Response to Reviewer

Dear Reviewer,

We would like to express our sincere gratitude for your detailed and insightful review of our revised manuscript. Your rigorous mathematical and statistical perspectives have been immensely helpful in refining our presentation and ensuring the scientific soundness of our retrieval framework. We fully agree with your comments regarding the statistical properties of the leave-one-out cross-validation (LOOCV) method and the operational implementation of the bias correction.

In this revised manuscript, we have addressed all your remaining concerns by adding new quantitative figures (Figures 10 and 11), clarifying the statistical metrics, and explicitly defining our operational static bias correction strategy. Please find our detailed, point-by-point responses below.

Reviewer Comment 1: *Line 330: 'These constraints are applied to all neighbouring retrieval levels and act mainly as smoothness controls under limited observational constraint'. Please see the later comment regarding this, as from what you present, it doesn't seem successful as a smoothness control.*

Response: Thank you for this penetrating comment. We apologize for the confusion caused by our previous presentation of the representative case (originally Figure 10, now Figure 12). We would like to clarify the physical rationale behind the selection of the adjacent-layer thresholds (8 K for temperature, 60% for relative humidity).

These values are purposefully selected to act as a **conservative physical baseline** rather than a strict smoothing operator. In dynamic marine boundary layers (such as Jiaozhou Bay), sharp vertical gradients like strong low-level inversions and abrupt humidity transitions (often associated with atmospheric ducting) are physically genuine. An overly stringent continuity constraint would artificially smooth out these real meteorological structures. Therefore, the constraint is designed only to filter out mathematically valid but physically impossible step-like numerical discontinuities during the NSGA-II optimization process, ensuring the algorithm retains full freedom to resolve genuine meteorological gradients.

Changes made in the manuscript: We have enriched the discussion in Section 3.3.3 to make this underlying physical logic explicit.

- **Lines 328–332:** Added explanation: *"The relatively large threshold values ensure that the constraint does not artificially smooth out real atmospheric features such as temperature inversions or sharp humidity gradients. These limits are introduced primarily to maintain basic physically reasonable profile continuity rather than to impose a strict thermodynamic lapse-rate condition."*

Reviewer Comment 2: *From line 358-387: ... As the 'leave one out' method uses the information of the biases for all other profiles to calculate the bias of the profile being corrected, when the statistics of the bias corrected plots is performed, because each profile is corrected by the 'leave one out' mean bias, averaging the corrected residuals over all 38 cases is identically zero at every level by construction. Indeed, this is what we see in plots c) and d)... Because of this, the plots of the average profile comparison are no longer informative. The RMSE at each level, however, would be the informative metric. Also, using the method you show, the standard deviation cannot decrease, but should increase by 1/37, contrary to what you state.*

Response: We sincerely appreciate your astute mathematical insight. You are entirely correct. By mathematical construction, the mean of the corrected residuals across all 38 folds is identically zero at every level. This is indeed a fundamental property of the LOOCV mean bias correction approach. Our original intention for Figure 9 was not to demonstrate a non-zero residual to prove correction effectiveness, but rather to visualize the *magnitude and vertical structure* of the systematic bias that existed prior to correction.

Furthermore, you are absolutely correct regarding the standard deviation. An increase in the standard deviation of the RMSE across the folds (from 0.52 K to 0.62 K for temperature) indicates a slightly wider statistical spread, which mathematically contradicts our previous flawed terminology of "improved consistency." Our intention was to highlight the robustness of the *overall* error reduction. To provide truly informative metrics, we have added a new figure (Figure 11) showing the RMSE distributions (boxplots) before and after correction, which clearly quantifies the substantial reduction in the mean RMSE.

Changes made in the manuscript:

Lines 503–510: Added explicit acknowledgement of the LOOCV mathematical property: *"We intentionally omit plotting the mean residual of the corrected retrievals here because, by*

mathematical construction of the LOOCV mean bias correction procedure, the mean residual of the corrected retrievals across all 38 folds is identically zero at every height level. This is an inherent property of the leave-one-out approach for mean bias correction, not a limitation of the method. For this reason, the mean residual cannot be used as evidence of correction effectiveness.Instead, the effectiveness of the bias correction is properly quantified by the reduction in root-mean-square error (RMSE), which measures the magnitude of the retrieval errors regardless of their sign. As shown in the newly added Figure 11, the overall RMSE is significantly reduced after correction, from 4.11 K to 2.13 K for temperature and from 24.09% to 21.42% for relative humidity."

- **Lines 542–546:** Added **Figure 11** and corrected the statistical description: *"Although the standard deviation of the RMSE across all folds increases slightly from 0.52 K to 0.62 K, the mean RMSE decreases significantly from 4.11 K before correction to 2.13 K after correction... Importantly, 34 out of 38 folds (89.5%) show improved retrieval accuracy for temperature after correction..."*

Reviewer Comment 3: *My other concern is that the whole bias correction method is no longer static- there are 38 different bias correction methods essentially, and it is not clear which one should be used if the algorithm were to be used operationally. I also wonder about how much variation there is between the different bias corrections. It would also be informative to plot the bias correction profile. With this in mind, I would ask that the authors a) explicitly say in the abstract that the statistics are post bias correction and also state the pre bias correction statistics b) either provide an explanation of why it is incorrect to say that the bias should equal zero through definition post correction, or replace these plots with plots of the RMSE before and after bias correction.*

Response: We completely agree with your perspective regarding operational deployment. To address this, we have plotted all 38 individual bias correction profiles derived from the LOOCV procedure in a new figure (**Figure 10**). As the figure demonstrates, the 38 individual profiles are remarkably consistent (with a standard deviation mostly < 0.3 K for temperature), which empirically justifies aggregating them.

For operational runtime applications, we recommend using the **arithmetic mean of all 38 samples as a single, static correction profile**. We have verified that this single static model

achieves nearly identical overall performance (RMSE difference < 0.02 K) compared to the dynamic LOOCV approach. However, we candidly acknowledge that this static model relies on localized campaign data and currently lacks blind validation against completely independent external datasets, a limitation we have now explicitly noted for future work.

Changes made in the manuscript:

- **Abstract (Lines 23–25):** We have updated the abstract to explicitly state the pre- and post-correction statistics: *"Under sparse-data conditions, the temperature RMSE is 4.11 K before systematic bias correction and 2.13 K after correction, while the relative-humidity RMSE is 24.09% before correction and 21.42% after correction."*
- **Lines 394–402 & 528–532:** Added **Figure 10** and detailed the operational static profile approach: *"For operational deployment over the study area, a single static bias correction profile is recommended. This static profile is calculated as the arithmetic mean of all 38 individual bias profiles..."*
- **Lines 651–656 (Discussion):** Added a rigorous disclaimer regarding independent validation: *"First, broader validation is required using marine observational datasets with extended temporal scales and environmental diversity. Specifically, while the single static bias correction profile proposed for operational use demonstrates highly stable performance matching the LOOCV results within our campaign, it has not yet been verified against a completely independent, external dataset. Future efforts will explicitly focus on deploying this static model under distinct offshore and open-ocean environments to rigorously evaluate its cross-regional transferability and eliminate any potential risk of local over-fitting."*

Reviewer Comment 4: *Figure 10: It seems from this plot that the continuity constraint adds more oscillations and increases the RMSE of the profile compared to the profile without the constraints. What was the logic in keeping it? I am also unsure from the continuity constraint and the level-dependent admissible intervals, how this could have increased the oscillations. Does the climatology profile also have large oscillations?*

Response: We apologize for the confusion caused by Figure 10 (now Figure 12 in the revised manuscript). We must clarify that this figure depicts a **representative special case**, not the average performance across the dataset. In this specific instance, the unconstrained solution

happened to fit the collocated radiosonde profile more closely (yielding a lower RMSE). However, across all 38 matched cases, the constrained retrieval achieves a slightly lower overall mean RMSE (2.13 K) compared to the unconstrained retrieval (2.27 K).

Regarding the "oscillations," our intention was to point out the unphysical layer-to-layer “jumps” (step-like discontinuities) seen in the unconstrained dashed line. The constrained profile actually maintains more gradual, physically reasonable vertical transitions. The climatology profile derived from our local radiosonde data does *not* exhibit large unnatural oscillations, and the continuity constraint is precisely what forces the NSGA-II algorithm to maintain this fundamental physical smoothness under limited observational constraints.

Changes made in the manuscript:

- **Lines 596–606 (including Figure 12 caption):** We have explicitly noted this is a special case and provided the global statistics to justify the constraint: *“Note: In this particular case, the unconstrained retrieval achieves a lower RMSE with respect to the collocated radiosonde. However, the unconstrained profile exhibits larger non-physical layer-to-layer jumps... Statistically, across all 38 matched cases, the constrained retrieval achieves a slightly lower overall RMSE (2.13 K for temperature and 21.42% for relative humidity) compared to the unconstrained retrieval (2.27 K... and 22.89%). More importantly, the continuity constraint eliminates unphysical discontinuities in 100% of the test cases...”*