Public justification (visible to the public if the article is accepted and published):
I appreciate the revised version of the manuscript; I am, in particular, grateful to the authors for pointing out their mistake in calculating some results that, in my view, change the conclusions reported in the paper significantly. The discussions on the ML methodologies applied and the physical background of the relation between PBL humidity and SNR measurements are important additions that make the paper considerably more useful for its audience.
Dear editor,

Thank you sincerely for reading through the revised manuscript and the response, and provide your suggestions. We have now modified Fig. 9 and 10 (formerly Fig. 7 and 8) according to both reviewers and your suggestions (see response #1 below). We add more in-depth discussions regarding each campaign/weather regime (see response #2 below). For Fig. 11, the extrapolated non-meaningful values beyond +1 and -1 due to python sns.violinplot artifacts have now been fixed. Please see response #3 for refinements related to the discussion of Fig. 11.

Nevertheless, I am convinced that the manuscript still needs major revisions, for the following reasons:

1) Both reviewers suggest changes to Figures 9 (formerly 7) and 10 (formerly 8); reviewer 1 made more specific suggestions to reduce the number of levels, enlarge the figure size, and maybe reduce the number of levels being shown. I fully agree with their assessment, even for the redrawn figures. For example, the ERA5 line is still not visible, nor are the mean values for each campaign visible; the figures are too busy. Increasing the size on a computer screen as suggested by the authors does not solve the problem. I can see no qualitative difference between the three figures for the 975, 950 and 925 hPa levels, nor between the 900 and 875 hPa levels.

Response #1: Following the suggestions by reviewers and the editor, we now doubling the line thickness of ERA5 result in Fig. 9 (gray thick solid line) and put it underneath the SNR result (black thick solid line), so readers can see the SNR-ML results track ERA5 closely, often overlapping each other, meaning that the SNR retrieval is comparable in quality with ERA5 reanalysis. The mean of each campaign is now not only plotted in a bigger symbol, but also encircled by a black boundary to make sure the mean and standard deviation of each campaign stands out from individual samples in both Fig. 9

and Fig. 10. The subset of ERA5 collocation samples to generate the gray line in Fig. 9 is now plotted as open symbols in Fig. 10 to better visual comparison. The black thick solid line connecting each campaign's mean in Fig. 10 is enhanced in thickness now.

We still believe it's important to show individual comparison samples in the background to demonstrate the variation and extremes in one campaign that standard deviation is not good enough to capture. For example, in Fig. 9 SNR retrievals during ARRecon_2020 (filled orange triangles) apparently outperform L2 (open orange triangles) for the extremely wet situations and relatively dry situations, even though their means are indifferentiable for this campaign at 900 – 850 hPa. In Fig. 10, we can see although the means from all 6 campaigns track the 1:1 line closely with slight dry-bias, there are many much-too-wet values in ERA5 during the EUREC4A campaign (blue dots that are away from the 1:1 line) in the lower levels. Also, although the mean comparison during the MARCUS campaign (cyan triangles) seems to show that ERA5 does a good job in the polar PBL, one can see the collocation samples deviate from the 1:1 line when mixing ratio is smaller than 3 g/kg.

In the text, the authors also do not discuss individual levels in detail; instead, they make qualitative statements such as "In general, better correlation are found when..." without pointing to any evidence in the figure. In the text, the authors also discuss only results from the ARRecon campaign in detail. The figure is clearly important, but the many details are neither discussed, nor can they be easily deducted from the figure alone. The authors may also consider other ways of representing their results, e.g. showing data for individual campaigns, as in Fig. 11

Response #2: Now the discussions around Fig. 9 and Fig. 10 have been revised substantially to add in more in-depth discussions and comparisons. It worths clarifying that Fig. 11 is the only figure that we identified a bug in coding, which changed some conclusions related to Fig. 11, but not to Fig. 9 nor Fig. 10.

Fig. 9 shows the ~~layer-by-layer~~ level-by-level comparison for all collocated samples from all campaigns. SNR-ML retrieval results are shown in filled color symbols while ~~Level-2~~wetPrf/wetPf2 retrievals are shown in open symbols. In addition, the averages from each campaign collocation subsets are connected together for better visual comparison against the 1 : 1 line (black solid lines for SNR-ML retrieval and black dotted lines for wetPrf/wetPf2 retrievals). We can see both SNR-ML retrievals and ~~Level-2~~wetPrf/wetPf2 retrievals demonstrate ~~close~~ general good agreement with ground "truth" for different weather regimes. For the SNR-ML retrieval results, better correlations are found ~~when the MABL is relatively dry or moderately-wet~~ for the

Southern Ocean (MARCUS campaign) and stratocumulus weather regimes (MAGIC campaign). Although the high correlations are comparable for the wetPrf/wetPf2 retrievals, the collocation samples are much sparser (Table3). This could be attributed to the frequent occurrence of super refractions in the stratocumulus region that causes a sampling bias of the wetPrf/wetPf2 results (Xie et al. (2010), Feng et al. (2020)). Spreads are slightly larger during the atmospheric river events (ARRecons). ~~Take the ARRecon campaign as an example,~~ SNR-ML retrievals show an overall better agreement compared to ~~Level-2~~ the wetPrf/wetPf2 retrievals at all 6 pressure levels, especially for the few extremely large specific humidity values ($> 12g/kg$). ~~are well captured by the SNR-ML retrieval but not the wetPrf/wetPf2 retrievals~~. The means of all ARRecon collocated samples also suggest that SNR-ML retrieval is the only one that does not produce a bias, while ~~the retrievals both ERA-5 and GNSS Level-2~~ wetPrf/wetPf2 are moderately (slightly) dry biased in atmospheric river scenarios at > 900 (< 900) hPa. ~~Such a close agreement appears to become noisier at 850 hPa, again demonstrated that signals at sharp boundaries (i.e., PBL top) are hard to retrieve.~~ ERA-5 from each campaign (only considering samples that SNR-ML retrieval collocation is found) exhibits good agreement to the ground truth too ~~if only mean values of each subset is considered~~ (gray solid lines). ~~However, the scatter plots are much noisier if all collocations are plotted (Fig. 10)~~. For the two deep tropics campaigns ATOMIC and EUREC4A, we can clearly see that none of the three datasets capture the humidity conditions in the MABL very well. They are all dry-biased, and ~~means from~~ ERA-5 reanalysis is slightly less dry-biased than GNSS retrieved values at 975 hPa and 950 hPa. SNR-ML method achieves overall comparable performance to ERA-5, which is expected because model is trained on ERA-5. ~~The operational wetPrf/wetPf2 product is noticeably dry-biased in the MAGIC campaign (i.e., the large deviation of the black dashed lines). As MAGIC campaign was carried out in the stratocumulus region off the California coast, frequent ducting-induced negative biases are probably the main reason that causes such a significant dry bias (Feng et al. (2020)).~~

For convenience in pinpointing ERA-5 MABL issues, we also make Fig. 10 as each valid radiosonde/dropsonde profile from all 6 campaigns can always collocate with an ERA-5 reanalysis data sample within 1.5° longitude, 1° latitude and 1 hr difference. Now we can clearly see ERA-5 frequently fails to produce the large variations in humidity in the trade-cumulus region (EUREC4A), the former of which tends to be always too wet. Otherwise, ERA-5 matches better than SNR-ML retrievals and wetPrf/wetPf2 retrievals in the deep tropics (EUREC4A and ATOMIC), albeit all of the three datasets contain persistent dry biases as also can be seen in Fig. 9. ~~didn't capture the MABL humidity change in the majority time during the EUREC4A campaign with large wet-biases~~. Another discernible bias happens in the Southern Ocean during the MARCUS campaign, where ERA-5 is consistently dry-biased when specific humidity is below $\sim 3g/kg$. The subset used to make the gray lines in Fig. 9 are overlaid in open symbols, so we can make straightforward and fare comparison between ERA-5 and SNR-ML retrievals. We can see that SNR-ML performs slightly better than ERA-5 in the atmospheric river scenarios (two ARRecon campaigns), and slightly worse than ERA-5 in the stratocumulus region (MAGIC campaign), both of which reflect in the correlation coefficient comparisons shown in Fig. 11 as well. Overall ERA-5 shows a small dry-bias globally at all levels, which agrees with early findings by Johnston et al. (2021) who used ~~Level-2~~wetPf2 GNSS-RO retrievals to identify such a dry bias. Note that some of the campaign profiles (e.g., ARRecon dropsondes) are actually assimilated in the ERA-5 data, so it is not a completely independent validation strictly speaking. However, it is also worth noting that some previous publications

employed ARRecon and EUREC4A radiosonde data as "ground truth" for evaluating ERA5 accuracy in capturing water vapor variabilities in the PBLs (e.g.,Cobb et al. (2021), Kruger et al. (2022)).

2) The correction of the calculation of correlation coefficients has changed Figure 11 considerably. I agree with the authors that the performance of all three data sets is

miserable for the tropical campaigns (EUREC4A and ATOMIC); and SNR-ML outperforms ERA5 and wetPrf/wetPf2 for ARRecon. However, the authors should also point out that ERA5 outperforms both SNR-ML and wetPrf/wetPf2 during the MARCUS and MAGIC campaigns, and that wetPrf/wetPf2 still outperforms SNR-ML during MARCUS. Thus, (page 19, line 331 onwards) it is not correct that "the quality of SNR-ML retrievals is comparable to ERA 5 and operational wetPrf/wetPf2 products", and that it even outperforms the other two in the ARRecon case. At best, the result is inconclusive; SNR-ML outperforms the other two in one case, underperforms in two cases, and is as useless as the others in another two cases. Also, the statement (page 24, line 425 onwards) that the results demonstrate "the real information content in the SNR signal is learnt" during the ARRecon campaign should be extended to say that in two other campaigns, the SNR-ML retrieval failed to extract this information from the data. A similar update is required in the abstract (page 1, lines 13 - 16); the SNR-ML also underperforms compared to ERA5, and the ML retrieval did not extract useful information from the SNR signal in these cases.
There is more to say on Fig. 11 b: Why do the violin plots indicate the presence of correlations with values > 1? The box plots inside the density estimates indicate no such data, but maybe a violin plot then is not a good way to show the characteristics of the data. Or the artefact should be mentioned in the discussion.

Response #3: The values beyond +1 and -1 originate from missing use a cutoff parameter in using the python sns.violinplot function. Now the cutoff has been added to cut out any artifacts beyond maximum and minimal values.

We partially agree with your interpretations above and had incorporated some into the discussions. Thank you very much. Specifically, we agree that “the performance of all three datasets is miserable for the tropical campaigns (EUREC4A and ATOMIC); and SNR-ML outperforms ERA5 and wetPrf/wetPf2 for ARRecon.” However, we believe your suggestion that “ERA5 outperforms both SNR-ML and wetPrf/wetPf2 during the MARCUS and MAGIC campaigns, and that wetPrf/wetPf2 still outperforms SNR-ML during MARCUS” is not 100% accurate. For the MARCUS campaign, ERA-5 is apparently dry-biased when specific humidity < 3g/kg (now mentioned the in the discussion of Fig. 10 in the revised manuscript). This bias seems to be slightly mitigated using the SNR-ML method. If you revisit Fig. 6, SNR-ML method produces much larger variations for very dry situations than ERA-5, which however also comes with large uncertainty to make it a robust retrieval. Moreover, we recently learnt from radiosonde payload

provider that the humidity sensor response time on radiosonde is significantly delayed at extremely low temperature (usually happens around tropopause, but could also happen over polar winters), in which case the radiosonde readings might not be trustworthy as the "ground truth". So we agree with your suggestion that results for MARCUS campaign (i.e., Southern ocean) is inconclusive. We have now included that in the discussion of Fig. 11.

For the MAGIC campaign, SNR-ML method actually outperforms the wetPrf/wetPf2 products in both the correlation coefficients (Fig. 11a) as well as the number of available samples especially at the lowest three pressure levels (Table 3). As super-refraction tends to occur more frequently at stratocumulus regions (e.g., Xie et al., 2010), SNR-ML method possesses unique advantage over wetPrf/wetPf2 products in providing unbiased PBL humidity retrievals.

Regarding editor's concern about violin plots, after fixing the artifacts, we believe this is our best plotting option to integrate multi-dimensional information into one figure. The skewness of the distribution of correlation coefficients, previously not discussed, is now included in the discussion as well. Fig. 11b in particular demonstrates the robustness of the SNR-ML retrievals across all 6 PBL pressure levels: although the highest positive correlations are always identified in ERA-5 and/or wetPrf/wetPf2 products, the medians of SNR-ML retrievals are consistently the highest with consistent top-heavy distribution except for 850 hPa, meaning that SNR-ML retrievals agree with radiosonde "truth" more consistently while ERA-5 and wetPrf/wetPf2 have more variations. Of course all these conclusions are limited by the smaller collocation samples (<=309 in total), and we for sure need more extensive evaluation for this research product before massive production.

The violin plots in Fig. 11 and number of collocated sample statistics in Table 3 help disentangle the merits/caveats of SNR-ML retrievals from multi-dimensional statistical metrics. ~~reveal more detailed difference in comparison statistics with respect to the radiosonde/dropsonde "truth", which more comprehensively demonstrate the values (and caveats) of the SNR-ML retrieval~~. Only correlation coefficients of all collocated samples collected from each campaign are displayed in Fig. 11. The ARRecon-2018 and ARRecon-2020 samples are further combined. From Fig. 11a, we can see ~~both SNR-ML retrieval and Level 2 retrieval agree much better with the ground truth across all extra-tropical campaigns compared to ERA-5~~ again that the MABL specific humidity is not well captured in the tropics by either of the three datasets (EUREC4A and ATOMIC), but SNR-ML ~~method~~

~~generated~~ retrievals perform slightly better than the operational wetPf2 products in the deep tropics and trade-cumulus regions. In the rest three campaigns in the mid- and high-latitudes, they all agree very well with the radiosonde/dropsonde ground truths. ERA-5 reanalysis does the best job at high-latitude southern ocean (MARCUS) as well as the stratocumulus region (MAGIC), while ~~surprisingly~~ in the atmospheric river regime, SNR-ML retrievals outperform the wetPrf/wetPf2 retrievals as well as the ERA-5 reanalysis. It is worth noting that SNR-ML retrievals perform slightly better than wetPrf/wetPf2 retrievals in the stratocumulus region (MAGIC) in both the medians and the top-heavy skewness of its distributions, which can be partially attributed to the scarcity of wetPrf/wetPf2 collocation samples in this weather regime and known bias in the Level 2 retrieved refractivity gradient (Xie et al. (2010)). For the polar region (MARCUS), although SNR-ML retrievals exhibit the lowest correlations among the three datasets albeit all correlations are statistically significant, it is inclusive at this point to say that SNR-ML method is not suitable for the polar region. As a matter of fact, SNR-ML method generates the largest variabilities among the three when the PBL is extremely dry (Fig. 6), but the SNR in this situation is generally too weak to generate a robust retrieval (i.e., uncertainty too large compared to retrieved value). The retrievals from the SNR-ML method at dry polar winters contain more potentials (e.g., Fig. 14 as an example) if future GNSS missions could improve the SNR.

~~The SNR-ML retrieval exhibits more robust correlation while Level 2 retrieval are really poor (negative correlations) at some levels. However, for the two deep-tropics campaigns, above statements do not work anymore. All three work really poor for the ATOMIC campaign with barely any correlation with the ground truth or even negative correlation for the Level 2 retrieval. For the EUREC4A campaign, things remain similar for RO retrievals no matter using Level 1 or Level 2 data, but ERA-5 works much better at capturing the humid MABL structure in this case albeit it's still dry-biased (not shown). We achieved 20-80% across-board more collocation samples using the SNR-ML retrievals versus the Level 2 wetPrf or wetPf2 retrievals, the latter of which is consistent with the general success rate shown in Fig. 1.~~

Fig. 11b demonstrates the robustness of the SNR-ML retrievals across all 6 PBL pressure levels. Although the highest positive correlations are always identified in ERA-5 and/or wetPrf/wetPf2 products, the medians of SNR-ML retrievals are consistently the highest with consistent top-heavy distribution except for 850 hPa, meaning that SNR-ML retrievals agree with radiosonde/dropsonde "truths" more consistently while ERA-5 and wetPrf/wetPf2 have more variations across different weather regimes. Of course all these conclusions are limited by the smaller collocation samples (<=309 in total), and we for sure need more extensive evaluation for this research product before massive production.

3) Neither reviewer discussed the ML approach in detail. If I understand correctly, the authors train the data on one period, perform the hyperparameter tuning on a test period, but then evaluate (or validate) the fitted model on the combine training and test data. Isn't this a case of data leakage?

You are correct. There is a data leakage in the prediction period, but the hyperparameter tuning strictly follows the standard ML procedure. The data leakage only affects the COSMIC-1 data and the MAGIC campaign comparison. We unfortunately didn't have enough disk array and computational powers at the time when performing the training/tuning. The reason for including both training and testing data for prediction is for generating robust enough samples for constructing statistically meaning climatology, especially the diurnal cycle (COSMIC-2 couldn't cover high-latitude). We now mentioned this issue in the revised manuscript.

In this work, we created a collocated and coincident ERA-5 - SNR training and validation dataset. The SNR records are from ~~both~~ four satellite series: COSMIC-1, COSMIC-2 METOP-A and METOP-B. The periods for training, independent testing, and prediction are listed in Table 1. Note that the ~~validation~~ testing period is independent from training period to avoid potential self-correlation using standard random splitting procedure. The prediction period however covers both training and validation periods mainly for generating enough samples to construct statistically robust climatology (e.g., diurnal cycles). This however creates an unfortunate data leakage concern for the comparison with the MAGIC campaign but not for the rest of other independent validation datasets (Table 2). The target variables are specific humidity at the aforementioned 6 pressure levels ($975hPa$, $950hPa$, $925hPa$, $900hPa$,$875hPa$ and $850hPa$). The input parameters are 52 levels of S_{RO}, 52 levels of σ_S^2, latitude, longitude, month and Rising/Setting flag.

In addition, the authors mention results from fitting other ML algorithms, such as gradient-boosted trees; it would be good to give references. The authors state that results obtained with these ML models were comparable with theirs; were hyperparameters also tuned, or would that provide an even better performance? I was also confused by "logistic regression" and "Support Vector Machines (SVM)" being mentioned. After all, they are classification algorithms, so why should they be applied here?

SVM, random forest and gradient boosting methods tested in this paper are all employed from the standard “scikit-learn” package. See https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html for the SVM regressor (should be called SVR indeed). We did minimal hyperparameter tuning for these methods, mainly just to check the performance metrics and the important factor ranking to assure physical consistency (see response letter Page 23). These are shallow ML models that are considered “outdated” in ML field these days, so we didn’t explore further of these models. I had a wrong memory about linear regression, which I never used. This has been crossed out in the revised version.

MLP regressor is employed from the pytorch library. We performed hyperparameter tuning for this one and the performance is slightly worse but overall quite comparable to CNN. There is really no preference of a ML model for this paper. We stated the reason why we didn’t perform extensive tuning and cross-ML model comparisons as this is not the focus of this paper. See below revised text:

We also tried some earlier ~~old-fashioned~~ ML models, e.g., random forest (RF), gradient boosting (GB), ~~logistic regression (LR),~~ support vector machine (SVM) and one deep learning model multilayer perceptron (MLP). The model performances are actually very close in terms of evaluating the RMSE except for the ~~LR and~~ SVM, the latter of which performed discernibly worse than the rest ML models. It is not a surprise finding as this is a relatively simple and straightforward task that ML models should handle easily, but not the case for multi-variable linear regression type of logistic models (hence, it explains the poor performance of ~~LR and~~ SVM). As the main focus of this paper is science and new information content embedded in SNR signals, we will not deviate the attention to spend more time discussing these model results. The semi-transparency of RF and GB models is appreciated by us though. We compared the feature importance rankings with Wu et al. (2022) findings, and find high consistencies (e.g., high ranking of SNR at $H_{SL} = -100\ km$ in the tropics, and SNR at $H_{SL} = -80\ km$ ranks the top in the polar region).

4) Table 1 (page 6) states that prediction intervals for Metop-A and -B were in "2012.01 - 2011.12". Apart from the wrong order of the orders, this is strange given that Metop-B was launched in late 2012 only. Please review the entries in this table carefully.

Thanks for identifying this typo. It should be 2012.01-2012.12. Typos are fixed now.

5) I believe Table B1 also contains wrong data. Excess phases increase towards the ground. Thus, increasing values of log(excess phase) correspond to data lower down in the atmosphere, and hence Hsl. However, the table claims that the largest excess values correspond to the highest Hsl.
Thanks for pointing that out! I went back checking the corresponding value and indeed the log(excess phase) should be reversed.

Finally, I strongly recommend to have a native speaker review the text before the resubmission of the manuscript. There are various leftovers from LaTeX code (e.g., "textcolorred" on page 7 line 160) and incomplete sentences that require a thorough review.

Thank you! We have scrutinized the cleaned version to make sure there are no leftover LaTex codes as well as no grammar errors (to the best we can).