Dear Reviewer #2,

We would like to express our deepest gratitude for your meticulous review and highly constructive comments on our manuscript. Your sharp insights, particularly regarding the depth of our literature review, the scientific rigor of our dataset splitting, the theoretical underpinning of retrieval performance under extreme conditions, and the methodology of our interpretability analysis, have been invaluable.

We have carefully reflected on each of your critiques and fully recognize that addressing them is crucial for elevating the scientific quality of our work. Following your guidance, we have thoroughly revised the manuscript, clarified our methodologies, and incorporated the highly relevant literature you recommended. Below, we provide a point-by-point response detailing how we have addressed each of your concerns.

Point-by-Point Responses

1) The retrieval of temperature and humidity profiles from MW and IR data is very large. The literature review is too sparse and this reflects on the paper maturity. More references are required. NN retrieval dates back from the end of the 90s.

Response:

We sincerely thank you for pointing out this significant omission. We completely agree that our original manuscript failed to adequately cover the rich history of neural networks in atmospheric retrievals and their mature applications in adjacent fields. Following your excellent suggestion, we have significantly expanded the Introduction section. We incorporated pioneering works from the early 2000s (e.g., Aires et al., 2001; Blackwell, 2005) to acknowledge the historical context of NN retrievals. Most importantly, we included a discussion on the successful implementation of CNNs in the precipitation community, specifically citing the highly relevant GPROF-NN paper (Pfreundschuh et al., 2022) recommended by you and the Editor.

Location in the revised manuscript: Lines 63-75.

2) The analysis is performed only in one month of data. This is very limited and not sufficient to compare to other retrievals. You mentioned that training, validation and testing has been chosen randomly, this means that there are no true independency between them because two neighboring pixels are almost the same. This is not standard practice.

Response:

We apologize for the confusing terminology used in our original draft regarding the dataset splitting. We completely agree that spatially random splitting would lead to severe data leakage due to the strong spatial auto-correlation of atmospheric fields. We would like to formally clarify that our dataset was NOT split randomly in space; rather, it was split strictly chronologically (by time). We used the first 80% of the timeline for training and validation, and the entirely unseen final 20% strictly for testing, guaranteeing true temporal independence. Any mention of “shuffle” in previous drafts referred solely to the standard shuffling of batches during the training loop.

Regarding the use of one month of data, we acknowledge that a longer period would be ideal. However, due to the extraction of 5×5 spatial patches for the CNN input, the dataset size expanded exponentially, resulting in over 15 million samples for a single month. Processing a larger multi-month dataset significantly exceeded our current hardware memory and computational constraints. Furthermore, because our proposed AR-CNN model features a relatively lightweight

architecture with a moderate number of parameters, the massive sample size generated from this one-month period is mathematically sufficient to effectively train the model, ensure convergence, and prevent overfitting. We have corrected the wording in the manuscript to make the chronological splitting unambiguously clear.

3) Some figures are not optimal. Figure 1 is useless, we can understand a 5x5 input window for a target as atmospheric profile. Figure 2 os tpp sùamm and almost cannot be read. I would like more comments on the architecture itself. Figure 4 is not necessary. If you want to comment on extremes, better graph can illustrate.

Response:

We sincerely thank you for these highly practical suggestions regarding the presentation of our work. We have carefully reviewed all the figures and made enhancements to improve their overall readability and resolution. Regarding Figure 1, we opted to retain it as a schematic visual aid, as we believe it helps readers from broader interdisciplinary backgrounds quickly grasp the spatial input design. Similarly, Figure 4 is retained as it serves as a preliminary, intuitive illustration of extreme cases, setting the stage for the deeper theoretical discussions that follow. We have also expanded the textual description of the network architecture as you suggested to provide more clarity.

4) Section 3.2 us useless

Response:

We deeply appreciate your perspective on this section. After careful consideration, we decided to retain Section 3.2, as we believe it provides an essential contextual bridge for readers to fully understand the physical logic behind the subsequent interpretability analysis. However, we have reviewed and refined the text to ensure it remains focused and smoothly integrated with the core message of the manuscript.

5) Comments about the extreme cases... you need to read the literature. For example the boucher et al recent paper about extremes, CNN and dampening effect.

Response:

We are extremely grateful for your recommendation of the works by Boucher et al. Reading these papers provided us with the exact theoretical foundation we were missing. We now understand that Deep Learning models, when optimized with standard loss functions like Mean Squared Error (MSE), mathematically tend to output the conditional expectation of the target distribution. As Boucher and Aires (2023) and Boucher (2024) demonstrated, this inevitably leads to a "dampening effect" where models pull predictions toward the statistical mean. We have integrated this profound theoretical insight into our discussion to scientifically explain the underestimation of extreme values in our results.

Location in the revised manuscript: Section 4.2 (Lines 300-307).

6) you don't compare your results with more simple, and older techniques like: linear regression, MLP, etc... I am quite sure that a MLP will give similar results than what you obtain with this complex model.

Response:

Thank you for raising this critical point. In fact, during our preliminary experiments, we thoroughly compared our AR-CNN against basic MLPs and standard CNNs. To maintain the manuscript's primary focus on physical interpretability, we placed these benchmark results in Appendix A (Table A1). As shown in Table A1, the AR-CNN (Test RMSE: 1.24 K for Temperature) significantly outperforms the simple MLP (Test RMSE: 1.77 K). We have added explicit text in the methodology section directing readers to this baseline comparison.
Location in the revised manuscript: Lines 147-149 and Table A1 in the Appendix.

7) The way to analyse the interpretation of the model is not correct. For instance, if you perturb one input channl by 1K random noise, it is not correct. Channels are highly correlated, and introducing incoherencies in inputs is not physical. If you want to measure the information content of the channels when using a MLP, you can add hierarchically the channels, or suppress them, to see the impact on the results.

Response:

We deeply appreciate your rigorous scrutiny and would like to clarify our approach. First, we did not apply "random noise" to the inputs. Instead, we applied a constant, deliberate impulse of +1K and -1K to specific channels. Mathematically, this is equivalent to calculating the physical Jacobian matrix ($\partial T/\partial TB$) using the finite difference method, which is a standard procedure in physical retrievals to assess specific spectral sensitivities, rather than breaking inter-channel correlations.

Secondly, we completely agree with your excellent suggestion that "suppressing channels" is the most robust way to evaluate information content. In fact, this is exactly what we conducted and presented in our Channel Ablation Experiments. We systematically removed individual channels and directly quantified the resulting deterioration in Bias and RMSE (via the Bias Difference Heatmap). We have strengthened the description of this ablation methodology in the text to prevent future misunderstandings.

8) Figure 5 right, indicate the channels frequency. Cannot identify the channels with the colours, need to change that. Left part, this is not reliable I think for the reasons I mentioned.

Response:

We have updated Figure 5 (right) to explicitly label the specific frequency for each channel on the Jacobian plot, resolving the identification issue. Regarding the left part of the figure, as clarified in our detailed response to Comment 7, the perturbation applied is a standard finite difference approximation for calculating physical Jacobians, not random noise. Therefore, the sensitivities derived remain physically reliable. We have updated the figure captions to make this methodology clearer.

9) Figures A1 and A2 cannot show differences considering the spread of the colorbar. They are not useful. You should rather do a bias and RMS map of the errors.

Response:

We completely agree that spatial Bias and RMSE maps are excellent tools for rigorous error evaluation, and we highly appreciate this constructive suggestion. However, Figures A1 and A2 are primarily intended to serve as illustrative snapshots of the global spatial distribution to intuitively demonstrate the overall consistency of the retrieval, rather than to conduct a minute

spatial error analysis. Therefore, we have retained the current format for these specific schematic figures, but we definitely plan to adopt your valuable suggestion of using detailed Bias and RMSE maps in our upcoming extensive global validation studies.