Response to Comments：#2

Comment 1: Further highlight the core innovation of the manuscript

The authors are encouraged to further highlight the core innovation of the manuscript. The current manuscript provides a relatively complete description of the research background and experimental results, but the main contributions of the study are somewhat scattered. It is suggested that the authors add a more focused summary of the contributions at the end of the Introduction. Specifically, the authors should clearly state that this study is not only intended to construct a GNSS-constrained FY-4A PWV calibration model, but more importantly, to quantitatively evaluate, through controlled experiments, the effects of GNSS training station number and spatial layout on model accuracy, spatial generalization capability, and temporal stability. The proposed station configuration principle, namely prioritizing spatial representativeness and coverage uniformity under limited station availability, should also be emphasized.

Reply: In the concluding section of the article's introduction, the main objective of this study is emphasized: to investigate the influence of the number and spatial distribution of stations on the reconstruction of FY-4A PWV, especially in terms of calibration. In the conclusion section, it is further highlighted that, under the constraints of limited station resources, priority should be given to spatial representativeness and the uniformity of coverage.

For specific details, please refer to lines 74 and 450 in the revised version of the article.

Line 74: 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.

Line 450: 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.

Comment 2: Clarify the resolution harmonization and scale consistency of auxiliary variables

In Section 2.2, some variables were subjected to resolution harmonization and scale transformation. The authors are advised to provide further details in the data processing or methodology section regarding how variables such as DEM and NDVI were spatially matched to the FY-4A PWV grid. Adding these explanations would help improve the transparency of the data processing workflow and enhance the rationality and reproducibility of the model input design.

Reply: Thank you for the valuable suggestions from the reviewers. To achieve strict alignment with the FY-4A PWV target grid, we performed spatial resampling on all variables with inconsistent resolutions. For static or quasi-static variables, such as DEM, NDVI, and LCT, bilinear interpolation was uniformly applied to resample them onto the FY-4A PWV grid. During the resampling process, we ensured that all input variables shared identical grid counts,

spatial extents, and grid cell center coordinates. This enabled precise pixel-by-pixel spatial matching and prevented errors caused by grid offset or boundary discrepancies.

Please refer to lines 163- 167 of the revised manuscript for specific details.

Lines 163-167: To ensure spatiotemporal consistency across all input data, a rigorous pre-processing procedure was applied. Temporally, the ERA5 data were aligned to match the exact observation hours of the full-disk FY-4A PWV product. Spatially, the FY-4A PWV data and the DEM over the study domain were resampled onto the ERA5's 0.25° × 0.25° grid using a bilinear interpolation method, thereby unifying the spatial resolution and grid alignment for all datasets prior to model training.

Comment 3: Provide more details on the construction of the three spatial layouts

The authors are encouraged to provide more details on how the three station spatial layouts were constructed. Section 3.3 mentions three station distribution patterns: random, surrounding, and clustered. However, it is still unclear how these layouts were generated. For example, it should be clarified whether the random layout was generated through repeated random sampling and whether the averaged results were reported, how the spatial clustering range of the clustered layout was determined, and whether the same number of training stations was strictly maintained among the three layouts. The authors are advised to supplement these methodological details and explain how the corresponding validation stations were determined, in order to enhance the reproducibility of the experiments and the reliability of the conclusions.

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.

Comment 4: Improve the presentation of figures, equations, and language

The overall structure of the manuscript is clear, and the figures, tables, and equations generally support the methodological description and experimental analysis. However, there are still some minor issues related to figure presentation, equation formatting, and formatting consistency. The authors are advised to further check and revise these aspects to improve the readability, standardization, and overall consistency of the manuscript.

(1) Some figures contain a large amount of information, especially multi-panel result figures. It is suggested that the authors further standardize the size and layout of the figures. For example, the aspect ratios of the subplots in Figures 6 and 7 could be appropriately adjusted to improve the standardization and consistency of the manuscript.

Reply: Based on the reviewer's suggestion, the size and layout of all figures have been further standardized and organized.

(2) In some figures, the text overlaps with graphical elements. The authors are advised to further check and standardize the font size, legends, color bars, axis labels, and subplot numbering, so that comparisons among different station numbers, model results, and spatial layouts can be presented more clearly.

Reply: Based on the reviewer's suggestion, the size and layout of all figures have been further standardized and organized.

(3) The authors should carefully check whether the equation formatting and variable descriptions are complete. For example, the explanation of the variables in Equation (4) should be further supplemented, and the symbol formatting in Equations (7)–(10) should be kept consistent. This would help avoid ambiguity when readers interpret the evaluation metrics.

Reply: The formula has been verified and the symbol descriptions have been supplemented.

Response to Comments：#3

1. The definition of the spatial layout experiments (clustered/surrounding/random) and the corresponding validation sets are unclear.

   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.

2. The meaning of the subplot labels (a/a1, b/b1) in Figures 3 and 4 needs to be explained more clearly in the main text. Furthermore, the color bar ranges and unit labels in Figures 5–7 suffer from overlap and occlusion; layout adjustments are recommended.

   Reply: Figures 3-7 have been revised, including adjustments to the meaning represented by each subplot and the color bar.

   Please refer to Figures 3-7 in the revised manuscript for the specific changes.

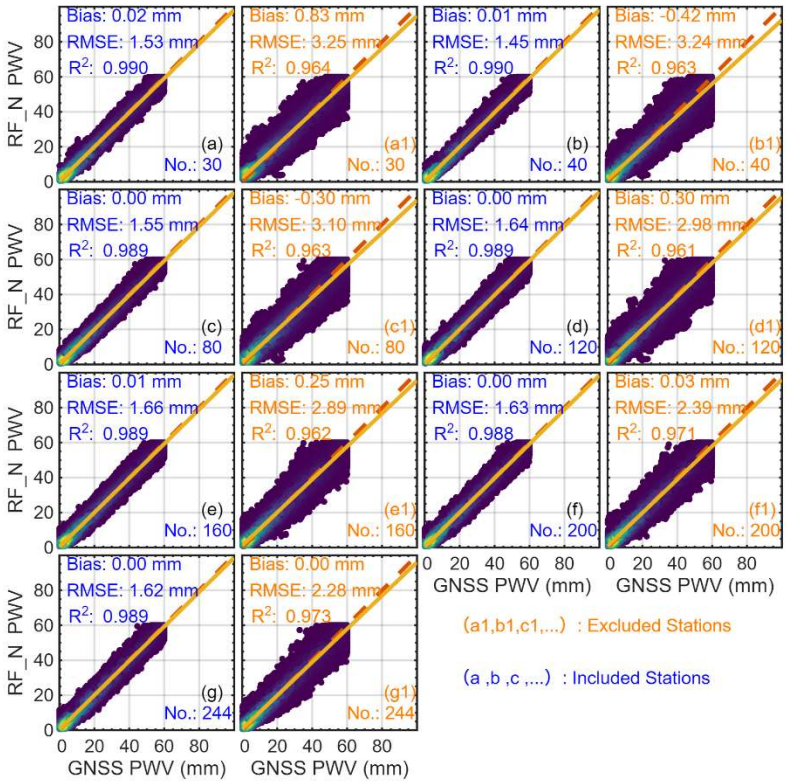

Figure 3. Validation Results of Training Models Based on Different Numbers of GNSS Stations. Among them, Figures (a, b, c, d, e, f, g) represent the model's internal consistency accuracy validated using stations that participated in model training, while Figures (a_i, b_i, c_i, d_i, e_i, f_i, g_i) represent the model's external consistency accuracy validated using stations that did not participate in model training.

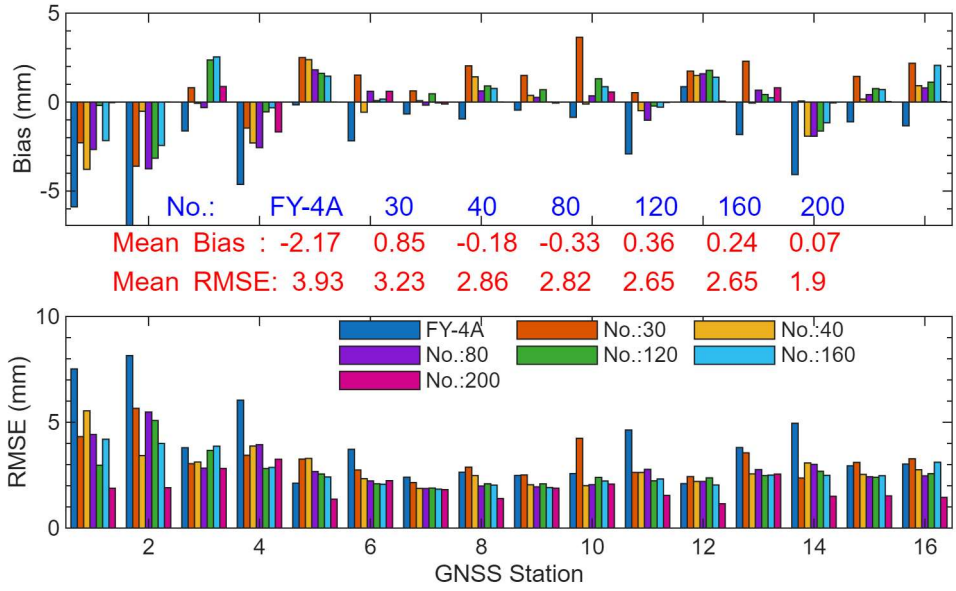

Figure 4: Validation Results of different Models Using GNSS Stations Not Involved in Model Training

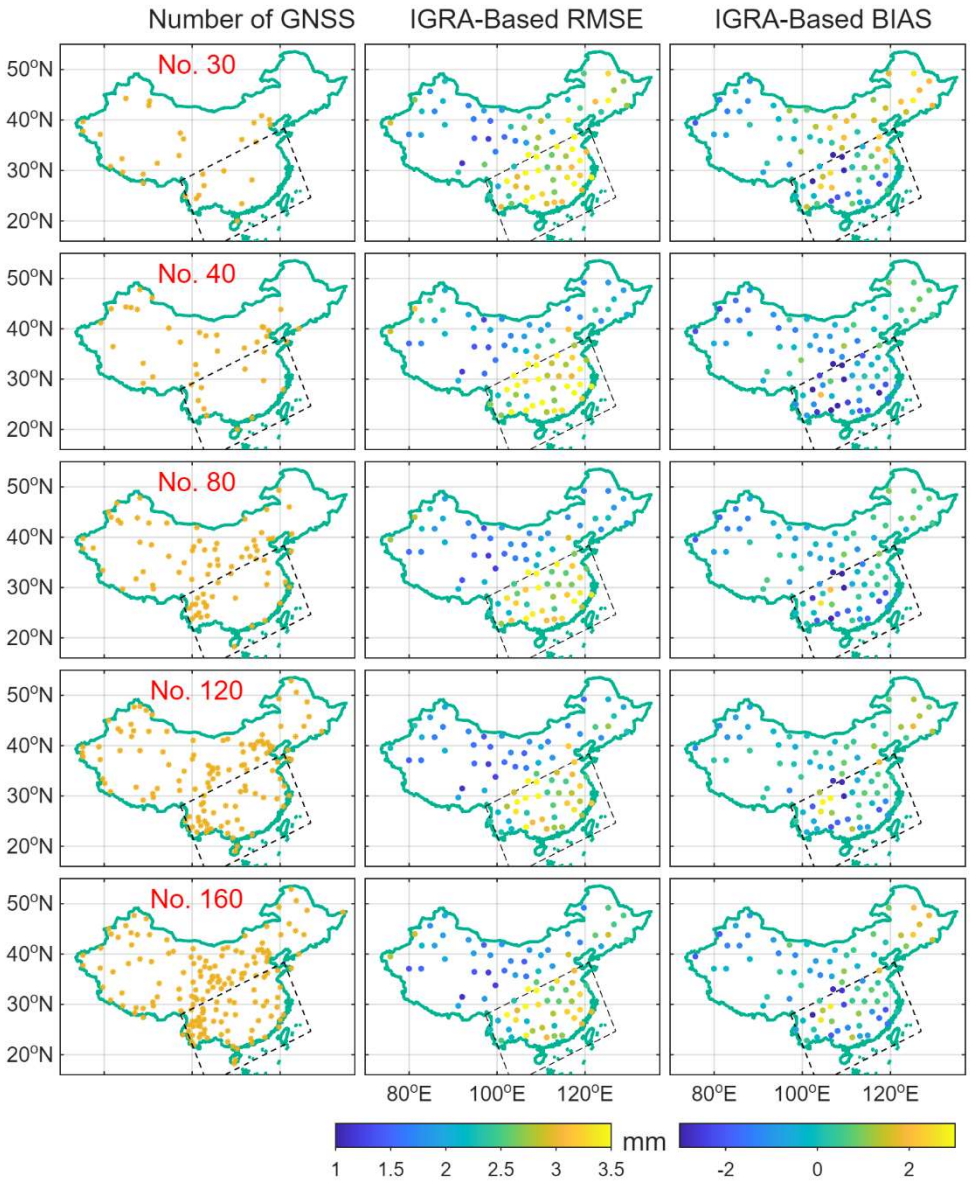

Figure 5. Spatial distribution of validation accuracy for different models using IGRA stations.

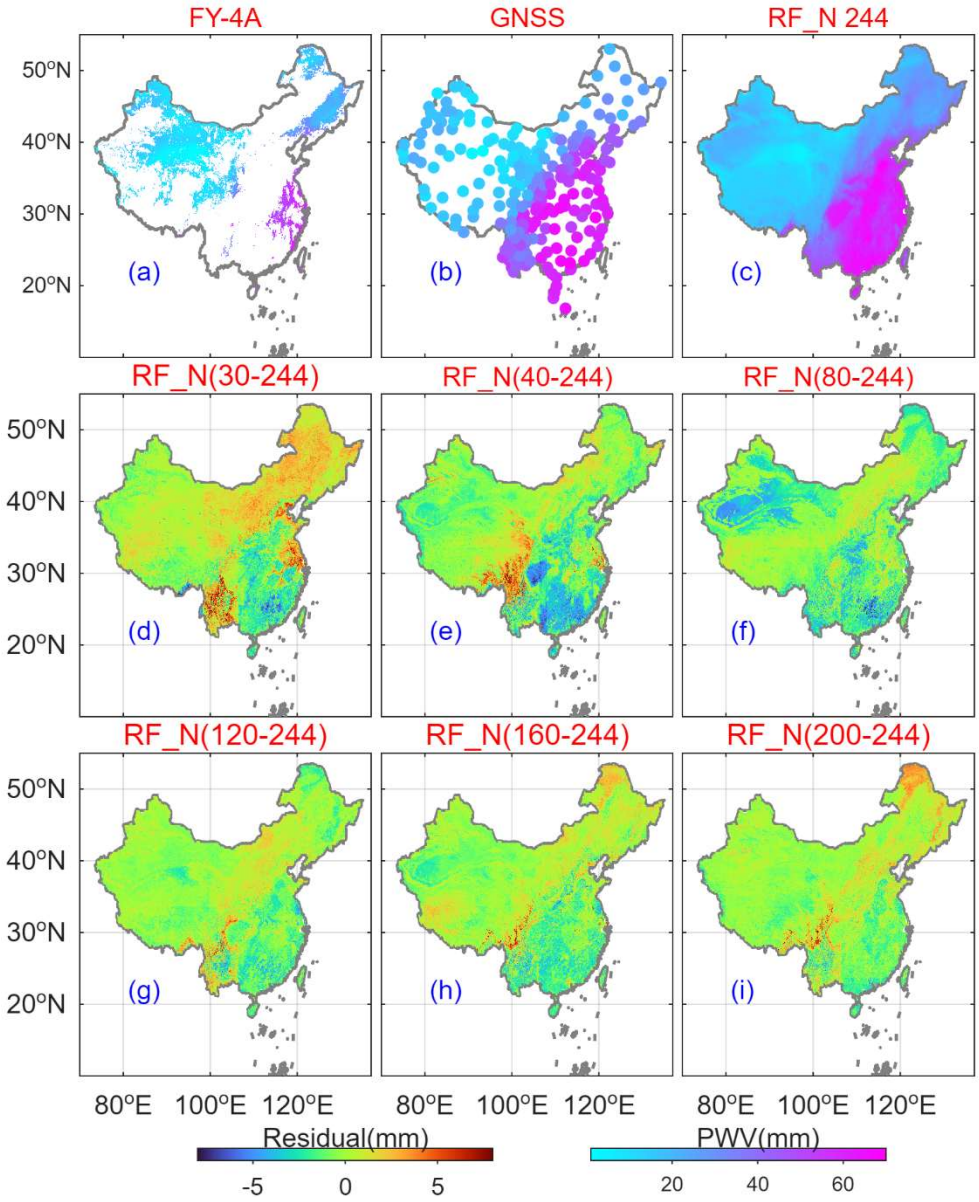

Figure 6. Comparison of retrieval results at 12:00 on the 232nd day of 2023 based on training models with different station configurations. Figure (a-c) respectively display FY-4A PWV, GNSS PWV, and RF_N244 model derived PWV. Figure (d-i) show the differences in retrieval results between each model and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

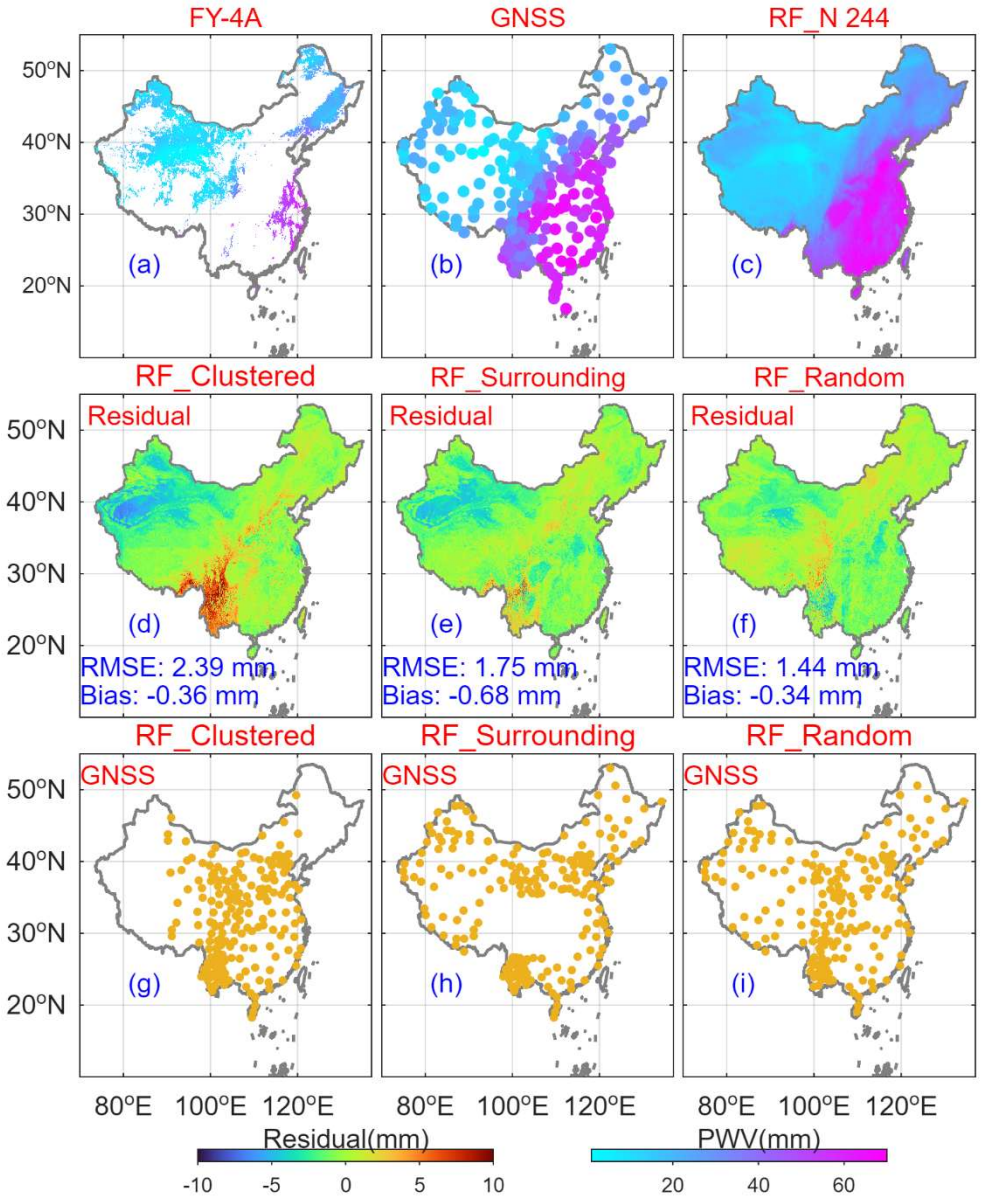

Figure 7. Comparison of retrieval results at 12:00 on the 232nd day of 2023 based on training models with different station distributions. Among them, Figure (a) displays FY-4A PWV, Figure (b) represents GNSS station-derived PWV, Figure (c) presents PWV retrieved from the model trained on 244 stations,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

3. Lines 13–16: "a spatially random layout consistently outperforms clustered or geographically biased distributions, reducing RMSE by up to 27%." It is suggested to clarify against which baseline (clustered or surrounding) this 27% reduction is measured, and to state that this was obtained under the same validation set.

   Reply: The relevant content has been revised based on the reviewer's suggestions. The modification is as follows: 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% under the same validation set.

4. Lines 7–60: The literature review mentions that "the number of GNSS stations is a key prerequisite," but lacks a review on whether the variable "spatial layout" has been discussed in existing studies. It is suggested to add 1–2 references on geostatistical sampling design (e.g., works by Wadoux et al., Brus & de Gruijter) or studies on the spatial representativeness of training samples in machine learning, to better highlight the novelty of this work.
   Reply: To the best of the authors' knowledge, existing research has predominantly focused on high spatio-temporal resolution water vapor inversion utilizing dense station networks. However, systematic investigations into the impact of station quantity and spatial distribution on the accuracy of water vapor inversion have not yet been reported. This perspective has been mentioned on line 70 of the manuscript.
5. Lines 74–78: The description of the research objective is lengthy. It is suggested to split it into two layers: the "scientific question" (how do site number and layout independently affect calibration performance) and the "engineering objective" (providing configuration strategies under sparse networks), to enhance logical clarity.
   Reply: This study preliminarily investigates the impacts of station number and spatial distribution on water vapor retrieval, particularly on the accuracy of water vapor calibration. While the question of how to provide configuration solutions under sparse station conditions is not the primary objective of the present chapter, it will be the focus of our subsequent work.
6. Lines 89–93: 244 GNSS stations are used as the training set, and 16 GNSS stations as the independent validation set—how were these 16 stations randomly selected? Was spatial stratification considered (e.g., ensuring representation from eastern, central, western regions, and the Tibetan Plateau)? Please explain the sampling strategy.
   Reply: 16 GNSS stations randomly selected within the study region cover different latitudes and elevations.

7. Lines 118–120: The reference for the Saastamoinen model should be Saastamoinen (1972), not Vedel et al. (2001). Vedel et al.'s work is an evaluation of the model, not its original proposal. Please verify and correct.
   Reply: This was an author's clerical error, and it has been corrected.
8. Lines 161–166: The input features for the RF reconstruction model (Stage I) include Lat, Lon, DEM, Time, T2m, SP, TCW, but not NDVI and land cover type (LCT); whereas the Stage II calibration model incorporates NDVI and LCT. Please explain the rationale behind this feature selection—why are surface parameters not needed in the reconstruction stage, but are required in the calibration stage?
   Reply: The first stage focuses on reconstructing FY-PWV at a coarse resolution by leveraging the intrinsic geographical and spatio-temporal coherence of atmospheric variables, prioritizing the overall completeness of the reconstructed field. The second stage introduces key land surface parameters to specifically address fixed systematic errors induced by localized surface attributes, thereby performing site-specific corrections.
9. Line 175: The calibration model formula uses the "reconstructed FY-4A PWV" as an input, but there is a lack of discussion on how the inherent uncertainty of this

reconstructed product (from Stage I) propagates to Stage II. Has the error accumulation of the two-stage RF been assessed? Please add relevant analysis or discussion.

Reply: Regarding the clarification that the first-stage reconstruction results exhibit accuracy close to that of FY-4A PWV, we understand your concern in further elucidating the contribution of this study and the necessity of the two-stage design. To address this, we provide the following additional explanation: The objective of the first stage is to reconstruct the data gaps in FY-4A PWV caused by cloud occlusion or other factors using machine learning methods, thereby obtaining a spatially complete water vapor field. Therefore, in areas with valid observations, the accuracy of the first-stage reconstruction is essentially consistent with that of the original FY-4A PWV product, which is an expected outcome of the model design and also validates the reliability of the reconstruction method in gap-free regions. However, the FY-4A PWV product itself still exhibits certain systematic biases under specific weather or surface conditions, which affect its accuracy for direct application. Precisely for this reason, we introduced the second-stage calibration model, the purpose of which is to further correct the systematic errors in the FY-4A PWV product, thereby enhancing the overall accuracy and reliability of the final water vapor product. The two stages serve distinct functions—"spatial reconstruction" and "accuracy calibration"—forming a complete processing chain together.

10. Lines 195–197 (Table 1): The table layout is confusing; entries like "200/Surrounding" are difficult to interpret. It is suggested to split it into two separate tables or redesign the column structure for better readability.

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		

11. Lines 231–235 (Figure 4): The FY-4A original product has a Bias of −2.17 mm and an RMSE of 3.93 mm. Please clarify the baseline for these metrics—specifically, which time period and what matching conditions they are based on—and state this in the main text.

Reply: Figure 4 illustrates the inversion performance of each model validated using data from 16 stations (not involved in training) for the year 2023. Additional explanations have been provided as suggested by the reviewer.
Please refer to line 241 in the revised manuscript for the modifications.

12. Lines 297–305: What is the basis for the north-south division into "black-box regions"? Is it based on geographical boundaries (e.g., the Qinling-Huaihe Line) or statistical differences in water vapor variability? Please clarify the division criteria and annotate the boundaries in the figure.
    Reply: The black border lines are drawn based on IGRA-Based RMSE values exceeding 2.5 mm, solely for the convenience of analysis.
13. Lines 329–344 + Table 4: This is the most critical concern in this review. In the three layout experiments, the validation sets are Area 1, Area 2, Area 3, and "remaining sites" respectively, resulting in three RMSE values that are not on the same comparison baseline. The conclusion that "Random is 27.08% lower than Clustered" needs to be recalculated using the same independent validation set (e.g., the 16 independent GNSS stations or the 80 IGRA stations). Please rerun this experiment and report the corrected results.
    Reply: The textual description was inaccurate and caused misunderstanding for readers. In this study, the validation of different distribution types was based on 16 GNSS stations that were not involved in the training process, all under the same reference framework. The content has been rephrased accordingly.
    The modifications are reflected in line 343 of the revised version of the article.
14. Lines 378–391 (Figure 8): Three GNSS stations are selected for time-series comparison. What are the latitude, elevation, and climate zone of these three stations? Are they representative? It is suggested to at least add a table listing the basic information of the selected sites.
    Reply: Based on the reviewer's suggestion, the latitude and longitude of the stations have been added to the figures.
15. Figure 6: The color bar label "ERRO(mm)" should be changed to "Error (mm)"; the "PWV(mm)" and "ERRO" color bars overlap and occlude each other; separate layout is suggested.
    Reply: The images have been adjusted according to the reviewers' suggestions.
16. Figure 7: The three-row, three-column structure is good, but the naming "RF_Area_1/2/3" is suggested to be changed to "RF_Clustered / RF_Surrounding / RF_Random" for better readability.
    Reply: Based on the reviewer's suggestion, I have changed the terms from "RF_Area_1/2/3" to "RF_Clustered / RF_Surrounding / RF_Random" accordingly.
17. Lines 55–56: "Ma et al. reported that when only 14 GNSS stations in the Tibetan Plateau were used…"—the citation is missing the year; please add it.
    Reply: This is a typographical error. The content has been corrected accordingly.
18. Line 230: Incorrect figure number citation ("Figs. 2 and 3 and Table 2" should be "Figs. 3 and 4"); please correct.
    Reply: This is a typographical error. The content has been corrected accordingly.

19. Throughout the text, the terms "all-weather" and "all-sky" are used interchangeably; it is recommended to unify them to a single term.
    Reply: I have replaced "all-sky" with "all-weather" throughout the entire text.

Response to Comments: #4

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.

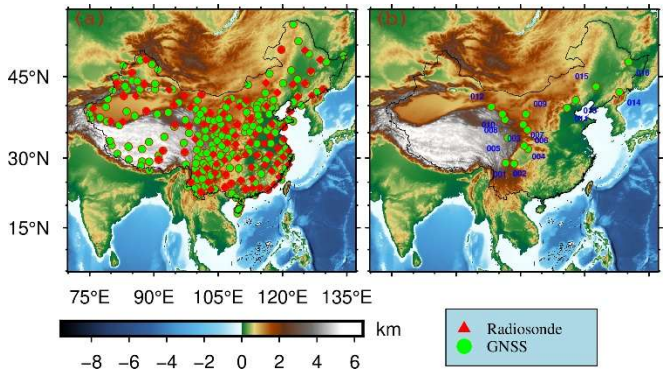

Figure 1: Spatial Distribution of GNSS and IGRA Stations in China.

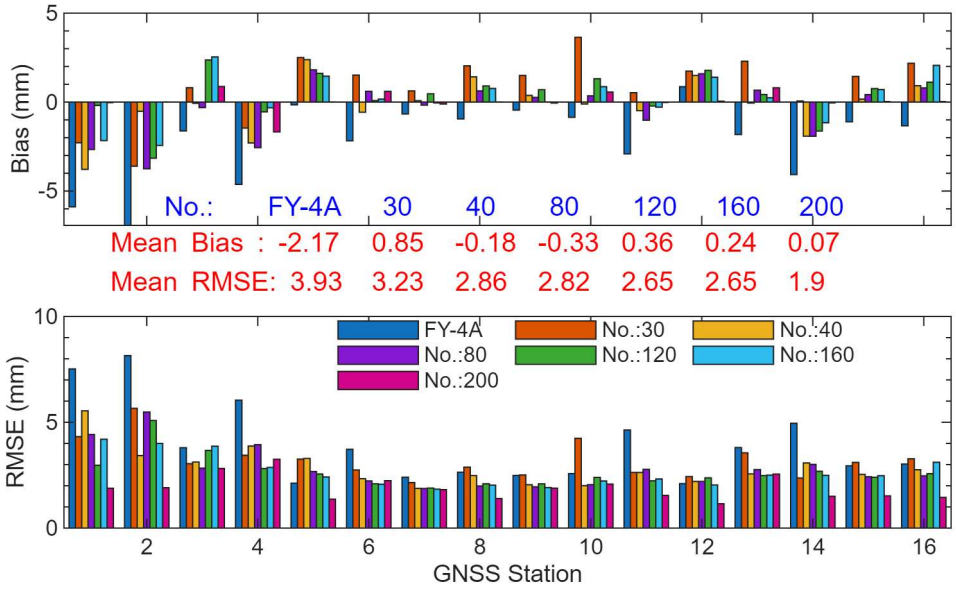

Figure 4: Validation Results of different Models Using GNSS Stations Not Involved in Model Training

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.

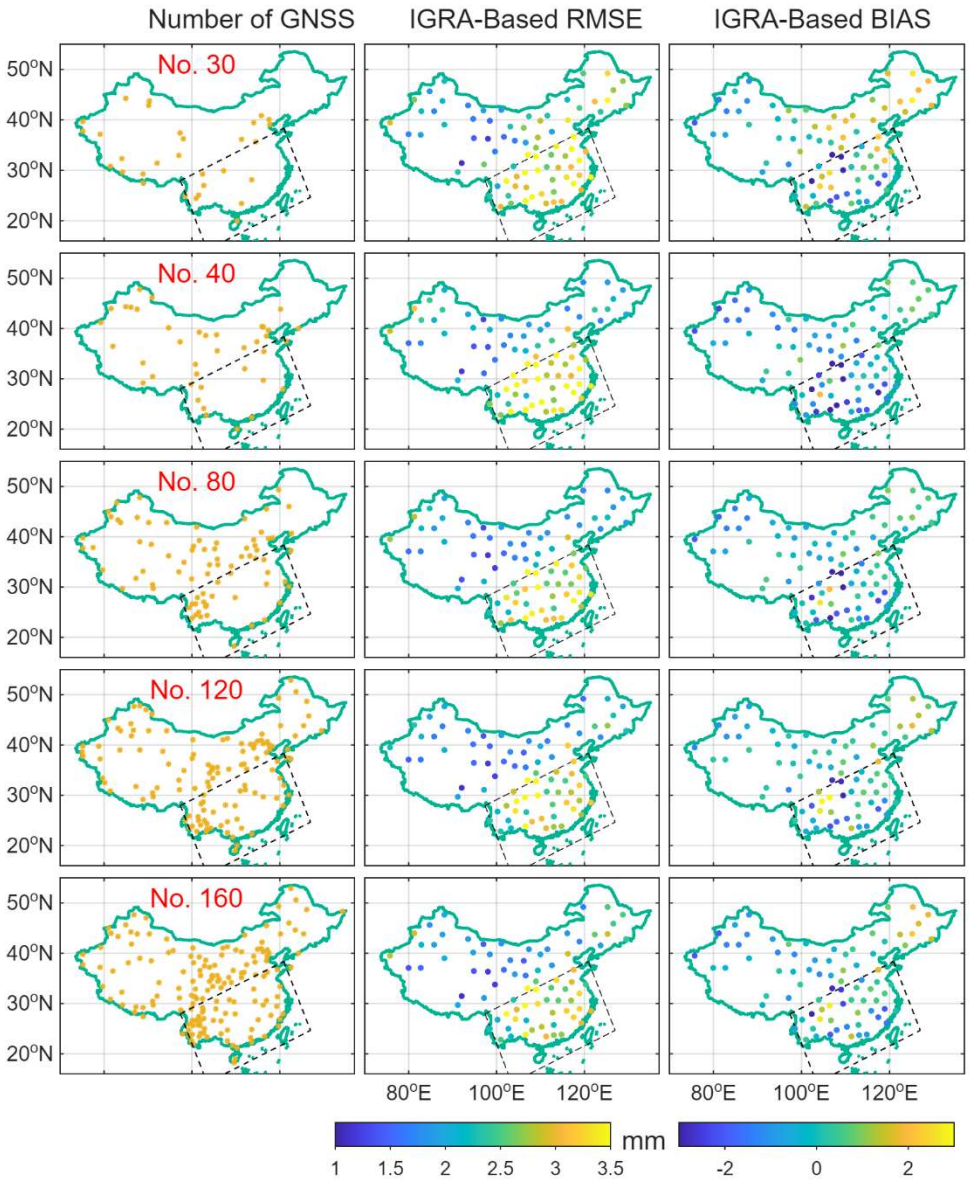

Figure 5. Spatial distribution of validation accuracy for different models using IGRA stations.

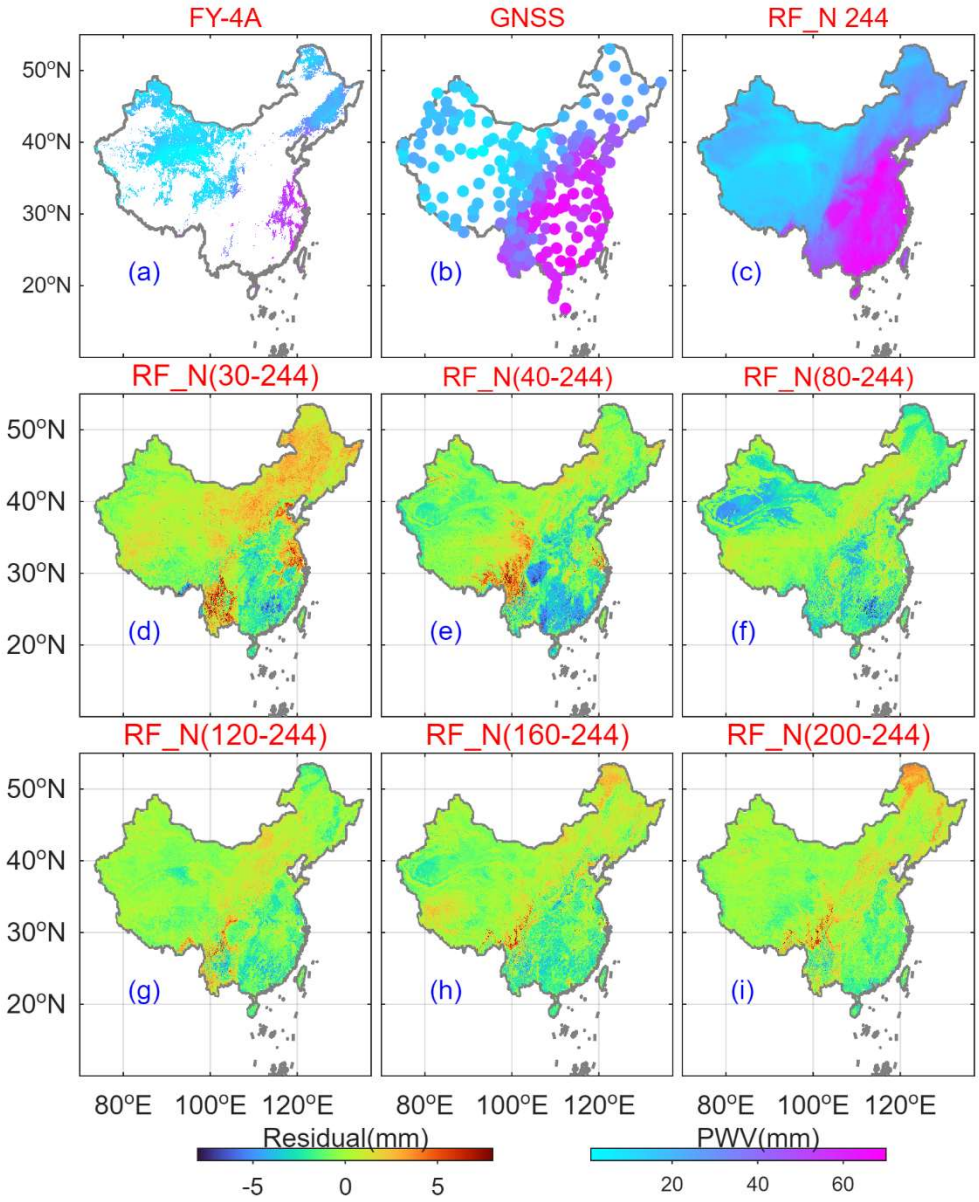

Figure 6. Comparison of retrieval results at 12:00 on the 232nd day of 2023 based on training models with different station configurations. Figure (a-c) respectively display FY-4A PWV, GNSS PWV, and RF_N244 model derived PWV. Figure (d-i) show the differences in retrieval results between each model and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

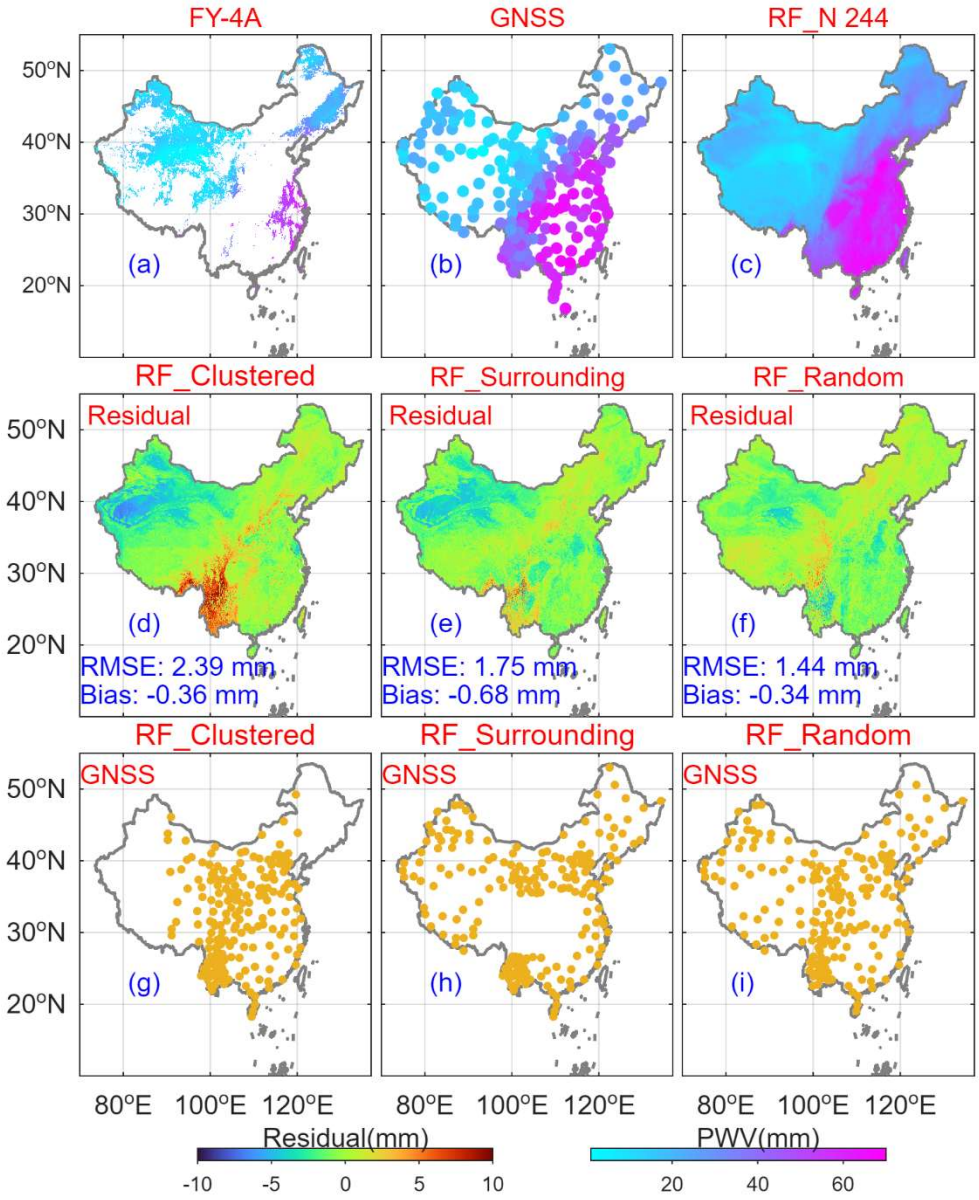

Figure 7. Comparison of retrieval results at 12:00 on the 232nd day of 2023 based on training models with different station distributions. Among them, Figure (a) displays FY-4A PWV, Figure (b) represents GNSS station-derived PWV, Figure (c) presents PWV retrieved from the model trained on 244 stations,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