Dear Editors and Reviewers:

We deeply appreciate your helpful comments and suggestions, which enabled us to improve the quality of our present study. We have made revisions and replied to all the comments. Please find the point-by-point responses to the comments below. The comments are given below in black, our responses are in blue, and proposed changes to the manuscript are in red.

Reply to Reviewer #2:

General Comments

1. Could you please include a few more details about the assimilation framework. Specifically:

-Timing of Observations: Do you select the observation that most closely matches the analysis time, or do you integrate over a broader time window?

-Data Processing Before Bias Correction: Aside from using the cloud mask, do you apply any additional procedures to improve measurement quality before performing bias correction?

-**QC**: Do you use any quality-control checks (e.g., discarding observations based on O−B thresholds), and if so, how many observations are rejected?

Response:

The observation nearest to the analysis time within ±10 minutes will be selected.

Before implementing bias correction, we sequentially applied cloud and precipitation detection, O−B (observation-minus-background) checks, and quality control using a whitelist. Cloud and precipitation detection were based on the FY-4B cloud mask and rain gauge data from MWRs. The O−B check served as a preliminary screening step that excluded observations with absolute O−B values greater than 15 K. The whitelist was established to identify reliable station observations based on the correlation between simulated and observed brightness temperatures (BTs). Specifically, we evaluated the correlation between observed and simulated BTs at each station. Stations where observed BTs did not vary consistently with simulated BTs—indicated by low correlation

coefficients—were classified as "problematic stations." Some of these problematic stations returned to normal after specific dates, and observations from those periods were retained.

After bias correction, additional quality control based on O−B was conducted, including a three-sigma observational error check. For the six-hour, three-hour, and one-hour assimilation intervals, the total numbers of observations were 3,724, 4,946, and 10,619, respectively. Following a three-sigma observational error check, 34 (0.91%), 70 (1.42%), and 76 (0.72%) observations were rejected. The corresponding changes we have implemented are as follows:

"Before implementing bias correction, clear-sky screening, first-guess departure check, and whitelist check were sequentially applied to improve measurement quality. Subsequently, a relative departure check was applied prior to minimization. For the 6 h, 3 h, and 1 h assimilation intervals, 34 (0.91%), 70 (1.42%), and 76 (0.72%) observations were rejected, respectively. The detailed procedure prior to a single assimilation cycle is as follows:

(1) Observation Selection: The observation nearest to the analysis time within ±10 minutes is selected.

(2) Clear-sky Screening: Clear-sky MWR observations were screened using the AGRI-based CLM, with background-simulated cloud liquid water path is zero.

(3) First-Guess Departure Check: Observations with (O−B) values greater than 20 K are excluded.

(4) Whitelist Check: Remove observations from stations identified as unreliable or displaying abnormal behavior.

(5) Bias Correction: a machine learning bias correction scheme was applied (see Section 3.2).

(6) Relative Departure Check: Applied when the absolute value of the O−B exceeds three times the standard deviation of the observational error, further rejecting questionable data."

2. In Chapter 3, the O−B statistics are examined in detail. From the scatter plots,

it appears that the K-band O−B distribution may be bimodal for both radiometer types—a potential issue for 3D-Var, which typically assumes unimodal (Gaussian) errors. After showing the initial scatter plots of BT(sim) vs. BT(obs), the paper primarily focuses on bias and standard deviation. It would be very interesting to see if the bias correction addresses this bimodality. I encourage you to present histograms or PDFs of the O−B errors before and after the bias correction is applied. I would also encourage you to comment briefly on the skew and kurtosis of the distributions.

Response:

The bimodality in the O−B distribution has been addressed to some extent through bias correction. The probability density functions (PDFs) of the O−B have been included in the manuscript, along with the skewness and kurtosis of the distributions. The corresponding changes we have implemented are as follows:

"From the scatter plots in Fig. 3, the O−B distribution appears bimodal—an issue that may affect 3D-Var, which typically assumes the errors to be unimodal (Gaussian). Similar to Fig. 3, channel 1 (K-band) and channel 13 (V-band) of HATPRO, as well as channel 1 (K-band) and channel 14 (V-band) of MP3000A, are selected for detailed analysis (Fig. 6). Results for the remaining channels are presented in Figures B1 and B2. The biases for HATPRO channel 1, HATPRO channel 13, MP3000A channel 1, and MP3000A channel 14 are 1.24 K, 2.21 K, 3.00 K, and –0.64 K, respectively, with corresponding STDs of 3.38 K, 2.90 K, 3.89 K, and 3.08 K. The differences between the test set and the full dataset (shown in Fig. 3) are negligible, with a maximum bias difference of 0.10 K and a maximum STD difference of 0.08 K, highlighting the strong representativeness of the test set. From the PDF distributions of O–B, both instruments exhibit a positive bias in the K-band with a unimodal distribution. In contrast, a bimodal distribution is observed in the V-band: the second peak appears on the right for HATPRO and on the left for MP3000A. These results are consistent with the scattering patterns shown in Fig. 3. After the bias correction is applied, both the bias and STD are reduced, and the O–B

distribution becomes more sharply concentrated around zero, accompanied by an increase in kurtosis. For example, in channel 1 of MP3000A, the bias and STD decrease from 1.24 K and 3.38 K to 0.03 K and 1.44 K, respectively, while the kurtosis increases markedly from 1.53 to 9.44. It is also noteworthy that the bimodal distributions in the V-band for both instruments become unimodal after the correction. Meanwhile, the skewness decreases from 1.04 and 1.54 to 0.55 and 1.05, respectively, indicating a more symmetrical O–B distribution. These results demonstrate that the proposed bias correction scheme effectively reduces bias and STD, addresses bimodal distribution, and shifts the O–B distribution closer to a Gaussian shape.

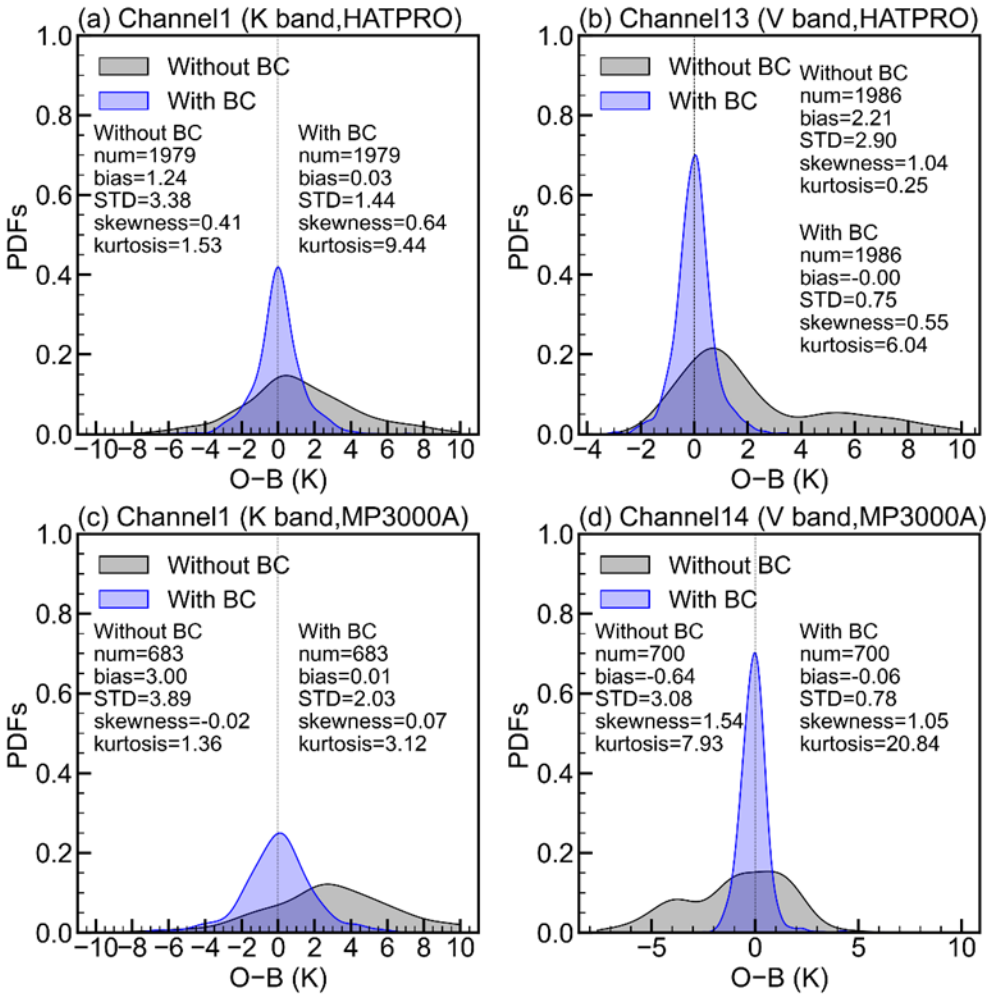

Figure 6: Probability density functions (PDFs) of the O−B distributions, based on a test set randomly selected from 30% of the three-month sample dataset collected from August to October 2023. The top and bottom rows correspond to the HATPRO and MP3000A sensors, respectively, while the left and right columns represent the K-band and V-band. Each panel displays the number of samples (num), the mean (bias),

standard deviation (STD), skewness, and kurtosis of the distributions."

3. The FSS improvements of rain rate seem impressive to say that the only change here is ground based radiometers. I suppose that the reason any impact is visible is only after 9 hours is because cloudy data is excluded. It may be interesting to show some statistics of the rainfall events in the 10-day period. Is one heavy precipitation event responsible for this improvement in forecast skill? Do you intend to repeat the experiment including cloudy data?

Response:

Thank you for your valuable suggestion. Upon re-checking the FSS calculation, we found that the improvement reported in the manuscript may not be as substantial as initially perceived.

The assimilation experiment period was selected to coincide with more frequent clear-sky conditions to maximize the availability of GMWR observations. During this period, precipitation was minimal, with no particularly intense precipitation events. Furthermore, during the phase of substantial FSS improvement (after 9 hours), observed precipitation was nearly absent. Although simulated precipitation exceeded observed amounts, it remained low overall, with false alarms being the primary source of error. The limited amounts of both observed and simulated precipitation led to a small sample size in the FSS calculation, resulting in the division of two small values. This mathematical characteristic amplified fluctuations in FSS scores, producing an apparent and pronounced improvement.

To mitigate the effect of small sample sizes in the FSS calculation described above, the thresholds for calculating FSS scores were lowered from 6–15 mm to 3–6 mm. The GMWR assimilation in this study, which is limited to clear-sky areas, does not utilize information from cloudy regions, resulting in reduced data usage—particularly during precipitation events. In future work, assimilation techniques for cloudy areas will be developed and focus on improving forecasts of heavy precipitation processes.

We revised the thresholds for FSS and provided clarification. The

corresponding text has been revised as follows:

“ The assimilation experiments were conducted during a period characterized by a higher frequency of clear-sky observations. Cloud cover and precipitation were limited throughout the 10-day period, resulting in the absence of frequent heavy rainfall events. Consequently, the FSS was calculated using small precipitation thresholds. In the CNTL experiment, the FSS for 3 h accumulated precipitation shows an initial decline followed by a subsequent increase with lead time, with relatively low FSS values observed around the 9 h forecast period. Moreover, the FSS generally decreases as the precipitation threshold increases. The time mean FSS values are 0.47, 0.45, 0.42, and 0.39 for thresholds of 3 mm, 4mm, 5mm, and 6 mm, respectively. Regarding the role of GMWR assimilation in precipitation forecasting, the results indicate that assimilating GMWR radiances enhances precipitation forecasts, with FSS differences increasing progressively at higher precipitation thresholds. Additionally, increasing assimilation frequency shows the potential to further enhance forecast performance. When assimilating GMWR data at a 1 h frequency, the time-averaged FSS improvements for 3 h accumulated precipitation are 0.02 (3.9 %) for the 3 mm threshold, 0.02 (4.7 %) for 4 mm threshold, 0.03 (7.3 %) for 5 mm threshold, and 0.04 (10.2 %) for 6 mm threshold precipitation. For 3 h accumulated precipitation with a threshold of 6 mm, the time-averaged FSS improvements are 0.01, 0.02, and 0.03 for GMWR_6H, GMWR_3H, and GMWR_1H, respectively. These findings are consistent with the above verification against radiosonde and surface station observations, suggesting that GMWR assimilation can improve forecasts and that higher-frequency assimilation leads to further enhancements.

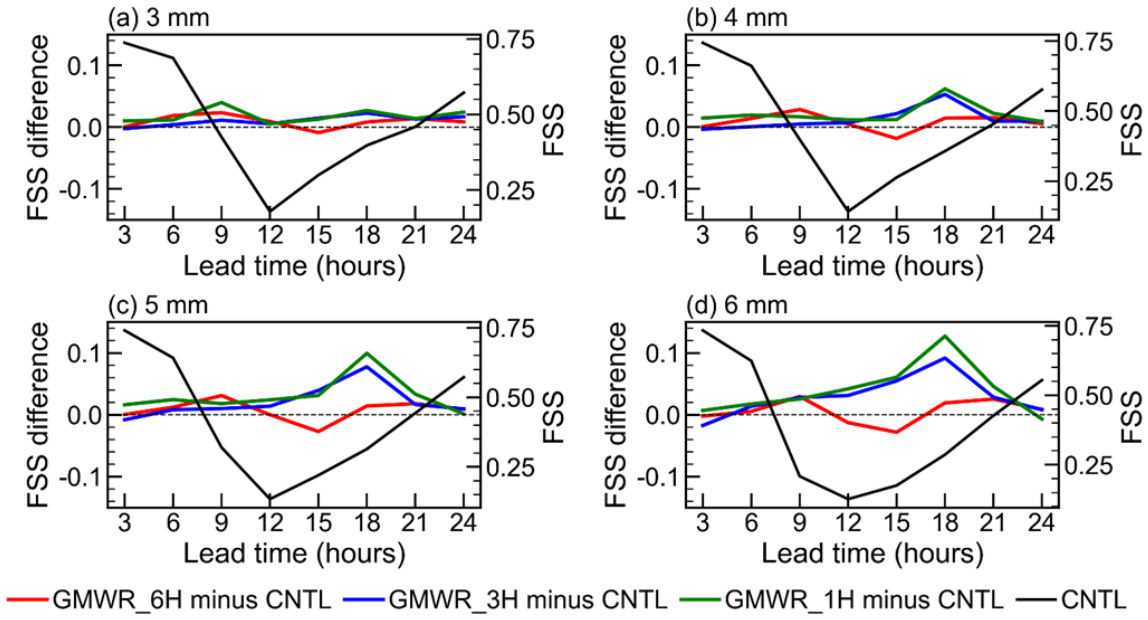

Figure 13: The time series of FSS (black line) for CNTL experiment and FSS differences (colored lines) between the assimilation experiments and the CNTL experiment. These experiments were conducted from 13 to 22 October 2023.The FSS was calculated for 3 h accumulated precipitation for thresholds of (a) 3 mm, (b) 4 mm, (c) 5 mm, and (d) 6 mm."

4. Could you add some information about how the background error covariance matrix was generated? Was any testing done to optimise this? Similarly, it would be helpful to know how you specified the observation-error covariances—for instance, the assumed observation-error standard deviations or correlations.

Response:

The background error covariance matrix was generated using the Generalized Background Error Covariance Matrix Model (GEN_BE v2.0) with the National Meteorological Center (NMC) method. Specifically, CV5, which is commonly used in WRFDA, was employed in this study, and the length scale was tuned using a single-observation experiment. Observation-error correlations are typically assumed to be zero in WRFDA, resulting in a diagonal observation-error covariance matrix. Observation errors were specified based on the standard deviation of O–B. More information about background error and observation error have been added in the manuscript as below.

"The static background error covariance for the variational experiments is estimated using the National Meteorological Center (NMC) method (Parrish and Derber, 1992), which uses the difference between WRF forecasts at lead times of 24 h and 12 h (T + 24 h minus T + 12 h) valid at the same time over a specified period. Control variables option 5 (CV5) is adopted for the background error covariance used in 3DVAR. CV5 is domain-dependent and therefore must be generated based on forecast or ensemble data over the same domain. It utilizes streamfunction, unbalanced velocity potential, unbalanced temperature, unbalanced surface pressure, and pseudo relative humidity. In this study, the background error covariance matrix was generated using the Generalized Background Error Covariance Matrix Model (GEN_BE v2.0) (Descombes et al., 2015) based on one month of WRF forecasts. Observation-error correlations are typically assumed to be zero in WRFDA, resulting in a diagonal observation-error covariance matrix. Observation errors were specified based on the standard deviation of O–B."

Specific Comments

5. Figure 1: Was the south-west China domain pre-defined in advance of the study? Were these conditions defined from the border of the computation domain or otherwise? Please justify.

Response:

"Similar to previous studies (Jiang et al., 2017; Nie and Sun, 2023), the target region of Southwest China in this study is defined as the area within the rectangular domain 22°–35°N, 93°–110°E (Fig. 1). This region encompasses the Hengduan Mountains, the Yunnan–Guizhou Plateau, and the Sichuan Basin, and is generally consistent with Chinese administrative divisions."

6. line 106: "However, MWR radiances are upward-looking microwave observations, which differ from the downward-looking observation of satellites." I find this sentence quite jarring. Before in the article you refer to "ground-based MWRs", so when you state that "MWR radiances are upward-looking" it seems

like you refer to the microwave radiometer instrument in general, not simply ground based microwave radiometers. I would change the sentence to refer simply to the platform (ground-based vs satellite) and not contrast microwave radiometer with satellite as this doesn't make sense.

Response:

To avoid confusion between ground-based microwave radiometers and microwave radiometers (MWRs) in general, the term "ground-based microwave radiometers (GMWRs)" is now used throughout this study. Accordingly, the sentence has been revised:

"RTTOV, a fast RTM, is widely used for assimilating satellite radiance data. However, GMWR radiances are upward-looking microwave observations, differing from the downward-looking measurements of satellite-borne microwave radiometers."

7. Line 134: Could you say (at some point in the paper, not necessarily here) when the three month training data for your bias correction algorithm was taken?

Response:

"To estimate the bias and develop a bias correction scheme for GMWR direct assimilation, a long-term experiment was conducted from August to October 2023, yielding a three-month sample dataset."

"The flowchart illustrating the training and evaluation process of the bias correction (BC) model is shown in Fig. 5b. The three-month sample dataset (described in Section 3.1) was randomly split into a training set (70 %) and a test set (30 %)."

8. Figure 2: Could you explain the plot axis label d Tans/ d ln P, I am not familiar with weighting functions of this type. Are the temperature/humidity increments representative of all data assimilation experiments? If not, please elaborate in the text. For plot c,d,e and f please label the colourbars and make the magnitude (1eX) more evident.

Response:

Transmittance τ varies monotonically with pressure p. The downwelling emission term can be expressed as:

$$L = \int_{\tau_t}^{1} B\,(T) d\tau = \int_{p_t}^{p_s} B\,(T) \frac{\partial \tau}{\partial lnp} dlnp$$

where L is the radiance at the ground, B is the Planck function, T is temperature, τ_t is the transmittance from the surface to the top of the atmosphere, p_t is the pressure at the top of the atmosphere, p_s is the surface pressure.

The weighting function, $\frac{\partial \tau}{\partial lnp}$, which is the derivative of transmittance with respect to height (pressure), describes the relative contribution of each atmospheric layer to the total radiation emitted to the surface (Cui et al. 2020; Thépaut, 2023). In figure 2, the axis label "d Tans/ d ln P" denote derivative of transmittance with respect to vertical coordinate, d(transmittance) / d(log(p)). The temperature and humidity increments are obtained from single-observation assimilation experiments, which are conducted to confirm that the GMWR direct assimilation module performs correctly. Although this experiment is not representative of all data assimilation experiments, they provide valuable insights into the characteristics of GMWR assimilation. Figure 2 and the corresponding text have been revised to improve clarity.

"According to Cui et al. (2020), WFs are calculated as the derivative of transmittance with respect to the natural logarithm of pressure.

It should also be noted that this experiment was conducted to verify the correct performance of the GMWR direct assimilation module and to provide valuable insights into the characteristics of GMWR assimilation. However, it is not representative of the subsequent multi-observation, multi-channel assimilation experiments.

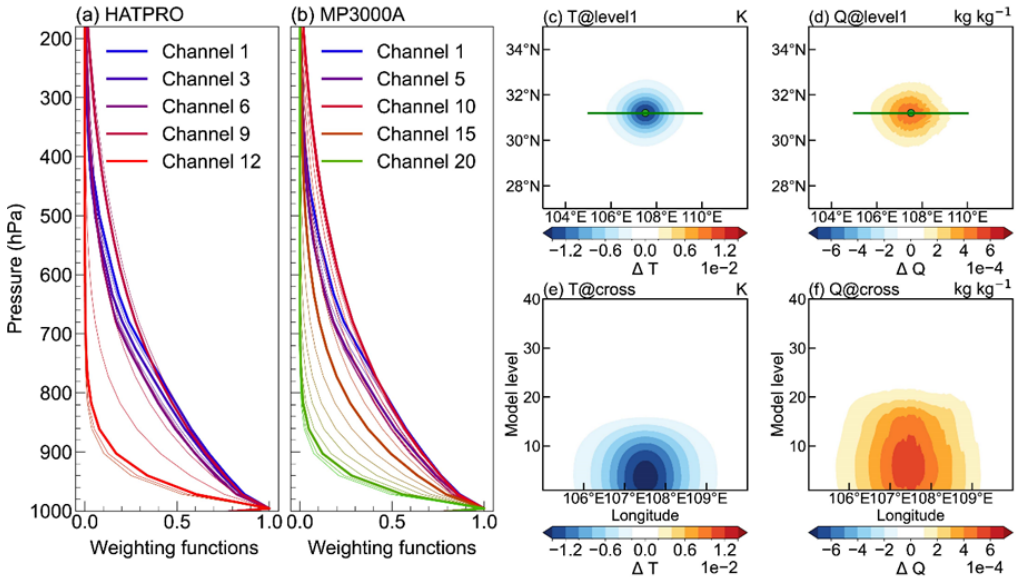

Figure 2: Normalized weighting functions of (a) HATPRO and (b) MP3000A calculated using the RTTOV-gb. The (c, d) horizontal and (e, f) vertical analysis increments for (c, e) temperature and (d, f) water vapor mixing ratio in single-observation assimilation experiment. The vertical increments are cross-sections along the green lines shown in the horizontal increments. The colorbar tick labels for temperature and water vapor mixing ratio are expressed in scientific notation as 1×10^{-2} and 1×10^{-4}, respectively."

Thépaut, J. N., 2003: Satellite data assimilation in numerical weather prediction: An overview. *Proc. ECMWF Seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean*, Reading, UK, ECMWF, 75–96.

Cui, X., Z. Yao, Z. Zhao, Y. Zhai, Z. Sun, W. Cheng, and C. Gu, 2020: Use of Double Channel Differences for Reducing the Surface Emissivity Dependence of Microwave Atmospheric Temperature and Humidity Retrievals. *Earth and Space Science*, **7**, e2019EA000854, https://doi.org/10.1029/2019EA000854.

9. Line 160: "For HATPRO, more than 6,000 samples are analyzed. The O−B biases are 1.25 K for the K band (channel 1) and 2.14 K for the V band (channel 13)". Are these statistics the average for the whole band or that particular channel? If they are for the whole band, make this explicit on Figure 3. If not, then these results should not be generalised to the whole band.

Response:

These statistics are for particular channel and the corresponding text have been revised.

10. Figure 3: Why were these particular channels selected. Could you comment in the text about whether these plots are representative of the whole band? It would also be nice if you could include the same plots for the other channels in the appendix.

Response:

These plots are not representative of the whole band. The same plots for the other channels are added in the appendix. And the corresponding text have been revised as following:

"A comparative scatterplot analysis of observed and simulated brightness temperatures was conducted. For most channels, the scatter points are closely aligned along the diagonal and exhibit high correlation coefficients, indicating strong agreement between the simulations and observations. However, the scatter for some channels forms two distinct clusters. To further investigate, representative channels from the K-band (water vapor absorption lines) and the V-band (temperature-sensitive oxygen absorption lines) were selected. Figure 3 presents scatterplots for channel 1 (K-band) and channel 13 (V-band) of HATPRO, and channel 1 (K-band) and channel 14 (V-band) of MP3000A. Results for the remaining channels are shown in Figures A1 and A2. For HATPRO, more than 6,000 samples are analyzed for channels 1 and 13. The O−B biases are 1.25 K for channel 1 and 2.14 K for channel 13, with standard deviations (STD) of 3.35 K and 2.82 K, respectively. Additionally, the scatter distribution for channel 13 is not centered, showing a cluster shifted to the right of the diagonal (Fig. 3b). For MP3000A, more than 2,000 samples are analyzed for channels 1 and 14, with O−B biases of 3.06 K for channel 1 and −0.54 K for channel 14. The O−B STDs are 3.94 K and 3.08 K, respectively. Similar to the results for HATPRO channel 13 (V-band), the scatter for MP3000A channel 14 (V-band) also shows a cluster offset from the diagonal, but to the left (Fig. 3d).

Based on these results, significant O−B biases are detected in GMWR observations, with their characteristics varying across different sensors and channels. However, the correlation coefficients between observed and simulated brightness temperatures are high, at least 0.95, suggesting that these biases can be effectively corrected.

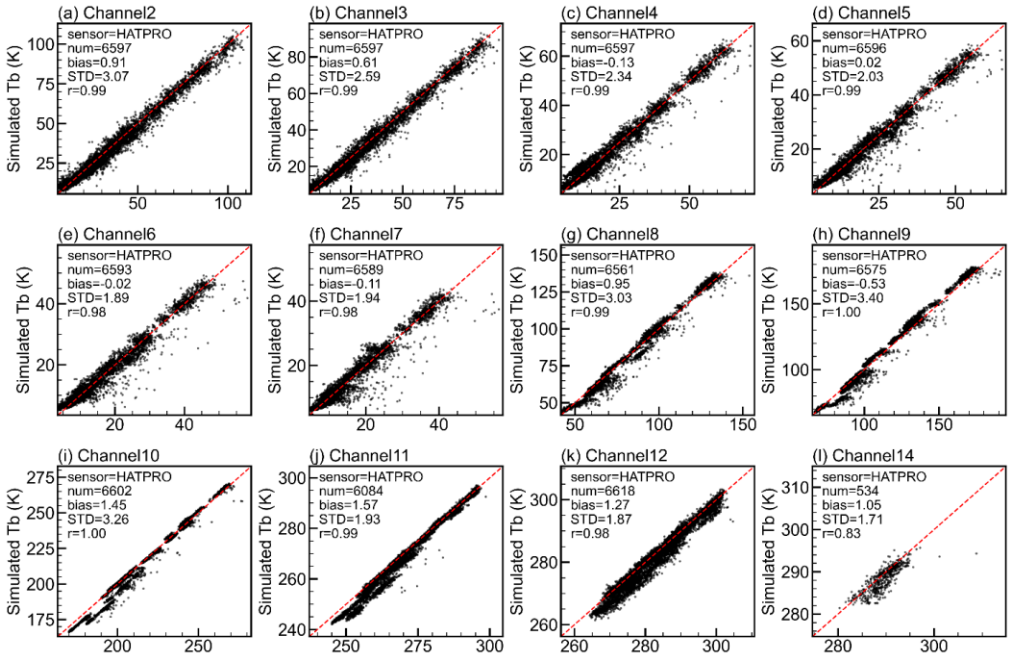

Figure A1: Scatter plot of observed brightness temperature (Tb) versus simulated Tb for HATPRO. Same as Fig. 3 but for additional channels.

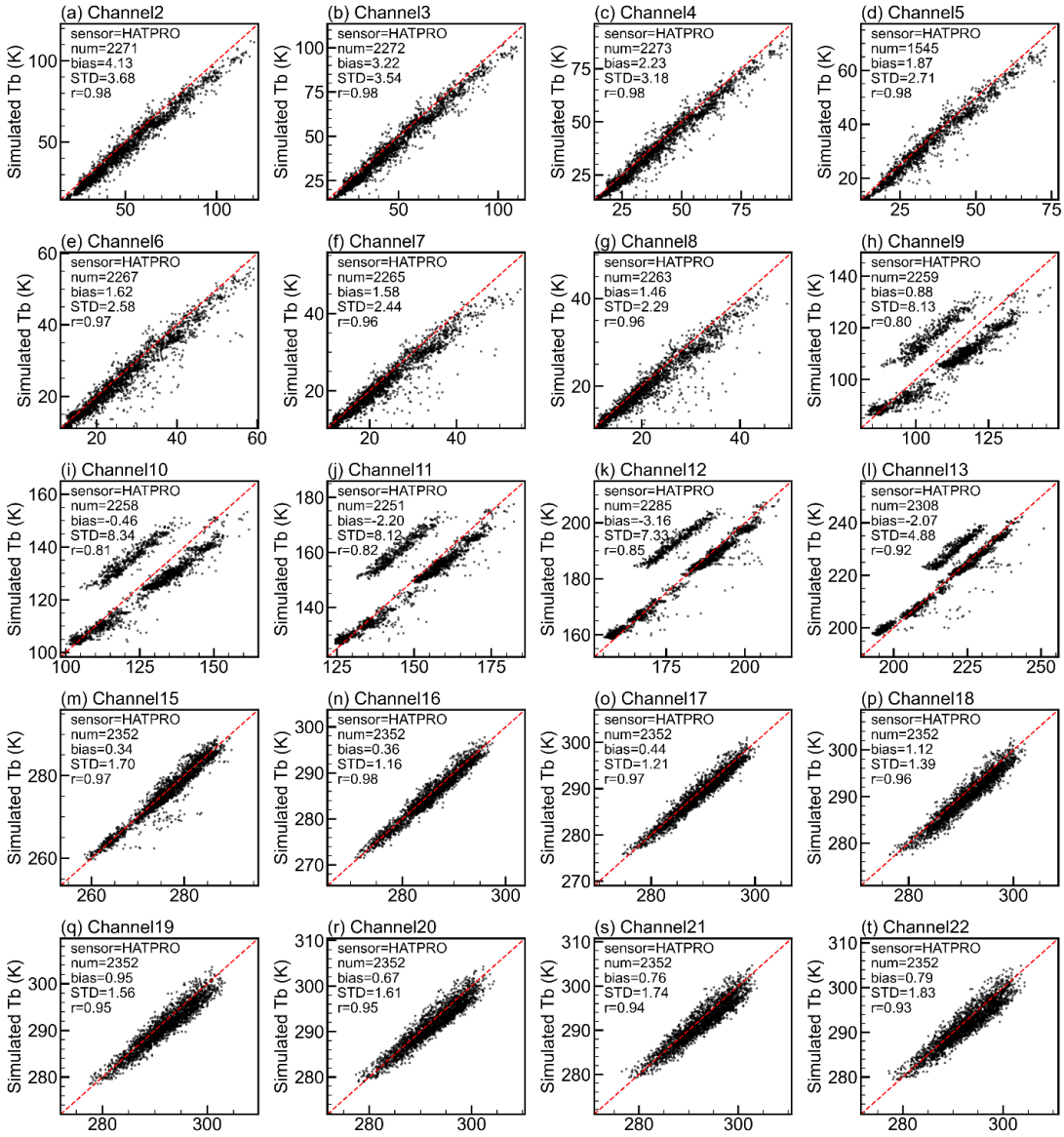

Figure A2: Scatter plot of observed brightness temperature (Tb) versus simulated Tb for MP3000A. Same as Fig. 3 but for additional channels."

11. With respect to the apparent two clusters in the V band, were the (clustered) offset values consistently from the same radiometers. Do the offsets also correspond to a particular time period (e.g. before vs after calibration) or weather condition (after rain, in direct sunlight)?

Response:

The apparent two clusters in the V band correspond to a cluster shifted to the right of the diagonal for channel 13 in HATPRO, and a cluster shifted to the left of the diagonal for channel 14 in MP3000A. These offsets correspond to

positive and negative O–B deviations, respectively. According to Figure 4, the positive O–B values originate from GMWRs at stations 56312, 56029, and 55664, while the negative O–B values are primarily from station 57461. The offset was consistently present and did not correspond to a particular time period.

12. line 182: “Regarding the correlation coefficients between observed and simulated brightness temperatures, the overall values are high but slightly lower for channels 4 to 9.” What is high? Please be more precise.

Response:

This sentence was revised as following.

“The correlation coefficients between observed and simulated brightness temperatures are high across all channels (typically above 0.90), although they are slightly lower for channels 4 to 9.”

13. Figure 4: Please add colourbar axis label and units for all sub plots. For HATPRO channel 14, values are mainly white, but this colour is not included in the colourbar- does that mean that the values are out of range, missing or otherwise?

Response:

The colorbar axis label and units have been added. At some stations, HATPRO did not observe channel 14, resulting in missing data that originally appeared as white areas in the figure and were not represented in the colorbar. These areas have been changed to grey, with an explanatory note added accordingly.

“

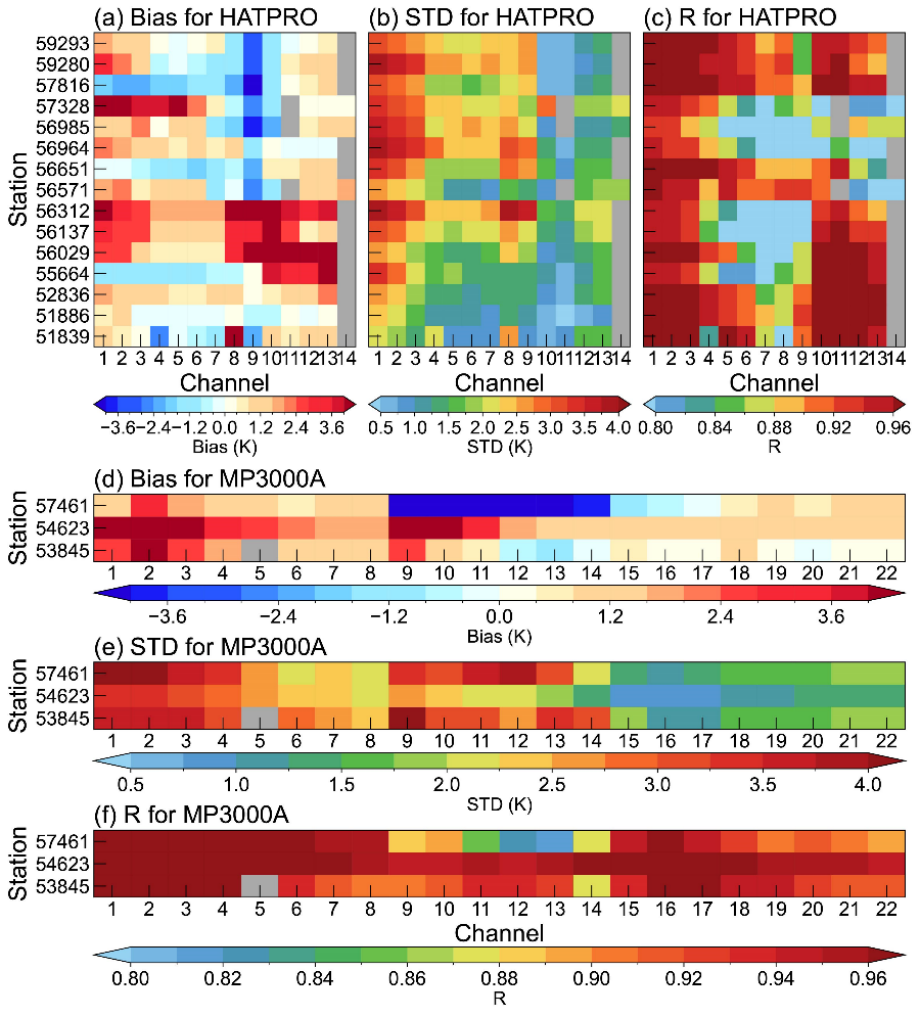

Figure 4: Statistics at each station based on samples collected from August to October 2023. O−B (a) bias and (b) standard deviations (STD) for HATPRO; (c) correlation coefficient (r) between observed and simulated brightness temperatures for HATPRO; (d–f) same as (a–c) but for MP3000A. Some stations did not provide observations for specific channels; the corresponding missing data are displayed in grey in the figure."

14. Line 215: "These hyperparameters were tuned using GridSearchCV"

Please properly reference scikit learn (or the library in question) when refering to this.

Response:

The reference for scikit-learn has been added.

15. Line 235: “. Furthermore, some individual channels, such as channel 10, display bimodal distributions.”

As stated above, it would be interesting to see these plots.

Response:

These plots have been added in appendix.

“For most channels, the PDFs exhibit a unimodal pattern, with peak positions deviating from zero, indicating that the O−B values are biased. For some channels, the distributions are multimodal, characterized by a secondary peak superimposed on the primary one. Although these issues are present in the original distributions, after bias correction, the PDFs approximate an unbiased distribution, and the secondary peaks are effectively suppressed, demonstrating the effectiveness of the correction. From the scatter plots in Fig. 3, the O−B distribution appears bimodal—an issue that may affect 3D-Var, which typically assumes the errors to be unimodal (Gaussian). Similar to Fig. 3, channel 1 (K-band) and channel 13 (V-band) of HATPRO, as well as channel 1 (K-band) and channel 14 (V-band) of MP3000A, are selected for detailed analysis (Fig. 6). Results for the remaining channels are presented in Figures B1 and B2.

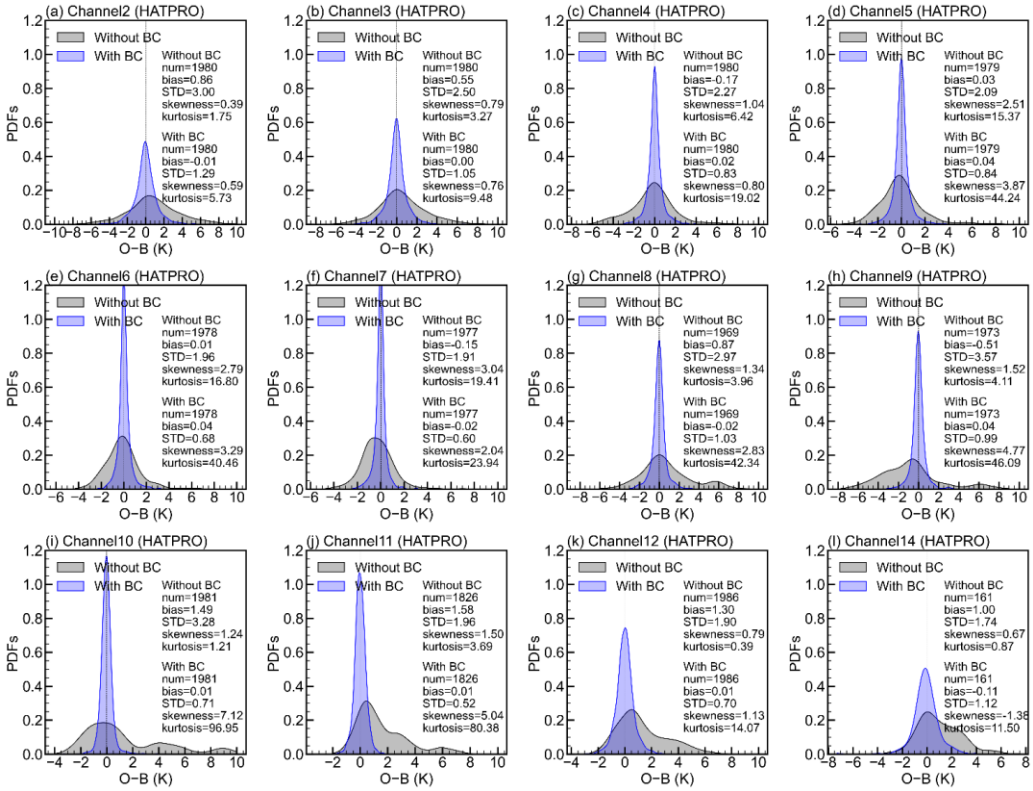

Figure B1: Probability density functions (PDFs) of the O−B distributions for

HATPRO. Same as Fig. 6 but for additional channels.

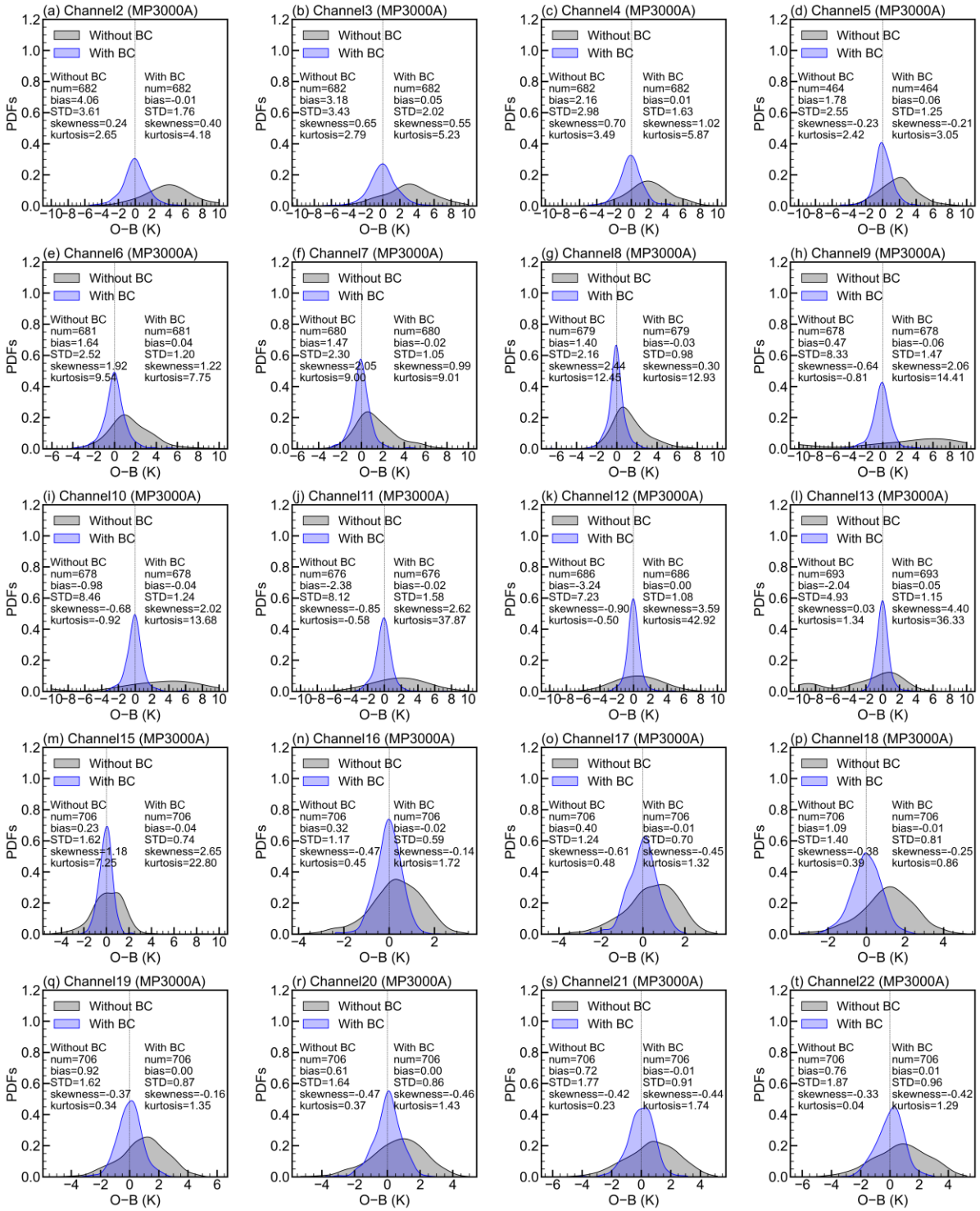

Figure B2: Probability density functions (PDFs) of the O−B distributions for MP3000A. Same as Fig. 6 but for additional channels."

16. Figure 6: temperatrure and water vapour bands are labelled the wrong way around. I'm not sure if a continuous line plot is the most appropriate for the unlinked variables (bottom plot).

Response:

Corrected

17. Line 415: "The RTTOV-gb coefficient files are trained on global profiles and are not tailored to the plateau region; consequently, their vertical coordinates extend up to 1050 hPa, while surface pressure in the plateau region typically exceeds 700 hPa."

Could you please clarify why having the RTTOV-gb coefficient files only extend down to 1050 hPa creates an issue at high-altitude stations? Conceptually, if the sensor is located in a region where the surface pressure is around 700 hPa, that site is "above" the portion of the atmosphere from 1050–700 hPa (which would presumably be below ground in the plateau setting). How does this mismatch in the vertical extent of the RTTOV-gb coefficient files lead to biases for sensors effectively well above 1050 hPa? In other words, why does the "upper pressure limit" become problematic when the actual surface is located at a pressure lower (i.e., a higher altitude) than the coefficient file's assumed maximum pressure?

Response:

RTTOV-gb is trained on global atmospheric profile datasets, where profiles are interpolated onto 101 pressure levels to derive regression coefficients between optical depths and predictors, thereby producing the coefficient files. These global profiles extend to high surface pressures, up to approximately 1050 hPa, enabling the vertical coordinate of the coefficient files to reach this level. However, in plateau regions, surface pressures typically fall below 700 hPa, indicating a substantial climatic difference between the training profiles and those required for simulations in high-altitude areas. This discrepancy may introduce simulation biases when RTTOV-gb is applied over plateaus. Indeed, in satellite data assimilation, efforts have already been made to enhancing fast radiative transfer model using local training profiles (Di et al. 2018). To clarify,

our intention is not to highlight a mismatch in vertical pressure coordinates, but to suggest that the applicability of globally trained coefficients in plateau regions may be limited due to the pronounced climatic divergence. This sentence has been revised accordingly to avoid misunderstanding.

"The RTTOV-gb coefficients are based on global atmospheric profiles, which may differ significantly from the climatic conditions of plateau regions, potentially affecting simulation accuracy."

18. Line 435: "It is noted that satellite-based microwave radiometers are primarily sensitive to the middle and upper atmosphere..."

This isn't technically correct. In atmospheric science, the middle atmosphere generally refers to the stratosphere + mesosphere, i.e. from the tropopause (roughly 8–17 km, depending on latitude) up to ~80–85 km. The upper atmosphere is often taken to be the thermosphere and above. Many operational satellite microwave sensors (e.g., AMSU, MHS, ATMS) retrieve temperature and humidity by sounding channels peaked in the troposphere and lower stratosphere, though they can extend somewhat upward. Their highest sensitivity is thus often in the mid- to upper troposphere, not solely above it.

Response:

This sentence has been revised as following:

"It is noted that GMWRs exhibit higher sensitivity and provide more valuable observations of the lower troposphere and planetary boundary layer compared to satellite-based microwave radiometers (Shi et al., 2023)."