

Response to Reviewer 1:

The data set presented contains measured microwave and infrared brightness temperatures, corresponding quality flags and retrieved integrated products on a minutely time resolution. In addition, retrieved temperature and relative humidity profiles are given on a 10 min time resolution. Both types cover the time period of Apr 2013 - Dec 2022.

In the presented article, the authors describe a multi-year data set of quality-controlled microwave radiometer observations, including comparisons to simulated brightness temperatures using atmospheric profiles from reanalysis. Two different retrieval approaches are presented, applied to cloudy and clear sky conditions, respectively, and compared to the retrieval approach provided by the manufacturer with respect to radiosonde data. Furthermore, a statistical analysis is carried out including a temperature trend and inversion analysis connected to air pollution.

The study presents a valuable data set with a high temporal resolution and the performed statistical analysis provides insights in long-term developments of the thermodynamic structure of the lower troposphere as observed from a ground-based remote sensing instrument. However, the argumentation for choosing different retrieval approaches for clear sky and cloudy conditions remains unclear. In addition, more information on the provided retrieval by the manufacturer is needed to ensure that a fair comparison is made in terms of underlying training data and radiative transfer methods. Instead the newly developed approaches could be compared for cloudy and clear sky cases. Also, the retrieval seems to be trained on observations rather than simulated data. Therefore the article / data set could be accepted after revision. All major and minor points, corresponding to the data set and manuscript, supporting this decision, are listed in detail below.

[We thank the reviewers for their constructive comments and suggestions. Detailed, point-by-point responses to all comments are provided. All corresponding revisions have been incorporated into the revised manuscript.](#)

Manuscript Major points:

1. How do the DNN and OE retrievals compare to radiosondes in all sky or clear sky / cloudy conditions? This comparison would add more insights rather than comparing it to the retrieval from the manufacturer (where additional information is missing). Also, the argumentation of using different approaches is unclear.

Reply: Thank you for this insightful comment. We agree that a direct comparison of the OE and DNN retrievals against radiosonde observations under different sky conditions is essential to better demonstrate the performance and applicability of the proposed methods.

In the revised manuscript, we have clarified this aspect and emphasized that both retrieval schemes are independently validated against radiosonde measurements, rather than relying solely on comparisons with the manufacturer-provided LV2 product.

Specifically:

(1) Clear-sky conditions (OE retrieval): The OE method shows strong agreement with radiosonde profiles throughout the 0–10 km layer. The mean bias remains within approximately ± 0.8 K for temperature and $\pm 6\%$ for relative humidity, with RMSE values generally below 3.6 K and 20%, respectively. These results demonstrate that the OE approach effectively constrains the retrieval using physical radiative transfer and prior information, leading to robust performance under clear-sky conditions.

(2) Cloudy-sky conditions (DNN retrieval): Under cloudy conditions, where the assumptions of the OE method are often violated (e.g., inaccurate cloud screening and forward model limitations), the DNN-based retrieval provides significantly improved agreement with radiosonde observations. The mean bias is close to zero for temperature across most altitudes, and the RMSE is substantially reduced, particularly in the boundary layer and upper troposphere. Similar improvements are observed for humidity profiles.

Importantly, this choice is not merely empirical but is motivated by the fundamental limitations of each method. The OE approach relies on an accurate forward radiative transfer model and well-defined prior information, which are generally valid under clear-sky conditions but become unreliable in the presence of clouds due to

uncertainties in cloud properties and radiative effects. In contrast, while a DNN model can flexibly learn complex nonlinear relationships from data, it does not explicitly enforce physical consistency and may be less constrained under clear-sky conditions where physical models are reliable.

Therefore, using a single retrieval approach for all atmospheric conditions may lead to suboptimal performance. The adoption of separate OE and DNN methods is a deliberate strategy to ensure that each method is applied within its domain of validity. To clarify this rationale, we have revised the manuscript (at the end of Section 3) to explicitly state that:

“From an application perspective, the OE and DNN approaches are employed as complementary rather than competing methods. The physically based OE method performs optimally under clear-sky conditions where forward radiative transfer modelling is highly reliable. Conversely, the DNN model demonstrates enhanced robustness under cloudy conditions by implicitly capturing the complex, nonlinear relationships between TBs and atmospheric states. By applying each method under its respective optimal conditions, this condition-dependent strategy yields retrieval results that are consistently closer to independent radiosonde observations than the generic operational RPG-LV2 product across all sky conditions.”

2. The analysis and data set should be separate for the different approaches to ensure a meaningful interpretation, and it should be made clear which retrieval approach and conditions (cloudy / clear sky) were used for the statistical analysis in Chapter 4 and in the final data set.

Reply: We fully agree with the reviewer that a clear distinction between the two retrieval approaches (OE for clear-sky and DNN for cloudy-sky) is essential for a meaningful physical interpretation. In the revised manuscript, we have addressed this as follows:

- (1) We have added a preamble to Chapter 4 (and specific notes in each subsection) to explicitly state which retrieval method and sky conditions were used for each analysis.
- (2) In Sections 4.2 and 4.3, for the seasonal diurnal variations (Fig. 9) and

atmospheric inversion studies (Figs. 10–11), we utilized the merged dataset (OE for clear-sky and DNN for cloudy-sky) to ensure continuous temporal coverage, which is critical for boundary layer statistics.

(3) For the final dataset structure, we have also clarified in the “Data Availability:” section that the final dataset provides both individual flags and the merged high-quality profile product, allowing users to separate them if needed.

3.1. 283: the retrieval training should be conducted with multi-year simulated TB (from ERA5), but observed TB from 2019 are mentioned. This needs to be checked and clarified.

Reply: We thank the reviewer for raising this important point. We confirm that the DNN model was intentionally trained using actual observed TBs rather than simulated TBs, and restricted to a single year (2019).

Unlike physical retrievals that rely on radiative transfer models, a key advantage of empirical machine learning models is their ability to implicitly learn and adapt to real-world instrument characteristics (e.g., specific noise patterns and minor calibration nuances). Training the DNN with actual observed TBs mapped to high-quality ERA5 profiles (Cao et al., 2026) allows the network to capture these real-world instrumental features, bypassing the uncertainties inherent in forward modeling (especially under complex cloudy conditions).

Using a single, representative year (2019, which contains a complete seasonal cycle) for training was a deliberate strategy to ensure rigorous model validation. By isolating the training phase to one year, we can apply the trained DNN to the remaining 9 years (2013–2018, 2020–2022) as a strictly independent, long-term test dataset. This temporal out-of-sample testing effectively prevents overfitting and robustly proves both the generalizability of our DNN model and the long-term stability of the MWR instrument.

We have clarified the rationale for this experimental design in the revised manuscript (L338-342).

4. The authors should add more information on the retrieval methods (e.g. utilized auxiliary data, noise) and on the radiative transfer method (e.g. absorption models). Was it tested against line-by-line models?

Reply: We sincerely thank the reviewer for these insightful questions. We have substantially expanded the description of our retrieval setup in the revised manuscript to address these points. Specifically, we have clarified the following:

(1) Forward Model: We described the radiative transfer model used in this study, namely the non-scattering microwave radiative transfer model nonScatMWRadTran (Löhnert and Crewell, 2003). The manuscript now explicitly states that the gas absorption parameterization is based on the (Rosenkranz, 1998) model, with updates to the 22.235 GHz water vapor line width (Liljegren et al., 2005) and water vapor continuum absorption (Turner et al., 2009).

Following the reviewer’s suggestion, we also added a benchmark evaluation against the MWRT model (Liu, 1998), which has previously been validated against line-by-line radiative transfer simulations (e.g., MonoRTM) for microwave radiometer observations at the XH site (Zou et al., 2021). Using 212 collocated radiosonde profiles with identical atmospheric inputs, the comparison demonstrated excellent agreement between nonScatMWRadTran and MWRT, with RMSEs below 0.12 K for the oxygen channels and below 1.5 K for the water vapor channels.

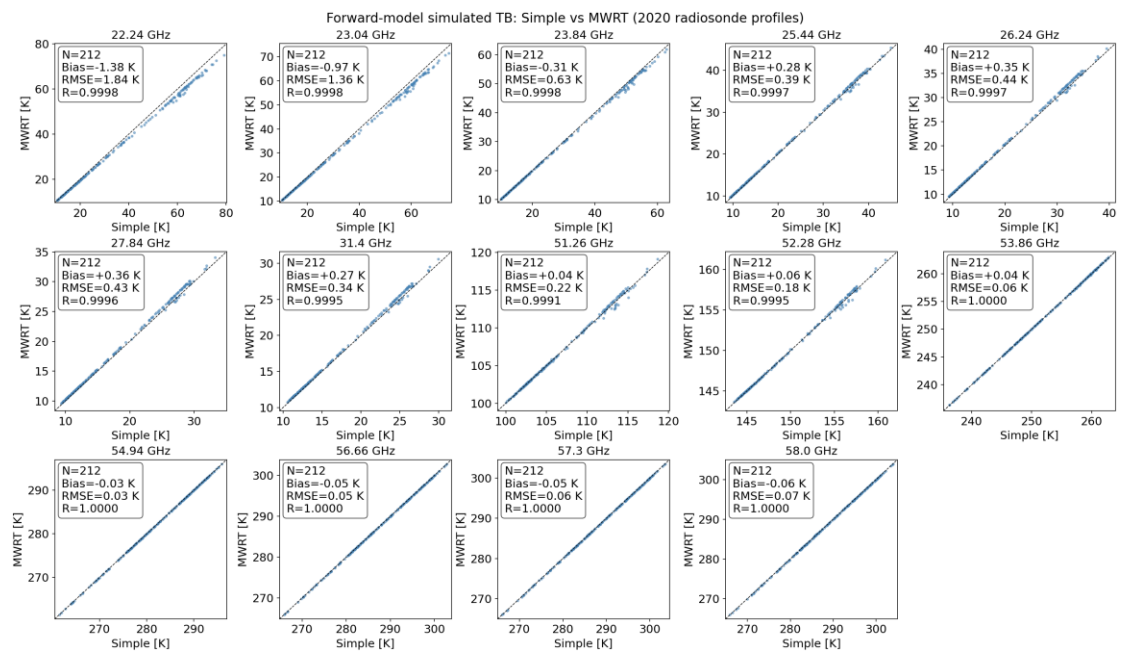


Figure R1. Comparison of simulated brightness temperatures between nonScatMWRadTran and MWRT using 212 collocated radiosonde profiles for the K-band water vapor channels and V-band oxygen channels.

(2) Retrieval Setup: The total effective observation error covariance matrix (S_e) is constructed as a diagonal matrix, assuming uncorrelated observation errors among channels. In this setup, the diagonal elements represent the total effective observation variance, which comprehensively accounts for instrumental thermal noise, calibration uncertainties, and forward-model errors. Based on the clear-sky observation-minus-simulation (O-B) residual statistics and the quality assessment of TBs detailed in Section 3.2.2, the prescribed observation error standard deviations (the square root of the diagonal elements of S_e) were specified as 2.0, 2.2, 1.8, 1.3, 1.4, 1.0, and 1.0 K for the seven K-band water vapor channels (22.24–31.40 GHz), and 1.5, 1.7, 1.2, 0.8, 1.0, 1.0, and 1.0 K for the seven V-band oxygen channels (51.26–58.00 GHz), respectively. In addition, the background error covariance matrix (S_a) was constructed separately for each season (DJF, MAM, JJA, and SON) using long-term statistics derived from collocated radiosonde observations during 2016–2019.

(3) Auxiliary Data: We clarified that local radiosonde data provided the necessary prior probability distribution (covariance matrix) for the OE retrieval.

These additions are now included in Section 3.3.1 of the revised manuscript.

Minor points:

1. More information on the site could be presented (e.g. map)

Reply: Done.

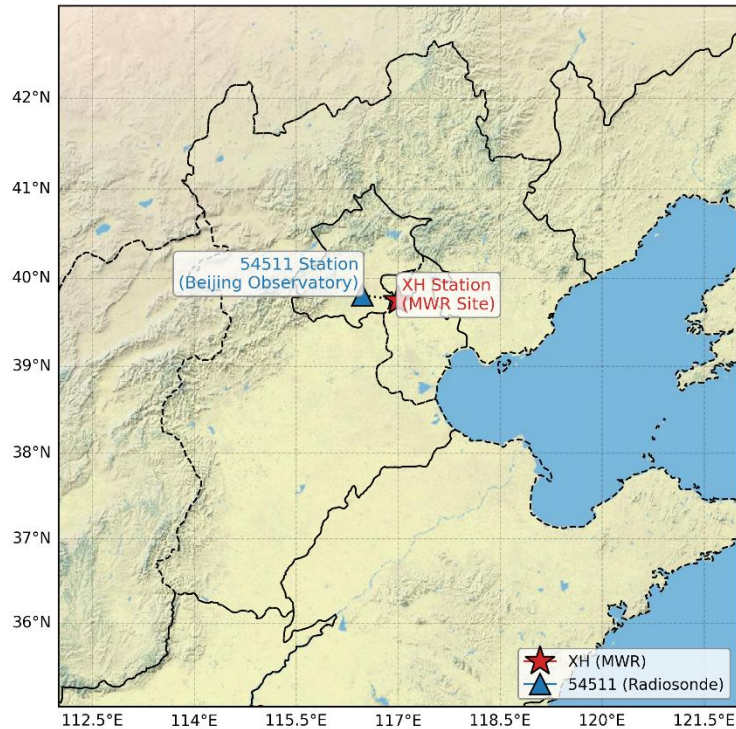


Figure 1. Geographical locations of the observation sites in the Beijing-Tianjin-Hebei region. The red star indicates the deployment site of the RPG-HATPRO ground-based microwave radiometer (MWR) at XH (39.75° N, 117.00° E). The blue triangle denotes the Beijing National Basic Meteorological Station (54511; 39.81° N, 116.47° E), which provides routine radiosonde profiles for data validation.

2. Were elevation scans available for improved temperature profiles?

Reply: We thank the reviewer for this insightful question. We acknowledge that the RPG-HATPRO MWR is capable of performing boundary-layer elevation scans, which can significantly improve the vertical resolution of temperature profiles in the lower troposphere. However, in this dataset, only zenith observations (90° elevation angle) were utilized.

3. The phrasing “weather type classification” is misleading, as it only contains a rain and cloudiness check and should be changed (e.g. to weather related information / flags) throughout the manuscript (e.g. l. 20).

Reply: Done.

4. 1.28 (and others): Change “MWR’s self-developed products” to “retrieval approach

provided by the manufacturer”. This also needs a more consistent naming, as "LV2" is too general.

Reply: Done. We have updated “MWR’s self-developed products” to “retrieval approach provided by the manufacturer”. Additionally, we agree that "LV2" is too general, and have adopted the more specific designation “RPG-LV2” for consistency.

5. 1. 70: Statistical retrievals can also rely on e.g. model data (as in this study; not only radiosondes).

Reply: We thank the reviewer for pointing out this inaccuracy. We completely agree that modern statistical retrievals extensively utilize model data (such as reanalysis datasets or numerical weather prediction outputs), rather than relying exclusively on historical radiosondes. We have revised the text in the introduction in L70-72:

“Statistical approaches, such as multivariate regression and neural network algorithms, rely on comprehensive profile datasets obtained from historical radiosondes or atmospheric models to establish empirical relationships between observed TBs and atmospheric parameters”.

6. 1. 77: cited paper doesn’t support statement and needs to be replaced to justify the decision of using multiple retrieval approaches for different atmospheric conditions.

Reply: We thank the reviewer for the correction. We have replaced the inaccurate citation (*Yan et al., 2020*) with *Maahn et al. (2020)* to properly justify the limitations of physical methods under cloudy/precipitating conditions due to hydrometeor scattering. The text is revised as follows in L76-79:

“While less dependent on large training datasets, physical methods are computationally intensive, especially under cloudy or precipitating conditions, as the complex scattering effects of large hydrometeors introduce severe non-linearity and convergence difficulties in the radiative transfer modelling (Maahn et al., 2020)”

7. 1. 81: add reference "Rose, T., Crewell, S., and Löhnert, U.: A network suitable microwave radiometer for operational monitoring of the cloudy atmosphere, Atmos.

Res., 75, 183–200, 2005." for the RPG-HATPRO system.

Reply: Done.

8. 1. 106: reference doesn't describe the specific retrieval developed for the presented study.

Reply: We have removed this reference.

9. 1. 108: change "microwave window" to "radome". Also, how often was the radome replaced?

Reply: We have revised "microwave window" to "radome". The radome of the ground-based microwave radiometer was replaced every 6 months according to the instrument maintenance manual. Monthly inspection and cleaning were also carried out during the observation period.

10. 1. 113: what is the distance to the radiosonde station?

Reply: The distance between the MWR site and the radiosonde station is approximately 50 km. We have clarified this in the revised manuscript.

11. chapter 3.1.1: this approach flags valid "extreme" events, whereas erroneous data with small deviations remain in the data set, which could cause a bias. This needs to be discussed.

Reply: We acknowledge that small-magnitude instrumental drifts or subtle erroneous data remain undetected by this 3σ check. However, these hidden biases are not ignored; they are systematically identified and eliminated in the subsequent, more rigorous QC steps—specifically, during the Observation-minus-Background (O-B) consistency checks using the radiative transfer model (as detailed in Section 3.1.3). We have added a dedicated paragraph at the end of Section 3.1.1 (L156-160) to explicitly discuss these limitations:

“While this statistical 3σ threshold serves as an efficient gross error check, it has inherent limitations. Genuine extreme meteorological events may be flagged as outliers;

however, such events typically involve precipitation and wet-radome conditions, which inherently invalidate standard MWR profiling. Conversely, undetected minor erroneous data (e.g., small instrumental drifts) remaining within the 3σ range are systematically identified and eliminated in the subsequent consistency checks to prevent long-term biases.”

12. 1. 143: 4 hours seem to be too long; was a blower/heater system deployed with the instrument?

Reply: Yes, the instrument is equipped with a continuously operating blower/heater system. While the radome typically dries within an hour, the severe sub-zero winters in Xianghe significantly slow down evaporation. Therefore, 4 hours was adopted as a strictly conservative maximum threshold to guarantee no residual moisture affects this decadal climate dataset. We have clarified this in the text (L164-167):

“Although an active blower/heater system is deployed, sub-zero winter temperatures can slow evaporation. Following cleaning, TB values generally return to normal within 1–2 hours, but this process can take and up to 4 hours in winter. Based on maintenance logs, TB observations recorded within 4 hours after radome cleaning are flagged as abnormal ($n1 = 1$ in Table 1) to conservatively ensure data purity.”

•13.1. 160: change "stability" to "variability"

Reply: Done.

14. chapter 3.1.3 b): how is CBH derived/retrieved from IRT?

Reply: We thank the reviewer for pointing this out. In this study, the CBH is not derived using a custom algorithm; rather, it is directly obtained from the standard data product provided by the RPG instrument’s built-in operating software. The manufacturer's proprietary algorithm estimates the CBH by comparing the downwelling infrared sky temperature measured by the IRT with the ambient atmospheric temperature profile to determine the condensation level. We have clarified the source of the CBH data in the revised manuscript to avoid any ambiguity (L193).

15. 1. 181: is 1 min the chosen output resolution, or are temporally higher resolved values averaged?

Reply: Thank you for this helpful comment. We agree that the temporal resolution of the TB data needed to be clarified. The original Level-1 TB observations were recorded at 1-s temporal resolution. However, the final released TB product was provided on a 1-min time grid. Specifically, each 1-min TB value was calculated by averaging the valid 1-s observations within a ± 5 s window around the nominal minute. We have added this clarification in the data description section and revised the final dataset description accordingly.

16. 1. 189: add "model uncertainties" to the argumentation

Reply: Done.

17. 1. 193: add "or bias from input data"

Reply: Done.

18. 1. 201: He et al. (2021) don't show a comparison of radiative transfer models

Reply: Done.

19. 1. 212: more details on the data matching should be presented

Reply: We thank the reviewer for this suggestion. We have expanded the text to explicitly describe the spatiotemporal matching process between the 1-minute MWR observations and the hourly model data. The temporal window and the alignment strategy are now clearly defined in L237-240:

“In the comparison study, a spatiotemporal synchronization approach was applied. Specifically, to match the 1-minute MWR observations with the hourly forward-modeled TBs, the single 1-minute MWR measurement closest to the exact hour of the simulation was selected. Subsequently, data pairs with anomalous deviations were removed using the 3σ criterion based on the Root Mean Square Error (RMSE).”

20. 1. 238: how are LWP & IWV retrieved? Retrieval approach from the manufacturer?
Was an often applied LWP offset correction carried out?

Reply: We thank the reviewer for the clarification request. The IWV and LWP values included in this dataset were directly obtained from the manufacturer's standard products. As the primary focus of this study is on the quality control of TBs and the retrieval of thermodynamic profiles (T and q), no additional post-processing, such as the LWP offset correction, was applied to these auxiliary parameters.

However, we did the IWV comparisons between MWR and CEMEL Sun photometer at XH, the results show that high precision of retrieved MWR IWV (Wang et al., 2025). Since both LWP and IWV are retrieved from MWR first 7 Channel covering the absorption of water vapor and liquid water. Under the quality assurance of TB observations from all microwave channels, the accuracy of the integrated water vapor (IWV) retrieval products can be used to indirectly verify the validity of the liquid water path (LWP) data from the microwave radiometer (MWR).

We have clarified the data source and processing in the revised manuscript (L265-266).

21. 1. 250-257: this part is repeated from the introduction and could be removed

Reply: Done.

22. Table 4: change to "Available samples after QC"

Reply: Done.

23. chapter 4.1: could the different occurrence of clouds and rain be linked to atmospheric conditions in the different years? Maybe using findings from other studies to provide more context.

Reply: We agree with the reviewer that linking the observed weather identifier frequencies to the regional climatic context adds significant value. In the revised manuscript, we have expanded the discussion in Section 4.1 (L460-467):

“Second, excluding 2013, the frequency of clear-sky occurrences at XH has shown a gradual decline in recent years, especially from 2018 to 2021, when the clear-sky frequency dropped to below 50%, with the corresponding frequency of cloudy conditions increasing to around 40%. This trend aligns with recent findings that summertime low-level cloud cover has significantly increased across China over the past two decades. Such an increase is closely associated with regional vegetation greening, which enhances surface evapotranspiration and provides local moisture supply for cloud formation (Zhang et al., 2024a). The occurrence of precipitation, except for the anomalously high value in 2013, remained relatively stable in other years, generally around 5%. In addition, uncertain samples have been consistently controlled at about 10%, except in 2017, where it approached 20%, likely due to insufficient instrument maintenance or supervision.”

24. Figure 7: the temperature from the AWS could be added

Reply: We thank the reviewer for this excellent and highly constructive suggestion. We carefully evaluated the inclusion of the AWS near-surface temperatures to validate the trends. However, after rigorous quality control, we found that the AWS record exhibits temporal inconsistencies and uneven data availability during parts of the study period, which would affect the robustness of long-term trend comparisons.

Furthermore, as part of our comprehensive revisions, we critically re-evaluated the entire long-term trend analysis section. Because the MWR observational data in certain early years were also relatively incomplete, we concluded that estimating a robust decadal trend might introduce unnecessary uncertainties. Therefore, to maintain methodological consistency and avoid potential sampling-related biases in the trend analysis, we have decided to completely remove the interannual trend analysis and the corresponding figure from the revised manuscript. Consequently, the AWS temperature series was not incorporated.

Nevertheless, the quality-controlled and temporally matched AWS observations were retained in the published dataset.

25. chapter 4.3: the generally “coarse vertical resolution” (limited information content) of MWR derived temperature profiles should be mentioned, especially regarding the EI layers. And what is the reason for increased STD in winter (Inversion Intensity)

Reply: We thank the reviewer for these important points:

(1) We have added a discussion at the beginning of Section 4.3 to acknowledge the relatively coarse vertical resolution and limited information content of MWR profiles, noting that this smoothing effect may lead to an underestimation of thin or sharp elevated inversion (EI) layers.

(2) We have also clarified the reason for the increased standard deviation (STD) of inversion intensity in winter. This high variability is attributed to the sharp contrast between extremely stable, stagnant periods (strong inversions) and well-mixed periods following frequent cold air outbreaks in the North China region during winter.

These changes are added in the revised manuscript in L502-504 and L522-524.

26. 1. 456: how many cases are included in each bin?

Reply: We thank the reviewer for the suggestion. To provide a robust statistical basis for the mean and standard deviation shown in the error bands, we have updated Figure 11 to include the number of identified inversion layers (N) for both SBI and EI within each $PM_{2.5}$ bin. The updated X-axis labels now explicitly show N_{SBI} and N_{EI} .

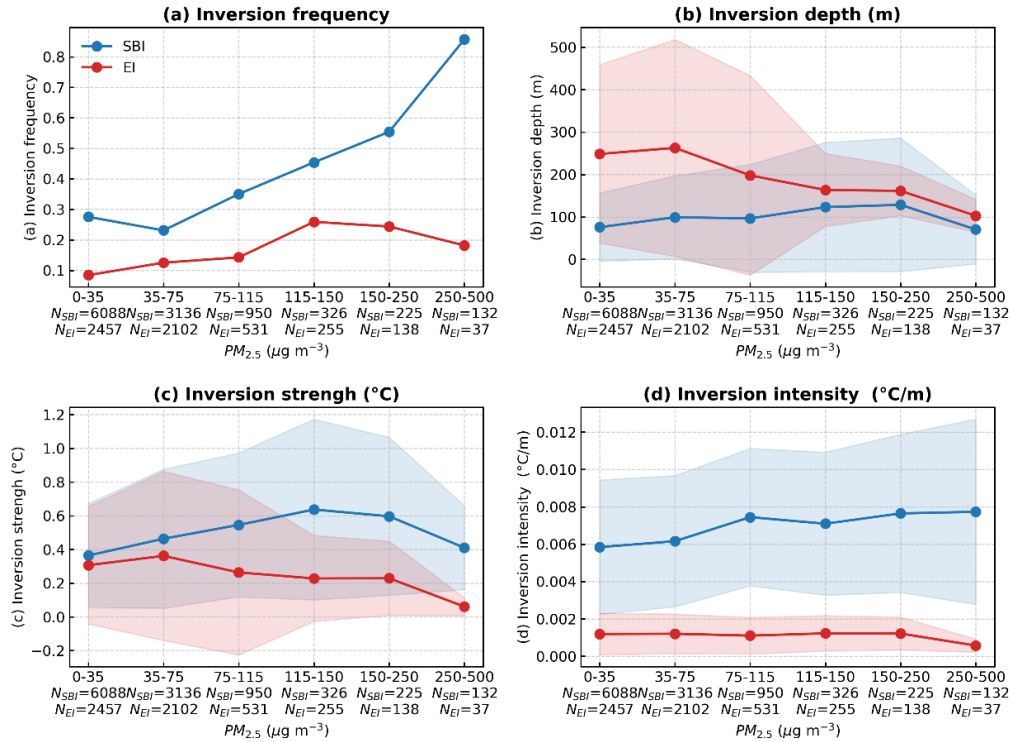


Figure 1: Statistics of inversion characteristics across different PM_{2.5} concentration bins at XH during 2018-2019. Panels show (a) inversion frequency, (b) inversion depth, (c) temperature strength (ΔT), and (d) inversion intensity ($\Delta T/\text{depth}$) for surface-based inversions (SBI, blue) and elevated temperature inversion (EI, red). The specific number of inversion layers used for statistical analysis (N_{SBI} and N_{EI}) is indicated on the x-axis for each bin. Error bands indicate ± 1 standard deviation.

Data set Major points:

1. All variables should have names and units

Reply: Thank you for this suggestion. We agree that all variables should have explicit names and units. In the revised data repository, we now provide this information in two dedicated README files, covering both profile products and brightness temperature/QC products:

(1) Profile datasets: README_profiles_with_method.md. This file documents variable names, definitions, units, time reference (UTC), height reference (m AGL), and retrieval-method labels for the merged profile datasets.

(2) Brightness temperature and QC datasets: README_BT_flag_units_v3.md.

This file documents variable meanings and units for the BT and flag files (including

nflag and IRT-related variables), and clarifies unit conventions.

Therefore, variable naming and unit definitions are now explicitly provided across the full released dataset.

2.It should be made clear which retrieval method was used

Reply: Thank you for this comment. We agree that the retrieval method used for each profile sample must be explicit. We have now clarified this in both the data files and the documentation: A dedicated column `retrieval_method` is included in the merged profile datasets. This field explicitly indicates which algorithm produced each timestamp: OE for Optimal Estimation and DNN for Deep Neural Network.

3.Uncertainty estimates (e.g. from the optimal estimation method) should be included

Reply: We appreciate this suggestion. Actually, in Fig.6 and Fig.7, the shaded regions represent the interquartile range (IQR, 25–75%) of difference, reflecting the estimated uncertainty at each level.

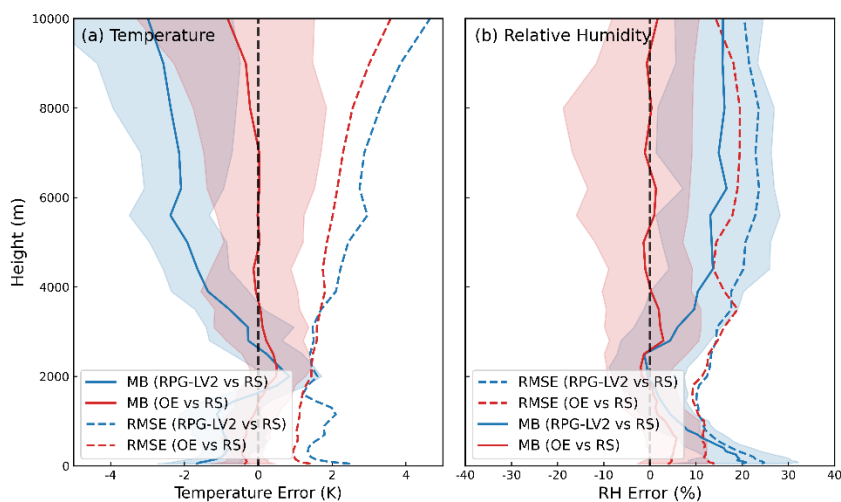


Figure 6: Vertical distributions of mean bias (MB) and root mean square error (RMSE) for RPG-LV2 and Optimal estimation retrieved temperature (a) and relative humidity (b) profiles compared with radiosonde profiles (RS) in 2020. Solid lines denote the median MB, dashed lines indicate the mean MB. The shaded regions represent the interquartile range (IQR, 25–75%) of difference, reflecting the estimated uncertainty at each level. A vertical dashed black line marks the zero-bias reference.

Minor points:

4. The data repository site lacks details about file contents (such as variable names, units, metadata), exact instrument location and information on how the products were derived

Reply: Done. We added a detailed README and variable dictionary to the repository. These files now include variable names, units, metadata, exact site coordinates (39.75°N, 116.96°E), and a concise description of the product derivation workflow (TB quality control, cloud classification, and retrieval strategy).

5. Time is encoded in two ways in the *cflag.csv files, which could be confusing

Reply: We removed the duplicated datetime column and kept only one time representation (Year/Month/Day/Hour/Minute/Second in UTC) in the revised files.

6. Is there a difference between "Rainflag" and "n2"? Could one of them be omitted?

Reply: Thank you for the comment on the precipitation flag definition. In the revised release, we keep both Rainflag and n2 in the TB+flag files. Rainflag is retained for direct consistency with the original MWR product stream, while n2 is retained for consistency with the multi-identifier QC framework used in this study. For samples where n2 is available, n2 and Rainflag are equivalent (0: no precipitation; 1: precipitation). For branch-terminated cases (e.g., n1=1, within 4 hours after cleaning), n2 can be NaN by QC design, whereas Rainflag remains available as the native precipitation indicator.

7. Are there two IR channels? IRT2 is not mentioned in the manuscript (and has a different unit than IRT1)

Reply: Yes, two IR channels are included. We clarified this in the metadata and explicitly distinguished their units: IRT1_K (Kelvin) and IRT2_K (Kelvin). We also updated the variable documentation accordingly.

8. "cflag" is not defined in the manuscript

Reply: We agree that this was an oversight. As “cflag” is an unnecessary variable for this specific data product, we have removed it entirely from the updated dataset.

9. Why do some quality flags show NaN

Reply: Thank you for the comment. We re-checked all yearly files (2013-2022) and confirmed that NaN values are intentional outputs of the QC decision tree, not random missing values. Importantly, NaN occurs in branch flags (n2/n3/n4), while nflag itself has no NaN values.

Specifically:

(1) When $n_1 = 1$ (within 4 hours after instrument cleaning), n_2 , n_3 , and n_4 are not evaluated and are set to NaN; $nflag = 3$.

(2) For precipitation cases ($n_2 = 1$), n_3 and n_4 are not evaluated and are set to NaN; $nflag = 2$.

(3) For non-precipitation cases with cloudy sigma-c condition ($n_2 = 0$, $n_3 = 1$), n_4 is not evaluated and is set to NaN; $nflag = 1$.

We have added this clarification to the dataset documentation and explicitly state that these NaNs mean “not applicable by design.”

“**Table 1.** Multi-Identifier for TB observations and the meanings of its values

| Identifier | Value | Checking Condition | Meaning |
|------------|-------|--|--------------------------|
| n_1 | 0 | $ TB(n) - \mu \leq 3\sigma$ and not within cleaning window | Normal |
| | 1 | $ TB(n) - \mu > 3\sigma$ or ($ TB(n) - \mu \leq 3\sigma$ and within 4 hours after cleaning) | Abnormal |
| n_2 | 0 | $Rainflag = 0$ | No precipitation |
| | 1 | $Rainflag = 1$ | Precipitation |
| | NaN | $n_1 = 1$ (branch terminated) | Not applicable by design |
| n_3 | 0 | $\sigma_c \leq a + b \cdot IWV$ | Clear-sky |
| | 1 | $\sigma_c > a + b \cdot IWV$ | Cloudy |

| Identifier | Value | Checking Condition | Meaning |
|------------|-------|---|--------------------------|
| | NaN | $n_1 = 1$ or $n_2 = 1$ (branch terminated) | Not applicable by design |
| n_4 | 0 | $CBH < 500\text{m}$ or $CBH > 8000\text{ m}$ | Clear-sky |
| | 1 | $500\text{ m} < CBH < 8000\text{m}$ | Cloudy |
| | NaN | $n_1 = 1$, or $n_2 = 1$, or $n_3 = 1$ (branch terminated) | Not applicable by design |
| $nflag$ | 0 | $n_1=0, n_2=0, n_3=0, n_4=0$ | Clear-sky |
| | 1 | $n_1=0, n_2=0, (n_3=1 \text{ or } n_4=1)$ | Cloudy |
| | 2 | $n_1=0, n_2=1$ | Precipitation |
| | 3 | $n_1=1$ | Other/uncertain |

Note: NaN values in n_2 - n_4 indicate branch-terminated checks and are not random missing data, while $nflag$ contains no NaN values.”

References:

Cao, Y., Yang, Y., Zhang, Y., Lv, X., Ma, Y., Liu, R., Ren, X., Hu, Y., and Xin, J.: Evaluation of FY-3E, CRA, and ERA5 temperature and humidity profiles over north China in summer, *Remote Sens.*, 18, 1058, <https://doi.org/10.3390/rs18071058>, 2026.

Liljegren, J. C., Boukabara, S.-A., Cady-Pereira, K., and Clough, S. A.: The effect of the half-width of the 22-GHz water vapor line on retrievals of temperature and water vapor profiles with a 12-channel microwave radiometer, *IEEE Trans. Geosci. Remote Sens.*, 43, 1102–1108, <https://doi.org/10.1109/TGRS.2004.839593>, 2005.

Liu, G.: A fast and accurate model for microwave radiance calculations, *J. Meteorol. Soc. Jpn.*, II, 76, 335–343, https://doi.org/10.2151/jmsj1965.76.2_335, 1998.

Löhnert, U. and Crewell, S.: Accuracy of cloud liquid water path from ground-based microwave radiometry 1. Dependency on cloud model statistics, *Radio Sci.*, 38, 2002RS002654, <https://doi.org/10.1029/2002RS002654>, 2003.

Maahn, M., Turner, D. D., Löhnert, U., Posselt, D. J., Ebell, K., Mace, G. G., and Comstock, J. M.: Optimal estimation retrievals and their uncertainties, *Bull. Am. Meteorol. Soc.*, 101, E1512–E1523, <https://doi.org/10.1175/BAMS-D-19-0027.1>, 2020.

Rosenkranz, P. W.: Water vapor microwave continuum absorption: a comparison of measurements and models, *Radio Sci.*, 33, 919–928, <https://doi.org/10.1029/98RS01182>, 1998.

Turner, D. D., Cadeddu, M. P., Lohnert, U., Crewell, S., and Vogelmann, A. M.: Modifications to the water vapor continuum in the microwave suggested by ground-based 150-GHz observations, *IEEE Trans. Geosci. Remote Sens.*, 47, 3326–3337, <https://doi.org/10.1109/TGRS.2009.2022262>, 2009.

Wang D., Gong Y., He W., Xia X., and Liu M.: Characterization of Atmospheric Water Vapor Using Ground-Based Microwave Radiometry, *气候与环境研究*, 31, 1–11, <https://doi.org/10.3878/j.issn.1006-9585.2025.24132>, 2025.

Zou, R., He, W., Wang, P., Mao, J., Chen, H., Li, J., Nan, W., and Chang, Y.: Assessment of Radiative Transfer Models Based on Observed Brightness Temperature from Ground-Based Microwave Radiometer, *Chinese Journal of Atmospheric Sciences*, 45, 605–616, <https://doi.org/10.3878/j.issn.1006-9895.2008.20134>, 2021.

Response to Reviewer 2:

Reviewer #2: The dataset contains around 10 years of ground-based microwave radiometer (MWR) observations at Xianghe Integrated Observatory in China. It consists of calibrated zenith brightness temperatures (TBs) in the K- and V-band channels, integrated water vapor (IWV), liquid water path (LWP) and a set of associated flags with a one-minute resolution. It also contains retrieved temperature and relative humidity profiles with a 10 minutes resolution computed using two different retrieval methods.

The associated manuscript describes the methodology used to construct the dataset. It details the quality control applied to the data, the retrieval processes and shows a validation of the data quality. It finally shows some case studies that exemplify how the data can be used. For the validation of their dataset, the authors compared the measured TB against radiative transfer simulation from model data and used radiosonde data to validate their retrieved temperature and humidity profiles. The authors also show some comparisons of their retrieved profiles against the manufacturer products.

In general, such high temporal resolution and long-term ground-based MWR observations are valuable and can provide interesting insights on the atmospheric boundary layer and its evolution. As exemplified by the authors, these measurements can be used in a variety of atmospheric studies (e.g. air quality, atmospheric trends, ...) and they can serve as validation data, e.g. for numerical weather predictions (NWP). The provided dataset is relatively easy to use but it does lack some documentation and some of the variables are not well defined in the manuscript nor in the data repository. The dataset and associated manuscript fits the scope of ESSD and my recommendation would be to accept it for publication after addressing the specific comments below.

Reply: We would like to express our gratitude to the Reviewer for the careful review and the encouraging overall assessment of our manuscript and dataset. We are glad that the reviewer recognizes the potential of this decadal dataset for diverse atmospheric and NWP validation studies. We have carefully considered your primary concern regarding the dataset documentation. In response, we have thoroughly updated the data

repository, provided comprehensive variable definitions, and ensured that all metadata meets the rigorous standards expected by ESSD data users. Please find our detailed point-by-point responses to your specific comments below.

Dataset:

1. The dataset is lacking some documentation, especially in explaining what is contained in the different files. Some of the variables are explained in the manuscript but some are undefined (e.g. cflag, IRT1 vs IRT2, ...). Also the units are generally missing from the dataset and should be specified for all the variables. Although most are obvious, some are not, e.g. datetime (is this UTC or LST?) or the altitude (I guess m.a.g.l. but this should be specified somewhere). Proper variable naming is missing in the 2013-2022final merged RH.csv

Reply: We thank the reviewer for their careful examination. We fully agree and have comprehensively updated the dataset documentation. We have added a README file explaining the file contents. All variables (including cflag, IRT1, and IRT2) are now clearly defined, and standard units have been assigned across all files. Specifically, we have clarified that datetime is in UTC and altitude is in m a.g.l.. Additionally, we have added the missing variable names to the header of the 2013-2022final merged RH.csv file.

2. To address these points, the provided dataset would highly benefit from an additional README file.

Reply: Done.

3. The origin of some quantities is unclear, e.g. LWP and IWV as they could in principle be retrieved (from either methods) or directly taken from the manufacturers retrieved values.

Reply: We thank the reviewer for pointing this out. To clarify, the Liquid Water Path (LWP) and Integrated Water Vapor (IWV) provided in our dataset are directly taken from the manufacturer's retrieved values (RPG-LV2), rather than being retrieved from

our OEM or DNN methods. We acknowledge that this distinction was not clear in the original text. Accordingly, we have explicitly stated the origin of these specific integrated quantities in the revised manuscript (L265) and README file in the data repository to avoid any future ambiguity.

4.The temperature and RH profiles should be flagged to identify which profile were retrieved from OEM or from the DNN method in the dataset. Also, the fact that the dataset is composed of merged profiles from two retrieval methods based on weather conditions should be clearly stated in the data repository.

Reply: We appreciate this valuable recommendation. We have updated the dataset to include specific flags that identify the retrieval method (OEM or DNN) used for each profile. In addition, the main description in the data repository have been updated to clearly clarify that the dataset consists of merged profiles from these two methods depending on the weather conditions.

5.Uncertainty estimates should be provided in the dataset.

Reply: We thank the reviewer for raising this important point. We completely agree that sample-specific uncertainty estimates are highly valuable. However, there is a fundamental methodological difference between the two algorithms used. While the physical Optimal Estimation (OE) method natively calculates the posterior error covariance to provide dynamic uncertainty for each individual retrieval, our standard Deep Neural Network (DNN) model is a deterministic approach that does not inherently generate sample-specific uncertainty estimates. Because the DNN cannot output per-profile uncertainties, providing observation-level uncertainty in the final dataset is not currently feasible. Instead, we have conducted a rigorous statistical evaluation against independent observations to quantify the overall empirical uncertainty of the entire dataset. These comprehensive overall uncertainty estimates (e.g., RMSE, bias) are detailed in Table 2 and Table 3 of the manuscript.

6.Adding the meteorological parameters measured by the AWS could also be beneficial

for a potential user of this dataset.

Reply: Thank you for this valuable suggestion. We agree that including meteorological parameters from the AWS can enhance the usability of the dataset for potential users. In the revised work, we have incorporated the available AWS meteorological variables into the data repository where quality-controlled observations exist. However, due to intermittent data gaps and quality issues in certain periods, the AWS record is not fully continuous throughout the entire dataset.

To ensure transparency and proper usage, we have documented the availability, temporal coverage, and quality considerations of the AWS variables in the README file of the data repository. This allows users to clearly understand the applicable periods and limitations when using these auxiliary meteorological parameters.

7. Also, is there a reason why the vertical grid is irregularly spaced? This could be specified somewhere in the manuscript, e.g. when giving more details on the retrieval methods.

Reply: Done. We thank the reviewer for pointing this out. The irregularly spaced vertical grid is fundamentally determined by the inherent characteristics of MWR observations, where the vertical resolution naturally decreases with altitude. To maintain comparability, our grid configuration generally aligns with the standard 39-level design of the RPG-LV2 product. However, we made specific modifications in the near-surface layer. While the standard RPG-LV2 product uses a denser and irregular spacing below 150 m (i.e., 0, 10, 30, 50, 75, 100, 125, 150 m), we adopted a uniform 50 m resolution near the surface (i.e., 0, 50, 100, 150 m) based on our specific retrieval requirements. This modification results in a total of 35 vertical levels instead of 39. Per your constructive suggestion, we have added a detailed explanation of this vertical grid configuration and its rationale to the retrieval methods section in the revised manuscript:

8. Although I don't think it's required from ESSD, sharing the retrieval code(s) could be a great asset and of interest for the MWR community

Reply: Done. We have prepared and uploaded the core Python scripts for our retrieval

approach to the data repository.

Manuscript

Measurement site:

Please add a short description of the measurement sites in Section 2. Especially the type of environment where it's located and the distance to the other measurement (RS, which is only mentioned in 1.266, or PM2.5).

Reply: We sincerely thank the reviewer for pointing this out. A clear description of the measurement environment and spatial collocation is crucial for understanding the representativeness of the dataset. In the revised manuscript (Section 2), we have added a comprehensive description of the Xianghe (XH) site, detailing its suburban environment and its geographical role in the North China Plain. Furthermore, we have explicitly clarified the distances to the other measurements: The PM2.5 observations are strictly co-located with the MWR at the XH site. The radiosonde station is located in an urban environment approximately 50 km northwest of the XH site.

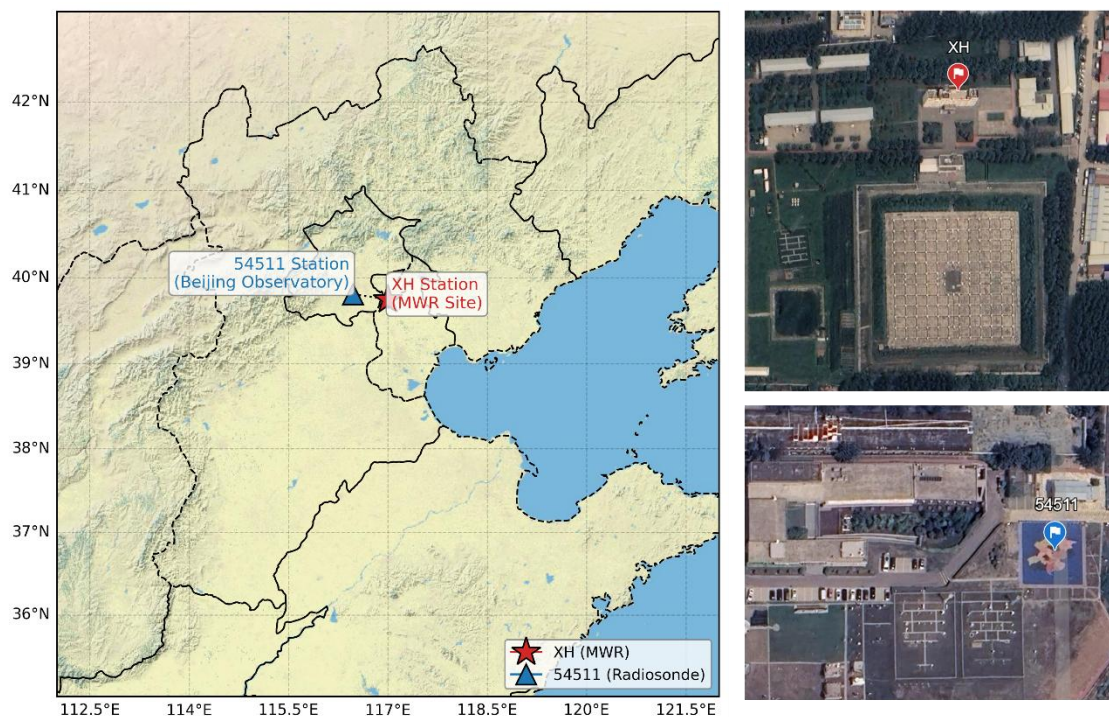


Figure 2: Geographical locations of the observation sites in the Beijing-Tianjin-Hebei region. The red star indicates the deployment site of the RPG-HATPRO ground-based microwave radiometer (MWR) at XH (39.75° N, 116.96° E). The blue triangle denotes the Beijing National Basic Meteorological Station

(54511; 39.81° N, 116.47° E), which provides routine radiosonde profiles for data validation.

Retrieval methods:

1. In general, the authors could add a bit more details on the retrieval methods used to compute the thermo dynamic profiles. For the OE retrieval, you should state which quantities were retrieved and add more information on the retrieval setup (e.g. noise covariance matrix) and maybe state the altitude range where the retrievals can actually be trusted and what vertical resolution can be expected. The uncertainty derived from the OE could be added as well as this, together with the vertical resolution will greatly impact any analysis conducted with this dataset.

Reply: We sincerely thank the reviewer for these insightful recommendations. We have substantially expanded the methodological details of our OE retrieval in the revised manuscript.

Specifically, we addressed the reviewer's points as follows:

(1) The OE method is mainly used for retrieving atmospheric temperature and humidity profiles in 0-10 km under clear-sky conditions.

(2) Vertical Resolution: To ensure seamless compatibility with the operational products and our DNN model, the OE retrieval adopted the identical vertical layering scheme as the LV2 product (39 layers in total, with 25 dense layers below 2 km having a resolution of 10–200 m, and coarser layers up to 10 km).

(2) Trusted Altitude Range: We explicitly stated that RPG-MWR profiling relies on the distribution of weighting function for each channel, meaning the highest vertical resolution and most reliable retrievals are concentrated in the lower boundary layer (0–2 km), while the information content and resolution degrade significantly in the mid-to-upper troposphere.

(3) Retrieval Setup: The total effective observation error covariance matrix (S_e) is constructed as a diagonal matrix, assuming uncorrelated observation errors among channels. In this setup, the diagonal elements represent the total effective observation variance, which comprehensively accounts for instrumental thermal noise, calibration uncertainties, and forward-model errors. Based on the clear-sky observation-minus-

simulation (O-B) residual statistics and the quality assessment of TBs detailed in Section 3.2.2, the prescribed observation error standard deviations (the square root of the diagonal elements of S_e) were specified as 2.0, 2.2, 1.8, 1.3, 1.4, 1.0, and 1.0 K for the seven K-band water vapor channels (22.24–31.40 GHz), and 1.5, 1.7, 1.2, 0.8, 1.0, 1.0, and 1.0 K for the seven V-band oxygen channels (51.26–58.00 GHz), respectively. In addition, the background error covariance matrix (S_a) was constructed separately for each season (DJF, MAM, JJA, and SON) using long-term statistics derived from collocated radiosonde observations during 2016–2019.

(4) Handling of Uncertainty: We appreciate the suggestion to include the analytical uncertainty derived from the OE posterior covariance matrix. Indeed, the theoretical uncertainty of the retrieval for each layer of the temperature and humidity profiles can be provided by the OE method. However, in this study, the product of this theoretical uncertainty was not separately output; instead, an independent comparison with the true values of radiosonde observations was adopted to quantitatively characterize the actual uncertainty of the retrieval results. It can be seen from the error analysis of the temperature and humidity profiles and the distribution of confidence shadow intervals in Fig. 6 that the retrieval uncertainty range is narrow and the precision is higher within the boundary layer below 2 km; as the height increases upward, the retrieval uncertainty gradually increases significantly.

2. It is unclear to me whether scanning observations were performed during the 10 year. I think it is a key information to provide in the manuscript (and also whether they were used in the retrieval) and if so, it should be either added to the dataset or specified that only the zenith observations are shared.

Reply: We thank the reviewer for pointing out this critical omission. During the 10-year operation at the XH site, the RPG-HATPRO MWR was primarily configured to perform zenith observations (90° elevation angle) to provide continuous, high-temporal-resolution (second-level) TB data. Then, a multi-angle boundary layer scanning (BLS) mode was carried out every approximately 20 minutes, with a duration of less than 2 minutes.

In this work, we mainly utilized the zenith observations for both the OE and DNN retrievals. We have explicitly clarified in the revised manuscript (Section 2 and the Data Availability statement) that the shared long-term TB dataset and the corresponding derived profiles are based solely on zenith observations. The BLS data would provide more valuable information for improving the currently retrieved profiles within 0-2 km in the future.

3.The origin of the IWV and LWP values should be clearly indicated in the manuscript and dataset as this could be taken either from the manufacturer or from the retrievals.

Reply: Done. The IWV and LWP values included in this dataset were directly obtained from the manufacturer's standard products. We have clarified the data source in the revised manuscript (L265-266).

4.Did you perform any kind of quality check on the retrieved profiles?

Reply: Yes, we performed basic physical rationality checks on the retrieved profiles prior to the final dataset integration. For instance, the retrieved RH was strictly constrained within the physical bounds of 0% to 100%. Any profiles containing unphysical outliers were excluded to ensure the reliability of the final dataset. We have added a brief note about this in the revised manuscript (L437-438).

5.The justification on why this merged retrievals approach was chosen should be more detailed as this remains quite vague for the reader, e.g. in Section 3.3.

Reply: We sincerely thank the reviewer for this constructive feedback. While the theoretical rationale for combining physically based and data-driven methods was introduced in the Introduction (Section 1), we fully agree that its specific application and justification needed to be explicitly reiterated and detailed at the beginning of Section 3.3 to guide the reader. In the revised manuscript, we have added an introductory paragraph to Section 3.3 to explicitly link the theoretical concepts from the Introduction to our methodological choices (L277-282):

“As discussed in Section 1, the justification for this condition-dependent strategy

is rooted in the varying physical complexities of microwave radiative transfer. Under clear-sky conditions, the physically based OE method is preferred as it ensures high thermodynamic consistency and avoids potential empirical overfitting. Conversely, for cloudy-sky conditions, a data-driven DNN is utilized to handle the complex non-linear scattering effects of clouds, which typically introduce significant uncertainties in traditional forward models.”

Comparisons of thermodynamic profiles:

1. It is an interesting approach to construct a the profile dataset out of 2 different retrieval methods but it would be interesting to compare how these 2 methods compare to each other. In fact, this comparison would allow the reader to judge whether the DNN is actually performing better under cloudy scene and provide some justification for the merged retrievals approach.

Reply: We thank the reviewer for this insightful comment. We completely agree that an empirical comparison between the two methods under cloudy conditions would be highly interesting. However, applying the OE method to cloudy scenes presents severe theoretical and practical challenges, which is precisely the fundamental reason we adopted this condition-dependent merged approach.

Theoretically, the accuracy of the physically-based OE method relies heavily on the precision of the forward radiative transfer model. Under cloudy conditions, microwave emission and scattering are highly sensitive to cloud microphysical parameters, particularly the vertical profile of Liquid Water Content (LWC). Without accurate *a priori* knowledge of LWC profiles (which are notoriously difficult to obtain continuously from standard observations), the forward model introduces massive uncertainties. Consequently, applying OE to cloudy scenes without precise cloud inputs typically leads to poor convergence rates and highly erroneous thermodynamic profiles. In contrast, the data-driven DNN effectively bypasses this explicit forward-modeling bottleneck by statistically learning the complex, non-linear relationships directly from large-scale historical datasets. Therefore, the decision to use DNN over OE in cloudy conditions is driven by fundamental physical constraints and the lack of accurate cloud

prior information, making a direct, fair empirical comparison unfeasible.

For the two methods themselves, we can make further comparisons for the retrieved profiles using both OE and DNN method under clear-sky conditions in the future.

2.I disagree with the 2 last sentences of Section 3 (line 353 to 356). We lack information on the manufacturer retrieval and it's likely that they are also using some kind of Neural Network approach. So the difference could be explained by the training dataset used as much as in the method. Also, the reader can not appreciate the last sentences as no comparisons are shown between the OE and the DNN approach, especially not under different sky conditions. The same can be said of the (2) of the Conclusions (1.492).

Reply: We completely agree with the reviewer's perceptive comments. We acknowledge that without detailed access to the proprietary manufacturer LV2 algorithm (which likely also incorporates neural networks), attributing the performance difference solely to the algorithmic architecture was speculative. The improvements over the LV2 product are indeed largely driven by our localized, condition-dependent strategy and the specific regional training datasets.

Accordingly, we have revised the end of Section 3.3 to remove the over-generalized statements about algorithmic superiority. The revised text now objectively focuses on the "application perspective" and the practical benefits of deploying OE and DNN under their respective optimal conditions (clear vs. cloudy skies). Furthermore, as per your suggestion, we have strictly aligned the corresponding statement in the Conclusions (point 2) to reflect this application-oriented perspective, avoiding unsupported comparisons:

"(2) A condition-dependent retrieval strategy is effectively implemented to optimize profile accuracy across different weather conditions. By applying the physically based OE method to clear-sky observations and the data-driven DNN model to cloudy skies, the retrieved profiles achieve substantial improvements over the generic operational LV2 product."

Technical correction

1.1.28-29: change “MWR’s self-developed products” by something more meaningful, e.g. “manufacturer’s retrieval”. Similarly, the term “LV2” should be replaced by something more meaningful (e.g. manufacturer, RPG, ...) throughout the manuscript.

Reply: Done. We thank the reviewer for pointing out this ambiguity. We fully agree that “self-developed” could be misleading to readers and that “LV2” is too generic without specifying the source. In the revised manuscript, we have systematically updated this terminology. We have performed a global search-and-replace for the term “LV2”. It has been replaced with a more specific and meaningful term as “RPG-LV2 product”.

2.1.33: SBI acronym should be defined in the line above.

Reply: Done.

3.1.48: here you could also add a reference to Thomas et al. 2025

Reply: Done.

4.1.96: Specify which version of the HATPRO was used.

Reply: The instrument deployed at the XH station is an RPG-HATPRO Generation 3 (G3) microwave radiometer. We have explicitly specified this version information in Section 2 of the revised manuscript.

5.1.126: maybe replace “electromagnetic” with “radio-frequency”?

Reply: Done.

6.1.135: can you maybe add a sentence justifying the “singular value check”? Were there a lot of data removed with this check and was this due e.g. to a specific issue, e.g. the radome wetting? Also maybe I’d rename this check as an “Outlier detection” step.

Reply: We appreciate the reviewer’s helpful comments. We have changed the term “Singular value check” to “Outlier detection” throughout the text.

To justify this step, we have expanded the explanation in the revised manuscript.

This specific filter aims to identify isolated outliers ($n1=1$) based on the condition: $|TB(n) - \mu| > 3\sigma$ or ($|TB(n) - \mu| \leq 3\sigma$ and within 4 hours after cleaning). It serves as a crucial first-pass safeguard. The data removed by this check account for only a small fraction of the dataset (approximately 10 %, Table 4). These specific outliers are generally not caused by natural atmospheric variations, but rather by transient hardware anomalies, unrecognized severe radome contamination, or artificial thermal instability introduced during the instrument's cleaning process.

7. Table 1: IRT threshold for the cloud checks should be specified and justified.

Reply: Sorry for our writing mistake. Actually, there is no `IRT_threshold` used here, for `n4`, it is derived mainly from CBH values which is obtained from IRT measurements. We have corrected the corresponding parts in Table 1 and L195-198.

8. Table 1: for `nflag=3`, the checking conditions should probably be `n1=1 OR n1=2` no?

Reply: We appreciate the reviewer pointing this out. We apologize for the confusion, which was caused by a typo in the original manuscript. The outlier detection algorithm is exclusively designed to filter out isolated anomalies, meaning the condition is strictly `n1=1`. We have corrected Table 1 and deleted the incorrect `n1=2` condition entirely.

9.1.200: probably the word “accuracy” and “efficiency” should be exchanged?

Reply: Done.

10.1.211: can you give more details on how the dataset was matched?

Reply: Done. We thank the reviewer for this suggestion. We have expanded the text to explicitly in L237-240 describe the spatiotemporal matching process:

“In the comparison study, a spatiotemporal synchronization approach was applied. Specifically, to match the 1-minute MWR observations with the hourly forward-modeled TBs, the single 1-minute MWR measurement closest to the exact hour of the simulation was selected. Subsequently, data pairs with anomalous deviations were removed using the 3σ criterion based on the Root Mean Square Error (RMSE).”

11. Figure 2: the colormap used in this figure does not comply to the journal guidelines, please consider changing it.

Reply: Done.

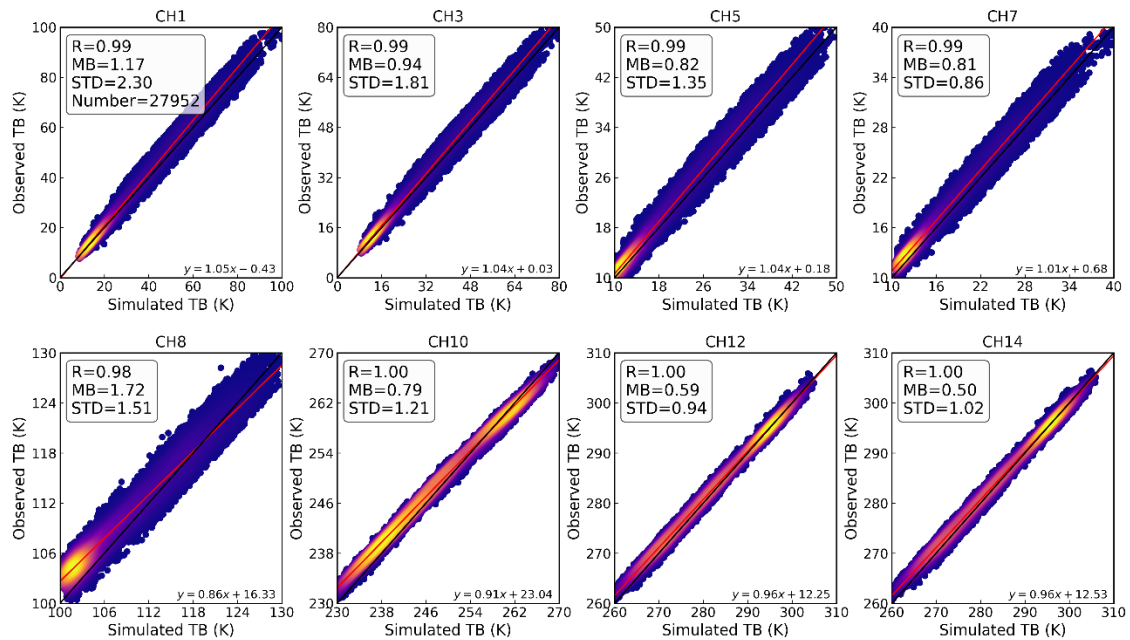


Figure 3: Scatter plots of multi-channel observed and simulated TB during 2013-2022.

12. Figure 3: the colormap design is somehow misleading as the zero bias is not centered in the colorbar.

Reply: Done.

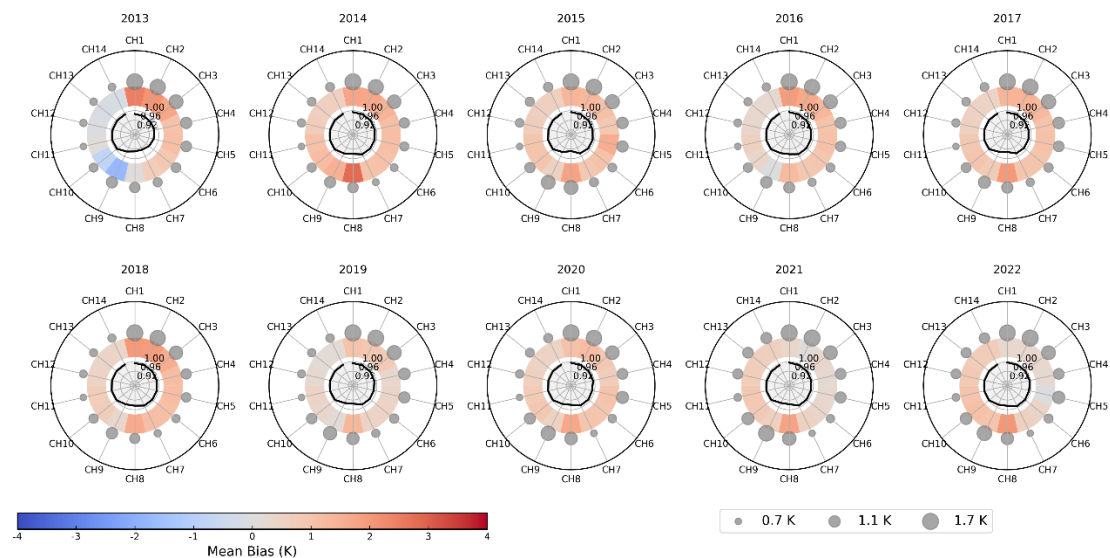


Figure 4: Ten-year (2013-2022) polar summary of the agreement between observed and ERA5-simulated brightness temperatures for all 14 channels. Each subplot shows one year, with the inner black solid curve representing the R, the middle colour shaded ring indicating the mean bias, and the circle marker sizes in the outer ring representing the standard deviation of differences. The bottom horizontal colorbar denotes the bias scale, and the adjacent legend explains the marker sizes for standard deviation

13.1.305: for the reference truth, why is ERA5 mentioned here? This is not used further for the profile validation (only for TB simulation) or did I miss it?

Reply: We thank the reviewer for pointing this out and have deleted the mention of ERA5 in this sentence in the revised manuscript.

14.1.339: although the temperature bias is small, it does look like it exist from Fig. 6.

Reply: We appreciate the reviewer pointing this out. We agree that a small temperature bias does exist. We have updated the manuscript to reflect this accurately. The revised text specifies that the DNN-retrieved temperature profiles have an overall mean bias of 0.38 K.

15. Figure 6: could be worth to state in the legend that the profiles shown here are now from the DNN retrievals, whereas Figure 5 showed the OE-retrieved profiles.

Reply: Done.

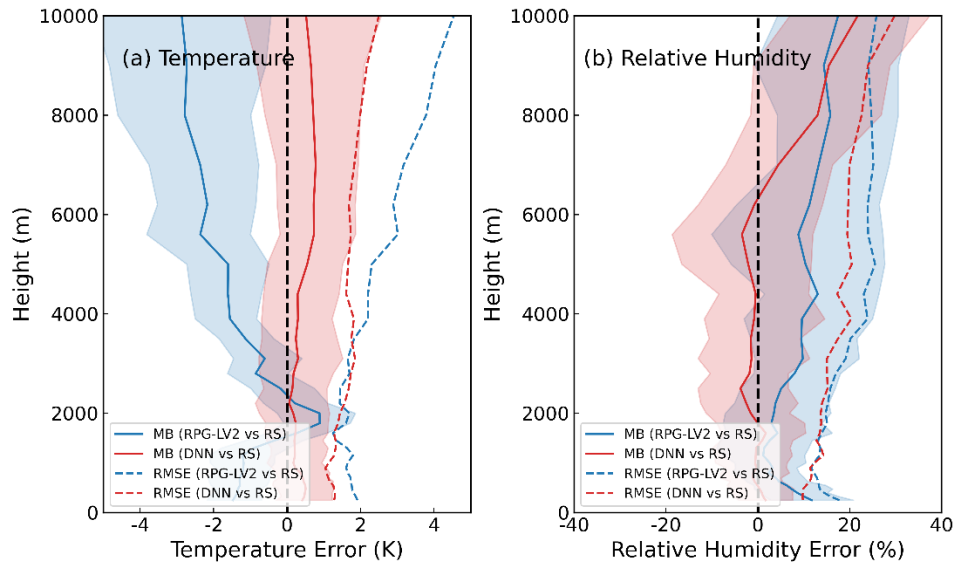


Figure 5: Same as Fig. 6, but for cloudy-sky conditions.

16.1.373-374 and Table 4: how do you define the “available” sample here? Does this refers to the good quality, unflagged data? Also do I understand correctly that the percentage refers to proportion of available data out of the total sample. Could have been interesting to have the proportion of available data out of the full year instead, the numbers would have been lower but the information equally valuable.

Reply: We thank the reviewer for this insightful comment and for allowing us to clarify this definition. Regarding the definition of “available” samples: it does not refer to strictly unflagged data. In our dataset, valid meteorological conditions are inherently flagged ($nflag = 0$ for clear-sky, 1 for cloudy, and 2 for rain). The “available” data refers to all observation samples that successfully passed the quality control checks. Specifically, it encompasses all data except those classified as $nflag = 3$ (which, as discussed in our response to Comment #6, represents instrumental outliers such as $|TB(n) - \mu| > 3\sigma$ or $(|TB(n) - \mu| \leq 3\sigma$ and within 4 hours after cleaning).

Regarding the percentage calculation: yes, you understood correctly. The percentage currently reported in Table 4 represents the proportion of these “available” data out of the total recorded observations (the total sample collected by the instrument). We agree with your perspective that reporting the availability relative to the full theoretical calendar year would also be interesting and yield a lower number by

accounting for power outages or downtime. However, for the purpose of this table, calculating the proportion based on the total recorded sample more directly reflects the effective yield rate of our quality control and retrieval algorithm on the actual measurements obtained.

To prevent any misinterpretation, we have added an explicit explanation in the revised manuscript to clearly define both the "available" sample and the denominator used for these percentages in L445-447:

“Specifically, the remaining “available” sample encompasses all data not flagged as instrumental errors ($nflag \neq 3$). The proportion indicates the ratio of these available samples out of the total recorded observations.”

17. Figure 7: the red and pink colors are not so easy to distinguish on this figure

Reply: We sincerely thank the reviewer for pointing this out and for their attention to the visual clarity of our figures. However, we have critically re-evaluated the long-term trend analysis in Section 4.2. Because the missing data in certain years (as shown in Table 4) could introduce uncertainties into the long-term trend estimation, we decided to completely remove the trend analysis to ensure the utmost scientific rigor of the manuscript. Consequently, this figure (the trend chart) has been entirely removed from the revised version.