

Response to Referee 1

General Comments

In this manuscript, the authors propose a machine learning retrieval method (AROMA) using a simple Multi-Layer Perceptron model to retrieve thermodynamic atmospheric profiles (pressure, temperature, and specific humidity) from GNSS-RO observations, primarily from COSMIC-2 and Spire. The retrievals are trained against CDAAC 1D-Var data and validated using three sources: the CDAAC products themselves (test set evaluation), reanalysis data (ERA5), and independent radiosonde observations (RAOB). While the general idea of using machine learning for RO retrievals is not new, the authors claim novelty through the exclusive use of CDAAC 1D-Var products as training targets, aiming to replicate retrievals in a more computationally efficient or independent way. However, there are several concerns regarding the methodology, terminology, completeness of the analysis, and scientific contribution.

Thank you very much for your review and feedback, as well as all the valuable comments and suggestions on how to improve the manuscript. We have tried to extend and strengthen the performance analysis carried out based on the suggestions made.

While the manuscript is generally well-structured and written in acceptable English, the scientific novelty is limited. The use of MLPs for GNSS-RO profile retrieval has already been explored in past studies, e.g. Lasota (2021), Hooda et al. (2023). The main difference here is the use of CDAAC 1D-Var as target values. However, the authors do not convincingly show that this leads to improved performance or independence from NWP models. On the contrary, the CDAAC retrievals themselves rely on ECMWF background fields, and ERA5, used for external validation, also assimilates many of the same observations. This compromises the claimed “independence” of the proposed method.

Although there are already studies available which use MLP architectures, we don't think that our work is not novel. There is only a small number of studies on similar approaches available, and our approach is substantially different from these in terms of target definition and spatial data coverage (through the inclusion of Spire data). Although it is true that training on operational 1D-Var results is not fully independent (which we never claimed in the manuscript), the level of independence is substantially higher than for approaches taken by other studies, e.g. using ERA5 data as targets. It is also practically impossible to show that our results are better than those of former studies, since the setup (target definition) is different. Therefore, it was not necessarily the goal to outperform former studies in terms of statistical error metrics but rather investigate how well operational 1D-Var results can be matched with our setup, without the need to include external meteorological data after training.

We will try to communicate these objectives and complications better in the revised manuscript.

The study focuses heavily on bias and standard deviation, while omitting or underusing the more informative RMSE in figures and discussion. RMSE is widely accepted as a more

representative metric in geophysical retrieval evaluation. Moreover, the use of inconsistent units (% , K, hPa) across tables and figures makes comparisons difficult. A more uniform and clearly explained presentation of metrics is required.

Bias and standard deviation are commonly used in RO studies (and other geophysical) research, since they allow for the separation and interpretation of both systematic and random errors. RMSE is basically a combination of these two measures. However, we are aware that RMSE is frequently used in ML studies and thus have adopted it as an additional metric in all plots in the revised manuscript. We will add a subsection introducing all metrics used, in case readers are not familiar with them.

The authors overuse of vague qualitative language. Terms such as "small", "good agreement", "slightly degraded", "negligible" appear throughout the text without being substantiated by numerical values or objective criteria. For a technically oriented audience, these expressions are insufficient. All such statements should be backed up with concrete values or quantifiable thresholds.

Claims such as "almost identical accuracy", "very satisfactory results", or "similar performance to CDAAC" are often too strong given the evidence provided. In particular, the AROMA method shows noticeably worse results for specific humidity, which is glossed over in the text. Additionally, the comparison to other studies (e.g., Lasota, 2021) lacks numerical transparency and fair contextualization (e.g., different latitude bands).

We will use more precise language in the revised manuscript. In terms of comparison with similar studies, we will extend and strengthen the discussion section. However, as mentioned earlier, our results are not comparable 1:1 to those of e.g. Lasota (2021). Furthermore, it is not the goal of this work to outperform any other study but to provide an alternative approach, which has a higher level of independence from external data than the existing ones. We will make this clearer in our motivation, discussion, and conclusions in the revised text.

Furthermore, the authors had access to longer observational periods (e.g., COSMIC-2 has been operating since 2019), but only used ~300 days of data. They also did not explore more modern and potentially better-performing DL architectures such as CNNs, RNNs, or Transformers, which could offer improvements in the reconstruction of vertical structures or better handling of sequential dependencies in profile data. This limits the potential impact of the study.

There is a trade-off between the amount of training data and the level of independence of the resulting retrievals. The larger our training data set, the more external data we will be using via 1D-Var targets. Therefore, we chose a rather small data set. Furthermore, with the currently used model architecture it is not guaranteed that more data equals better results. This is likely different for more complex architectures such as transformers or CNNs. We are currently investigating such setups too. However, these results are out of scope of this work. Nevertheless, we plan to expand our training/validation/testing datasets by using COSMIC-2, Spire as well as GeoOptics profiles spanning 2021-2023.

The "Outlook" section is vague and does not offer concrete suggestions or rationale for the chosen next steps. Given the availability of longer time series and more diverse RO data, it

is unclear why these were not used in the current study. Moreover, no consideration is given to improving the model architecture, hyperparameter tuning strategies, or integration with other ML advances.

We will try to extend the Outlook section for some more future ideas. We have explained the reasoning behind limiting the training data set already above.

Figure captions and table descriptions are often imprecise or incomplete. X-axis scaling is often poorly chosen - in many plots, the data curves are compressed or overlapping, making interpretation difficult. Several figures lack panel labels (e.g., a, b, c), which makes it hard to reference specific subplots from the text.

We will adjust the axis-scaling and add panel labels to all relevant figures.

These issues, while not invalidating the study, significantly reduce its clarity, reproducibility, and impact. Addressing them would considerably strengthen the manuscript.

In the following, we have answered your specific comments one-by-one.

Specific comments:

Line 118: The authors use a maximum horizontal distance of 500 km for RO–RAOB pairing. This threshold is very large and may introduce considerable representativeness errors, especially in dynamically active regions or the lower troposphere. Please justify this choice and consider sensitivity testing with smaller radii (e.g., 200 km), which are more common in literature.

The main motivation was to have a larger sample size for the chosen period. We know this threshold is large and might introduce representativeness errors. However, these errors would identically impact both the 1D-Var and AROMA error statistics. The actual absolute errors are not the most important information communicated in these comparisons. The focus of our interpretation is more on the relative differences between 1D-Var and AROMA, which are small to non-existent. Identical error levels as 1D-Var is, by definition, the best possible result to be achieved with our target setup.

Still, the representativeness errors may be overwhelming the actual differences between 1D-Var and AROMA, so we plan to extend the amount of RAOB observations included in the validation by extending the validation period. Then we will be able to have an equal (or larger) sample size while reducing the collocation distance to 200-300 km.

Line 132: The term "refractional radius" is used in the manuscript, but "impact parameter" is more widely recognized and standard in the GNSS-RO and atmospheric remote sensing communities. For consistency and clarity, consider switching to the standard term.

This will be corrected in the revised manuscript.

Line 136: The coefficients k_{1k_1} and k_{2k_2} are introduced in line 136 without a reference. Please cite a standard source (e.g., Smith and Weintraub, 1953 or ITU-R recommendations) to support these values and aid reproducibility.

This will cite Smith and Weintraub (1953) in the revised manuscript.

Line 165: The authors describe ANNs as supervised models "by definition". This is misleading, as neural networks can also be used in unsupervised (e.g., autoencoders, SOMs), semi-supervised, or reinforcement learning settings. Please revise to: "In this study, a supervised ANN is used..."

This will be corrected in the revised manuscript.

Line 174: The authors mention hyperparameter tuning but do not explain what it entails. Since this is a key concept in building neural networks, a short clarification would benefit readers unfamiliar with machine learning terminology. Additionally, the transition from a description of network architecture to applications in atmospheric science feels abrupt. Introducing a linking sentence would help maintain logical flow and improve readability.

This section will be reorganized. Context on ML in atmospheric science will now be included in the introduction.

Line 182: The manuscript mentions that hyperparameters were tuned, but does not describe how. Did the authors use grid search, random search, or another method? Please include at least a brief description and justify the chosen approach.

We have used a small grid search using the configurations that are mentioned in the manuscript. For the revised manuscript, we plan to extend this approach for a few more configurations and describe it in more detail.

The authors report testing only a few configurations (e.g., 1000, 2000, 2500 neurons). Why were smaller and potentially more efficient architectures (e.g., 32–512 neurons) not tested? Consider expanding the hyperparameter search space and reporting performance for smaller models. This could improve generalizability and reduce overfitting.

As mentioned in the answer to the last comment, a few other hyperparameter setups will be tested. The corresponding RMSE results of validation performance using these setups will be provided in the revised manuscript. Extensive hyperparameter tuning is not doable in this study because of computational and time resources.

Moreover, early stopping or validation monitoring is not mentioned. Using a fixed number of epochs without regularization can result in overfitting - especially with limited training data.

We will refine our strategy for the revised manuscript. We plan to set a maximum number of epochs and employ monitoring of the validation RMSE. Based on this metric, we will apply early stopping if values did not improve over the last 5 epochs (patience value of 5). We add information on this procedure in the revised manuscript.

Line 190: The manuscript claims similar performance across input feature combinations, yet no quantitative comparison is given. Please report error metrics (e.g., RMSE) for each tested combination to support this claim.

We will do so in the revised manuscript for all tested setups.

Furthermore, consider testing the inclusion of metadata features (e.g., satellite mission ID, time of day/month), which might capture climatological variability. Was this tested?

This was not tested yet. However, we will add several of these features to the new feature setup(s) investigated.

Lastly, feature importance analysis (e.g., permutation importance, Random Forest, or SHAP) could support the chosen input configuration and identify redundant inputs.

For the revision, we plan to carry out a SHAP analysis for the full feature setup. These will be presented and discussed in a newly added section in the manuscript. This will also increase the level of novelty of our work, since no other study has yet provided such insights.

Line 211: It is unclear whether a validation set was used separately from the test set. Using the test set for both evaluation and tuning can result in biased performance estimates. Please clarify this split.

This will be clearly explained in the revised manuscript. We monitor performance on the validation data set (15% of the total data set) along with performance for the training set (70% of data) to prevent overfitting. The test set (another 15%) was completely spared out of this process and the model's performance using this test data was evaluated independently afterwards.

Also, what does "random split" mean? If RO profiles are temporally correlated (e.g., from the same day), a random split may lead to data leakage. A time-based or mission-based split would improve robustness.

In the original version of this work, we used a random split between all profiles available. We now realized that this might be problematic and thus changed to a temporal split which we detail in the revised manuscript. A mission-based split would not be ideal in our opinion, since the model's ability to generalize over instrumental/orbital differences between missions would be weakened. However, we will still use PlanetIQ profiles only during testing, which gives a good idea of the model's performance for missions, whose data has not been seen in training.

Line 225: The method used for interpolating ERA5 to RO locations is not explained. Please specify the horizontal (e.g., bilinear) and vertical interpolation method (e.g., linear, log-pressure).

We used bilinear horizontal interpolation and linear vertical interpolation. We will adopt exponential/logarithmic vertical interpolation for pressure and humidity (which is a better approach for these variables) in the revised study and add this information in the revised manuscript.

Line 231: Also, Min–Max scaling is sensitive to outliers. Did the authors test Z-score normalization, especially for variables like LSW or SNR?

Although the RO profiles provided by CDAAC are generally high quality and undergo a specific quality control procedure themselves, we realized that outliers might still be an issue. Therefore, we plan to change to z-score based scaling in the new version.

Line 235: Moreover, why were targets (pressure, temperature, humidity) scaled? Many regression models perform well without target scaling. Was the effect tested?

Yes, this was tested and scaling the targets improves the results and efficiency of the training process. One good reason for adopting this strategy for our problem are the significantly different magnitudes of the target variables, as already explained in the original manuscript. For pressure targets specifically, we will now apply logarithmic scaling, to put equal emphasis on all vertical levels, instead of the lowest altitude regions which dominate in terms of magnitude of pressure values.

Line 241: The manuscript uses RMSE, bias, and STD in the results, but these metrics are not introduced in the methodology section. Please define all evaluation metrics in Section 3 and explain why they were chosen.

A specific section introducing the metrics has been added in the revised manuscript.

Line 247: The so-called “internal validation” refers to performance on the test set (i.e., hold-out data). In ML, “internal validation” is uncommon terminology — consider renaming to “test set evaluation” for clarity.

This will be corrected in the revised manuscript.

Figure 1: The figures show bias \pm STD, but not RMSE, even though RMSE is arguably a more comprehensive and common error metric. For consistency with Table 3, RMSE profiles should also be plotted.

We will show bias, STD, RMSE and the Pearson correlation coefficient for all evaluations.

Table 3: Table 3 presents RMSE, bias, and R, but it is unclear how these were aggregated - per level, per profile, or vertically averaged? Please clarify. Also, define R explicitly in the caption - likely Pearson correlation, but currently ambiguous.

These are vertical averages, and R is the Pearson correlation coefficient. We will add this information to the manuscript.

Section 4: Since data from both COSMIC-2 and Spire were used, it would be valuable to show performance separately for each mission. Differences in instrument characteristics may affect retrieval quality and generalizability.

This separate evaluation will be added, thanks for the recommendation. These problems are also the reason why we think a mission-based split is not ideal as mentioned above.

Line 275: The figures (2–4) only show bias \pm STD, despite RMSE values being reported in Table 4 . Since RMSE is a more informative metric (combining both systematic and random errors), vertical RMSE profiles should be plotted alongside, or instead of, bias \pm STD for clarity.

We will add RMSE to these figures.

Consider plotting the vertical error profiles for all three missions (C2, Spire, PlanetiQ) in a single figure per variable (e.g., three lines per plot). This would simplify comparison and reduce redundancy across Figures 2–4.

As mentioned above, this separate evaluation will be added.

Table 4: Please explicitly reference Table 4 in the main text. Also, revise the caption: "Results of external validation of C2 profiles using ERA5" could be misread as validation of ERA5. Suggest: "External validation of GNSS-RO retrievals using ERA5 as reference".

We will correct this in the revised manuscript.

Figure 6: Figure uses average differences rather than RMSE, which can mask true error magnitudes. RMSE better captures both bias and spread. Including it would make the performance evaluation more meaningful.

We will include RMSE here, as for all the other comparisons in the manuscript.

The x-axis scale in most figures in the manuscript is poorly chosen. It compresses the spread between profiles, making the plots less informative. Dynamically adjusting the scale to fit actual data ranges would help.

We will change the axis-scaling in the revised manuscript.

Please clarify in the caption or legend what each of the four lines represents. It seems they indicate mean \pm STD for CDAAC and AROMA, but this must be clearly stated.

Yes, they represent mean (solid) and STD (dashed) for CDAAC/1D-Var (blue) and AROMA (pink), exactly as for the other figures before (e.g. for the validation using ERA5). We will add this information to the caption.

Line 299: The text states: "negligible differences in relative errors for temperature and pressure", yet differences reach 2 hPa - non-negligible in many RO applications. Consider rephrasing or quantifying what is meant by "negligible" in this context.

We will rephrase the text accordingly.

Figure 7: How were the three RAOB-RO profile pairs in Figure 7 selected? Random sampling, specific regions, or performance extremes? Clarifying selection criteria would aid interpretation.

The profiles were selected to cover different latitude regions (tropics to higher latitudes). We will add this information to the revised manuscript.

The figure (and also other figures in the manuscript) lacks subpanel labels (e.g., a, b, c), which makes textual references ambiguous. Please label each panel clearly.

The revised manuscript will include subpanel labels for all figures.

Line 315: The text says "large differences... in the lowest 2 km", but Figure 7 shows significant deviations up to 8 km, exceeding 6 K. Recommend clarifying the vertical range and associating differences with the specific method (AROMA).

Section 4: Include actual numerical error values (e.g., max pressure error = 2.3 hPa, temperature = 6.3 K) to substantiate claims like "slightly degraded performance". Avoid vague descriptors like "some larger differences".

For both of the last two points, we will correct the wording in the revised manuscript. However, since we will retrain our model on an extended data set, the results might change.

Line 321: The method is presented as "novel", but MLPs have been previously used for RO profile retrieval (e.g., Lasota, 2021). The novelty lies more in the use of CDAAC 1D-Var as ground truth. For genuine innovation, deeper architectures (CNNs, RNNs, Transformers) could be tested.

As mentioned earlier in our response to the general comments, we are currently testing some of these architectures but including them here would go beyond the scope of this study. We will emphasize the novelty of this work relative to prior studies.

Line 323: Calling the test set evaluation "internal validation" is misleading. This is standard ML procedure. Please refer to it as "test set evaluation" for clarity.

We will adopt this wording in the revised manuscript.

Line 326: Although bias is low, RMSE remains high (see Table 3). Avoid claiming "good agreement" if total errors are still significant.

We will reformulate the relevant sections in the revised manuscript

Section 5: Replace generic descriptors like "small", "slight", "significant", and "good agreement" with quantified metrics or thresholds. This improves transparency and enables comparison with literature.

As mentioned earlier in our response to the general comments, we will reformulate the relevant parts of the manuscript accordingly.

Line 331: "Instabilities in the retrieval" needs clarification. Does this refer to signal loss, high noise, or unphysical values? Specify the cause and manifestation.

Although we have not investigated this in detail, we suspect that it is observation noise learned by the model, in an altitude region where on average only small amounts of moisture are present. We will clarify this in the revised text.

Line 326: Be consistent with error metrics – mix of K, %, hPa hinders interpretation. Prefer using physical units consistently (RMSE in hPa/K), and clarify if % is used (relative to what?).

For the revised manuscript, we will reevaluate our approach and decide if we still continue to use relative error metrics at all (due to suggestions made by another reviewer). We have clarified what relative means in this context in lines 243-245 of the old manuscript: “*Relative differences are calculated by normalizing the absolute differences with the average vertical profile of each retrieval parameter over all profile samples of the entire test set*”. Should we still continue to use relative errors, then we will try to emphasize even more how they are calculated.

Line 347: Only 300 days of data were used, though C2 has been active since 2019. Please justify the short training period - limited computational capacity, data availability?

As mentioned earlier, this is due to the trade-off between size of the training data set and the level of independence of the retrieval method.

Line 358: Drawing conclusions from only 3 RAOB–RO pairs is insufficient. Clearly state the illustrative nature of these examples and consider showing a “failure case” to discuss model limitations.

These are just examples of profiles observed at different latitudes, and we are not drawing ‘global’ conclusions from them. For this, we provide the results presented in Figure 6 and Table 6, which show differences to 1D-Var results of about 0.3 hPa/K for pressure/temperature as well as 0.02 g/kg in specific humidity. We will try to outline these results and their interpretation further in the revised manuscript.

Line 362: Direct comparison with Lasota (2021) is needed. Present numerical values side-by-side and discuss why AROMA performs worse for humidity (Q).

Since Lasota (2021) focused on tropics/subtropics, your RMSE (2.1/1.9/1.0) vs hers (1.9/1.9/0.5) should be broken down by latitude to make a fair comparison.

We try to answer the last two comments here together since both are related to a comparison to Lasota (2021). This study is important and provided inspiration for this work, but there are several differences which make a ‘fair’ comparison difficult to impossible. The most important difference is the target definition (1D-Var for this study, ERA5 for Lasota(2021)). Therefore, the error metrics retrieved from validation in the two studies are not comparable 1:1. Furthermore, as noted by the reviewer, the spatial coverage is different since we also include Spire profiles. We note that they are of similar magnitude (maybe apart from specific humidity) which is encouraging. Given our new plans to break out the analysis by RO mission, this comparison between AROMA and Lasota (2021) will become more meaningful but will still require caveats. We will include discussion on these points as appropriate in the revised text.

Line 380: As in earlier sections, the conclusions emphasize low bias but omit consistent RMSE reporting. Replace vague claims (“good agreement”, “very satisfactory results”) with quantitative metrics and exact values.

We will revise these formulations, as for all the other parts that were mentioned earlier.

Line 402: The "future work" section mentions obvious extensions (longer period, more missions) without explaining why they were not done here. No new architectures, tuning strategies, or validation steps are proposed.

We will extend this section for a discussion on using new architectures, which might be promising for the RO retrieval problem. At the moment, we are not aware of new tuning or validation strategies which might help, but we might do some further literature research on that. We will try to include some of the 'obvious' extensions, which we already proposed, in the revised version. As mentioned before, testing completely new architectures such as transformers or CNNs is out of scope for this work, but we plan to investigate some of these in a future study.

Line 410: Critical limitations are not addressed:

- Use of CDAAC 1DVar (based on ECMWF) may contradict the goal of independence from NWP.
- Performance in lower troposphere is not isolated.
- No insight into model robustness in sparse-data regions.

The first point we explained in detail for several of the previous comments. The use of CDAAC 1D-Var definitely contradicts independence from NWP, but far less than other target definitions such as ERA5 profiles. For sure, training a model just using collocated radiosonde observations might be possible, but the resulting data set would be rather small even for a long time period. Furthermore, the spatial coverage of radiosonde soundings (a lot of them launched over land) is not ideal for building a retrieval algorithm that works well on the global scale. We expect that specific hybrid approaches for the target definition might be feasible to test in future studies. We already hinted that in the original manuscript (Line 409-412). We will add some further discussion on this point in the revised manuscript.

We realized that the second point might originate from a non-ideal title definition on our side. We will rename the study to "*Retrieval of thermodynamic profiles in the lower **atmosphere** from GNSS radio occultation using deep learning*", since this is more accurate to the actual altitude levels we are including in our setup.

The third point will be covered by analyzing performance for different latitude regions/bands in the revised manuscript.

Recommendation

While the manuscript touches on an important topic - data-driven retrieval of RO profiles - it falls short in several key areas: originality, clarity of methodology, proper evaluation practices, and rigorous analysis. The method presented here does not outperform existing ones and the contribution is incremental at best. If the authors address the methodological gaps, improve evaluation rigor, and provide a more compelling future outlook, the paper could become a valuable contribution. In its current form, however, major revisions are needed.