https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

Measurement Report: Quantifying the Trade-off Between Station Number and Spatial Layout in Sparse GNSS Networks for Calibrating All-Weather FY-4A Precipitable Water Vapor

Yongchao Ma[1], Zhengsheng Chen[1], Tong Liu[2], Zhibin Yu[3], Zhihao Wang[4]

[1] Department of Automation, Rocket Force University of Engineering, Xi'an, China;
[2] Department of Land Surveying and Geo-Informatics, the Hong Kong Polytechnic University, Hong Kong, China
[3] School of Aerospace Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China;
[4] Institute of Geospatial Information, Information Engineering University, Zhengzhou, China;

Correspondence to: Tong Liu (tong2.liu@polyu.edu.hk)

Abstract. Integrating satellite-derived precipitable water vapor (PWV) provides data with high spatiotemporal resolution, which is crucial for monitoring and forecasting extreme weather. However, current fusion and calibration methods typically relies on dense GNSS networks, hindering application in data-sparse regions. It remains unclear whether improving calibration under sparse conditions depends more on increasing station numbers or optimizing their spatial placement. To address this, we developed a machine learning-based calibration framework for FY-4A all-weather PWV and conducted controlled experiments across China. Our key finding is that for a fixed station budget, a spatially random layout consistently outperforms clustered or geographically biased distributions, reducing RMSE by up to 27%. While increasing station density improves spatial generalization, with RMSE at independent stations dropping from 3.24 mm to 2.28 mm and bias converging near zero, performance gains saturate beyond approximately 120-160 stations. Spatially, errors under sparse, non-uniform networks concentrate in regions with strong humidity gradients or complex terrain; a uniform layout distributes errors more evenly. Temporally, all calibrated models capture seasonal cycles, with residual errors peaking in summer due to convective activity. This study demonstrates that in sparse network design, maximizing spatial coverage uniformity is more critical than simply adding stations. We thus provide a transferable framework and a quantitative principle for generating reliable satellite PWV products where GNSS observations are limited.

1 Introduction

Atmospheric water vapor is a key component of the climate system, and its accurate monitoring via precipitable water vapor (PWV) is essential for weather and climate studies (Rocken et al. 1997; Trenberth et al. 2005). High-accuracy, spatiotemporally continuous PWV fields are therefore critical for understanding hydrological processes and improving meteorological forecasts (Lu et al. 2016; Chen and Liu 2016).

At present, PWV observations are mainly derived from ground-based measurements and satellite remote sensing. Ground-based techniques, including radiosondes, microwave radiometers, and Global Navigation Satellite Systems (GNSS), generally

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

provide high accuracy and high temporal resolution, making them suitable for capturing rapid variations in atmospheric water vapor (Ware et al. 2000; Durre et al. 2006; Namaoui et al. 2017). However, their spatial representativeness is strongly constrained by station density and distribution, limiting their ability to provide spatially continuous PWV fields over large regions. In contrast, satellite remote sensing provides regional to global coverage, complementing the spatial limitations of ground networks (Kaufman and Gao 1992; Zhang et al. 2019; Zhao et al. 2024). Nevertheless, near-infrared and thermal infrared PWV products are often affected by cloud contamination, viewing geometry, and retrieval assumptions, leading to insufficient accuracy, missing data, and poor spatiotemporal continuity. These issues are particularly pronounced over complex underlying surfaces or in regions with strong water vapor gradients (Jiang et al. 2024), thereby limiting the applicability of satellite-derived PWV in high-resolution moisture monitoring and process-oriented studies.

To improve the quality of satellite PWV products, existing studies have generally followed two main technical pathways. One approach focuses on post-processing calibration of satellite PWV using ground-based observations to reduce systematic biases and enhance consistency across different observing systems. The other approach aims to improve retrieval algorithms and parameterization schemes at the source, thereby strengthening physical consistency and reducing retrieval uncertainty (Merrikhpour and Rahimzadegan 2017; He and Liu 2020).

In post-calibration studies, GNSS-derived PWV has been widely adopted as a reference “truth” for satellite PWV calibration. Investigations based on relatively large GNSS networks consistently demonstrate that multi-station constraints can substantially improve the accuracy and stability of satellite PWV products. For example, Bai et al. (2021) constructed a linear MODIS PWV calibration model using 260 GNSS stations over China, achieving an overall accuracy improvement of approximately 20%. Subsequently, Researchers developed machine-learning-based calibration models for MODIS and FY-3A PWV using 238 and 214 GNSS stations over China, respectively, significantly enhancing the long-term performance of satellite PWV under all-sky conditions (Ma, Yao, Zhang, and He 2022; Xu and Liu 2023a). Xu and Liu (2023b) further validated the effectiveness of machine learning approaches in improving all-weather satellite PWV quality using 453 ground stations in Australia. To address the limited spatiotemporal continuity of PWV maps, Ma et al. (2023) proposed a distributed ensemble framework integrating multi-source heterogeneous data based on 207 GNSS stations in New South Wales. Ma et al. reported that when only 14 GNSS stations in the Tibetan Plateau were used to construct a calibration model, the sparsity of stations prevented machine learning approaches from achieving high calibration accuracy (Ma, Yao, Zhang, and Du 2022). The consistent conclusions drawn from these studies across different regions and satellite products indicate that the number of GNSS stations is a key prerequisite influencing satellite PWV calibration performance. However, most existing studies treat the available station network as a given condition, and whether increasing the number of stations necessarily leads to linear or unlimited performance gains remains insufficiently understood.

Meanwhile, dense GNSS PWV observations have also been used to develop multi-parameter satellite PWV mapping models. Researchers incorporated land-cover information using 173 GNSS stations in North America and achieved more than a 66% improvement in MODIS near-infrared PWV retrieval accuracy through multi-channel coupling (Ma, Yao, Zhang, Qin, et al. 2022). In addition, progress has been made in high-accuracy PWV retrieval from Fengyun satellites(Zhao et al. 2024), as well

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

as in all-weather PWV retrieval through the fusion of infrared, near-infrared, and microwave observations (Sun et al. 2024; Du et al. 2025). Despite these advances, station number and spatial configuration are typically regarded as background conditions rather than explicitly treated as independent variables governing model performance.

In summary, existing satellite PWV reconstruction and calibration studies primarily aim to establish mapping relationships between satellite observational information—including products, spectral features, spatiotemporal attributes, and environmental variables—and high-accuracy ground-based PWV, particularly GNSS-derived PWV. However, most studies have concentrated on regions with relatively dense GNSS station coverage, while quantitative understanding of model performance, stability, and generalization behavior under sparse station conditions is still lacking. How the number of training stations and their spatial configuration jointly constrain the spatial consistency and extrapolation capability of calibration models has not been systematically evaluated. To address this gap, we develop a machine learning-based calibration framework for the all-weather FY-4A PWV product over China. Through controlled experiments, we systematically isolate and quantify the impacts of training station density and spatial layout on model accuracy, spatial generalization, and temporal stability. The objective is to provide quantitative guidance on station configuration and methodological support for satellite PWV reconstruction under sparse observational conditions.

The remainder of this paper is organized as follows. Section 2 describes the study area, datasets, and preprocessing procedures. Section 3 presents the all-weather PWV model construction and evaluation metrics. Section 4 reports the experimental results and discussion. Finally, Section 5 summarizes the main conclusions and outlines future research directions.

2 Study Area and Data

2.1 Study Area

This study focuses on mainland China as the study region (18°–53° N, 73°–135° E; Fig. 1). The region exhibits pronounced topographic heterogeneity characterized by a stepwise terrain pattern, including the Tibetan Plateau, inland basins, and eastern plains. Such complex terrain leads to strong regional contrasts in climatic background and water vapor transport and convergence mechanisms, resulting in pronounced spatial heterogeneity in PWV.

To systematically evaluate the influence of training sample size and spatial configuration on the generalization capability of all-weather PWV models, a total of 244 GNSS stations were selected as modeling samples. In addition, 16 GNSS stations were randomly excluded from the training process and reserved as an independent validation set. Furthermore, 80 radiosonde stations from the Integrated Global Radiosonde Archive (IGRA) were selected as an external independent validation dataset. The spatial distribution of all stations is shown in Fig. 1, where Fig. 1(b) highlights the randomly selected 16 independent GNSS validation stations.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

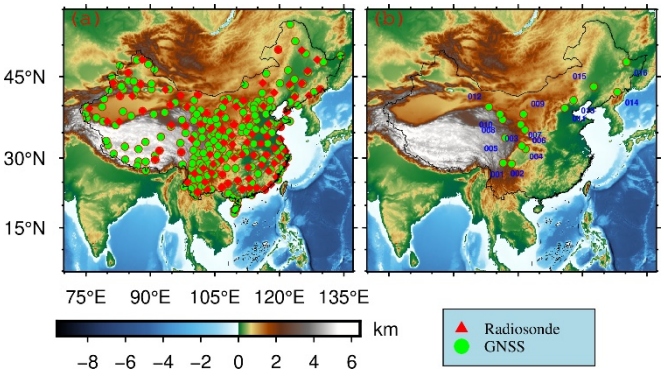

Figure 1: Spatial Distribution of GNSS and IGRA Stations in China.

2.2 Data and Processing

2.2.1 FY-4A

The Fengyun-4A (FY-4A) satellite is China's second-generation geostationary meteorological satellite, operating in geostationary orbit and carrying the Advanced Geostationary Radiation Imager (AGRI). AGRI provides high-temporal-resolution observations; in addition to full-disk scanning, it performs regional observations over China at a 15-min interval. FY-4A PWV products are freely available in near real time from the official data portal of National Satellite Meteorological Center (http://satellite.nsmc.org.cn). Because near-infrared channels cannot penetrate thick cloud cover, cloud detection is required prior to PWV retrieval. In this study, cloud-contaminated PWV observations were identified and excluded using the MERSI cloud mask product.

2.2.2 GNSS

GNSS-derived PWV data were obtained from the China Crustal Movement Observation Network (CMONOC), which provides continuous, high-precision, and high-temporal-resolution observations across mainland China. CMONOC has been widely used for monitoring crustal deformation, gravity field variations, tropospheric water vapor, and ionospheric electron content (Li et al. 2012). In this study, CMONOC observations for 2023 were processed using the GAMIT/GLOBK software to generate hourly zenith total delay (ZTD) time series. Zenith hydrostatic delay (ZHD) and zenith wet delay (ZWD) were separated, and PWV was derived using standard conversion factors. The basic relationship can be expressed as

$$ZWD = ZTD - ZHD \quad (1)$$

$$PWV = ZWD \cdot \frac{10^6}{\rho_w \cdot R_w \cdot \left(\frac{k_3}{T_m} + k_2' \right)} \quad (2)$$

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

$$Tm = \frac{\int \frac{e}{T} ds}{\int \frac{e}{T^2} ds} \quad (3)$$

where k_3 = 377600.0 K²/Pa, k'_2 = 16.52 K/hPa, T is temperature (K), ρ_w is the density of liquid water, R_w is the specific gas constant for water vapor (461.5 J kg^{-1} K^{-1}), P is water vapor pressure (hPa), and ds denotes the vertical layer thickness. Surface pressure and weighted mean temperature required for GNSS PWV estimation were provided by ERA5.

ZHD was computed using the Saastamoinen model,

$$ZHD = \frac{0.0022768 \cdot P_s}{1 - 0.00266 \cdot \cos(2 \cdot \varphi) - 0.00028 \cdot h_s} \quad (4)$$

which achieves sub-millimeter accuracy under standard atmospheric conditions (Vedel et al., 2001). The resulting GNSS PWV was treated as one of the reference "truth" datasets for all-weather PWV model training and validation.

2.2.3 Radiosonde

Radiosonde-derived PWV data were obtained from the Integrated Global Radiosonde Archive (IGRA), which is freely accessible at https://www.ncei.noaa.gov/products/weather-balloon/integrated-global-radiosonde-archive. PWV derived from 2023 radiosonde observations was used as an external independent validation dataset to assess model consistency across different observation systems. To ensure temporal representativeness and continuity, stations with severe data gaps were excluded. Only stations with valid observations for no less than one third of the year (i.e., at least 120 days per year) were retained. After quality control, PWV data from 80 radiosonde stations in 2023 (Fig. 1) were selected for final model evaluation.

2.2.4 Topographic and Land Cover

Digital elevation model (DEM) data from the Shuttle Radar Topography Mission (SRTM) were used as the elevation reference for FY-4A PWV. The SRTM DEM was jointly produced by NASA and the National Geospatial-Intelligence Agency (NGA) and is available at http://www.resdc.cn/. In this study, SRTM DEM data with a spatial resolution of 90 m were resampled to 4 km to match the spatial resolution of the reconstructed FY-4A PWV. Land cover information was obtained from China's first annual land cover dataset derived from Landsat imagery. The dataset includes nine land cover types: cropland, forest, shrubland, grassland, water bodies, snow/ice, bare land, impervious surfaces, and wetlands. Considering the 4-km spatial footprint of FY-4A PWV, a circular buffer with a radius of 2 km was constructed around each GNSS station. The area fractions of different land cover types within each buffer were calculated and used as model input variables to enhance the representation of surface heterogeneity and regionally non-uniform errors.

2.2.5 NDVI

Vegetation conditions and evapotranspiration processes play an important role in near-surface moisture sources and recycling. Therefore, the normalized difference vegetation index (NDVI) serves as an effective indicator for characterizing land-surface

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

contributions to atmospheric water vapor and regional variability (Ma, Yao, Zhang, and He 2022). NDVI data were obtained from the MODIS vegetation index product MOD13A2, which is generated from atmospherically corrected daily bidirectional surface reflectance and provides NDVI at a spatial resolution of 1 km and a temporal resolution of 16 days (https://ladsweb.modaps.eosdis.nasa.gov/). Given its temporal characteristics, a fixed 16-day cycle was adopted, and a constant NDVI value was assigned within each cycle to ensure data consistency and avoid additional uncertainty introduced by excessive temporal interpolation.

3. Methods

To reduce the impact of cloud contamination and missing observations in FY-4A PWV products on statistical analysis and spatial comparison, an all-weather FY-4A PWV reconstruction was first performed to generate spatiotemporally continuous PWV fields. Subsequently, GNSS-derived PWV was introduced as an external constraint to calibrate reconstructed FY-4A PWV and reduce systematic biases. The stability and generalization of the all-weather PWV model were then evaluated under different training station numbers and spatial configurations. The overall workflow is illustrated in Fig. 2.

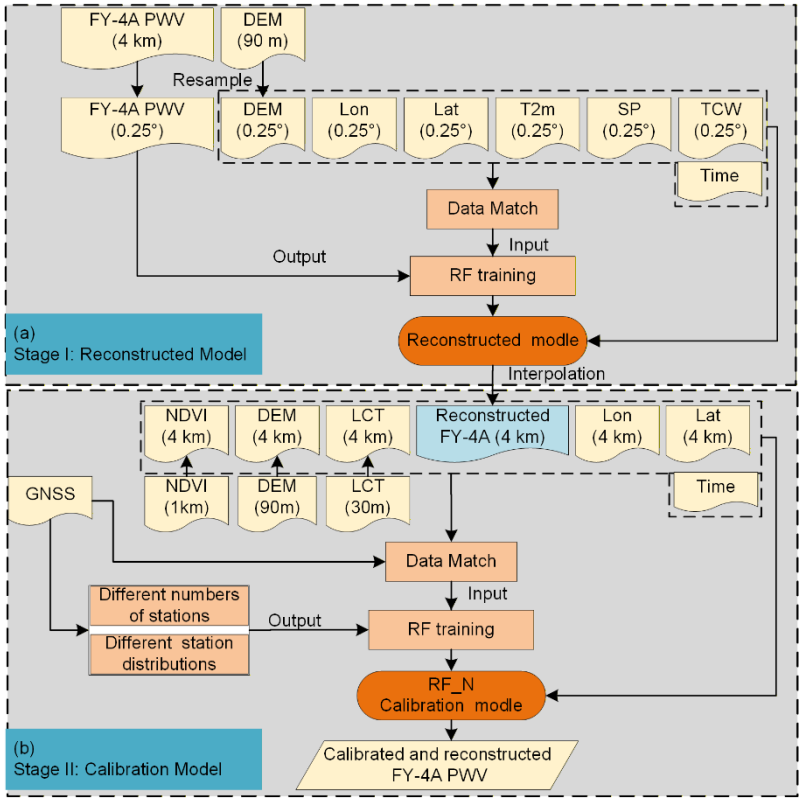

Figure 2. Flowchart of the all-weather PWV map reconstruction and calibration.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

3.1 All-weather FY-4A PWV Reconstruction

The reconstruction aims to restore PWV continuity under cloudy or missing-data conditions by establishing a nonlinear mapping between FY-4A PWV and multiple auxiliary variables, including meteorological parameters, topography, time, and spatial information. Prior to modeling, FY-4A PWV, ERA5 variables, and DEM data were temporally aligned to hourly observations and spatially resampled to a common grid over China.

A random forest (RF) model was employed to describe the nonlinear relationship between PWV and auxiliary variables:

$$PWV_{FY-4A} = f_{RF}\left(Lat, Lon, Dem, Time, T2m, TCW, SP\right) \quad (5)$$

where *T2m*, *SP*, and *PWV* denote near-surface temperature, surface pressure, and water vapor–related parameters, respectively; *Dem* represents elevation; *Lat* and *Lon* denote spatial location; and *TCW* is total column water. After training, the model was applied to generate spatiotemporally continuous reconstructed FY-4A PWV fields. Detailed validation of the reconstruction has been reported in previous studies (Wang et al. 2026) and is not repeated here.

3.2 GNSS-constrained FY-4A PWV Calibration

Although reconstruction improves data completeness, regional systematic biases remain in FY-4A PWV. Therefore, GNSS PWV was used as a reference to calibrate reconstructed FY-4A PWV. To address spatial mismatches between point-based GNSS observations and gridded satellite PWV, bilinear interpolation was applied to project GNSS PWV onto the FY-4A grid. Temporal mismatches were handled by averaging GNSS PWV at the two nearest epochs before and after satellite overpass times.

An RF-based calibration model was constructed to map reconstructed FY-4A PWV, spatiotemporal information, NDVI, and land cover characteristics to GNSS PWV:

$$PWV_{GNSS} = f_{RF}\left(Lat, Lon, Elv, Time, NDVI, P_{LCTi}, PWV_{FY}\right) \quad (6)$$

where P_{LCTi} denotes the area fraction of the i-th land cover type within a 2-km buffer around each GNSS station.

3.3 Station Configuration Experiments

To quantitatively assess the effects of training station density and spatial structure on model generalization, two groups of comparative experiments were designed (Fig. 2b).

(1) Training station number experiment. Different numbers of GNSS stations were randomly selected from the available 244 stations to construct training datasets. Seven training sample sizes were considered: No. = 30, 40, 80, 120, 160, 200, and 244. The remaining GNSS stations were used as independent validation samples, and IGRA radiosonde PWV data were further introduced for external validation.

(2) Spatial structure experiment. Under the same training station number, three typical spatial configurations were constructed, including random, surrounding, and clustered distributions. These configurations were designed to evaluate how spatial coverage uniformity and representativeness influence spatial error patterns and extrapolation capability.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

Through this experimental design, the relationships among sample size, spatial coverage, and error response were systematically examined, enabling the identification of major limiting factors governing model performance improvement and the corresponding saturation thresholds.

3.4 Model Parameter Optimization

Key RF parameters, including the number of trees and leaf nodes, were optimized using out-of-bag (OOB) error as the evaluation criterion. Bayesian optimization was applied to automatically search the parameter space, and five iterative optimization cycles were performed to reduce random variability. The final optimal parameter settings are summarized in Table 1.

Table 1. Random Forest Hyperparameter Values Based on Different Station Configurations

Number of Station	Leaf Nodes	Decision Trees	Distribution Type	Leaf Nodes	Decision Trees
30	1	150	200/Surrounding	1	200
40	1	150	200/Clustered	1	150
80	1	150	200/ Random	1	200
120	1	150			
160	1	200			
200	1	200			
244	1	200			

3.5 Evaluation Metrics and Validation Strategy

Model performance was evaluated at three levels, including training GNSS stations, independent GNSS validation stations, and IGRA radiosonde stations, to distinguish fitting capability from spatial extrapolation performance. The coefficient of determination, RMSE, and mean bias were adopted as the primary evaluation metrics:

$$Bias = \frac{1}{m}\sum_{i=1}^{m}\left(Data^{model} - Data^{ref}\right) \quad (7)$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(Data^{model} - Data^{ref}\right)} \quad (8)$$

$$R^2 = 1 - \frac{\sum\left(Data^{ref} - Data^{model}\right)^2}{\sum\left(Data^{ref} - \overline{Data^{ref}}\right)^2} \quad (9)$$

where $Data^{model}$ denotes model-derived PWV and $Data^{ref}$ represents reference PWV observations.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

In addition to overall statistical metrics, model robustness was further assessed in terms of spatial error patterns, PWV time series at representative stations, and monthly aggregated statistics. These analyses were conducted to evaluate model stability, seasonal dependence, and spatial consistency across different observational networks.

4. Results and Discussion

4.1 Overall Performance

A central finding of our experiments is that in-sample fitting accuracy is remarkably insensitive to the number of training stations, whereas spatial generalization (validation at independent sites) critically depends on it. As shown in Figure 3, for training stations, the accuracy of all-weather PWV model remains consistently high as the number of training stations increases from 30 to 244. The coefficient of determination R^2 remains stable within the range of 0.988–0.990, while RMSE varies only slightly between 1.45 mm and 1.66 mm. Mean bias is close to zero throughout all configurations ($|Bias| \leq 0.02$ mm). These results suggest that the random forest model is capable of sufficiently learning the regional-scale mapping relationship between GNSS PWV and FY-4A PWV, and that further increases in training samples yield only marginal improvements in fitting accuracy.

In contrast, the number of training stations has a pronounced effect on model performance at independent GNSS validation stations. As training station numbers increase, retrieval accuracy at independent stations improves steadily, with R2 increasing from approximately 0.96 to 0.97 and RMSE decreasing markedly from 3.24 mm to 2.28 mm. Meanwhile, Bias gradually converges from 0.42 mm toward near-zero values. These results demonstrate that enlarging the training dataset enhances the model's ability to represent regional PWV heterogeneity and effectively suppresses systematic errors during spatial extrapolation. Notably, when the number of training stations exceeds approximately 160, improvements in independent-station accuracy become marginal, indicating a saturation behavior in model performance and suggesting that the dominant statistical characteristics of PWV over the study region have been adequately captured.

This saturation identifies a cost-effective range for station numbers. Beyond this range, merely adding stations yields diminishing returns. To dissect the drivers of this saturation and explore how to optimize performance within a fixed station budget, we now turn to a spatial analysis of errors.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

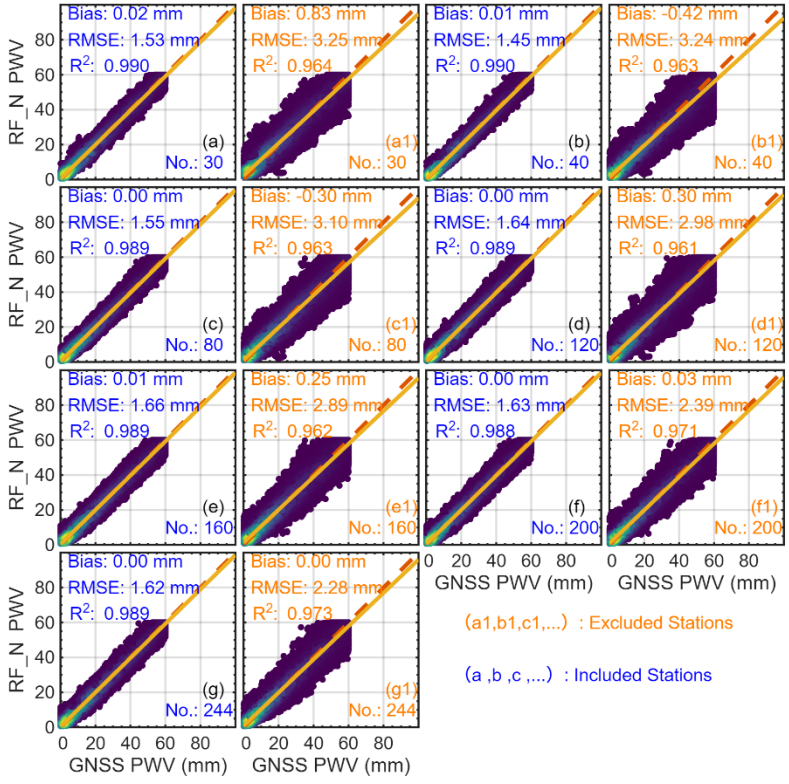

Figure 3. Validation Results of Training Models Based on Different Numbers of GNSS Stations

Figure 4 further illustrates the effect of training station number on retrieval performance at independent validation stations. Uncalibrated FY-4A PWV exhibits substantial systematic underestimation (Bias: −2.17 mm) accompanied by large random errors (RMSE: 3.93 mm). With increasing training station numbers, calibration errors are significantly reduced and spatial performance becomes more stable, with particularly pronounced improvements observed in low-latitude regions. When the training station number reaches 200, the mean Bias decreases to 0.07 mm, while RMSE shows a monotonic decline from 3.93 mm to 1.90 mm.

The consistency of these improvements across independent stations indicates that increasing training sample size effectively mitigates systematic bias and enhances model generalization. From a mechanistic perspective, when training data are limited, the model cannot adequately represent the spatial heterogeneity of regional PWV, leading to error accumulation and amplified random errors in regions with distinct climatic backgrounds, surface conditions, and moisture variability. As training station numbers increase, the diversity of PWV scenarios represented in the training data expands, enabling the model to learn more representative regional statistical features and thereby improving both accuracy and stability at independent locations.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

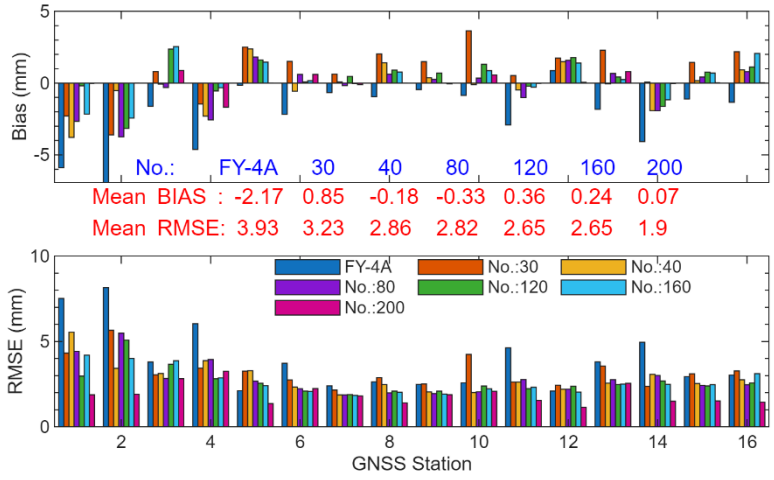

Figure 4: Validation Results of different Models Using GNSS Stations Not Involved in Model Training

Independent validation using IGRA radiosonde PWV is summarized in Table 2. When the number of training stations increases from 30 to 120, model errors decrease substantially, with RMSE reduced from 2.68 mm to 2.36 mm and Bias converging to near-zero values (0.02 mm). This indicates that a moderate training sample size is enough to achieve a favorable balance between accuracy and robustness. Further increasing the number of training stations to 160–244 results in only minor fluctuations in RMSE (approximately 2.37 mm–2.39 mm) and Bias, confirming that model performance gradually approaches saturation.

Table 2. Validation Results of different Models Using IGRA Stations.

Model	Bias [mm]	RMSE [mm]
RF_30	0.29	2.68
RF_40	-0.86	2.72
RF_80	-0.47	2.46
RF_120	0.02	2.36
RF_160	-0.14	2.38
RF_200	-0.39	2.39
RF_244	-0.14	2.37

Overall, the consistent patterns revealed by Figs. 2 and 3 and Table 2 demonstrate that increasing training station density is essential for improving the spatial generalization capability of GNSS PWV calibration models. However, larger sample sizes do not necessarily translate into proportional performance gains. From the perspective of station number and random spatial distribution alone, a training station scale of approximately 120–200 provides an effective compromise between accuracy improvement and computational efficiency for regional-scale PWV calibration over China.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

4.2 Spatial Dependence of Model Performance

While the analysis above confirms the role of station number and its saturation, a more pressing practical question emerges under resource constraints: Can the spatial layout of a fixed number of stations serve as a lever to enhance performance beyond what the number alone dictates? To answer this, we analyze the spatial patterns of model errors.

Figure 5 presents the spatial distributions of RMSE and Bias over China based on IGRA radiosonde validation under different training station numbers, together with the spatial distribution of training GNSS stations. Overall, both RMSE and Bias exhibit pronounced spatial heterogeneity, reflecting the complex spatial variability of atmospheric water vapor and its controlling factors.

When the number of training stations is limited (e.g., No. = 30 and No. = 40), spatial discrepancies in model errors are particularly pronounced. RMSE is generally higher in regions with more active moisture variability, especially in eastern and southern China, where RMSE at some stations exceeds 3.0 mm. At the same time, Bias shows a clear pattern of alternating positive and negative values, indicating strong regional dependence during spatial extrapolation. These results suggest that insufficient training samples with limited spatial coverage hinder the model's ability to capture regional-scale PWV heterogeneity, leading to increased systematic errors in specific regions.

As the number of training stations increases to 80 and 120, the spatial error structure improves substantially. Overall RMSE levels decrease, the extent of high-error regions shrinks markedly, and RMSE at most stations concentrates within approximately 2.0 mm–2.5 mm. Meanwhile, the spatial distribution of Bias becomes smoother, with extreme positive and negative deviations significantly reduced. This improvement indicates a substantially enhanced capacity of the model to adapt to regional differences in water vapor variability. When the number of training stations further increases to 160, the improvement in RMSE and Bias becomes relatively limited, suggesting that the dominant spatial variability patterns of PWV have largely been captured under the current data and regional setting, and that further increases in training samples yield diminishing returns in terms of spatial error reduction.

Overall, these spatial patterns confirm that small training datasets with insufficient spatial coverage are unable to represent regional PWV heterogeneity, resulting in localized systematic biases. In contrast, increasing training station numbers enhances the model's adaptability to diverse climatic backgrounds and moisture regimes, thereby effectively reducing spatially non-uniform errors. Once the number of training stations reaches approximately 120–160, model performance gradually approaches saturation, consistent with the statistical results discussed previously. This consistency further emphasizes the importance of jointly considering training sample size and spatial coverage in regional-scale PWV calibration.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

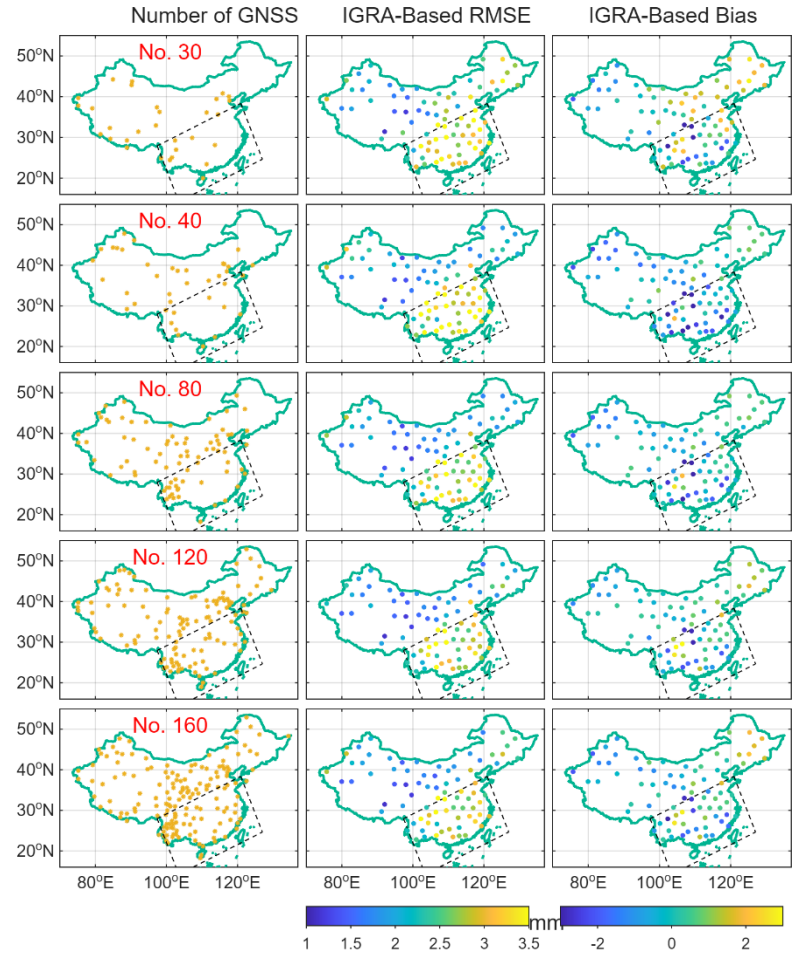

Figure 5. Spatial distribution of validation accuracy for different models using IGRA stations.

To further examine the role of spatial density, Table 3 summarizes the relationship between calibration accuracy and station spatial resolution, quantified by the mean inter-station distance derived from a Delaunay triangulation. RMSE shows an overall positive correlation with mean station spacing, indicating that denser station networks provide stronger spatial constraints on the model. When the number of training stations exceeds approximately 120, improvements in accuracy become notably weaker, again exhibiting a diminishing marginal benefit. This behavior can be attributed to the effective reduction in inter-station distance with increasing station numbers, which enhances the spatial representativeness of training data and improves the model's ability to characterize PWV spatial variability. These results demonstrate that, beyond sample size, the spatial configuration of training stations is a critical factor limiting further improvements in calibration performance.

Due to the combined influence of land–sea contrast, temperature fields, and large-scale circulation, southern China (black-box region) exhibits stronger spatiotemporal variability in water vapor and more pronounced short-term fluctuations than northern China. Under limited station coverage, the model struggles to capture fine-scale variability in this region, resulting in generally

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

lower accuracy compared with northern China. For example, for the RF model, increasing the training configuration from RF_N30 to RF_N160 reduces mean station spacing by 46.40% in southern China and leads to a 13.04% reduction in RMSE, whereas in northern China, mean spacing is reduced by approximately 55% and RMSE decreases by 17.10%. These comparisons further demonstrate that reducing inter-station distance and improving spatial coverage density can significantly enhance calibration accuracy, although the magnitude of improvement is constrained in regions with more intense moisture variability.

Table 3. Comparison of GNSS Spatial Resolution and Model Accuracy Using IGRA Station

Model	Outside the black box			Inside the black box		
	Average Distance [km]	Bias [mm]	RMSE [mm]	Average Distance [km]	Bias [mm]	RMSE [mm]
RF_N30	821.81	0.52	2.30	626.84	0.01	3.16
RF_N40	646.35	-0.52	2.12	784.79	-1.28	3.45
RF_N80	481.59	-0.32	1.95	473.30	-0.66	3.08
RF_N120	441.37	0.08	1.90	381.58	-0.05	2.91
RF_N160	369.75	-0.10	2.00	336.01	-0.18	2.86

Figure 6 further compares the spatial distribution of FY-4A PWV retrieved by different models under varying training station numbers. A comparison of Fig. 6 (a–c) indicates that the all-weather model can effectively fill in regional missing values. At large spatial scales, the RF_N244 retrieval exhibits strong consistency with the GNSS-derived PWV field, characterized by a gradual increase from northwestern to southeastern China. To better illustrate the influence of training sample size, RF_N244 is used as a reference benchmark, and spatial differences between other models and RF_N244 are analyzed, as shown Fig.6 (d-i). When training station numbers are small (e.g., RF_N30 and RF_N40), deviations relative to RF_N244 are large and spatially heterogeneous, with local differences exceeding 6 mm in some regions. These discrepancies are particularly pronounced in areas with complex terrain or strong moisture gradients, indicating elevated uncertainty in representing local PWV structures under sparse training conditions.

As training station numbers increase, spatial differences relative to RF_N244 decrease markedly, with patterns transitioning from scattered to more continuous and smoother distributions. Extreme difference regions shrink substantially, and when training station numbers reach 160–200, only minor small-scale fluctuations remain. Notably, although station numbers increase beyond 120, the overall spatial coverage pattern does not improve substantially, reinforcing the conclusion that station layout and coverage uniformity are as important as sample size in constraining model performance.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

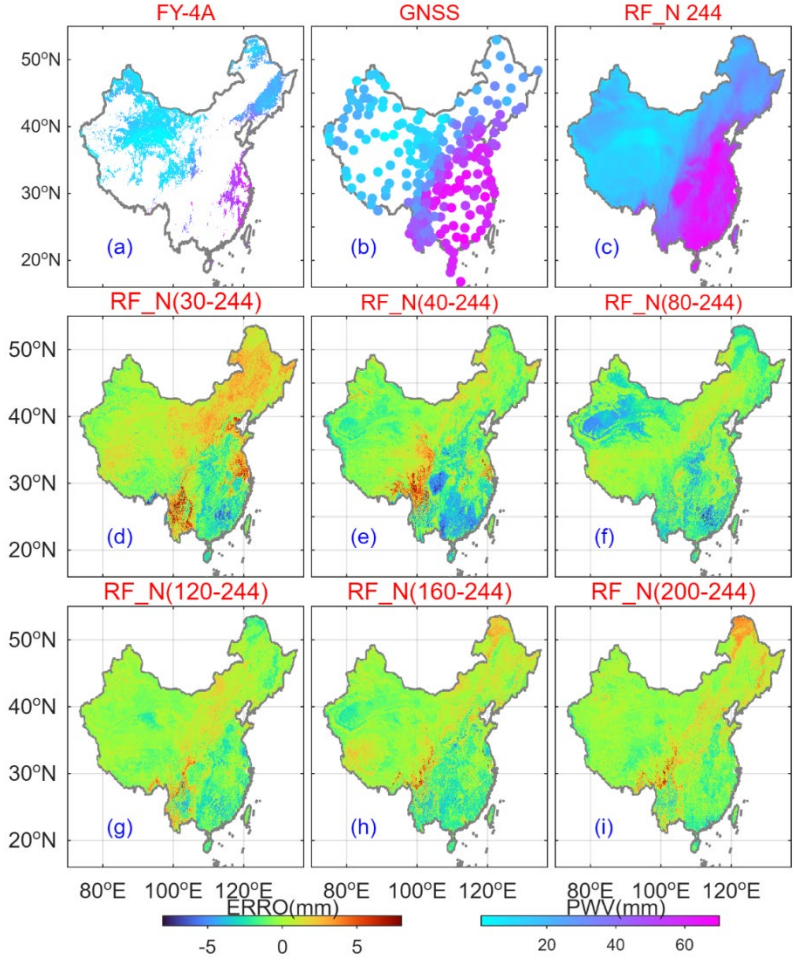

Figure 6. Comparison of retrieval results at 12:00 on the 232nd day of 2023 based on training models with different station configurations. (a-c) respectively display FY-4A PWV, GNSS PWV, and RF_N244 model derived PWV. (d-i) show the differences in retrieval results between each model and the RF_N244 model.

The persistent spatial error patterns under sparse conditions suggest that the spatial representativeness and coverage uniformity of stations may be a more critical performance controller than sheer quantity. To test this hypothesis directly, we designed a controlled experiment isolating the effect of spatial configuration while holding the station number constant.

To explicitly isolate the effect of spatial configuration, calibration models were trained using GNSS stations arranged under three different spatial distribution structures, and their performance was independently evaluated using the remaining GNSS stations. The validation results are summarized in Table 4.

As shown in Table 4, the randomly distributed configuration consistently achieves the best calibration performance. Compared with the clustered and surrounding configurations, the random distribution reduces RMSE by 27.08% and 26.59%, respectively, and reduces Bias by 65.96% and 15.69%, respectively. These results indicate that a spatially uniform and representative training set is more effective in constraining regional PWV variability and suppressing systematic errors.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

In contrast, the surrounding configuration exhibits the poorest performance among the three layouts. This degradation is likely attributable to the spatial mismatch between training and validation stations: most of the remaining validation stations are concentrated in the mid- and low-latitude regions, where water vapor variability is stronger and more dynamic. The lack of representative training information for these regions limits the model's ability to characterize intense moisture variability, leading to increased random errors and systematic bias during spatial extrapolation.

Overall, Table 4 show that, for regional-scale PWV calibration, spatial representativeness and coverage uniformity of training stations are more critical than merely increasing station numbers. Local clustering or surrounding layouts do not guarantee improved accuracy and may even degrade spatial extrapolation capability, whereas uniformly distributed random configurations offer the most robust and transferable solution for GNSS-constrained satellite PWV calibration.

Table 4. Accuracy of models trained with different station distribution structures across the remaining GNSS stations

Remain GNSS	RF_Area1 (Clustered)		RF_Area2 (Surrounding)		RF_Area3 (Random)	
	RMSE [mm]	Bias [mm]	RMSE [mm]	Bias [mm]	RMSE [mm]	Bias [mm]
Area 1	2.77	-0.47	--	--	2.02	0.16
Area 2	--	--	3.61	-0.51	2.65	-0.43
Area 3	--	--	--	--	2.86	0.04

Figure 7 compares model performance under different station layouts while keeping the training station number fixed (No. = 200). Fig. 7 (d-f) visualize the spatial difference patterns of RF_N244 retrievals under distinct station layouts, while Fig. 7 (g-i) depict the actual geographical arrangements corresponding to clustered, surrounding, and randomly distributed stations, respectively. Even with identical sample sizes, significant performance differences emerge among spatial configurations: random distributions perform best (RMSE: 1.44 mm, Bias: -0.34 mm), followed by surrounding configurations (RMSE: 1.75 mm, Bias: -0.68 mm), while clustered configurations perform worst (RMSE: 2.39 mm, Bias: -0.36 mm). This finding highlights the decisive role of spatial coverage uniformity and representativeness in determining model generalization capability.

Further inspection reveals that performance differences among configurations are relatively small over low-elevation plains, whereas the largest discrepancies occur in regions characterized by strong topographic and climatic gradients, such as plateaus, basins, and transition zones. A more balanced station layout provides more comprehensive spatial constraints, thereby improving the representation of water vapor variability in complex terrain. For instance, compared with a surrounding configuration, a more uniformly distributed configuration yields markedly improved performance over the Tibetan Plateau and the Yunnan–Guizhou Plateau, indicating reduced systematic bias at high elevations.

It is noteworthy that although clustered configurations increase station density locally, they result in the poorest overall performance, as shown Fig. 7 (d). In regions with strong spatial heterogeneity in water vapor and surface conditions, locally clustered observations can introduce spatial bias in the training dataset, making it difficult for a single global mapping function to simultaneously represent diverse environments such as plateaus, mountains, plains, lakes, and urban areas. Consequently,

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

fitting errors accumulate and spatial extrapolation capability deteriorates. These results demonstrate that blindly increasing station density through localized clustering may be counterproductive, and that optimized spatial distribution is essential for robust regional calibration.

Taken together, the spatial analysis confirms that while training station number is important, optimizing spatial layout and coverage uniformity under a given sample size is even more critical for improving calibration accuracy and robustness in practical applications.

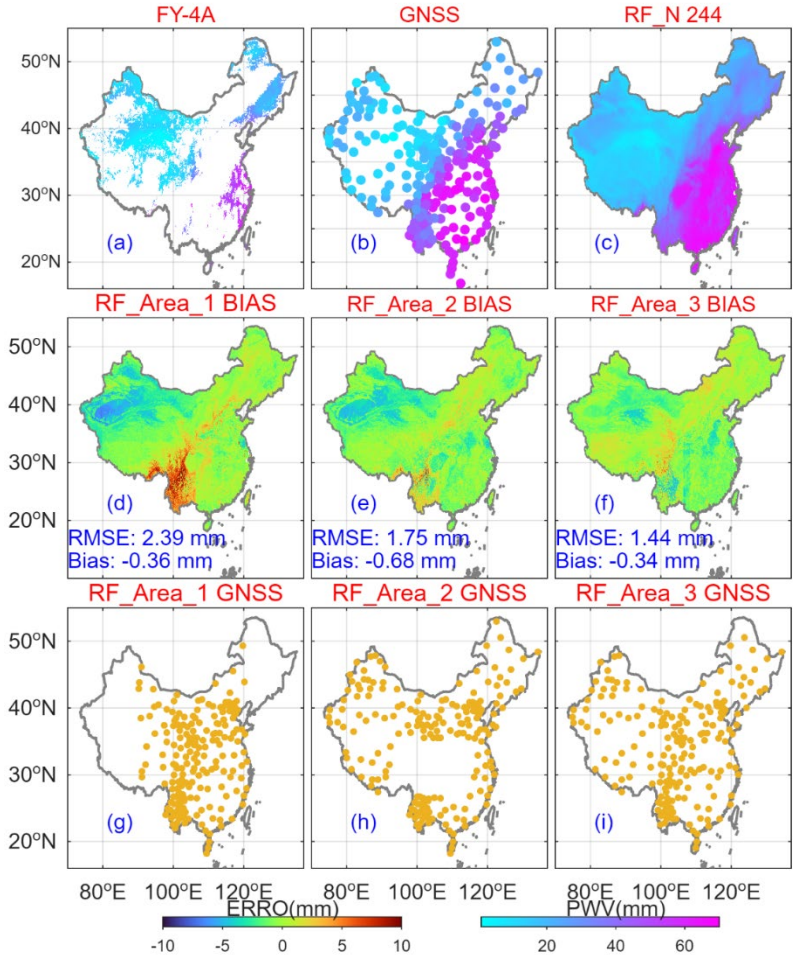

Figure 7. Comparison of retrieval results at 12:00 on the 232nd day of 2023 based on training models with different station distributions.

4.3 Temporal Stability of Model Performance

Finally, we evaluate temporal stability to assess whether optimizations in station number and layout yield robust performance under dynamic atmospheric conditions. Figure 8 presents PWV time series comparisons at three GNSS stations located at different latitudes. All three stations exhibit pronounced seasonal cycles, characterized by high PWV in summer and low PWV

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

in winter. Across different training configurations, RF-based models generally capture the seasonal evolution of GNSS PWV well; however, differences remain in amplitude representation and short-term variability.

When training station numbers are small (e.g., RF_No. = 30), the model tends to underestimate peak PWV during high-moisture periods, with evident smoothing effects in some intervals. As training station numbers increase to 80 and 200, agreement with GNSS observations improves substantially, particularly at mid- and low-latitude stations. This improvement indicates that enlarging the training dataset enables the model to better learn both the amplitude and temporal structure of PWV variability.

Analysis of error time series further reveals that, although errors generally fluctuate around zero, their magnitude and dispersion strongly depend on training station number. Sparse training configurations exhibit larger positive and negative oscillations and higher error dispersion, whereas increasing training station numbers leads to more concentrated error distributions and a marked reduction in extreme errors. These results demonstrate that increasing training station numbers enhances not only mean accuracy but also temporal stability and robustness. Nevertheless, occasional synchronized biases across different training configurations persist during some high-PWV periods, suggesting that retrieval errors are not solely controlled by training sample size, but may also be influenced by rapid moisture transport, strong convective activity, and observation or matching uncertainties.

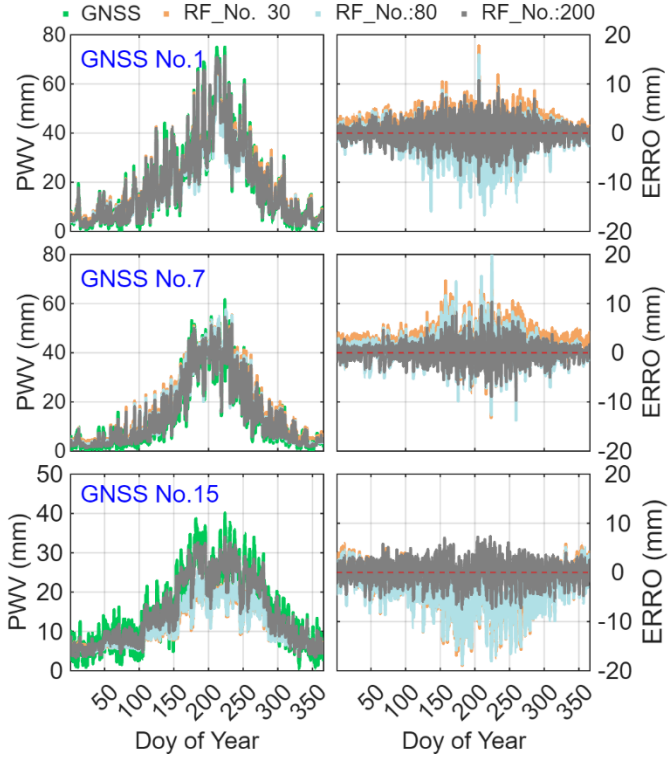

Figure 8. Time-series Comparison of PWV Retrievals from Models Trained with Different Numbers of Stations.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

Monthly-scale variations in Bias and RMSE under different training station numbers are shown in Figure 9, with the original FY-4A PWV product included for comparison. Overall, calibrated models exhibit substantially lower Bias and RMSE than the original FY-4A product throughout the year, while also displaying a clear seasonal dependence, with larger errors in summer and smaller errors in winter.

In terms of Bias, monthly mean values fluctuate around zero for all calibrated models, but their amplitudes depend on training station number. Models trained with fewer stations show more pronounced positive and negative deviations in certain months, particularly during seasonal transitions or under high-PWV conditions. As training station numbers increase to 80–120 and beyond, monthly Bias variations converge markedly toward zero, indicating effective suppression of systematic errors and improved monthly stability.

RMSE exhibits a consistent seasonal pattern across all models, with maximum values occurring during June–August and minimum values in winter months. Under sparse training conditions, RMSE during high-PWV months can reach 4 mm–5 mm. As training station numbers increase, overall RMSE decreases significantly and monthly variability becomes smoother. When training station numbers reach approximately 120–200, RMSE differences among months are substantially reduced, demonstrating enhanced robustness under varying seasonal moisture regimes.

It should be noted that even with increased training station numbers, RMSE remains relatively high during summer months characterized by intense moisture variability. This suggests that monthly-scale errors are not solely determined by training sample size, but are also influenced by enhanced convective activity, rapid moisture transport, and increased observational uncertainties. Nevertheless, the convergence of Bias and RMSE with increasing training station numbers provides a solid basis for further investigations into seasonal dependence and extreme moisture processes.

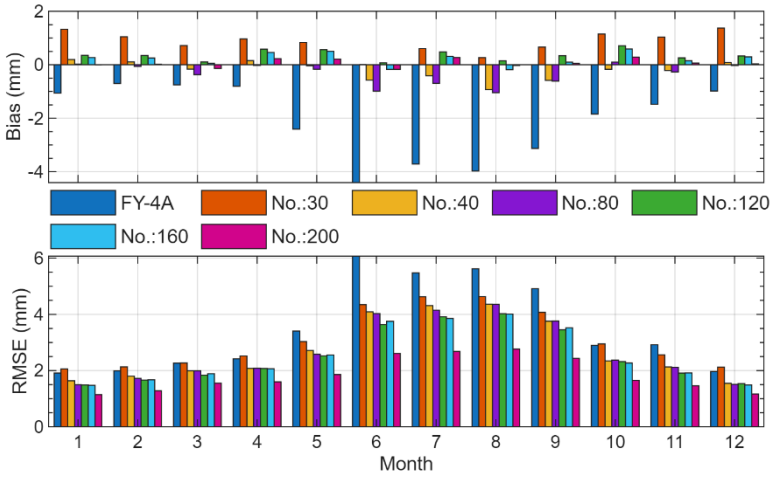

Figure 9. Comparison of the Accuracy of FY-4A PWV and Their Calibrated Versions Using Different Models Based on GNSS Validation.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

5. Conclusions

This study developed a GNSS-constrained calibration framework for FY-4A PWV and systematically investigated how training station number and spatial configuration influence all-weather PWV model's accuracy, spatial generalization, and temporal stability at the regional scale over China. By first reconstructing all-weather FY-4A PWV fields and subsequently applying a random forest–based calibration using GNSS PWV as an external constraint, the proposed approach effectively reduced systematic bias and random errors in satellite-derived PWV. The results demonstrate that, while high in-sample fitting accuracy can be achieved with relatively limited training data, the spatial generalization capability of the calibration model is strongly controlled by the density and spatial representativeness of GNSS training stations.

Comprehensive analyses across spatial, temporal, and seasonal dimensions reveal a clear performance saturation behavior. Increasing the number of training stations leads to substantial improvements in independent validation accuracy and spatial error homogeneity, particularly when training sample sizes increase from sparse to moderate levels. However, once the dominant spatial variability of PWV is sufficiently represented, further increases in training station numbers yield diminishing returns. For regional-scale applications over China, a training station number of approximately 120–200 provides an effective balance between accuracy, robustness, and computational efficiency, assuming a quasi-uniform spatial distribution. Moreover, experiments isolating spatial configuration effects demonstrate that station layout and coverage uniformity are at least as important as sample size, and that locally clustered station distributions may even degrade model performance in regions with strong moisture heterogeneity.

Despite the overall improvements achieved, residual uncertainties remain, particularly during summer months characterized by intense water vapor variability and rapid moisture transport. These errors indicate that calibration performance is not solely governed by training sample characteristics, but is also influenced by atmospheric dynamics, surface–atmosphere interactions, and observation or matching uncertainties. Future work will focus on incorporating additional dynamic predictors related to moisture transport and convection, refining season-dependent calibration strategies, and extending the proposed framework to other satellite platforms and regions with sparse or unevenly distributed GNSS networks.

Beyond the specific case of FY-4A PWV over China, this study elucidates a generalizable principle for calibrating satellite geophysical products under sparse ground constraints: when observational resources are limited, prioritizing spatially representative and uniformly distributed stations is more effective than pursuing a higher station count alone. This 'layout-over-density' principle stems from a fundamental requirement for spatial generalization—the model must capture the spatial heterogeneity of the target variable. A uniformly distributed network maximizes the spatial representativeness of the training data with a minimal number of stations.

Therefore, this work not only delivers an optimal station configuration strategy for operational FY-4A PWV calibration but also provides a quantifiable and transferable design framework for enhancing a wide range of satellite-derived products in regions constrained by sparse ground truth observations.

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

Financial support

This study was financially supported by the Science and Technology Program of Guangdong Province (Grant No. 2025B1212050001), in part by Shenzhen Science and Technology Program (Grant No. JCYJ20240813105116022), in part by Rocket Force University of Engineering Foundation for Young Scientists (Grant No. 2024QN-B037).

Author contributions

CRediT: Yongchao Ma: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft; Zhengsheng Chen: Data curation, Project administration; Tong Liu: Methodology, Validation, Writing - review & editing; Zhibin Yu: Resources, Supervision; ZhiHao Wang: Methodology, Resources, Validation, Writing - review & editing.

Competing interests

No potential conflict of interest was reported by the author(s).

Data availability statement

The generated FY-4A PWV calibration model are available from Zenodo, at DOI: https://doi.org/10.5281/zenodo.18751647 (Ma 2025). The FY-4A water vapor data are openly and freely available at http://satellite.nsmc.org.cn. The radiosonde data can be accessed through https://www.ncei.noaa.gov/products/weather-balloon/integrated-global-radiosonde-archive. The NDVI data can be download at https://ladsweb.modaps.eosdis.nasa.gov/. The SRTM DEM is available at http://www.resdc.cn/. Any additional information or data used in this study are available from the corresponding author upon reasonable request. The raw GNSS data are proprietary and cannot be shared publicly due to privacy and confidentiality restrictions imposed by the data provider. Access to the restricted data can be requested by contacting Zhihao Wang at zhihaowang1997@outlook.com. Requests are subject to approval by the data owner and may require a data-use agreement; the data will be provided for non-commercial academic research only.

References

Bai J, Lou Y, Zhang W, Zhou Y, Zhang Z, Shi C. 2021. Assessment and calibration of MODIS precipitable water vapor products based on GPS network over China. Atmospheric Research. 254:105504. https://doi.org/10.1016/j.atmosres.2021.105504

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

Chen B, Liu Z. 2016. Assessing the performance of troposphere tomographic modeling using multi-source water vapor data during Hong Kong's rainy season from May to October 2013. Atmos Meas Tech. 9(10):5249–5263. https://doi.org/10.5194/amt-9-5249-2016

Du Z, Zhang B, Yao Y, Zhao Q, Zhang L. 2025. Integrating near-infrared, thermal infrared, and microwave satellite observations to retrieve high-resolution precipitable water vapor. Remote Sensing of Environment. 318:114611. https://doi.org/10.1016/j.rse.2025.114611

Durre I, Vose RS, Wuertz DB. 2006. Overview of the Integrated Global Radiosonde Archive. Journal of Climate. 19(1):53–68. https://doi.org/10.1175/JCLI3594.1

He J, Liu Z. 2020. Refining MODIS NIR atmospheric water vapor retrieval algorithm using GPS-derived water vapor data. IEEE Transactions on Geoscience and Remote Sensing. 59(5):3682–3694. https://doi.org/10.1109/TGRS.2020.3016655

Jiang N, Wu Y, Li S, Xu Y, Wang Y, Xu T. 2024. First PWV Retrieval Using MERSI-LL Onboard FY-3E and Cross Validation With Co-Platform Occultation and Ground GNSS. Geophysical Research Letters. 51(8):e2024GL108681. https://doi.org/10.1029/2024GL108681

Kaufman YJ, Gao B-C. 1992. Remote sensing of water vapor in the near IR from EOS/MODIS. IEEE Trans Geosci Remote Sensing. 30(5):871–884. https://doi.org/10.1109/36.175321

Li Q, You X, Yang S, Du R, Qiao X, Zou R, Wang Q. 2012. A precise velocity field of tectonic deformation in China as inferred from intensive GPS observations. Sci China Earth Sci. 55(5):695–698. https://doi.org/10.1007/s11430-012-4412-5

Lu C, Li X, Li Z, Heinkelmann R, Nilsson T, Dick G, Ge M, Schuh H. 2016. GNSS tropospheric gradients with high temporal resolution and their effect on precise positioning. JGR Atmospheres. 121(2):912–930. https://doi.org/10.1002/2015JD024255

Ma X, Yao Y, Zhang B, Du Z. 2022. FY-3A/MERSI precipitable water vapor reconstruction and calibration using multi-source observation data based on a generalized regression neural network. Atmospheric Research. 265:105893. https://doi.org/10.1016/j.atmosres.2021.105893

Ma X, Yao Y, Zhang B, He C. 2022. 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

Ma X, Yao Y, Zhang B, Qin Y, Zhang Q, Zhu H. 2022. An Improved MODIS NIR PWV Retrieval Algorithm Based on an Artificial Neural Network Considering the Land-Cover Types. IEEE Transactions on Geoscience and Remote Sensing. 60:1–12. https://doi.org/10.1109/TGRS.2022.3170078

Ma Y. 2025. Calibration model for FY-4A PWV based on different GNSS station network. [Data set] Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.18751647

Ma Y, Liu T, Yu Z, Jiang C, Xu G, Lu Z. 2023. All-weather precipitable water vapor map reconstruction using data fusion and machine learning-based spatial downscaling. Atmospheric Research. 296:107068. https://doi.org/10.1016/j.atmosres.2023.107068

https://doi.org/10.5194/egusphere-2026-1149
Preprint. Discussion started: 8 May 2026

Merrikhpour MH, Rahimzadegan M. 2017. Improving the Algorithm of Extracting Regional Total Precipitable Water Vapor Over Land From MODIS Images. IEEE Trans Geosci Remote Sensing. 55(10):5889–5898. https://doi.org/10.1109/TGRS.2017.2716414

Namaoui H, Kahlouche S, Belbachir AH, Van Malderen R, Brenot H, Pottiaux E. 2017. GPS water vapor and its comparison with radiosonde and ERA-Interim data in Algeria. Adv Atmos Sci. 34(5):623–634. https://doi.org/10.1007/s00376-016-6111-1

Rocken C, Anthes R, Exner M, Hunt D, Sokolovskiy S, Ware R, Gorbunov M, Schreiner W, Feng D, Herman B, et al. 1997. Analysis and validation of GPS/MET data in the neutral atmosphere. J Geophys Res. 102(D25):29849–29866. https://doi.org/10.1029/97JD02400

Sun Q, Ji D, Letu H, Ni X, Zhang H, Wang Y, Li B, Shi J. 2024. A method for estimating high spatial resolution total precipitable water in all-weather condition by fusing satellite near-infrared and microwave observations. Remote Sensing of Environment. 302:113952. https://doi.org/10.1016/j.rse.2023.113952

Trenberth KE, Fasullo J, Smith L. 2005. Trends and variability in column-integrated atmospheric water vapor. Climate Dynamics. 24(7–8):741–758. https://doi.org/10.1007/s00382-005-0017-4

Wang Z, Chai H, Zhu C, Ma H, Zheng N, Chen P. 2026. Reconstruction of high-resolution precipitable water vapor of FY-4A based on GNSS and remote sensing data. Meas Sci Technol. 37(2):025803. https://doi.org/10.1088/1361-6501/ae2b8d

Ware RH, Fulker DW, Stein SA, Anderson DN, Avery SK, Clark RD, Droegemeier KK, Kuettner JP, Minster JB, Sorooshian S. 2000. SuomiNet: A Real–Time National GPS Network for Atmospheric Research and Education. Bull Amer Meteor Soc. 81(4):677–694. https://doi.org/10.1175/1520-0477(2000)081%3C0677:SARNGN%3E2.3.CO;2

Xu J, Liu Z. 2023a. 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 Trans Geosci Remote Sensing. 61:1–14. https://doi.org/10.1109/TGRS.2023.3300880

Xu J, Liu Z. 2023b. Improving the Accuracy of MODIS Near-Infrared Water Vapor Product Under all Weather Conditions Based on Machine Learning Considering Multiple Dependence Parameters. IEEE Trans Geosci Remote Sensing. 61:1–15. https://doi.org/10.1109/TGRS.2023.3252024

Zhang B, Yao Y, Xin L, Xu X. 2019. Precipitable water vapor fusion: an approach based on spherical cap harmonic analysis and Helmert variance component estimation. J Geod. 93(12):2605–2620. https://doi.org/10.1007/s00190-019-01322-1

Zhao Q, Ma Z, Yin J, Yao Y, Yao W, Du Z, Wang W. 2024. General method of precipitable water vapor retrieval from remote sensing satellite near-infrared data. Remote Sensing of Environment. 308:114180. https://doi.org/10.1016/j.rse.2024.114180