Reviewer #2:

Summary:

We thank the two reviewers gratefully for their helpful and constructive suggestions. Please see our responses in **blue letters** below every comment, and **red letters** highlight the changes made in the revised manuscript. Following the reviewers' suggestions, we had made some major updates in the revised manuscript. In particular:

(1) we didn't perform the SNR retrieval for Metop-A and Metop-B in the original submission after observing slightly different non-linear relationships (Fig. A1 and A2) compared to COSMIC-1 (Fig. 2). In the revised manuscript, 2 additional ML models are trained for Metop-A and Metop-B, respectively, and are used for generating predictions (updated Table 1) to enhance the robustness of subsequent climatology and diurnal variation studies. Note that adding in Metop-A and Metop-B retrievals only add 14 additional samples to the collocated SNR-ML retrieval – radiosonde/dropsonde data samples during the MAGIC campaign.

(2) Meanwhile, please accept our apology if there's any confusion in the wording in the original submission. The key purpose of this work is to **prove that SNR can be used to generate profile-by-profile MABL (marine atmosphere boundary layer) specific humidity retrievals, and retrievals generated by the SNR-ML method have comparable or better quality to the operational wetPrf/wetPf2 products and ERA5 reanalysis in different weather regimes globally with 20 – 80% more successful retrievals than wetPf2 products in the lowest MABL.** Although producing a harmonized multi-year program-of-record (PoR) using the current algorithm is our ultimate goal, it is beyond the scope of this current work and we cannot do it in this paper with the limited funding and time.

(3) We added a new Fig. 3 to demonstrate the signal coherence at profile-by-profile level so to prove the retrievability using SNR. We can also observe the non-linear response is mission-dependent, which justifies why we need to build 3 individual ML models for each mission series.

(4) We added Section 2.3 to recap the underlying physical mechanisms discussed in Wu et al. (2022) and Fig. 3 to justify the reason of using ML instead of physical models for realizing the operational retrieval with the SNR measurements.

(5) We added a new Fig. 5 to illustrate the model internal architecture and deleted terminologies in the text that might be just jargons to readers.

(6) We redraw Fig. 2, Fig. 4, Fig. 9, Fig. A1 and A2 to improve the readability.

(7) We fixed a code bug in plotting Fig. 11 (previous Fig. 9). We also found a bug that the total number of collocated wetPrf/wetPf2 product is pressure level dependent. Therefore, we added a new Table 3 to reflect the correct collocated sample sizes at each pressure level for each campaign. The entire Section 3.3 is largely rewritten to reflect the changes of major findings.

The manuscript presents an interesting idea to harness other available information of a radio occultation observation to improve the retrieval of water vapour in the lowest boundary layer (marine here). I do have though major concerns regarding how data is selected, presented, and conclusions drawn. It is unclear whether the different retrieval / re-analysis comparisons are actually using the same data, or whether different data is entering each retrieval, and thus comparisons are misleading. I'd also like to see or at least discussed, whether a more advanced retrieval setup can be used, that uses SNR/amplitude and bending angles to improve the data in the lowermost atmosphere. A physical based retrieval will certainly be preferred by the NWP and Climate community, rather than one that needs per satellite tuning. There is also no real discussion on the higher SNR available with e.g., COSMIC2, and whether that improves the SNR retrieval (in fact, it is actually fairly unclear where C1 and where C2 is used, except that C2 is not available at higher latitudes and for certain periods). Metop is only included in the appendix, and it is unclear why that data wasn't used as well.

Thanks for your constructive comments. We apologize for unclarity in many places. Now as also suggested by Reviewer #1, we trained two additional ML models for Metop-A and Metop-B respectively. The periods and mission names for training, independent testing and prediction are listed in Table 1. The evaluation in Section 3.1 and 3.2 used the "testing periods" following standard ML procedures, while from Section 3.3 and onwards, data in the "prediction periods" are fed into the computation.

We didn't train a separate ML model for COSMIC-2, but rather apply the ML model trained solely on COSMIC-1 data to the COSMIC-2 SNR profiles for prediction. Then we used collocated radiosonde and the COSMIC-2 predictions during the 2020 ARRecon, EUREC4A and ATOMIC campaigns for validation. This is to test the robustness of the so-called "transferred learning", which seems to be good based on the consistent performance. However, we had to sacrifice the higher-frequency sampling rate for COSMIC-2 and to down-sample the SNR data to 1 Hz in order to use the COSMIC-1 ML model. In the future we might want to develop one ML model for one mission.

Regarding the joint bending angle + SNR retrieval, this is a great idea. I would picture it being especially useful when refractive index is negative and where the bending angle is biased (Feng et al., 2020). They can complement each other in different PBL scenarios. However, how to realize that in practice has a long way to go. At least for

now, we can see two major issues: (1) bending angle is in height (or pressure) coordinate, while deepSNR signal is below surface so it is difficult or maybe impossible to mirror the signal to a specific above-surface coordinate as the deepSNR signal probably comes from multi-paths (a toy model is given in the appendix of Wu et al., 2020). (2) as now discussed in Section 2.3, the underlying physical mechanisms for deepSNR to carry PBL information are not fully understood. Multiple scenarios can co-exist to cause the information to reemerge from below surface. As a physical deepSNR model is not readily available, we do not see this product to be assimilated in the NWP in the near future, but can be rather used as another observational reference to evaluate NWP forecasts. This discussion is now included in the conclusion part as outlook to the usefulness of this new product.

One thing to clarify is that “bending angle” is a level-2 product from GNSS-RO. The satellite-dependent “tuning” (a.k.a., calibration) happens behind the scenes during Level-1 to Level-2 data processing, and many “bending angle” profiles have to be excluded during “quality control” step when its SNR is below certain threshold. However, for these profiles, although “bending angle” cannot be used anymore, the SNR signal reemerged from below surface still contain the physical PBL information. If the SNR profile could be assimilated to NWP models in the future, they would be processed at the front-end and NWP centers do not need to worry about satellite-dependent tuning.

Text based comments (with Page/Line):

P1L10: Is this Monte Carlo anywhere used in the manuscript?

Yes. The dropout method’s full name is “Monte Carlo dropout method”. We didn’t realize this is an unfamiliar terminology to some readers, and now have modified words in Section 2.2 “Machine learning model selection” to clarify. It now reads as:

“In the prediction step, 30 predictions were carried out given each input set of variables, the mean and standard deviation of which were used as the final prediction and errorbar. It is worth highlighting that in each convolutional and fully-connected layer, a dropout rate of 0.25 is applied to generate the variation, which is then used to calculate the standard deviation of the “ensemble prediction” as a way to measure the retrieval uncertainty. This so-called "Monte Carlo" dropout method was ~~Note that the dropout layers were~~ designed in ML as a standard technique to regularize model over-fitting (Srivastava et al. (2013)), but were also employed widely as a Bayesian-approximation to quantify model uncertainties (Gal and Ghahramani (2016)). Admittedly the current method only provides a quantification for ML model errors. There is no consideration of SNR measurement errors nor propagation of the error to

the final retrievals at this moment, although this is certainly some procedure to be in place in the future works."

Figure 1: Maybe add the approximate hPa as well?

Done. Added.

P5L24: The table referred to seems to be B1. And, are the entries in B1 the right way round? And is excessive phase the same as excess phase (which is commonly used in RO)?

Table B1 in this paper is an updated version of Table A1 in Wu et al. (2022). Thanks for spotting this cross-referencing error. And we have corrected the misspelling. It should be "excess phase".

P5L114: Using RO data to validate reanalysis performance in the PBL and citing it here is misleading on the RO capabilities. Even Johnston stated in the abstract: "Negative C2 moisture biases are evident within the boundary layer, so we focused on levels above the boundary layer in this study.".

Yes. This is a great point. Indeed this is a caveat of using GPS-RO data especially in marine stratocumulus region where negative refractivity index is found frequently and causing bending angle biases (e.g., Feng et al., 2020). However, since there is no absolute "best global truth" observation available for MABL moisture vertical structure, we stretched the result shown in Fig. 3 of Johnston et al. (2021) paper in the boundary layer as an indication of possible systematic biases of humidity in MABL in ERA5 data. In fact, some other studies (e.g., Virman et al., 2021) suggested the opposite. Based on our radiosonde comparison shown in Fig. 10 (previous Fig. 8), there seems to be a slight but systematic dry bias in ERA5 in the MABL at all 6 pressure levels. Nevertheless, since some of these campaigns' radiosonde data are likely having been assimilated in ERA-5, it is nearly impossible to identify some independent dataset to evaluate ERA-5 biases.

Coming back to your point here, we acknowledge our citation was too stretchy, and now add a new sentence at the end of this paragraph, read as: "However, it is also warned in Johnston et al. (2021) that GPS-RO retrievals tend to have its own biases especially in MABL, and in fact some other research suggested wet biases in certain regions (e.g., Virman et al. (2021))."

P5L121: Is there any further restriction in the used C1, C2 data set? Not all occultations / SNR values will go down to the required hPa levels.

For C2 dataset, we took all wetPf2 retrievals (wetPrf for 2012 and 2013) provided on the UCAR data portal and performed the collocation. Therefore, the bending angle profiles that do not pass the SNR threshold should have been filtered out before C2 product is processed (Wee et al., 2021).

For C1 dataset, we didn't describe in detail about our quality control procedures. Now it's added in: "In practice, we use averaged SNR between 35 and 65 km altitude range as the SNR0, and any profile with SNR_0 < 200 or $\sigma^2_{SNR0} > 0.05$ is considered "low-signal" and is filtered out." ... "In practice, we also filtered out bad open-loop profiles, profiles with data gap greater than 2 km, and profiles with outlier S_{RO} or σ^2_S values."

Table 1: What is the validation period? And why not name the prediction period too in the title? And is there any reason for this limited data set use?

The "testing" period is the validation period, and "prediction" is now added to the table caption. There's a terminology confusion between ML field and geoscience field. In our geoscience field, "validation" means the "independent testing" samples in ML terminology, and the "validation" in ML field means the samples used during the training (here it's the randomly picked 10% samples out of the entire training sample).

Downloading hourly ERA-5 reanalysis was extremely slow when we carried out this research, so the training period was limited. Having said so, we got 211,425 training samples for COSMIC-1 data, which we believe is large enough for such an easy job (Fig. 3c). The Metop-A and Metop-B training datasets are relatively small due to the time constraints of the revision deadline. As this is a research work, we feel it is a good starting point to prove the feasibility and merits of this product. Making it operational certainly takes much more effort if it is going to happen in the future.

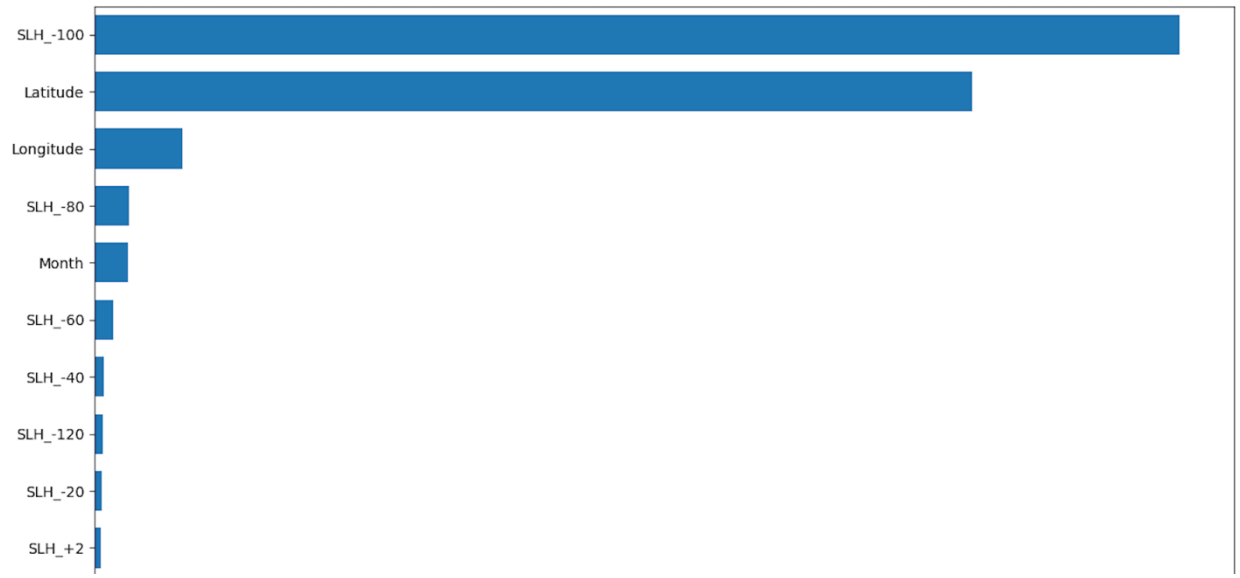
Figure 2, Caption: The figure does not only show grid levels, but also hPa ones.

Please see the new Fig. 2. Non-useful cross-correlation information has been removed. We hope this new figure is easier to read and more concise in delivering the information content.

P6L125: Is there actually any correlation visible between your 6 pressure levels and latitude? Some of these levels will be always above the PBL, e.g. at higher latitudes with low PBL heights. And is that impacting the retrieval quality, as fewer pressure levels are contributing? Is there an improvement possible; the low water vapor at high latitudes will also lead to limited bending, further complicating the retrieval?

That's exactly why we chose to use a ML method, and include latitude as one of the input variables, because we found the SNR-humidity correlation is the highest at HSL =

-100 km at tropics and mid-latitude, but at HSL = -80 km at high latitudes. Below is the top 10 most important parameters that contributes to the 950 hPa retrieval when we employed the random forest model and the gradient boosting model (those models have to make one prediction at one time):



P6L126: Aren't other levels right above also correlated with the same magnitude?

Please see the new Fig. 2 for a closer look of the details. We can certainly extend to another pressure level above, but based on independent shipborne radiosonde data comparison (Fig. 9), we can see the performance of SNR-ML method degrades at 900 hPa or above. This is reasonable as SNR-ML should work the best when the PBL has a very sharp gradient if our physical explanation of the working mechanism is correct (Section 2.3).

P7L132: One of the highly relevant advantages of RO data is the independence on the instrument. Having an instrument dependent contribution here will limit any data use for e.g. climate significantly. And, by the way, how does COSMIC-2 look like? And is there a constellation dependent factor too, as C2 observes GLONASS?

Please see the detailed response to your general concerns at the beginning of this response letter. The reason why users see RO data to be independent of instrument is because the calibration has been done for each instrument when processing the received data. The SNR thresholds for filtering out bad RO data are also instrument dependent (Ganeshan et al., 2024).

For the SNR-ML method, it works the same way. We have published the input training data on zenodo, which has been tuned for each individual mission, and interpolated to the excess phase coordinate so to homogenize the signal across different missions (see Fig. 5 in Wu et al., 2022 for an example of its importance). Yet, the non-linear correlation pattern is still instrument-dependent, but COSMIC-1 and COSMIC-2 look nearly identical in this correlation pattern (Fig. 2), and METOP-A and METOP-B look nearly identical in this correlation pattern too (Fig. A1 and Fig. A2).

P7L136: Are all these radio- and dropsondes comparable in their accuracy/sensitivity?

This is a great question but I'm afraid we couldn't provide a knowledgeable answer. Please refer to the citations in Table 2 if you are interested in details of the data quality evaluation for each campaign.

Figure 3: add that the legend numbers are the total number of sondes (I assume this) in campaign.

Added in the caption. And Fig. 3 (now Fig. 4) had been replotted into three sub-panels to show the sonde locations more clearly.

P9L171: "Fig 4, but were otherwise look..." Not sure what this means. Maybe without the "were"?

We meant to say the correlation patterns look extremely similar if we break down Fig. 4 (now Fig. 6) into 6 panels for each pressure level separately.

P9/L176: Ducting would lead to a disappearance of the signal, as it is bended towards the surface. So you need to have strong gradients, maybe getting close to ducting conditions (albeit I doubt that, as we do see signals below -100km regularly, and even further than -200km, likely caused by close to ducting conditions).

We believe ducting could be one of the several causes of the retrievability of MABL moisture information using reemerged SNR signal (Sokolovskiy et al., 2014). A schematics is shown below for the ducting condition:

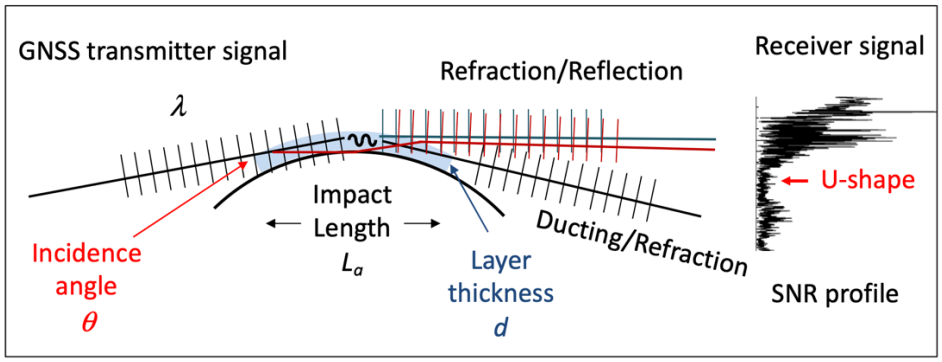


Fig. 3 from Wu et al. (2022), showing GNSS-RO wave optics as refracted by an atmospheric layer and reflected by the surface in multi-path interference.

P9L182: So you are primarily using only GPS, from COSMIC1 here? What about COSMIC2, and the GLONASS constellation it observes? And maybe make the naming consistent, either COSMIC-1 or COSMIC1.

Here (Section 3.1) includes the standard ML independent testing. Since COSMIC-2 data were not used for training, we shouldn't use that for independent testing per ML standard procedure. The ML model trained on COSMIC-1 data is then applied to the COSMIC-2 SNR profiles to test the robustness of transferred learning (which is robust based on independent radiosonde validation).

Thanks for spotting the inconsistency of abbreviations. We have now use COSMIC-1 and COSMIC-2 throughout the manuscript for consistency.

Figure 4: Suggest to add correlation coeff./std dev to plot.

Added.

Figure 5 a,c: (1) why are these going up to about 22g/kg, and Figure 4 only up to about 18g/kg? Is there some filtering on-going? (2) Keep using the same units, either g/kg or kg/kg in the manuscript.

Thanks for spotting this! During data pre-processing for ML training, every variable is normalized, and we divided every specific humidity value by 22 g/kg to make sure the input values are between 0 – 1 because the largest value among all data points is 22 g/kg. Since Fig. 4 and Fig. 5 (now Fig. 6 and Fig. 7) uses the 3 months of 2018 data for independent testing, the original plotting code for Fig. 4 uses "q_max" as the axis boundary, and the largest specific humidity value for these three months are 17.8 g/kg, so that's why Fig. 4 (now Fig. 6) has a different axis range. We now have updated Fig. 4

(now Fig. 6) with the same axis range with Fig. 7 (previous Fig. 5), and changed the colorbar units to g/kg in Fig. 7 (previous Fig. 5) to be consistent with other figures.

P11L193: Estimating the SNR contributions to the total uncertainty should be further elaborated, e.g. what contributors do you expect? Horizontal inhomogeneity, ducting, reflections, etc?

This is a very sharp question. The answer is we don't know what exact physical mechanisms cause the uncertainty, and by how much. As summarized in Section 2.3, a couple of atmospheric conditions could co-exist to enable deepSNR retrievals, which we don't have a good physical model to simulate so far. However, as a common sense, we do know when the signal is very weak (i.e., small SNR values), it is hard for any models (physical or machine learning) to separate the signal from the background noise. Hence, retrievals based on small SNR values should come along with a relatively large uncertainty. This is what Fig. 6 (now Fig. 8) and related text tells us.

Figure 6: Please mention the percentage you remove with the 50% uncertainty filter. Maybe also include some further info here on the distribution, e.g. percentage of data below 20% uncertainty?

There are only 2.23% of independent testing data for COSMIC-1 that has uncertainty beyond 50%. If we use 20% uncertainty threshold, about 16.14% retrievals do not pass the uncertainty filter.

P12L209: Please be consistent, if you use only GPS, then don't use GNSS sometimes. 5 lines below, you use GPS again.

We should use GNSS because we use both GPS and GLONASS. The inconsistency has been corrected.

P12L221: Is Johnston using GPS only? Not too clear from the paper.

GNSS.

Figure 7: (1) there are several L2 retrieval symbols with no SNR symbol next to it. Is this really a fair comparison? What happens if only collocated data is used?; (2) I assume these are g/kg units shown?; (3) these "bigger symbols", are those the vertical lines that are sometimes visible, sometimes not? Please improve readability here. (4) again, consider adding some further info, e.g. number of data points, overall correlation stats, e.g. bias, std dev, correlation coefficient.

Do you mean some orange samples from ARRecon_2020 campaign? The fact that filled orange triangles (SNR-ML retrievals) are closer to the 1:1 line than open triangles (RO product) indicates that SNR-ML method works better for the atmospheric river scenario in the MABL. Some other studies used the ARRecon, GNSS-RO and ERA5 collocation data also reported refractivity negative bias in GNSS-RO (COSMIC-2, SPIRES) and dry bias in ERA5 in MABL (Murphy and Haase, 2022).

Murphy, M. J., J. S. Haase: Evaluation of GNSS Radio Occultation Profiles in the Vicinity of Atmospheric Rivers, Atmosphere, **https://doi.org/10.3390/atmos13091495**.

We now modified Fig. 7 (now Fig. 9) to incorporate suggestions from both reviewers. Please note that both Fig. 9 and Fig. 10 are provided in huge image size with very high dpi resolution considering the rich information that has to be contained in these two figures. Interested readers can enlarge the plots to check more details.

Figure 9: I am unsure if this comparison is fair towards ERA5! There are e.g. 549 sondes for Magic (Figure 3), but about 10% or less are shown for L2 and SNR, while I guess ERA5 shows the correlations for all. Same with the other campaigns, or rather even worse, EUREC4A has 1349 sondes, but only 53 are shown for SNR.

We are sorry for the misunderstanding here. We had claimed in the Fig. 9 (now Fig. 11) caption that the correlations were made for the sub-samples for ERA-5 where a SNR-radiosonde collocation is identified, and that's why we didn't write the number of samples above ERA-5 violins. It is indeed very hard to find a radiosonde-GNSS collocation. Even though we added Metop-A and Metop-B predictions in this revised version, it only added two more collocation samples among all 6 campaigns.

Most importantly, we identified two bugs in our plotting codes for Fig. 9 (now Fig. 11). For ERA-5, we didn't filter out missing values when compute the correlation, causing the correlation coefficients to be very small. We have fixed the bug and now all three datasets look quite comparable in the updated Fig. 11. We also found that the success rate for both SNR-ML and wetPrf methods are pressure-level dependent. Hence, we now provide a new Table 3 to illustrate the robust success rate using the SNR-ML method. The entire Section 3.3 has been largely rewritten with the new updated results.

P16/L256: What is the reason to exclude sea ice (above it talks of AIRS issues)? The MERRA model? Or the L2 retrieval? The SNR one?

Because we worry sea ice induced reflectometry signal could contaminate our deepSNR based retrievals.

Figure 10: Any reason why these plots do not cover the same respective area? Are you comparing the same numbers/locations/times here? Is the SNR based retrieval maybe picking primarily wet conditions and having trouble with dryer situations, thus leading to this higher humidity? How is topography taken care of, e.g. for Norway in SNR retrievals?

Because we used 2.5 X 2.5 degree resolution for MERRA-2, but have to use 4 X 4 degree resolution due to the sparsity of GNSS data during the boreal winter season.

P17L258-262: I think these last sentences would need much further assessment, e.g., you mention topography limitations when discussing Figure 5, I am still not sure whether you are comparing like with like, the SNR retrieval has several quality controls implemented (that seem to filter out dryer observations, e.g. discussion with Figure 6), etc.

Are you saying that because we filtered out weak SNR profiles, our retrievals tend to biased toward wet-conditions? If that is the case, how to explain the dry-bias in the southern ocean for those very small values? Another evidence that this SNR-ML retrieval product is not inherently wet-biased is the comparison between Fig. 10 and Fig. 11 (now Fig. 12 and Fig. 13). In the Antarctic region, MERRA-2 doesn't show systematic dry bias.

P17/L265: Again, topography might have an impact on your sampling, also the dry quality control filtering, etc. Thus SNR needs to show that it is doing a better job against an independent data set. Re-analysis models might just also get it wrong there (or the sampling is not consistent).

True. I agree with you.

Figure 12: Again, is this comparing the same data at the same location and time? Or just taking all available SNR retrievals in the area, and comparing it to all ERA5 data in that box?

We took all available retrievals during Dec.-Feb. period listed in Table 1 in a given region to construct the diurnal cycle. Since these campaigns were all carried in the pre-COSMIC era, the campaign data cannot be used to verify

Table A1: What is the ascending / descending orbit? Is that of any relevance? Or is this the occultation setting / rising?

You are correct. It's occultation setting and rising. For typical satellite community users, ascending/descending is a more familiar terminology, but we had added the clarification in the Table caption now. The caption now reads:

Table A1. Summary of GNSS-RO instrument noise (σ) used in this work. "Ascending" and "descending" here are equivalent meaning of occultation rising and setting.

Figure A1: Why are you using different labels and titles here, compared to the COSMIC one you show? And is this really Metop, as the caption talks of COSMIC? And why is this the only Metop result shown?

We have reedited Fig.2, Fig. A1 and Fig. A2 to show the correlation situation for COSMIC-1, METOP-A and METOP-B for SNR and sigma^2_SNR. The reasons to show the three is to demonstrate that (1) the correlation is not highly non-linear, so a ML model is better to handle this rather than a single-level regression or multi-variable regression model; (2) the correlation pattern is highly instrument dependent, so we need to build individual ML models for each mission separately.

I noted that the other reviewer was very thorough also regarding typos, textual improvements, thus I mostly did not include those here. But a thorough re-editing is needed!
Thanks again for your constructive suggestions. If you have time, please feel free to also read the responses to reviewer #1's questions. With compiling the suggestions from you two, we hope the manuscript now is improved for publication. Thanks!