML), has led to the widespread adoption of ML by the remote sensing community” is missing.

R: Thank you for your careful review. We have replaced line 83~84 with: “With advancements in computer science, particularly the proliferation of machine learning (ML), ML has been widely adopted by the remote sensing community”. At the same time, grammatical and expression errors elsewhere in the manuscript are being progressively corrected.

Specific：

1. Line 216-217, The full names of “WAT, WET, ENF, EBF, DNF, DBF, and MF” should be provide when they first appear in the manuscript.

R: Thank you for your careful review. We added the full names in Line 243~246 with: “(Water Bodie (WAT) and Permanent Wetlands (WET) are 4.43 mm and 3.69 mm, respectively. In forested regions (Evergreen

Needleleaf Forest (ENF), Evergreen Broadleaf Forest (EBF), Deciduous Needleleaf Forest (DNF), Deciduous Broadleaf Forest (DBF) and Mixed Forests (MF)), the RMSE ranges from 2.90 to 5.49 mm". We also added the full names of MODIS IGBP in Figure 1. (line 582) with: "(Water Bodie (WAT), Evergreen Needleleaf Forest (ENF), Evergreen Broadleaf Forest (EBF), Deciduous Needleleaf Forest (DNF), Deciduous Broadleaf Forest (DBF), Mixed Forests (MF), Closed Shrubland (CLS), Open Shrubland (OSH), Woody Savanna (WSA), Savanna (SAV), Grassland (GRA), Permanent Wetlands (WET), Croplands (CRO), Urban and Built-up Lands (URB), Natural Vegetation Mosaic (NVM), Permanent Snow and Ice (SNW), Barren (BDR)"

2. Figure 5: It is clear that the amount of data when MWRI PWV is compared to IGRA2 PWV is much smaller than when it is compared to SuomiNet GPS PWV and enGPS PWV. This should be supplemented in the manuscript as well as giving possible reasons for this discrepancy.

R: Thanks. For the amount of data MWRI PWV compared to IGRA2 PWV is much smaller than when it is compared to SuomiNet GPS PWV and enGPS PWV, the main reason is that IGRA-2 provided only twice a day (00:00 and 12:00 UTC), while GNSS can provide PWV with higher temporal resolution. Given the fact we match the satellite and ground-based PWV data points if the temporal difference between them should be less than 15 minutes, we can obtain more MWRI-GPS pairs over the same time period. We added further explain in line 238 with: “Limited by the frequency of IGRA-2 measurements of PWV, we obtained a small sample size of MWRI and IGRA-2 matches.”

3. Figure 8: The different colors of the solid dots in the figure should be clearly explained. Please include a color bar to indicate what each color represents for better clarity.

R: Thank you for your comments. Given that the use of a density scatterplot is not appropriate for a single site with a small number of samples, we have replaced Figure 10 with the following scatterplot to allow the reader to understand it more intuitively.

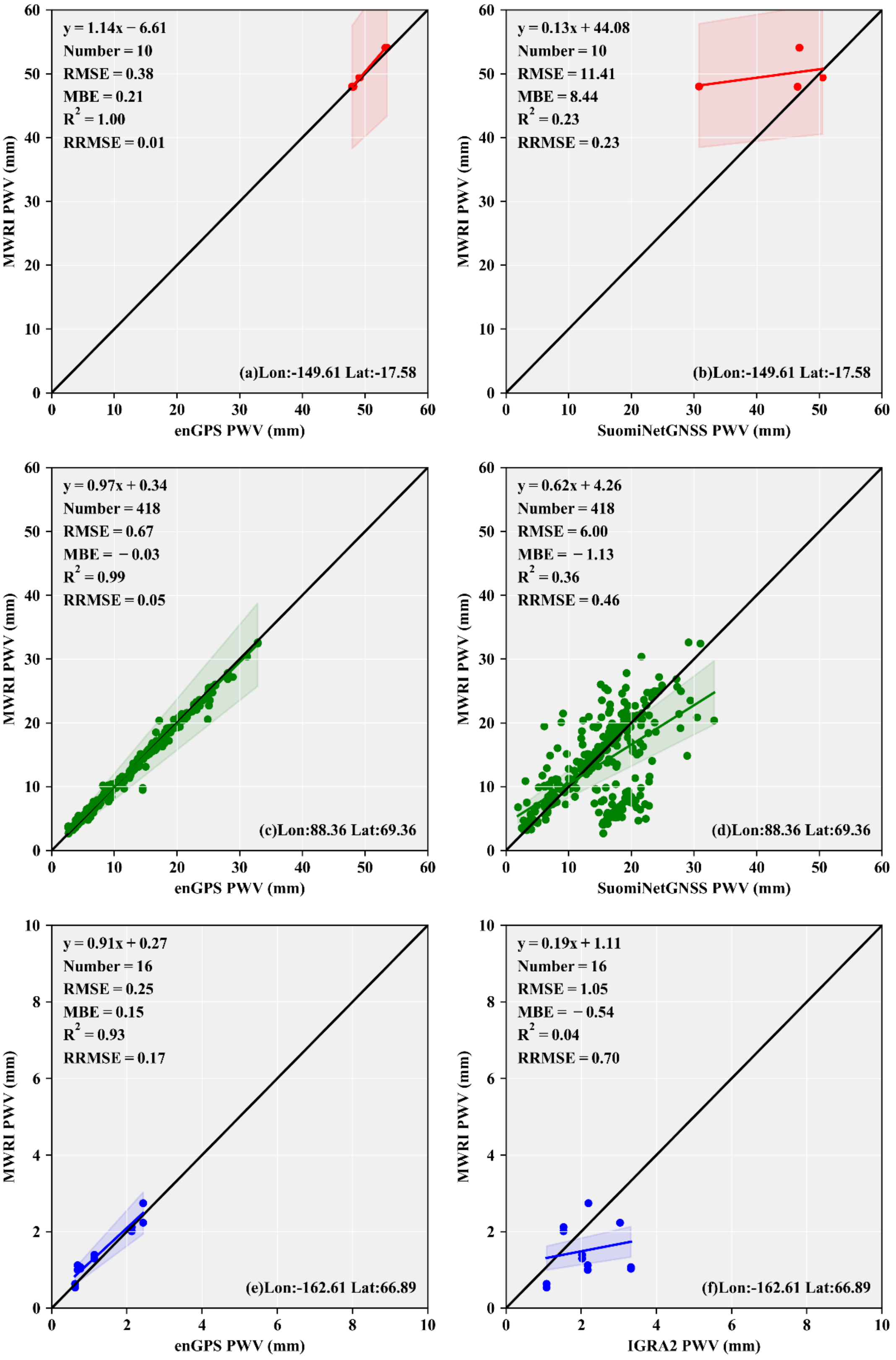

Figure 10. Comparison of PWV from MWRI against enGPS (right) for stations with abnormal differences between MWRI and SuomiNet PWV or IGRA2 PWV (RMSE > 7 mm or RRMSE > 0.4).

4. The manuscript states that "the MWRI PWV exhibits a wet bias at low PWV values and a slight PWV underestimation at high PWV values." Could you provide possible explanations for these observed biases? Discussing potential reasons, such as instrument limitations, atmospheric conditions, or retrieval algorithm issues, would help clarify this point.

R: Thank you for your valuable advice. The reason why the MWRI PWV shows a wet bias at low PWV values and a slight PWV underestimation at high PWV values can be caused by the following factors: 1) We currently lack a feature fully describing extreme PWV conditions (generally associated with rainfall) in our machine learning model. 2) The training samples with extremely lower or higher PWV values are still limited. These reasons would lead to an overestimation or underestimation of extreme dry or wet PWV events in our trained models. In the future, we will try to include more representative PWV samples (e.g. droughts, exceptionally heavy rainfall) to improve the accuracy of the model when more training data points are available.