Response to Comments：

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.

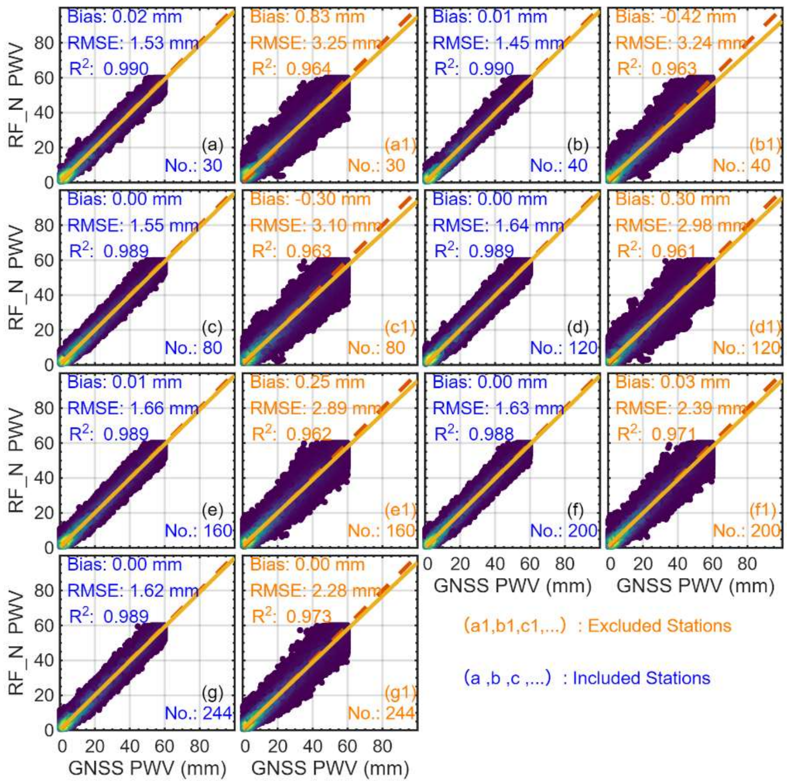

Figure 3. Validation Results of Training Models Based on Different Numbers of GNSS Stations. Among them, Figures (a, b, c, d, e, f, g) represent the model's internal consistency accuracy validated using stations that participated in model training, while Figures (a_i, b_i, c_i, d_i, e_i, f_i, g_i) represent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

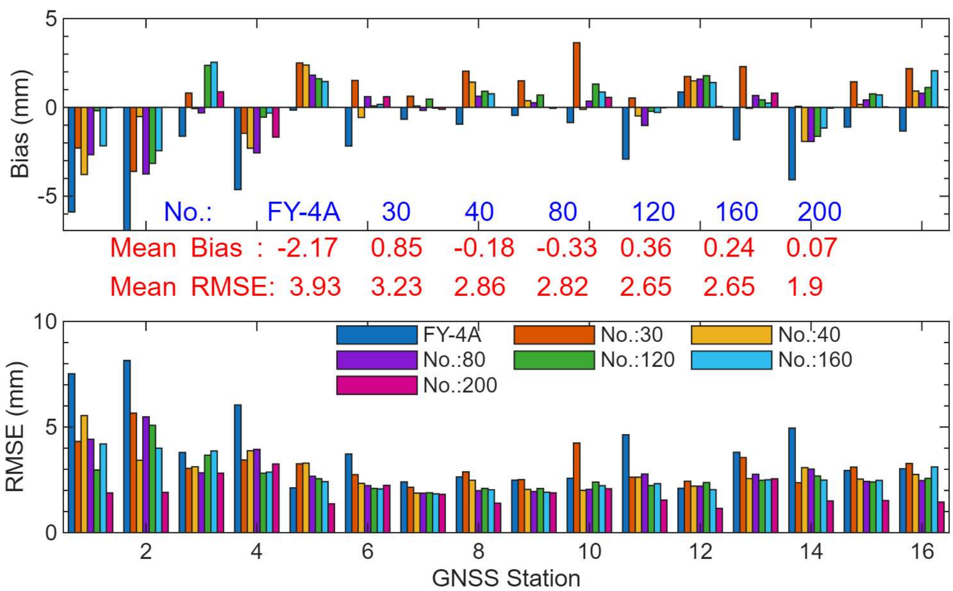

Figure 4: Validation Results of different Models Using GNSS Stations Not Involved in Model Training

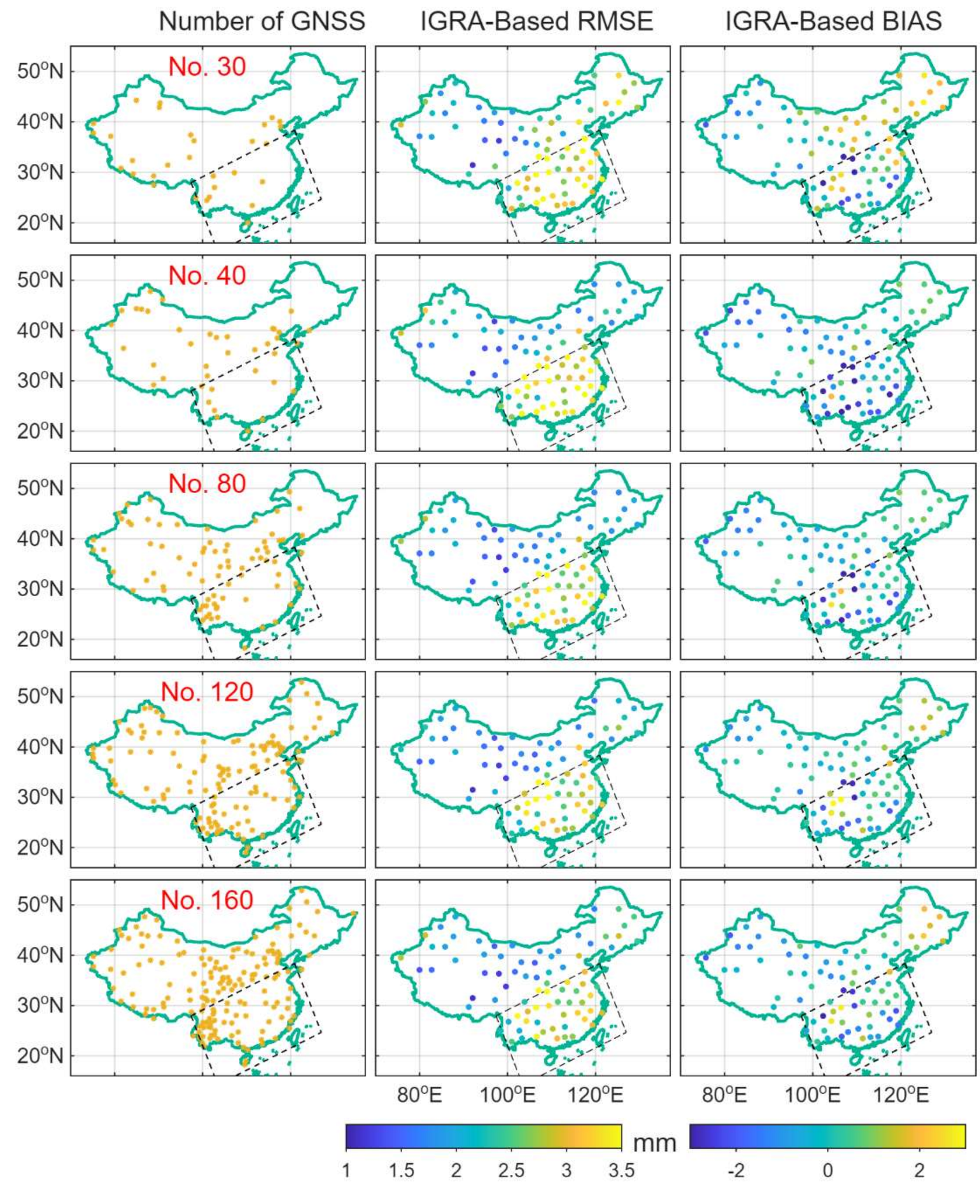

Figure 5. Spatial distribution of validation accuracy for different models using IGRA stations.

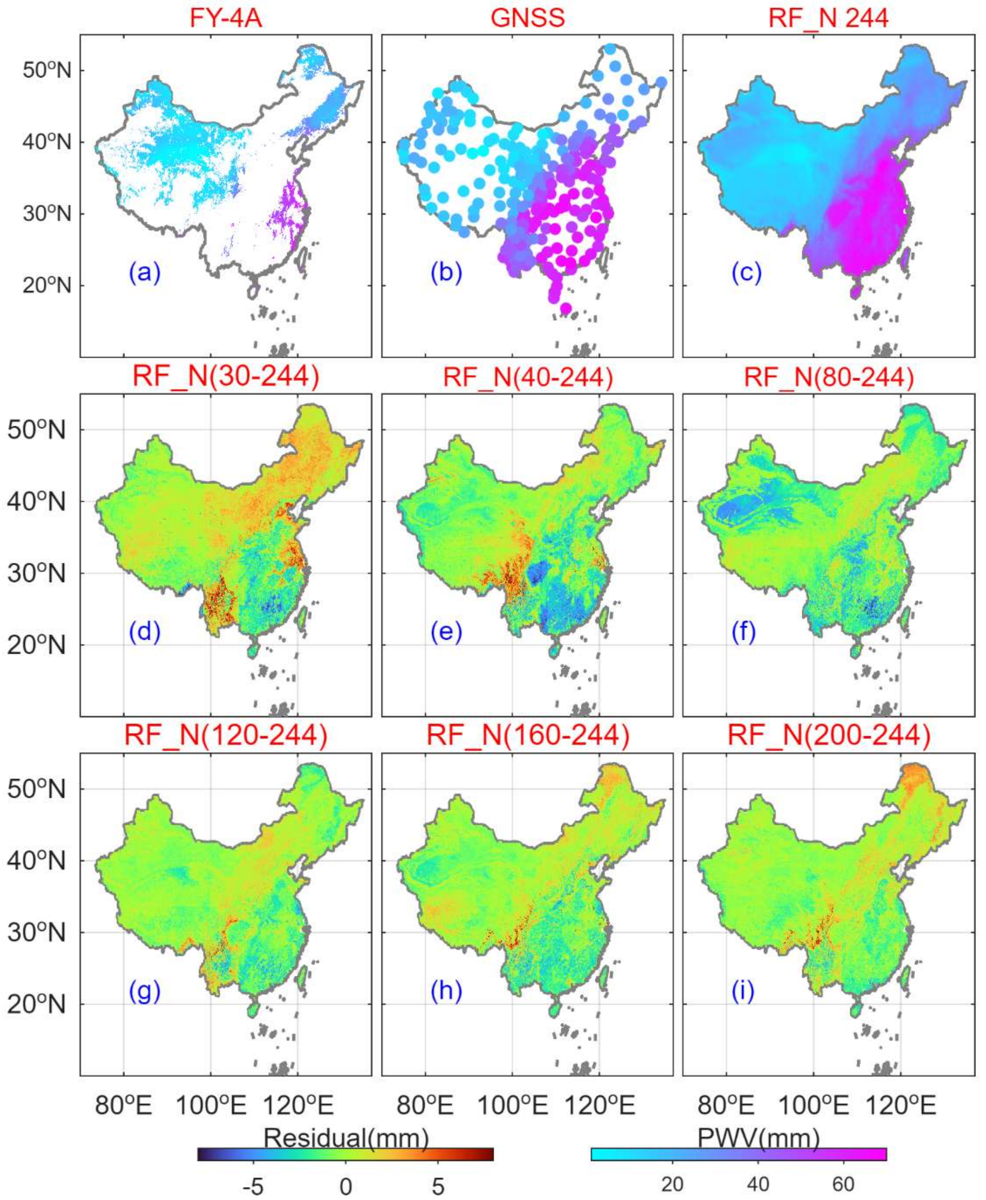

Figure 6. Comparison of retrieval results at 12:00 on the 232nd day of 2023 based on training models with different station configurations. Figure (a-c) respectively display FY-4A PWV, GNSS PWV, and RF_N244 model derived PWV. Figure (d-i) show the differences in retrieval results between each model and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

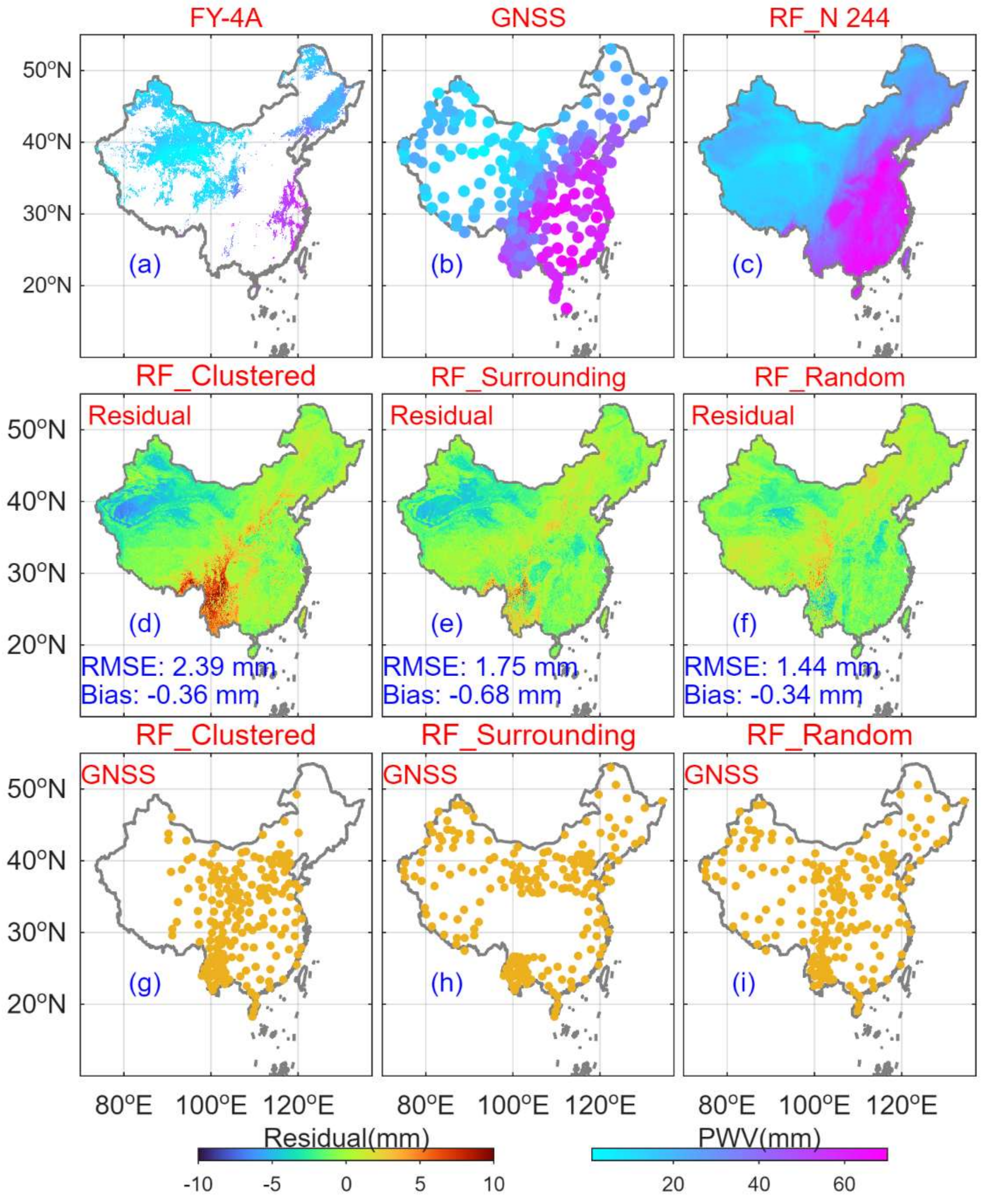

Figure 7. Comparison of retrieval results at 12:00 on the 232nd day of 2023 based on training models with different station distributions. Among them, Figure (a) displays FY-4A PWV, Figure (b) represents GNSS station-derived PWV, Figure (c) presents PWV retrieved from the model trained on 244 stations,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