Response to Comments:

The geometric definitions of the three layouts in the spatial arrangement experiment (does “surrounding” refer to surrounding the validation station or the study area?) are unclear. Whether the number of stations is strictly identical across the three layouts is not specified. The reconstruction stage directly cites Wang et al. (2026) without providing necessary accuracy information, affecting the interpretability of the calibration stage results.

Reply: Thank you for the valuable comments from the reviewer. In response to the point raised regarding the unclear definition of the "spatial layout experiments (clustered/surrounding/random)" and the corresponding validation sets, we have carefully reviewed the issue and made revisions and improvements to the original description. The specific clarifications are as follows:

In this study, we aimed to investigate the impact of different spatial distribution patterns of a 200-station network on PWV calibration accuracy. Our baseline network consists of 244 stations, among which 16 stations covering diverse geographical locations and climatic environments are reserved as the validation dataset and are not involved in the modeling process. To construct three representative spatial layouts, the following methods were employed:

Random distribution: A network composed of 200 stations is selected completely at random from all stations. This method simulates a scenario where station locations have no specific spatial pattern.

Surrounding distribution: To simulate the pattern of stations distributed around a central area, the geographic center of all stations is first identified. Then, 28 stations are randomly removed from the central area. The remaining 200 stations, primarily located in peripheral regions, constitute the training set.

Clustered distribution: To simulate the pattern of stations densely concentrated in a certain area, the opposite strategy is adopted: 28 stations are randomly removed from the peripheral areas of the station network, retaining 200 stations in the central and relatively concentrated regions.

In the revised manuscript, we have clearly defined the specific construction methods for these three spatial layouts. It is noted that the "random distribution" serves as the baseline scenario, while the "surrounding distribution" and "clustered distribution" are used to examine the model’s performance under two extreme cases of non-uniform sampling: missing central data and missing edge data, respectively. We believe this additional explanation clearly delineates the different experimental setups and clarifies the rationale behind the composition of their respective validation sets.

The content of this study represents a further extension of previous work, with a focus on investigating the impact of varying numbers and spatial distributions of ground stations on water vapor calibration. Due to space limitations, the preliminary reconstruction results were not analyzed in detail; interested readers may refer to the earlier work for specifics. However, the reviewer’s suggestion was highly pertinent, and therefore validation results have been added to the revised manuscript.

For details, please see line 173 of the revised version.

1. There are errors in figure citation (e.g., line 253: “Figs. 2 and 3” should be “Figs. 3 and 4”), and some English expressions are not sufficiently accurate.
   Reply: This was an author's clerical error, and it has been corrected.
2. In line 166, the authors state: “Detailed validation of the reconstruction has been reported in previous studies (Wang et al. 2026) and is not repeated here.” However, reconstruction accuracy directly affects the input quality of the calibration stage and consequently all downstream conclusions. It is recommended to at least report the overall RMSE/Bias/R^2 of the reconstructed FY-4A PWV relative to GNSS PWV in a table.
   Reply: The content of this study represents a further extension of previous work, with a focus on investigating the impact of varying numbers and spatial distributions of ground stations on water vapor calibration. Due to space limitations, the preliminary reconstruction results were not analyzed in detail; interested readers may refer to the earlier work for specifics. However, the reviewer’s suggestion was highly pertinent, and therefore validation results have been added to the revised manuscript.
   For details, please see line 173 of the revised version.
   Line173: The model performance was evaluated using the test set, and the results show that the reconstructed model achieved an RMSE of 1.12 mm and a Bias of 0.51 mm. The reconstructed PWV has been verified to effectively retrieve the spatial distribution characteristics of water vapor over the entire region, and its distribution is consistent with that of ERA5 PWV. Detailed validation of the reconstruction has been reported in previous studies (Wang et al. 2026) and is not repeated here.
3. The current table mixes hyperparameters for different station counts (30–244) with three spatial layouts in two columns, resulting in poor readability. It is recommended to split this into two separate tables or adopt a clearer multi-column structure.
   Reply: The content of the table has been revised in accordance with the reviewer's suggestions.
   Please refer to Table 1 in the revised manuscript for the specific changes.

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

Training station number experiment		Spatial structure experiment	
Number of Station/ Distribution Type	Number of Decision Trees	Number of Station/ Distribution Type	Number of Decision Trees
30/ Random	150	200/Surrounding	200
40/ Random	150	200/Clustered	150
80/ Random	150	200/ Random	200
120/ Random	150		
160/ Random	200		
200/ Random	200		
244/ Random	200		

4. Line 192 mentions using Bayesian optimization with 5 iterations for hyperparameter search, but the search space (e.g., range of tree numbers, maximum depth, minimum samples per leaf node) is not specified. Five iterations are generally insufficient for Bayesian optimization. Please justify this choice or extend the iteration count and report hyperparameter convergence.

Reply: The text description contained an issue, and the content has been adjusted accordingly. In fact, we repeated this process 5 times.

Please refer to line 210 of the revised version.

5. Figure 4 uses bar charts to display Bias and RMSE for 16 independent validation stations and 6 models, resulting in extremely high information density. Moreover, the negative bias bars of the raw FY-4A data obscure the details of the calibrated models. Suggestions: (i) Plot the raw FY-4A results separately or use a logarithmic scale; (ii) Add a cross-reference between station numbers and geographic locations (can be annotated in Fig. 1b); (iii) Supplement the standard deviation across the 16 stations for each model to characterize spatial consistency.

Reply: The adjustments to Figure 4 have been completed, and station numbers have been added to the measurement locations in Figure 1b.

Please refer to Figure 1 and Figure 4.

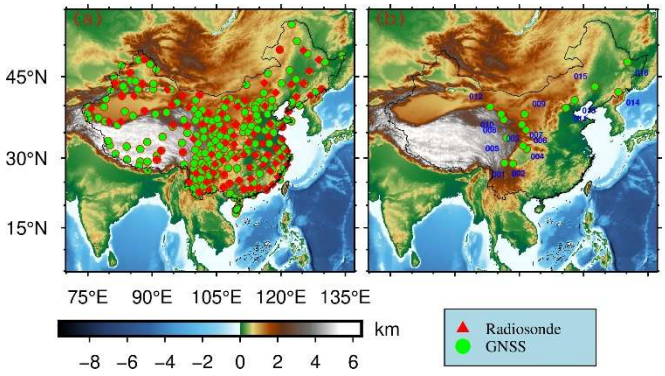

Figure 1: Spatial Distribution of GNSS and IGRA Stations in China.

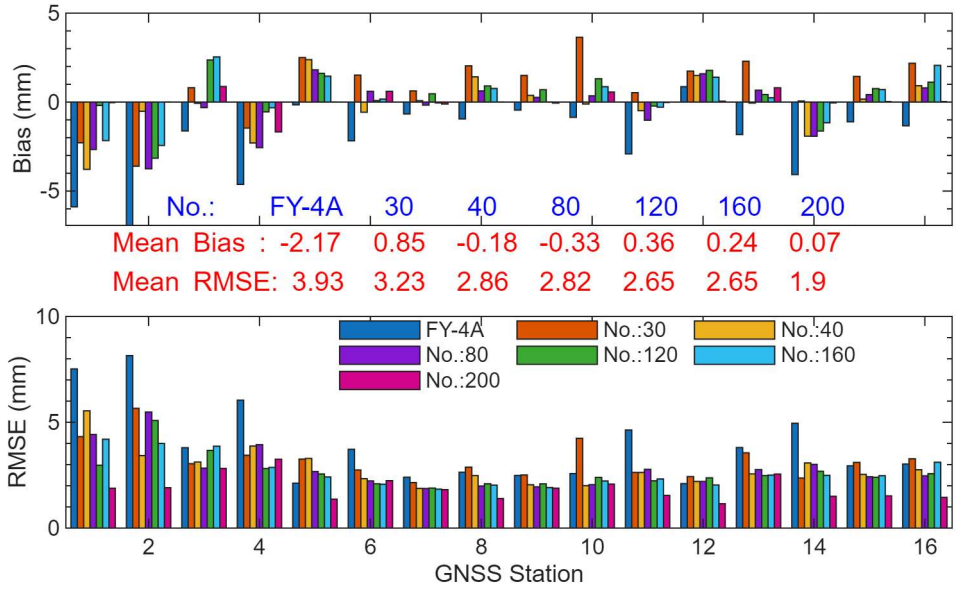

Figure 4: Validation Results of different Models Using GNSS Stations Not Involved in Model Training

6. Table 2 shows that RF_120 achieves better Bias (0.02 mm) and RMSE (2.36 mm) than RF_244 (-0.14 mm, 2.37 mm). This phenomenon contradicts the intuition that “more

training samples lead to better performance." A discussion of this phenomenon is recommended at the end of Section 4.1.

Reply: Following the reviewers' suggestions, a discussion on this phenomenon has been provided. it was found that the model trained on 120 stations achieved higher accuracy than those trained on a larger number of stations. This phenomenon can be attributed to two main reasons. First, as the number of training stations increases, the model is compelled to learn more generalizable physical patterns, which may slightly reduce its fitting accuracy to the training data (i.e., a shift from "overfitting" toward "generalization"). Second, redundant stations fail to provide additional useful information; instead, they introduce a smoothing effect that dilutes the critical spatial features present in the original data.

7. The future outlook in lines 436–438 is too general. It is recommended to add specific directions, such as: (i) transfer learning validation in truly sparse regions (e.g., Tibetan Plateau, Sahara, ocean islands); (ii) comparison with satellite–radiosonde data assimilation methods.

   Reply: Thank you for the valuable suggestions from the review experts. We have adopted and incorporated them into the revised manuscript.

   Please refer to line 459 of the revised article.

   Line 459: 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. Furthermore, further efforts will be directed toward achieving optimal calibration of satellite-derived water vapor using existing ground-based measurements in station-sparse regions, such as the Tibetan Plateau and oceanic islands.

8. In the abstract, "hindering application" should be "hindering its application"; "relies" should be "rely."

   Reply: The modifications have been made.

9. The manuscript mixes terms such as "training station number," "training sample size," and "station density." It is recommended to unify the terminology.

   Reply: We understand your recommendation regarding the clarity and consistency of terminology. Upon careful consideration, we believe that the three terms used in the original text—"training station number," "training sample size," and "station density"—serve distinct and specific purposes in the context of this study, as they describe different dimensions of the model training data. Unifying them may compromise the precision of the descriptions. The specifics are as follows:

   "Training station number": This term specifically refers to the number of meteorological observation stations with independent geographical locations used for model training. It emphasizes the quantity of spatial points, which forms the basis of

the spatial sampling framework. For instance, when comparing "200 stations" versus "244 stations" in the spatial layout experiment, this concept is precisely what is applied.
"Training sample size": This term refers to the total number of data records used for model training. It is calculated as the "training station number" multiplied by the number of observations per station over the time series. It emphasizes the total volume of the dataset, which directly impacts the statistical fitting capability of the model. For example, this term is more relevant when discussing whether the model is overfitting or underfitting.
"Station density": This term is a spatial statistic, typically defined as the "number of stations per unit area." It describes the spatial distribution density of the stations and is a key metric for evaluating the representativeness of the observation network and analyzing the spatial heterogeneity of the results. This differs fundamentally from the mere "number of stations."
In the study, we deliberately distinguish among these three terms:
When discussing the impact of spatial sampling design on the model, we use "station number" or "station density."
When discussing the effect of data volume on model training outcomes, we use "sample size."
We will follow your recommendation by providing clearer definitions for these terms when they are first introduced in the revised manuscript and ensuring consistency in their usage throughout the subsequent text to avoid confusion. We believe that this precise distinction will better facilitate the clear communication of the scientific design of this study.

10. The titles of Figures 5, 6, and 7 are too brief and lack key information. It is recommended to clearly state in each title what each subfigure (a, b, c, etc.) corresponds to, so that the figures can be understood independently.
Reply: Revisions have been made to the titles of Figures 5-7, with clarifications added regarding the content corresponding to each subfigure.

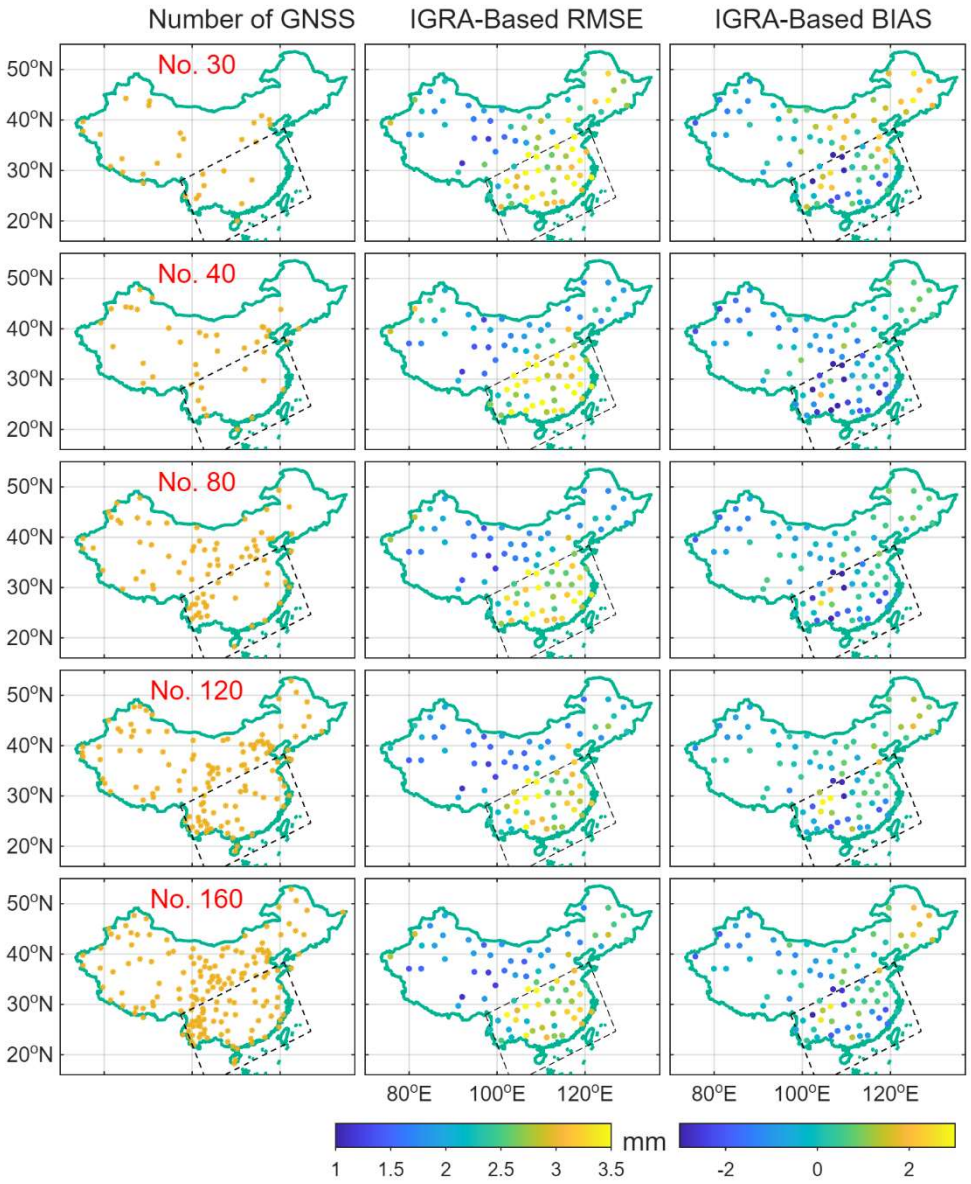

Figure 5. Spatial distribution of validation accuracy for different models using IGRA stations.

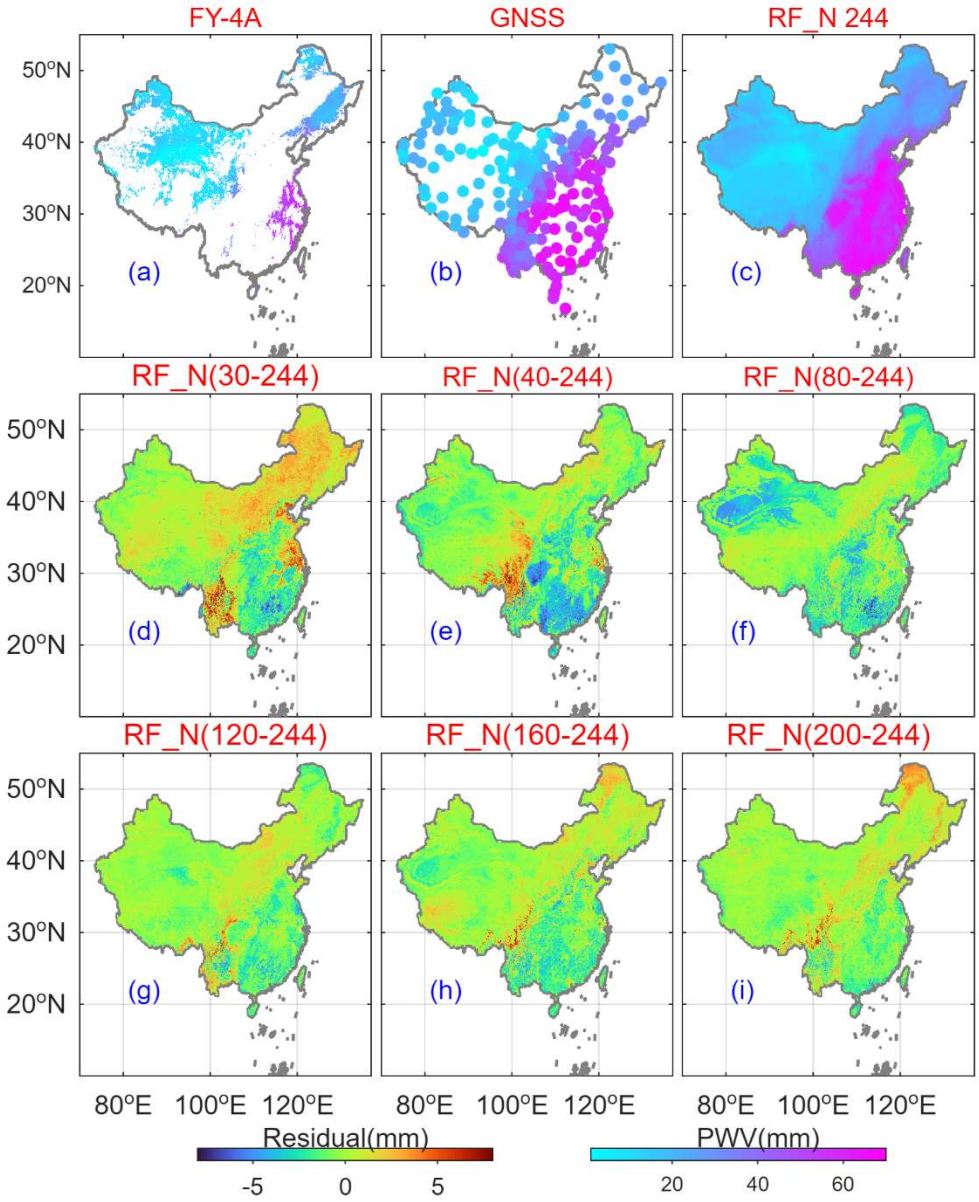

Figure 6. Comparison of retrieval results at 12:00 on the 232nd day of 2023 based on training models with different station configurations. Figure (a-c) respectively display FY-4A PWV, GNSS PWV, and RF_N244 model derived PWV. Figure (d-i) show the differences in retrieval results between each model and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

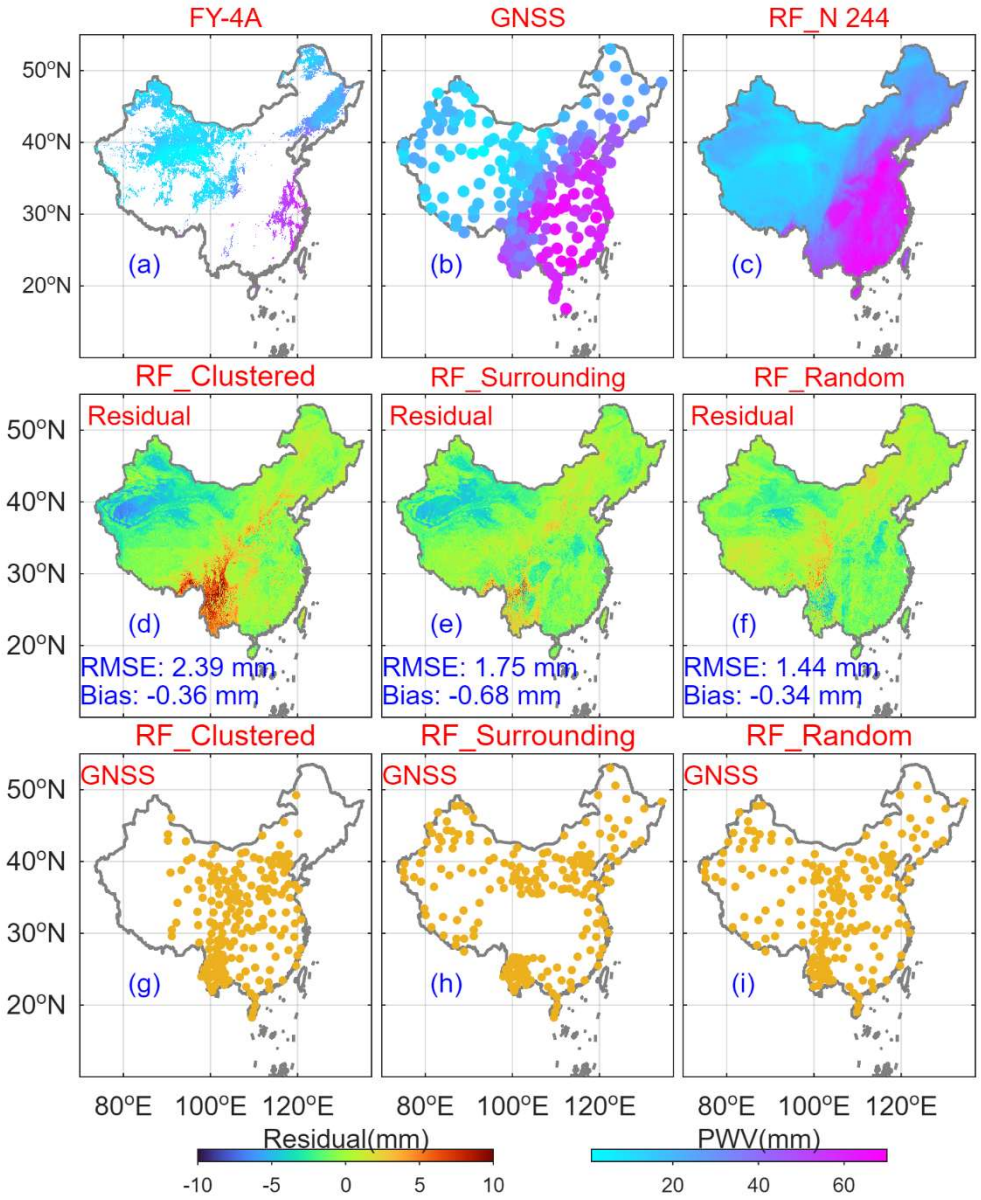

Figure 7. Comparison of retrieval results at 12:00 on the 232nd day of 2023 based on training models with different station distributions. Among them, Figure (a) displays FY-4A PWV, Figure (b) represents GNSS station-derived PWV, Figure (c) presents PWV retrieved from the model trained on 244 stations, Figures (d)-(f) show PWV retrieved from models trained with different station distributions, and Figures (g)-(i) illustrate the distribution of GNSS stations.

11. In Table 1, expressions such as “200/Surrounding” do not conform to table formatting standards. It is recommended to rename the “Number of Station” column to “Configuration” and change “200/Surrounding” to a more standard format like “200 (Surrounding).”

    Reply: The content of the table has been revised.

    Please refer to Table 1 in the revised manuscript for the specific changes.