A Machine-learning Based Marine Atmosphere ~~Planetary~~ Boundary Layer (MABL) Moisture Profile Retrieval Product from GNSS-RO Deep Refraction Signals

Jie Gong[1,*], Dong L. Wu[1], Michelle Badalov[2], Manisha Ganeshan[3,1], and Minghua Zheng[4]

[1]Climate and Radiation Lab, NASA Goddard Space Flight Center
[2]Dept. of Computer Science, Univ. of Maryland College Park
[3]GESTAR-II/Morgan State University
[4]University of California at San Diego

Correspondence: Jie Gong (Jie.Gong@nasa.gov)

Abstract. Marine Atmosphere ~~Planetary~~ Boundary layer (MABL) water vapor amount and gradient impact the global energy transport through directly affecting the sensible and latent heat exchange between the ocean and atmosphere. Yet, it is a well-known challenge for satellite remote sensing to profile MABL water vapor, especially when cloud or sharp vertical gradient of water vapor are present. Wu et al. (2022) identified good correlations between Global Navigation Satellite System (GNSS) deep refraction signal-to-noise-ratio (SNR) signal and the global MABL water vapor specific humidity when the radio occultation (RO) signal is ducted by the moist planetary boundary layer (PBL), and they laid out the underlying physical mechanisms to explain such a correlation. In this work, we apply a machine-learning/artificial intelligence (ML/AI) technique to ~~realize~~ demonstrate the feasibility for ~~pixel-level~~ profile-by-profile MABL water vapor ~~profiling~~ retrieval using the SNR signals. Three convolutional neural network (CNN) models are trained using ~~20 months of~~ multi-months of global collocated hourly ERA-5 reanalysis and COSMIC-1, METOP-A and METOP-B 1 Hz SNR observations between 975 – 850 hPa with 25 hPa vertical resolution. The COSMIC-1 ML model is then applied to both COSMIC-1 and COSMIC-2 in other time ranges for independent retrieval and validation. Monte Carlo Dropout method was employed for the uncertainty estimation. Comparison against multiple field campaign radiosonde/dropsonde observations globally suggests SNR-ML method retrieved water vapor consistently outperforms ~~ERA-5 reanalysis and~~ the ~~Level-2~~ wetPrf/wetPf2 standard retrieval product at all six pressure levels between 975 hPa and 850 hPa, and either outperforms or achieves similar performance against ERA-5, indicating real and useful information is gained from the SNR signal albeit training was performed against the reanalysis. ~~The only exception is in the deep tropics where the fundamental assumption for SNR-ML method to work is invalidated frequently by interactions among ocean surface, MABL and shallow convections.~~ Climatology and diurnal cycle of MABL structure constructed from the SNR-ML technique is studied and compared to the reanalysis. Disparities of climatology suggest ERA-5 may systematically produces dry biases at high-latitudes, and wet biases in marine stratocumulus regions. The diurnal cycle amplitudes are too weak and off-phase in ERA-5, especially in Arctic and stratocumulus regions. These areas are particularly prone to PBL processes where this GNSS SNR-ML water vapor product may contribute the most.

1 Introduction

As a key component of Earth's lower atmosphere, the planetary boundary layer (PBL) water vapor plays a pivotal role in Earth's energy budget, exerting profound influence on weather and climate processes. It is an essential factor of the Earth's energy budget, influencing radiative forcing and consequently climate variability and long-term changes. Furthermore, PBL water vapor is instrumental in modulating local and regional weather patterns by affecting cloud formation, precipitation and temperature. Therefore, study of PBL water vapor stands as a vital element in advancing our comprehension of Earth's atmosphere and its broader implications for our planet's climate system.

70% of the Earth's surface is covered by water. The sensible and latent heat exchange between ocean boundary and the marine atmosphere boundary layer (MABL) happens at different spatial and temporal scales, which is determined by not only ocean surface properties (e.g., wind speed, sea surface temperature) but also MABL thermodynamic structures. For example, under the context of ~~polar proneness~~ proneness of polar area to the climate change, Boisvert et al. (2015) found Arctic PBL humidity and temperature biases in the reanalysis are the major error sources for the evaporation estimation compared to satellite observations. Cloud-cimate feedback is another motivation highlighted by NASA's PBL incubation study (Teixeira et al. (2021)). As another example, Milan et al. (2019) found strong correlation between MABL cloud top height and below-cloud water vapor amount using two joint satellite retrieval products.

Data sparsity is a critical problem for advancing MABL science. Satellite remote sensing undoubtedly provides the best solution in terms of global coverage, but it is very difficult to retrieve MABL WV and its vertical distribution when cloud or sea ice are present. When clouds are present in the scene, emissions from clouds often overwhelms the emission signal from the MABL water vapor and prevents passive instruments sensing the below-cloud atmosphere. When sea ice is present ~~in the scene~~, scattering or surface emission from the sea ice are often inseparable from water vapor emission signals and distort the retrieval result. Taking the aforementioned two research as examples, Boisvert et al. (2015) uses Level 2 AIRS water vapor and temperature retrieval products, which are only available for clear or partially-cloudy sky situations, so it inherently contains a sampling bias. Milan et al. (2019) derived MABL total WV amount from subtracting MODIS above-cloud water vapor from AMSU-A total column water vapor, which ~~couldn't be used to gain~~ still lacks the vertical information of WV in the MABL.

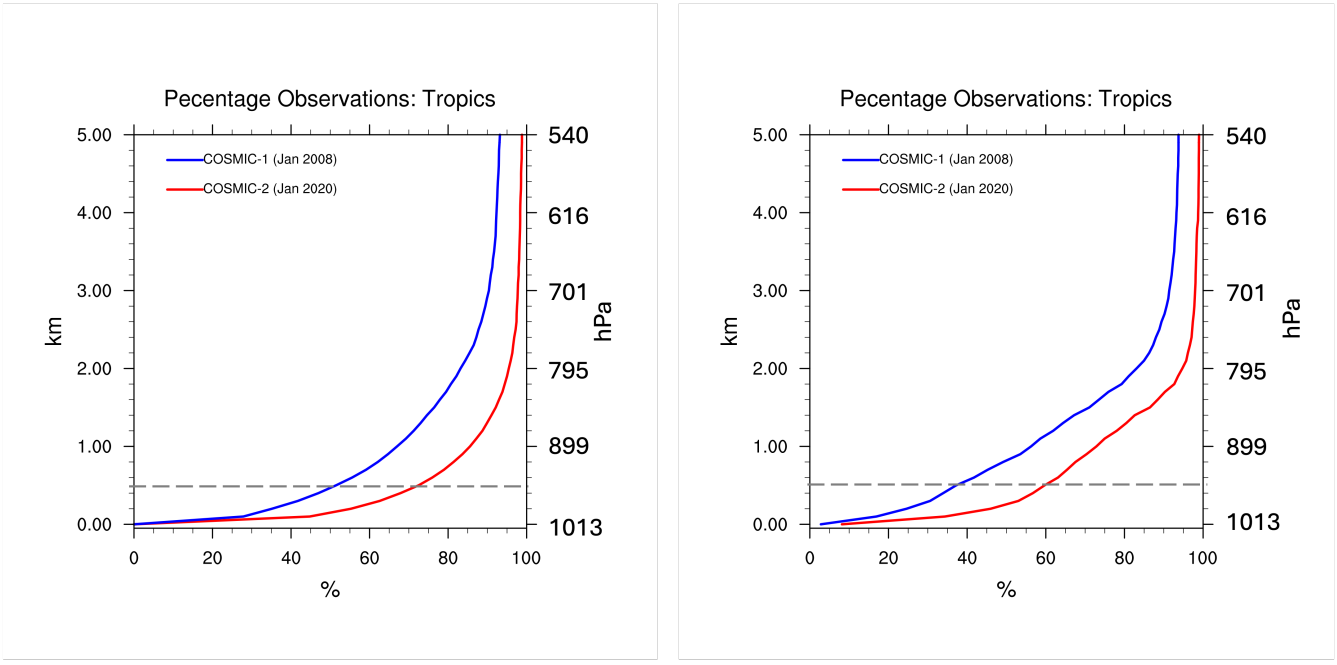
Using low-frequeny ~~MW~~ microwave L-band to transmit signals along the limb path, the Global Navigation Satellite System (GNSS) satellite overcomes the two above difficulties and provide ~~superb~~ high vertical resolution of the MABL water vapor under all-sky conditions (100 - 200 m). GNSS Radio Occultation (GNSS-RO) retrieves temperature and water vapor profiles using the 1D-Var approach routinely from the Level 2 bending angle product (referred as "standard L2 product" or "operational L2 product" hereafter), the latter of which is used operationally in numerical weather data assimilation systems to improve weather forecasts (e.g.,Kuo et al. (2000)). Because of the rapid growth of SmallSat/CubeSat constellations from both the commercial and non-profit sectors, GNSS-RO technique provides a promising future for the needed global spatial-temporal sampling of MABL WV and its variability. Like other limb sounders, the disadvantage of GNSS-RO is its relatively coarse horizontal resolution (several hundred kilometers) that smears out horizontally inhomgeneous signals. This is typically not a

big ~~conern~~ concern in MABL as vertical gradient is much sharper than horizontal gradient and harder to characterize ~~if not using in-situ measurements (e.g., shipborne radiosonde)~~.

However, GNSS-RO WV retrieval profiles have excessively high failure rate in the MABL. That is because the GNSS-RO signal-to-noise ratio (SNR) decreases with decreasing altitude ~~height~~ due to the atmospheric defocusing effect, and the Level-2 radio-occultation (RO) signal hence often does not meet the SNR threshold near the surface. ~~refractivity retrievals which require a high SNR drops subsequently in the MABL as the signal becomes noisy~~. As a result, ~~the GNSS-RO water vapor retrievals that rely on the refractivity profiles often fail to converge~~ the GNSS-RO 1D-Var based retrievals fail in the MABL due to weak RO signal. Fig. 1 gives an example of the success statistics (%) as a function of height for temperature (Fig. 1a) and water vapor (Fig. 1b) over the tropical ocean. Using $0.5km$ and $1km$ above the ocean surface as the reference lines, we can see although the COSMIC-2 (Constellation Observing System for Meteorology, Ionosphere, and Climate-2) has significantly improved its SNR compared to its predecessor COSMIC-1, the success rate is about 60% at $0.5km$ and slightly over 70% at $1km$ for the GNSS-RO WV retrieval, while this number is only ~~25% and 60%~~ ~~40% and 55%~~ for COSMIC-1 at respective altitudes. The low SNR widely exists for commercial GNSS satellites especially in the lowest 500 m above the surface (Ganeshan et al. (2024)). Moreover, even passed the SNR threshold, some bending angle profiles are significantly biased in the PBL when ducting happens because the refractivity index becomes negative, which leads to biases in the operational water vapor retrievals (Feng et al. (2020)).

Wu et al. (2022) found out that the Level-1B deep SNR from the straight-line height (H_{SL}) is statistically significantly correlated with the MABL water vapor amount in the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) after averaging over a month at $2.5°X2.5°$ grid resolution. The averaging is necessary to effectively beat down the random noise. This paper attributed such a positive correlation to the strong refraction from a horizontally stratiform and dynamically quiet MABL water vapor layer that acts to enhance the SNR amplitude at deep H_{SL} through ducting and diffraction/interference (a summary recapitulation of the physical mechanism can be found in Section 2.3). Some caveats of this work limit its application to weather phenomena. First, it builds upon a single level regression statistics, the correlation coefficient of which was found the highest at $H_{SL} = -100km$ in the tropics, and $H_{SL} = -80km$ at high latitudes. Hence, any simple linear regression-based retrieval algorithm will ~~understandably~~ suffer from arbitrary latitudinal discontinuities. As a matter of fact, SNR at different H_{SL} levels are found correlated with MABL water vapor with different signs and magnitudes (e.g., Fig. 2), which should be used together to enhance the information content. Secondly, the robust relationship is only found for monthly averages in Wu et al. (2022), because the profile-by-profile noise is usually too high ~~for~~ to yield~~ing~~ a meaningful retrieval from SNR, and only through averaging large amount of profiles can the noise be lowered down to the level where signal stands out. These are all caveats of traditional statistical approach. Machine learning approach, however, is suitable at picking up weak signals through large amount of training data. As such, the scopes of this paper are to demonstrate the feasibility of using the ML method to extract MABL WV information from the GNSS SNR signals, and to demonstrate the scientific value of this new product over the existing ~~Level-2~~ operational water vapor retrievals.

Artificial Intelligence/Machine learning (AI/ML) ~~technique~~ applications in remote sensing field is trending in the past decade. ~~It is more and more appreciated by the satellite remote sensing field in recent years.~~ It has been increasingly used



(a) AtmPrf success rate

(b) WetPf2 success rate

Figure 1. Level 2 atmPrf (temperature) and wetPf2 (water vapor) successful retrieval rate (%) as a function of height above sea-level from COSMIC-1 during January 2008 (blue) and COSMIC-2 during January 2020 (red). Success rate is calculated by dividing number of valid ~~Level-2~~ retrieval files over number of Level-1B file at a certain height. The grey dashed lines mark the reference at 0.5 km from the tropical ocean surface.

in remote sensing fields in recent years. Traditionally physics-based radiative transfer (RT) theories and modelings are used to link the remote sensing measurements (e.g., GNSS radio occultation signal) to the physical quantities (e.g., temperature and water vapor profiles). They are often highly non-linear, computationally expensive and involve many explicit or embedded ssumptions/simplifications, which may or may not propagate properly into part of the retrieval errors eventually ~~subsequently~~. Given the fact that satellite measurements usually contain large amount of data, and the association is highly non-linear between the measurement space and the physical space, the retrieval process becomes an ideal testbed for ML capabilities. Some pioneer works had attempted this approach to retrieve PBL atmosphere profiles and achieved notable success. For example, Ye et al. (2021) used the routine radiosonde measurement at a Atmospheric Radiation Measurement (ARM) site as the ground truth to train a ground-based infrared spectrometer to predict the PBL height. The capability is limited to only the stations where both observations are routinely available. Milestein and Blackwell (2016) employed a neural network (NN) framework on retrieving the temperature and water vapor profiles from the spaceborne Atmospheric Infrared Sounder (AIRS) observations (AIRS Version 7 product). The training "truth" was from the ECMWF analysis fields. It is worth mentioning that Milstein (2022), as a follow-up work, pointed out the ML-only retrieval framework tends to smooth out sharp gradient features in prox-

imity to the PBL top. To mitigate this caveat, Milstein et al. (2023) employs the 3D deep neural network training on the AIRS granule image against ERA-5 reanalysis that helps PBL height recognition from passive imagers.

In this paper, we will ~~thoroughly~~ explore the ML capability at retrieving the MABL WV information from the deep SNR signal at profile-by-profile basis (i.e., Level-2 standard). Section 2 introduce the training and validation datasets as well as the model structure; Section 3 presents the retrieval results and independent validation; Section 4 will expand the discussion to the usage of this product in studying MABL water vapor climatology and diurnal variabilities; Section 5 summarizes the major findings and shortcomings of the current work that may be improved in the future.

2 Data and Model

This section introduces the training, validation and independent validation datasets, as well as the ML model architecture and the underlying physical foundations for the ML technique to work upon.

2.1 Training and Validation Datasets

The definition of SNR follows Wu et al. (2022) which uses the normalized SNR (S_{RO}):

$$S_{RO} = (SNR - \sigma)/(SNR_0 - \sigma) \tag{1}$$

$$\sigma_S^2 = VAR(S_{RO} - \overline{S_{RO}}) \tag{2}$$

SNR_0 is the free atmosphere SNR. In practice, we use averaged SNR between 35 and 65 km altitude range as the SNR_0, and any profile with $SNR_0 < 200$ or $\sigma^2_{SNR_0} > 0.05$ is considered "low-signal" and is filtered out. σ is the instrument-specific noise determined for each individual instrument from very deep H_{SL}. The value for σ used in this work is an updated version from Table A1 in Wu et al. (2022) and shown in the Appendix A (Table B1). Wu et al. (2022) also found an instrument-dependent shift of the mean S_{RO} profile as a function H_{SL}. Luckily, such an issue can be resolved to use the excess phase at L1 (ϕ_{L1}) as the vertical coordinate. In practice, the raw calculated S_{RO} and σ_S^2 are mapped to a fixed 52-level $Log_{10}(\phi_{L1})$ vertical grid. It is roughly linearly correlated with H_{SL}. The value for the vertical grid is listed in Table ~~A2~~B1 in the Appendix A. 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_S^2 values.

The ERA-5 reanalysis is so far the best global reanalysis dataset in terms of PBL water vapor amount and distribution. Johnston et al. (2021) compared specific humidity from ERA-5 and MERRA-2 reanalysis against collocated and coincident GNSS-RO ~~Level-2~~ wetPf2 specific humidity retrieval profiles, and found ERA-5 outperforms MERRA-2 everywhere in the PBL. They both exhibit consistent dry biases with larger bias from mid-high latitudes. However, ERA-5 percentage bias is roughly 1/2 of that from the MERRA-2 reanalysis in the PBL and tropopause regions. Given that many previous works used ERA-5 reanalysis or ECMWF analysis for training or validating the satellite retrievals for water vapors (e.g., Milestein and Blackwell (2016), Milstein et al. (2023)), especially some recent ones using it as the standard to evaluate recent GNSS-RO missions (e.g., Chang et al. (2022), Zhran (2023), Ganeshan et al. (2024)) , it is well justified to use ERA-5 hourly reanalysis

as the "training" dataset to create a large sample globally. However, it is also warned in Johnston et al. (2021) that GNSS-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).)

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. 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.

Table 1. Training, testing and prediction periods

Training (90% and 10% random-splitting)	COSMIC-1	2012.01-2012.12, 2016.01-2016.03, 2017.01-2017.03
	METOP-A	2017.01-2017.03
	METOP-B	2017.01-2017.03
Testing	COSMIC-1	2018.01-2018.03
	METOP-A	2018.01-2018.03
	METOP-B	2018.01-2018.03
Prediction	COSMIC-1	2012.01-2012.12, 2013.01-2013.12, 2016.01-2016.03, 2017.01-2017.03, 2018.01-2018.03
	COSMIC-2	2020.01-2020.12
	METOP-A	2012.01-2011.12, 2013.01-2013.12
	METOP-B	2012.01-2011.12, 2013.01-2013.12

Fig. 2 elucidates the linear correlation between COSMIC-1 S_{RO} at each of the 52 levels and ERA-5 specific humidity at 975, 950, 925, 900, 875 and 850 hPa over global ocean. The largest positive correlations are found around Level #40 to Level #45, which roughly correspond to $H_{SL} = -100km$ to $-80km$ (Table ~~A2~~B1). Based on the monthly averages, Wu et al. (2022) found the highest correlation at $H_{SL} = -100km$ in the tropics and at $H_{SL} = -80km$ for the polar regions, which is consistent with our profile-by-profile correlation as well. But Fig. 2 also shows positive or negative correlations at different $Log_{10}(\phi_{L1})$ levels, which impede methods like multi-variable regression from working. $\sigma_S^2 NR$ also exhibits non-linear patterns with ~~generally~~slightly weaker correlations with MABL water vapor that are opposite in sign compared to that

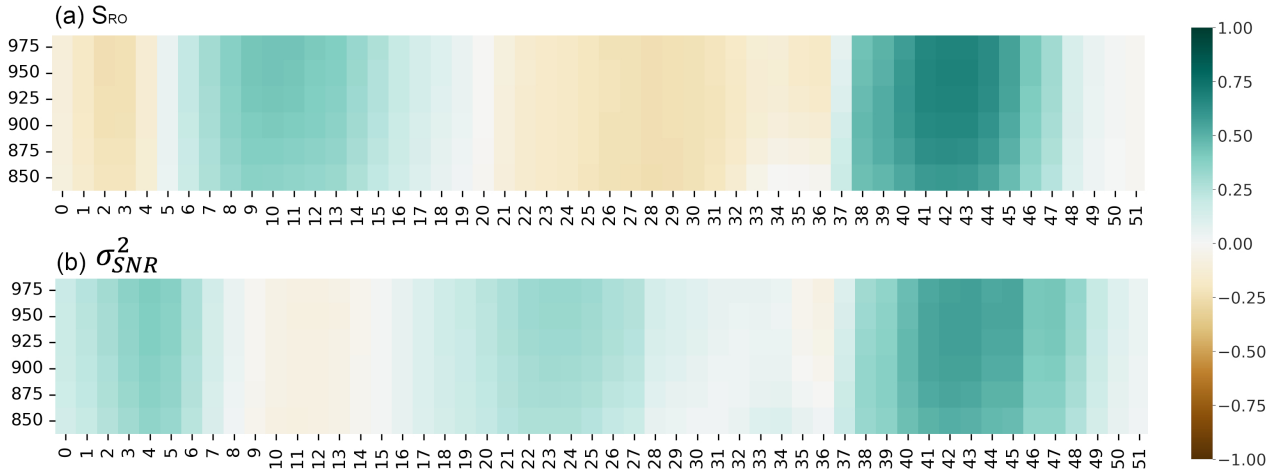


Figure 2. Correlation between collocated ERA5 specific humidity at 975 - 850 hPa and S_{RO} (top) and σ^2_{SNR} (bottom) at various excess phase levels from the training COSMIC-1-ERA5 dataset). Only grid indices are shown in the axis titles, and the corresponding $Log_{10}(\phi_{L1})$ values can be found in Table ~~A2~~B1

.

of S_{RO}. It is worth noting that these relationships are also instrument dependent, as can be clearly seen in the S_{RO} cross-correlation for Metop-A and Metop-B in the Appendix Fig. A1 and A2. Considering the instrument-dependent correlation patterns, three ML models are developed separately for COSMIC, METOP-A and METOP-B satellites. For the COSMIC series, we observed similar pattern from COSMIC-2 compared to Fig. 2 after downsampling the frequency to 1 Hz (not shown). Therefore, the ML model developed using COSMIC-1 observations is applied directly to the downsampled COSMIC-2 SNR observations. Through this practice we can also test the transferred learning among similar satellite series for the hope of stitching them together for longer record in the future research.

The correlation holds with the same slope at piece-wise level using individual profiles. For example, between SNR at $H_{SL} = -100km$ and ERA-5 specific humidity at $950hPa$, Wu et al. (2022) observed the near linear correlation with monthly averaged and gridded data, while we can see that the same slope is preserved at ~~Level-2 scale~~ textcolorredprofile-by-profile level in Fig. 3. While this robust correlation proves that developing a Level-2 MABL specific humidity retrieval product using SNR profiles is feasible, the discernible larger noise at individual profile level versus month averages (Fig. 3d) suggests it is a challenging task. ML method is hence introduced to tackle this highly complex regressional problem.

GNSS-RO operational water vapor retrieval product provided by the University Corporation for Atmospheric Research (UCAR) is employed to evaluate the quality of the SNR-ML retrievals. This operational product is called "wetPf2". Compared to an old 2013 processed "wetPrf" version, "wetPf2" has better penetration depth (Wee et al. (2022)) and is used for constructing Fig. 1, but "wetPrf" product is used for the MAGIC campaign comparison because of data availability constraints at the time when this research was conducted. We've compared the success rate in the MABL between wetPrf and wetPf2 during

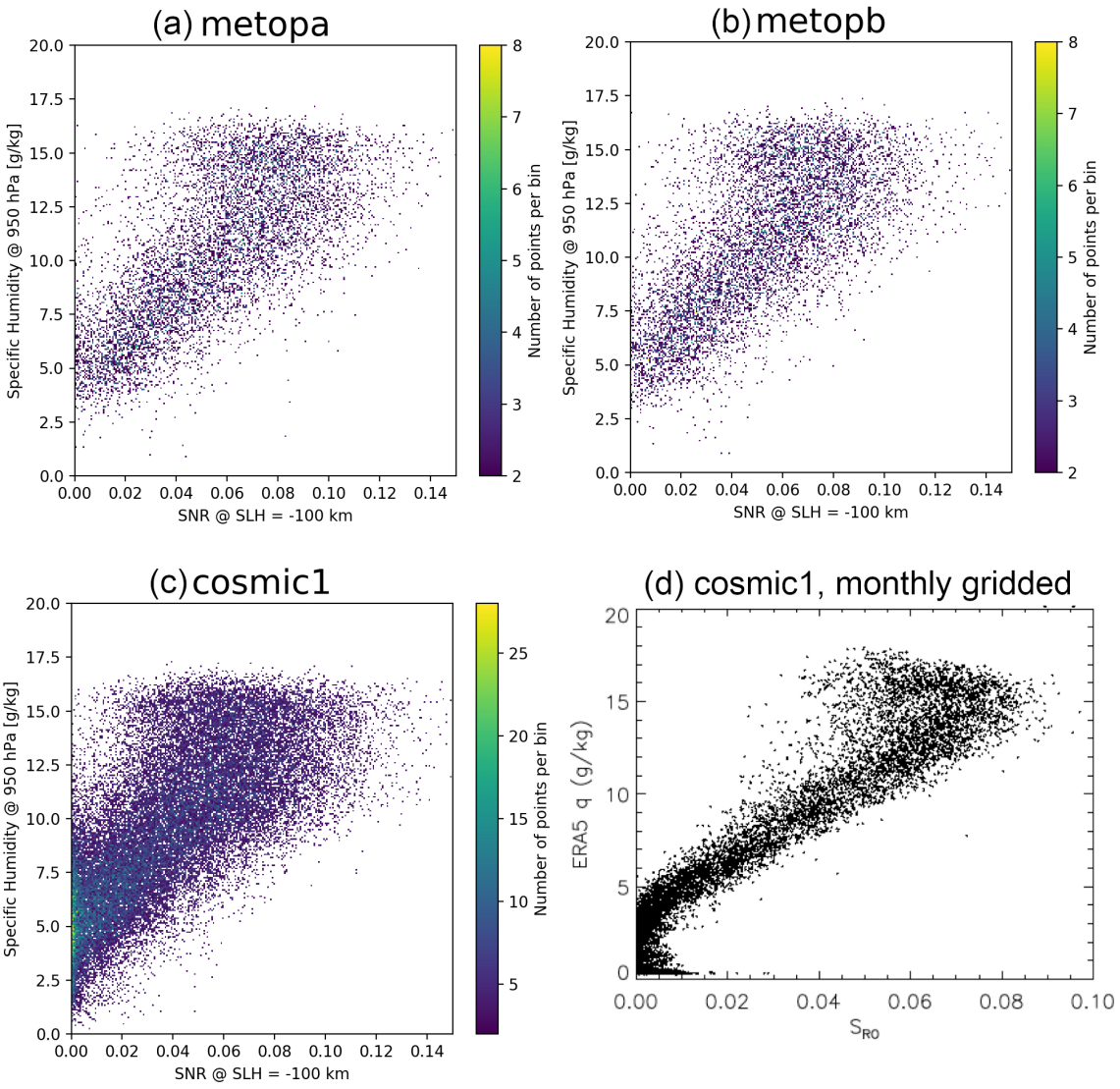


Figure 3. Density plots of the SNR-specific humidity relationship for (a) Metop-A, (b) Metop-B, (c) COSMIC-1 constructed from the entire training dataset between 45 and 45. The SNR value is taken from $H_{SL} = -100km$ while the specific humidity value is taken at $950hPa$. Fig. 9c from Wu et al. (2022) is reproduced here as (d) to demonstrate that the same relationship with the same slope holds at individual profile level.

Jan. 2008 (Fig. 1b) and only found very marginal improvements for COSMIC-1. Note that the key Level-2 profile to enable
170 the 1D-VAR retrieval used by the wetPrf/wetPf2 product is the bending angle, which is assimilated in the ERA-5 reanalysis. Therefore, this is not an independent evaluation dataset. The purpose of this comparison is to identify the merits and caveats of the SNR-ML retrievals against an existing mature product.

In addition to the independent testing which is a standard procedure for ML/AI training and evaluation against the wetPrf/wetPf2 operational product, a handful of shipborne radiosonde campaigns and airborne dropsonde campaigns data are collected for further independent assessment. The campaign names, location and total number of valid profiles are presented in Fig. 4 and Table 2. We can see from the summary of weather scenarios during each campaign that this independent validation dataset comprehensively covers major marine weather regimes from extremely dry Southern Ocean (MARCUS), mid-latitude stratocumulus region (MAGIC), tropical trade cumulus region (EUREC4A, ATOMIC), to episodically wet atmospheric river events (ARRecon). This exercise is critical for assessing the quality of ERA-5, Level 2 retrieval, and Level 1 SNR-based retrieval under different weather scenarios. Moreover, as the ML model trained solely on COSMIC-1 SNR data are then applied to the COSMIC-2 data, the independent validation using the three campaigns in 2020 (ARRecon-2020, EUREC4A and ATOMIC) provides some solid evidences to evaluate the robustness of the "transferred learning".

Table 2. Campaign Information

Campaign Name	Period used for validation	Location	Weather Regime	Type	Reference
MARCUS	2017.11-2018.03	Southern Ocean	Mixed-Phase PBL cloud	Radiosonde	Evan et al. (2022)
ATOMIC	2020.01-2020.02	Tropical North Atlantic	Tropical trade wind zone	Radiosonde and Dropsonde	George et al. (2021)
EUREC4A	2020.01-2020.02	Tropical North Atlantic	Tropical trade wind zone	Radiosonde	Stephan et al. (2021)
MAGIC	2012.10-2013.09	Eastern North Pacific Ocean	Subtropical MABL	Radiosonde	Evan et al. (2022)
ARRecon	2018.02; 2020.01-2020.02	Northeast Pacific off the coast of California	Atmospheric River	Dropsonde	Zheng et al. (2024)

2.2 Machine Learning Model Selection

The Convolutional Neural Network (CNN) model (LeCun et al. (2015)) is chosen as our regression ML model. The model internal architecture is illustrated in Fig. 5. There are a total of 109 input parameters, including one dimensional array of S_{R0} of 52 elements, one dimensional array of σ_S^2 of 52 elements, both interpolated to a fixed excess phase grid (Table B1), and latitude, longitude, month and Rising/Setting flag. The output parameters are specific humidity at 6 pressure level between 975 and 850 hPa with cadence of 25 hPa.

Compared to some earlier ~~old-fashioned~~ ML models (e.g., random forest, gradient boosting), CNN learns also the vertical cross-correlation within the 52-layer input SNR profiles, as well as within the targeted 6-layers of specific humidity profiles.

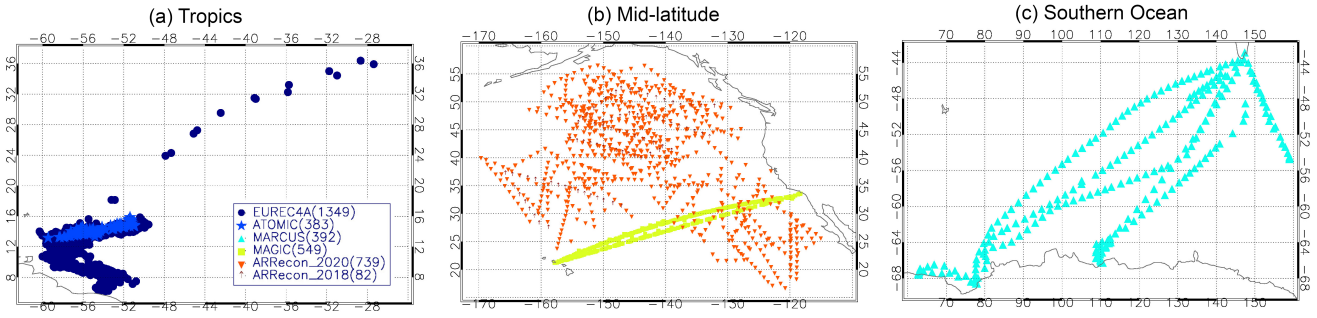


Figure 4. Maps for radiosonde/dropsonde locations from different shipborne or airborne campaigns in (a) tropics; (b) mid-latitudes; (c) southern ocean. Detailed campaign information can be found in Table 2. The total number of valid radiosonde/dropsonde profiles are listed in the parentheses in the legends.

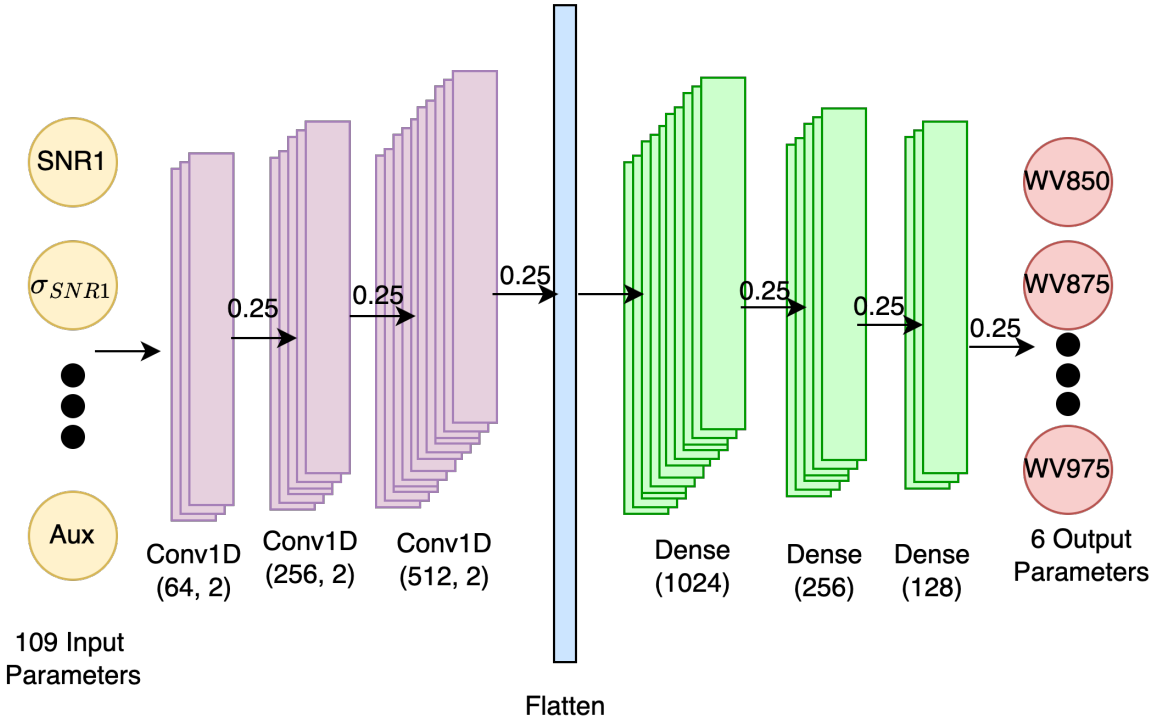


Figure 5. Convolutional Neural Network (CNN) model internal structure for this work. The numbers on top of the right-pointing arrow is the Monte Carlo dropout value applied between each layer. Numbers inside the parentheses of Conv1D layer indicate filter size and pool size, while numbers inside the parentheses of the Dense layer indicate number of the fully-connected nodes. The training takes 100 epochs, which suffice the needs for quick convergence.

We conducted a comprehensive search of best hyperparameters using the root-mean-square-error (RMSE) as the loss function. ~~The final selected architecture contains one 1-D convolutional block consisting of a Conv1d layer, a MaxPool1d layer, a BatchNorm1d, two dense layers, and a ReLU activation function. The Kernel size is selected to be 5, and a dropout layer of rate = 0.1 is added between layers. 100 epochs suffice the needs for quick convergence for both training and validation.~~

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.

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).

2.3 Underlying Physical Mechanisms

It is necessary to provide a summary of the underlying physics to emphasize the solid physical ground for this product, so readers would not misunderstand this as a pure statistics-based ad-hoc finding. The underlying physical mechanisms to explain the observed high-correlation between MABL water vapor and the GNSS SNR signal remain to be an active research area. Wu et al. (2022) articulated that the diffractive effect on the RO signal under the condition of limb sounding through a sharp MABL can extend the signal below the sharp edge of the obstacle with a limited depth. Both diffractive and refractive processes are required to happen along the radio wave propagation to produce the RO signal at deep H_{SL}. Another example (Sokolovskiy et al. (2024)) found enhancement of SNR when super-reflection happens. In reality, complex MABL can produce a mixed effect in the soundings from a combination of conditions that include normal bending, grazing reflection, super-reflection, ducting or diffraction (**??**). As a result, sophisticated physical radiation transfer models (e.g., radiohologram, canonical transform) can in principle be used but at the expense of high computational costs and hence impractical operationally. Moreover, the retrieval

itself is essentially still an under-constraint problem, which commonly occur for satellite retrievals and assumptions (no matter physically making sense or not) need to be made to fully constrain the physical model. As the quasi-linear relationship is preserved at profile-by-profile level with larger noise compared to the monthly gridded and smoothed data (Fig. 3), and the height-dependency of the regression coefficient is highly non-linear (Fig. 2), a ML model is simply the best choice to extract the signal.

3 Results

3.1 Retrieval Performance Evaluation

As the first comparison, Fig. 6 and Fig. 7 showcase the statistical closeness to the $1:1$ line and the resemblance of geographical distributions for the three independent testing months: January - March, 2018, for COSMIC-1. All 6 pressure levels are compiled together to make Fig. 6, but were otherwise look extremely similar if plotting layer-by-layer. The only deviation from $1:1$ line occurs at very small specific humidity values (ERA-5 specific humidity $< 1g/kg$), i.e., very dry conditions, normally occurs at high-latitudes.

Such a discrepancy reveals itself more clearly when we map out the percentage difference (Fig. 7b). The largest percentage differences indeed are shown at polar regions as well as near the coastal lines with SNR-retrieved humidity tends to be larger than ERA-5. Note that to satisfy ducting conditions in order to use SNR signal at deep H_{SL}, the surface is required to be flat. Therefore, the discrepancies around the coastal line are believed to be related to issues with SNR-ML retrievals when topography starts to play a role. However, as we will show later in Fig. 10, ERA-5 indeed shows consistent dry-bias at high-latitudes compared to independent radiosonde measurements. So SNR-ML retrieval might produce a closer-to-truth results as will be seen later as well. Moreover, one can visually discern discrepancies in the tropical deep convection/ITCZ zones with ERA-5 in general wetter than SNR-retrieved values. Such a discrepancy is not conspicuous in Fig. 7b simply because of the large value in the denominator. We will also show later that none of the three datasets we will evaluate (SNR-ML retrieval, GNSS-RO Level 2 retrieval and ERA-5 reanalysis) capture well the tropical MABL structures. For the SNR-ML method, it is probably because the ducting assumption is easily and frequently violated in the tropical MABL.

3.2 Uncertainty Quantification

Unfortunately, for the very dry conditions, SNR-retrieved specific humidity also inherently comes along with large uncertainties, as can be clearly seen in Fig. 8. The SNR signal is too weak in this situation to yield any robust retrievals even with powerful ML models. Although we still believe the SNR-ML retrievals might be "more correct" than ERA-5 for very dry conditions, in practice we mark any retrieval with greater than 50% uncertainty with a quality flag in the published product, and those data do not pass the the quality control to be used later in this paper for independent validation nor constructing the climatologies. This threashold only filters out about 2% of the data with very weak SNR signals. If we would apply a threshold of 20%, about 16% of data would be filtered out. In the later section when the diurnal cycle is compared using multi-year

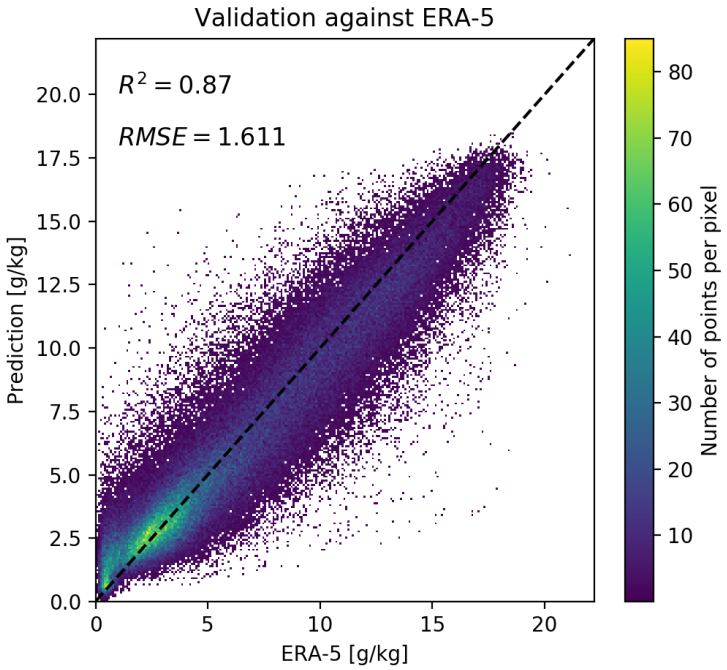


Figure 6. Heatmap for independent validation from January - March 2018 for COSMIC-1 combining all 6 levels together.

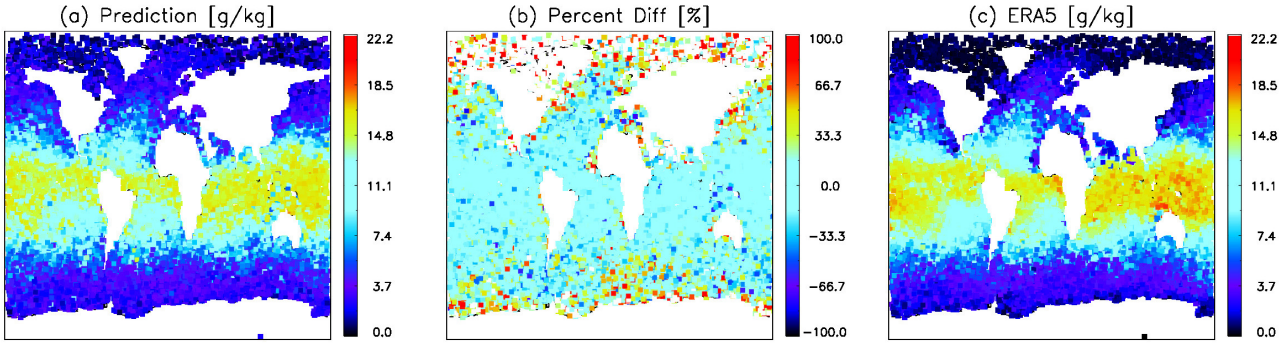


Figure 7. Geographic distribution of the (a) predicted values using COSMIC-1 SNR observations versus (c) ERA-5 validation values at 950 hPa for January - March, 2018. The middle panel is the percentage difference between (a) and (c).Only ERA-5 samples that collocate and coincident with COSMIC-1 SNR-ML retrievals are selected for this comparison

regional averaged data, we found that heavy-averaging effectively beat down the noise so to reveal a visible diurnal signal in the extremely dry polar region, whereas ERA-5 is essentially a fixed value (Fig. 14). We can also see from Fig. 8 that almost

all SNR-ML retrievals greater than $2g/kg$ passes the quality control. Readers should keep in mind that our current uncertainty estimation approach under-estimates the real uncertainty because it does not take into account SNR errors.

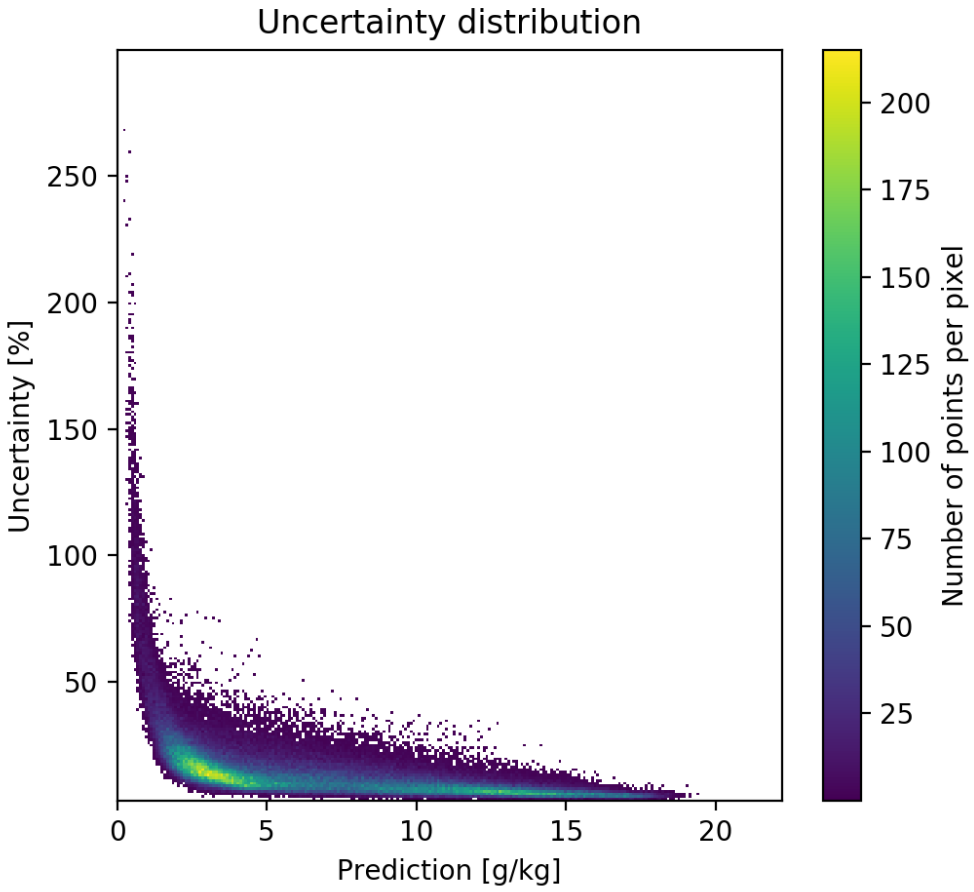


Figure 8. Percentage uncertainty distribution as a function of predicted value.

3.3 Comparison to Independent Radiosondes

In order to find collocation samples in every campaign, the collocation criteria are slightly different given the consideration of (1) the abundance of radiosonde/dropsonde profiles; (2) the typical spatial and temporal homogeneity of the local weather regime; (3) the availability of daily COSMIC-1, COSMIC-2, Metop-A and Metop-B profiles. In practice, for EUREC4A and ATOMIC, collocation is defined as longitude difference within $2°$, latitude difference within $1.5°$, and time difference within $1\ hr$. For the two Southern Ocean campaign, the thresholds become $4°$,$2.5°$ and $2\ hr$ correspondingly. For ARRecon and MAGIC, the thresholds are $4°$,$1.5°$ and $2\ hr$.

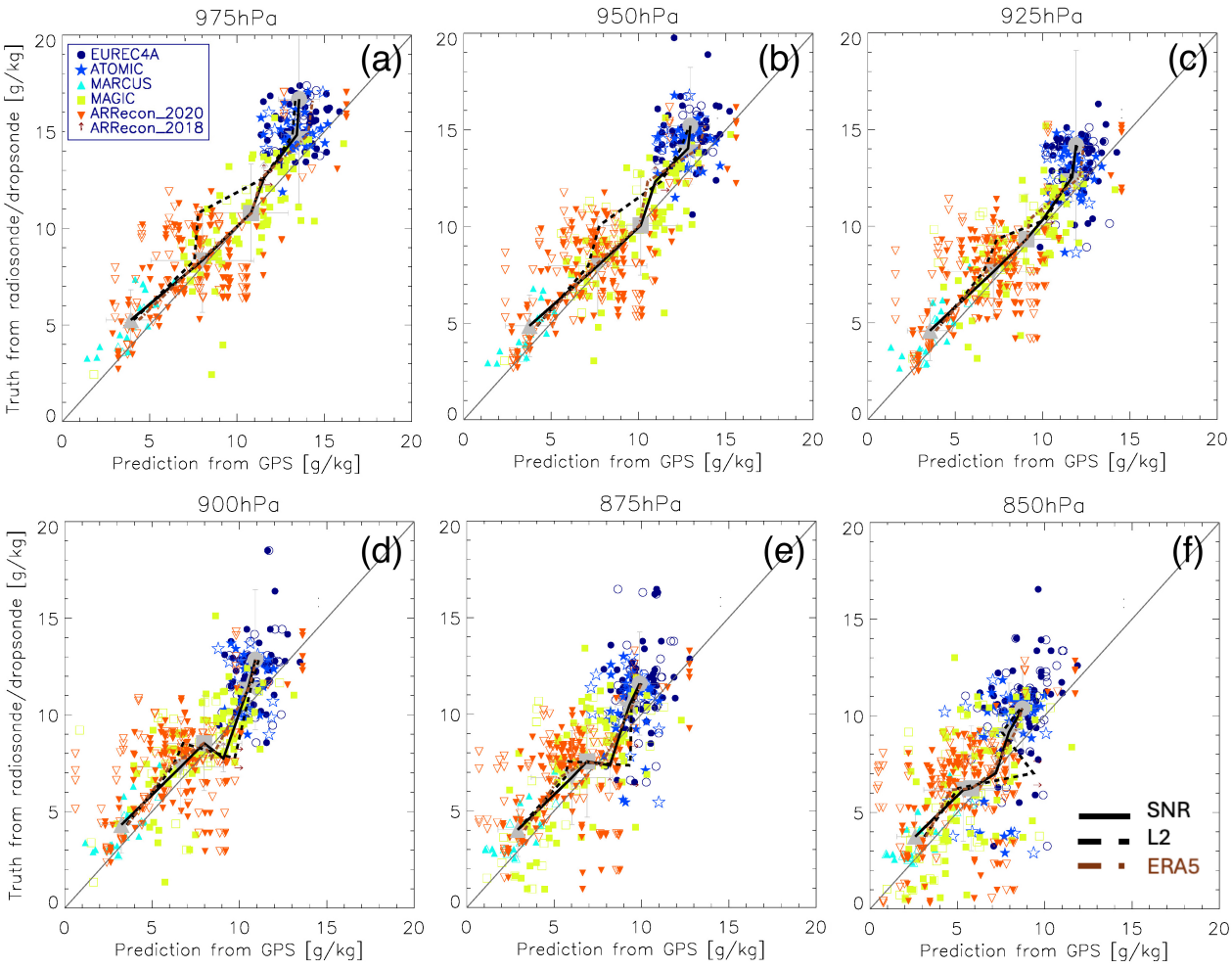


Figure 9. Scatter plots of collocated specific humidity [g/kg] comparison between radiosonde "truth" and retrievals from SNR (closed symbols) and ~~Level-2~~wetPrf/wetPf2 standard retrieval (open symbols) for each pressure level. Black thin diagonal lines are the 1:1 lines for reference. The mean and standard deviation from the SNR-ML retrieval from each campaign are shown as bigger grey symbols. In addition, the mean retrieved values from each campaign as opposed to the mean from radiosonde "truth" are shown as the bold black lines for SNR-ML retrievals, bold black dash-dotted lines for wetPrf/wetPf2 retrievals, and ~~grey~~ brown dashed-dotted lines for ERA-5 from the subset where collocations are found for SNR-ML and radiosonde data samples.

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 colors 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. We can see both SNR-ML retrievals and ~~Level-2~~wetPrf/wetPf2 retrievals demonstrate close agreement with ground "truth" for different weather regimes. In general, better correlation are found when the MABL is relatively dry or moderately-wet. Taken the

ARRecon campaign ~~specifically~~as an example, SNR-ML retrieval has an overall better agreement compared to ~~Level-2~~the wetPrf/wetPf2 retrievals (black lines versus black dash-dotted lines, and orange solid triangles versus orange open triangles), and the few extremely large specific humidity values ($> 12g/kg$) are well captured by the SNR-ML retrieval but not the ~~Level-2~~wetPrf/wetPf2 retrievals. The mean of all ARRecon collocated samples also suggest SNR-ML retrieval is the only one that doesn't produce a bias, while ~~both ERA-5 and~~ GNSS ~~Level-2~~wetPrf/wetPf2 retrievals are slightly dry-biased in atmospheric rivers. 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) exhibit good agreement to the ground truth too ~~if only mean values of each subset is considered~~ (brown dash-dotted 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 none of the three datasets capture the humidity conditions in the MABL. They are all dry-biased, and means from ERA-5 reanalysis are 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 the convenience of 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 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. 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)).

The violin plots in Fig. 11 and number of collocated sample statistics in Table 3 reveal more detailed difference in comparison statistics with respect to the radiosonde/dropsonde "truth", which more comprehensively demonstrate the values (and ~~cons~~ 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. SNR-ML method generated retrievals perform slightly better than the operational wetPf2 product. 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

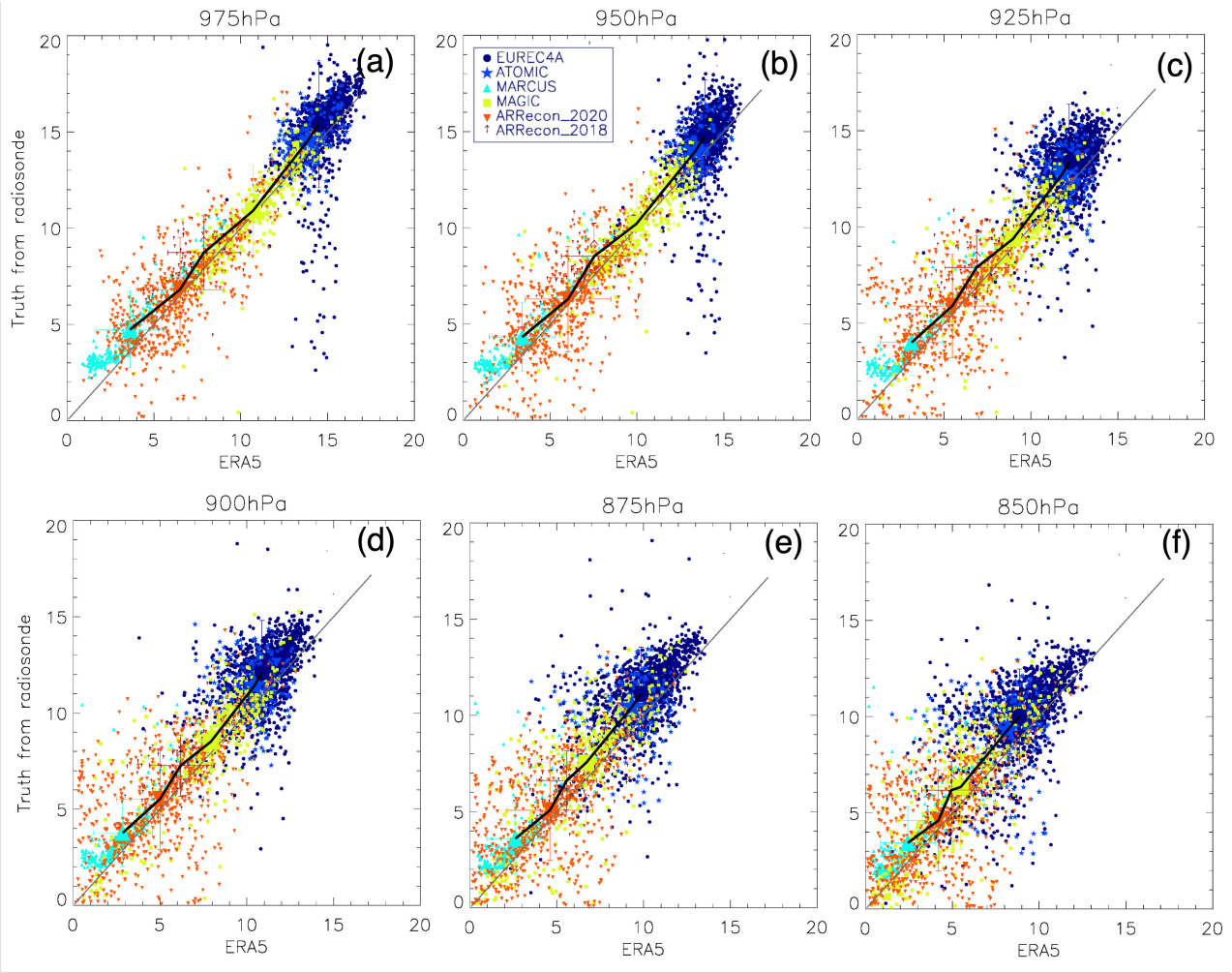


Figure 10. Same with Fig. 9, except for all available radiosonde/dropsonde samples in all these campaigns with collocated ERA-5 specific humidity. The bold black solid line connects the mean values from each campaign.

at high-latitude southern ocean, while surprisingly in the atmospheric river regime, SNR-ML retrievals outperform the GNSS operational retrievals as well as the ERA-5 reanalysis.

~~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~~
310 ~~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.~~

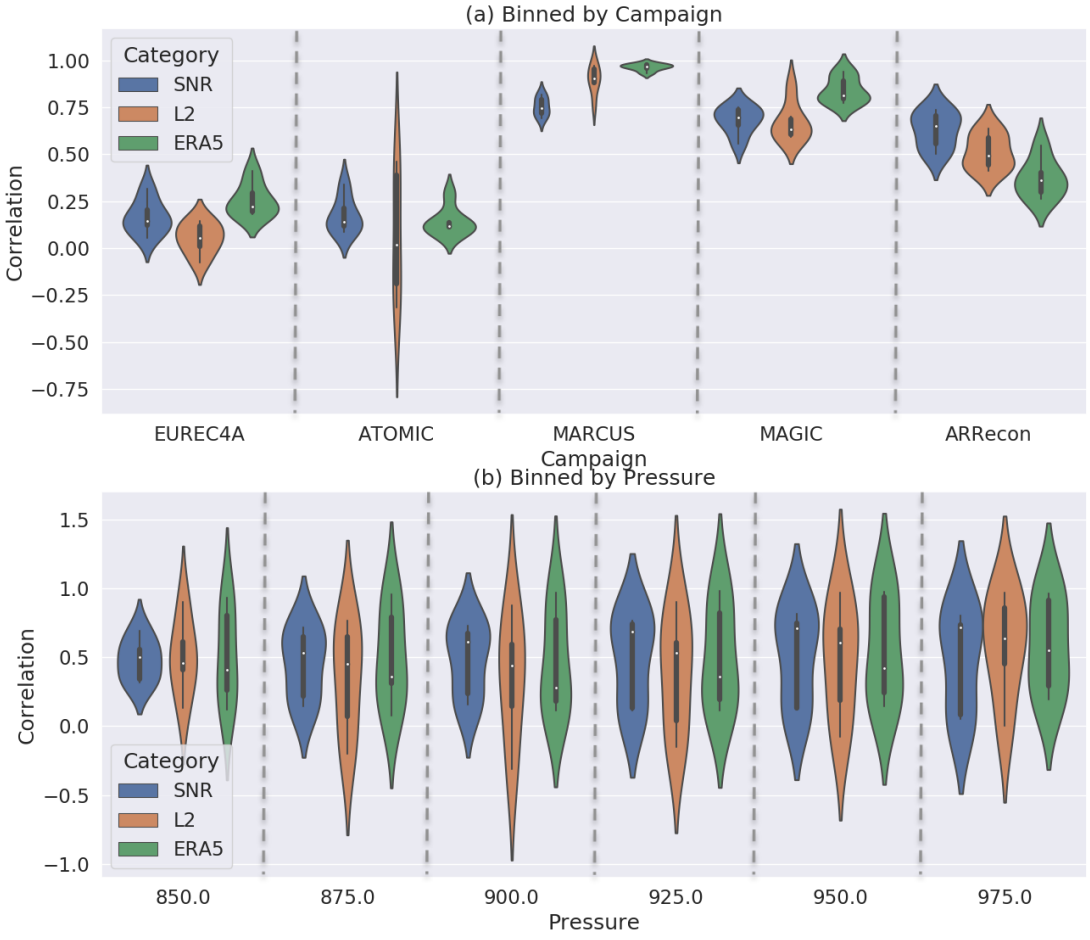


Figure 11. Violin plots of the correlation coefficients calculated from collocated samples for SNR-ML retrievals (blue), Level 2 retrievals (orange) and ERA-5 (green). (a) is all-level statistics for each campaign; and (b) is all campaign but binned by different pressure levels. Medium, standard deviation and minimum/maximum values are shown as the white dots, black box and extended vertical thin lines in each violin, respectively. The number of total samples are listed on top of each violin. For ERA-5, only the subset of samples that SNR-ML retrieval collocations available are selected to calculate the statistics.

315 In Fig. 11b, the correlation statistics are binned by different pressure levels. We can see the SNR-ML retrieval and ~~Level-2~~ wetPrf/wetPf2 retrieval are in general comparable in robustness in capturing the entire MABL specific humidity vertical structure, both of which outperform the ERA-5 ~~significantly~~ slightly in terms of the averaged correlation coefficient magnitude. The correlation coefficients using the SNR-ML retrievals are consistently slightly higher than that using the ~~Level-2~~wetPrf/wetPf2

retrievals, but the agreement between ground truth and SNL-ML retrievals decrease with increasing height, suggesting that the useful information that deep SNR signal carries is confined in the MABL. ~~975 hPa is the only exception, where ERA-5 seems to outperform the Level-2wetPrf/wetPf2 retrieval slightly, but SNR-ML retrieval is significantly better across-board than ERA-5.~~ Considering the spread is the smallest for the SNR-ML retrievals, we can see that this method does produce stable robust results at all 6 pressure levels..

Another big advantage of the SNL-ML retrieval is its consistently higher success rate in the MABL compared with the wetPrf/wetPf2 product. This is clearly seen in Table 3, where the percentage difference between the two are listed in the parantheses for each campaign at each pressure level. For stratocumulus region (MAGIC campaign) that ducting or super-refraction happens frequently, the success rate of SNR-ML method can go up to 700% more than using the wetPrf product at the lowermost altitude. Although the superiority of success rate of the SNR-ML retrievals gradually vanishes when getting closer to the MABL top, they are still across-the-board more than wetPrf/wetPf2 products.

Table 3. Number of collocated GNSS-radiosonde/dropsonde samples in each campaign. Two numbers in each cell are from SNL-ML method and wetPrf/wetPf2 product, respectively, and their percentage differences are shown in the parentheses.

Campaign Name	975 hPa	950 hPa	925 hPa	900 hPa	875 hPa	850 hPa
EUREC4A	50, 19 (160%)	50, 23 (117%)	51, 29 (76%)	51, 31 (65%)	51, 34 (50%)	51, 38 (34%)
ATOMIC	49, 23 (113%)	49, 27 (81%)	49, 29 (69%)	49, 29 (69%)	49, 35 (40%)	49, 44 (11%)
MARCUS	13, 5 (160%)	13, 7 (86%)	13, 7 (86%)	13, 7 (86%)	13, 8 (63%)	13, 9 (44%)
MAGIC	72, 9 (700%)	72, 25 (188%)	72, 34 (112%)	72, 40 (80%)	72, 43 (67%)	72, 46 (57%)
ARRecon	120, 84 (43%)	120, 101 (19%)	120, 101 (19%)	120, 106 (13%)	120, 106 (13%)	120, 106 (13%)

To summarize the major findings for comparisons against the limited independent ~~comparison against~~ radiosonde/dropsonde datasets available over the open ocean, we can draw the following conclusions. Firstly, The quality of the SNR-ML retrievals is comparable to ERA-5 and the operational wetPrf/wetPf2 product. In atmospheric river weather regime, SNR-ML method even outperforms the other two. The robustness and stable performance of SNR-ML retrievals remain the best within the MABL, although its advantage gradually vanishes with increasing height. ~~SNR-ML retrieval are the best at capturing the MABL vertical structure from the surface (975 hPa) up to 850 hPa globally compared to the ERA-5 reanalysis except at the deep tropics. The reason for the poor performance of the SNR-ML method in the deep tropics is probably due to the breakdown of the original assumption: turbulence and mixing in the tropical MABL by frequent shallow convections constantly disrupt the ducting condition, causing SNR reemerging at the deep H_{SL} blending other information and hence are not useful for MABL water vapor retrieval.~~ Secondly, compared to the operational ~~Level-2~~ retrievals, the SNR-ML method can achieve $10-700\%$ more samples in the MABL, especially over stratocumulus regions where ducting and super-refraction frequently occur that cause failure of operational retrievals. This suggests some unique value that the SNR-ML method can bring to the science community in facilitating understanding the water vapor-stratocumulus coupling mechanisms. ~~over high-latitudes or in the deep tropics~~.

Although some of the "independent validation dataset" is not completely independent as they may have been assimilated in the ERA5, the fact that SNR-ML retrieval statistics outperform ER-5 at all 6 pressure levels in diverse weather regimes prove that real physical information from SNR observations is learnt and kept by the ML model for prediction, admittedly it is impossible to quantify how much the real observed information contributes without accurate physics-based modeling simulations.

4 Discussions

In the section, we present and discuss some use case examples in order to demonstrate how to use this SNR-ML MBPL specific humidity product to identify and even quantify model or reanalysis issues.

4.1 Climatology

~~As suggested in Johnston et al. (2021),~~ Several previous studies suggest that MERRA-2 reanalysis has larger dry-biases in the polar regions compared to ERA-5 (Johnston et al. (2021), Ganeshan and Yang (2019)), while some other studies using in-situ campaign data suggested smaller dry-bias in the MERRA-2 reanalysis (e.g.,Seethala et al. (2021)). Here we map out the climatological distribution of ~~SNR-retrieved water vapor~~ specific humidity retrieved using the SNR-ML method to track down geographical discrepancies in the Arctic (Fig. 12) and Antarctic (Fig. 13) with respect to MERRA-2. The coldest months were not selected because of the concern that sea ice induced reflectometry signal might contaminate our SNR-ML retrieval results, but we didn't exclude retrievals over possible glaciers that MERRA-2 do not produce a valid value at 925 hPa because we used a fixed terrain map. Therefore, direct comparison should not be considered whereever MERRA-2 value is blank.

Overall again we can see the SNR-ML method retrieved polar MABL is much more humid than that from MERRA-2 in the Arctic during early spring and late fall seasons ($> 100\%$ in most areas). If we neglect sampling induced geographical inhomogeneities in the SNR-ML retrievals, we can actually see in Fig. 12 that the ~~general patterns~~ geographic distribution of highs and lows and their gradients are in general agreeable. The largest differences are that the wet intrusion along the Bering strait seems to be too weak during both April and November in MERRA-2, which could account for the dry-bias in the deep Arctic ocean. Meanwhile, the wet intrusion associated with the North Atlantic overturning circulation seems to be too strong during November in MERRA-2. These discrepancies connect possible root causes down to the ocean circulation, and up to the Arctic front, and should be further investigated in a whole Earth-system point of view.

Although Southern Ocean and South Pole seem ~~more boring~~ lacking geographical variations (Fig. 13), we can actually observe some interesting potential issues related to topographies. For example, the tip of the Andes mountain effectively blocks MABL water transport across the mountains, but such a local effect on humidity appears further downstream in MERRA-2. The gradient of water vapor amount from north to south is apparently much weaker compared to MERRA-2, which impacts the latent heat and sensible heat flux quantification when considering global energy transport.

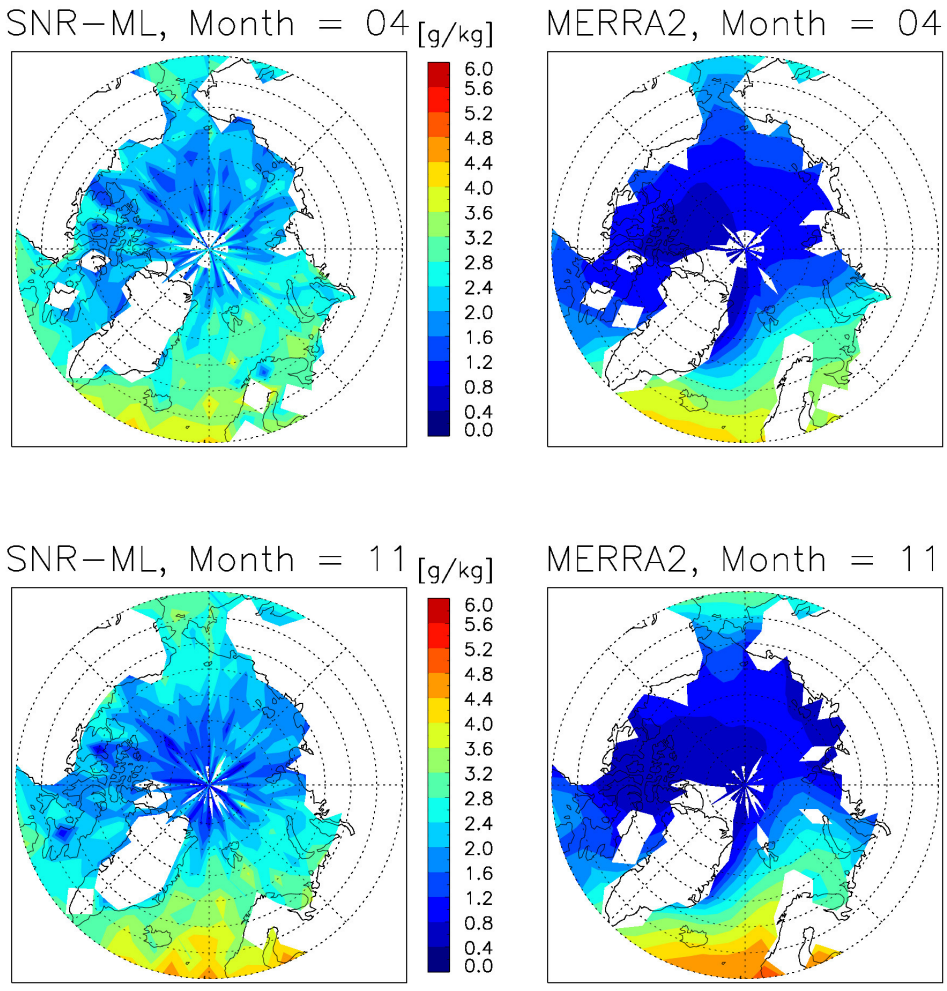


Figure 12. Monthly averages of 950 hPa specific humidity from COSMIC-1 SNR retrieval (left) compared to MERRA-2 reanalysis (right) for Arctic during April (top) and November (bottom), 2012 and 2013.

4.2 Diurnal Variation

It is well-known that global climate models (GCMs) have serious issues at reproducing the cloud, precipitation and convection diurnal cycles (e.g., Tian et al. (2004), Yin and Porporato (2017)). Although such a ~~notorious~~ problem is mostly attributed to the issues with cumulus parameterization schemes, we argue that the diurnal cycle of MABL water vapor also plays a nontrivial role as it ties closely to the shallow cumulus and stratocumulus, the latter, for example, is also closely related to the MABL height diurnal variation (e.g.,Liu and Liang (2010), Chepfer et al. (2019), Teixeira et al. (2021)). Ground truth

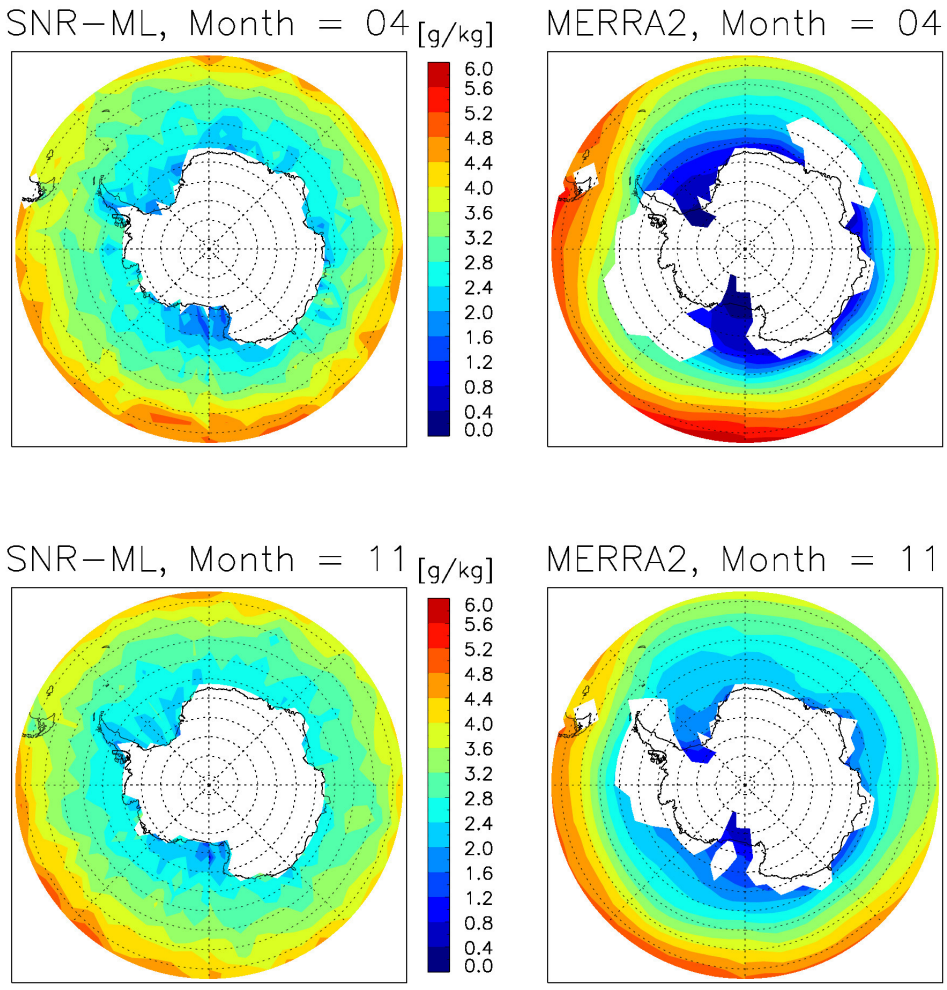


Figure 13. Same with Fig. 12, except for the Antarctic/Southern Ocean.

of the diurnal variation of MABL water vapor structures is extremely rare, probably because of the high cost associated with long-duration shipborne campaign that often only launches radiosondes twice daily and hence cannot capture the diurnal variabilities. Therefore, here we only aim at showing the discrepancies between ERA-5 and our SNR-ML retrieval generated diurnal cycle rather than determining which is right or which is wrong (Fig. 14).

In addition to the Southern Ocean MARCUS campaign and the atmospheric river regime ARRecon campaign that we have ground truths to compare with, several additional campaign regions and corresponding months are selected motivated by the observed diurnal variations of the MABL height established in Liu and Liang (2010). These 4 two additional regions include ~~tropical east Pacific (EPIC campaign), stratocumulus region off the coast of Chile (VOCALS campaign),~~ South Indian ocean

(INDOEX campaign, representing deep tropics), and the Arctic open ocean, representing polar winter conditions. The last one was added for the sole interest to check if there is any diurnal cycle in the coldest season.

The averaged specific humidity at 875 hPa agrees well between two datasets in the MARCUS and ARRecon campaigns, but the diurnal cycles in ERA-5 are too weak compared to the ground truths (red asterisks), while the SNR-ML method retrieved stronger diurnal cycles. It is worth noting that neither SNR-ML nor ERA-5 reproduced a strong peak below 900 hPa around 10 AM local time that both MARCUS and SOCRATES campaigns observed. The latter is another research campaign in the vicinity of MARCUS ship routes and season (Vomel and Brown (2018)), but was not employed for independent validation because of lack of collocations with GNSS observations. This peak is probably associated with the shallow mixed-phase cloud pocket precipitation that is spatially so small and inhomogenous in scale (Alessandro et al. (2021)) that neither GNSS nor ERA-5 are able to capture or reproduce. The under-estimation of the diurnal variability in the ARRecon campaign region is probably associated with the sampling bias, because the campaign "truth" was sampled only during AR events, while the SNR-ML and ERA-5 samples the climatology background.

Although we have no ground truth to assess the diurnal cycles of MABL humidity in other two regions, we can tell that ERA-5 is wetter in the South Indian ocean, and significantly drier in the Arctic ocean. Compared to the SNR-ML method retrieved diurnal cycle, MABL water vapor diurnal cycle in ERA-5 is too weak in 3 areas but not the INDOEX campaign region. ~~They seem to be slightly off-phase as well.~~ To put into context of the diurnal cycle of PBL height (Liu and Liang (2010)), ~~in the EPIC campaign region, the SNR-retrieved water vapor show a dip in the morning (8-9 am local time), coinciding with the dip of PBL height in this region. Then the water vapor continues to build up throughout the day, while the PBL height reaches maximum at local noon and then drops down decreases. In the VOCALS campaign stratocumulus region, the water vapor diurnal cycle derived from the SNR-ML method echos the PBL height change very well (Liu and Liang (2010), Fig. 10), indicating the important role water vapor plays in driving the diurnal cycle of stratocumulus, but ERA-5 barely shows any diurnal signature.~~ in the INDOEX campaign region, the diurnal cycle from ERA-5 and SNR-ML method agrees reasonably well, both anti-correlated with the diurnal cycle of PBL height change observed during that campaign. In the Arctic ocean, ERA-5 apparently has set some arbitrary threshold to keep the water vapor at a constant low level, while SNR-ML retrievals suggest a weak diurnal variation.

Overall ~~from this exercise~~, we can see the diurnal coupling between MABL water vapor, PBL height and clouds are vastly different from area to area. However, ERA-5 likely under-produced the diurnal cycle of MABL water vapor globally. For SNR-ML retrievals, day-to-day variability ~~is overwhelmingly large, making it harder to interpret or trust a diurnal cycle~~ often overwhelms the signal of diurnal cycle, yet the amplitude of diurnal cycle is still stronger and matches better with the limited ground truth. Ultimately, the lack of MABL water vapor ground "truth" measurements will ~~keep this topic foggy~~ continuously make observing and verifying the true diurnal cycle difficult. Other shipborne measurements, e.g., upward pointing radiometers, might be helpful to disentangle this mystery in the future.

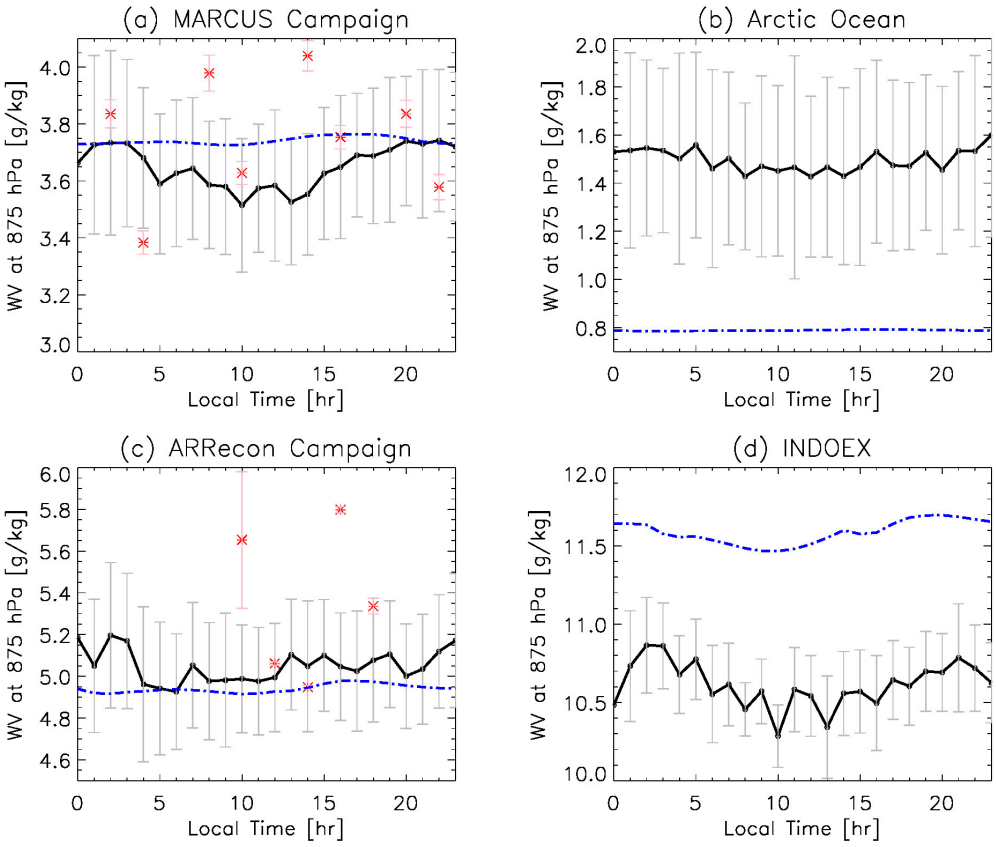


Figure 14. Multi-year mean diurnal variation of 875 hPa specific humidity retrieved from ~~COSMIC-1 and COSMIC-2SNR~~ all four missions (black with errorbars in grey) and from ERA-5 hourly reanalysis (dash-dotted blue) during ~~December - February~~ November - March for (a) ~~EPIC~~ MARCUS campaign region, $60°E - 150°E$, $60°S - 40°S$; (b) Arctic ocean, $180°W - 180°E$, $70°N - 90°N$; (c) ~~VOCALS~~ ARRecon campaign region, $160°W - 120°W$, $20°N - 50°N$; (d) INDOEX campaign region, $55°E - 75°E$, $25°S - 15°S$. The MARCUS radiosonde and ARRecon dropsonde "truths" are overlaid in (a) and (c) as asterisks with standard deviations shown in pink vertical bars.

5 Conclusions

Marine planetary boundary layer (MABL) water vapor amount and vertical gradient are among the key factors to couple the ocean and atmosphere cloud, precipitation and convection together, but meanwhile it is also among the hardest object to retrieve from satellite remote sensing perspective. Given the penetration capability of GNSS signal through clouds, we proposed a novel way in Wu et al. (2022) to utilize the GNSS signal-to-noise (SNR) ratio in the deep H_{SL} to retrieve MABL water vapor profiles. In this paper, we demonstrated it is workable at profile-by-profile level, leveraging the power of machine learning (ML) in capturing weak and non-linear signals. The surprising and novel findings in this paper, is that the ML-trained model can make better predictions that outperforms the training dataset (i.e., ERA-5) in some places, which demonstrates the real

information content in the SNR signal is learnt which would otherwise not be harnessed using traditional statistical methods. The new SNR-ML retrievals has more stable performance against the operational ~~Level-2~~ wetPrf/wetPf2 GNSS-RO retrievals, and it can produce $20-700\%$ more successful retrievals in the lowest $1\ km$ where observations are critical to understand ocean-atmosphere coupling.

We then showed two use cases to demonstrate possible ways to use this dataset. There is no conclusive results because of lack of ground "truth" to validate, but we do find both reanalysis tend to systematically produces dry biases at high-latitudes, and too weak diurnal cycles over global oceans. This SNR-ML retrieval dataset also has its own caveats. Whenever the "ducting" condition is violated (e.g., coastal topography, convective tower, mixing and turbulence in the MABL), the fundamental assumption breaks down, resulting in poor performances. More extensive comparisons and validations against other high-quality ground measurements are needed in the future.

Based on results from this work, one can see that deepSNR can complement the current GNSS-RO operational bending angle product for retrieving PBL information for different PBL conditions. A merged product is certainly of interest to future investigations, but fully understanding the physical mechanisms behind the reemerged deepSNR signal is the foundation for other downstream applications (e.g., data assimilation). Right now this can be considered as a stand-alone observational product for independent comparison or validation against model simulations or other observations.

Data availability. The Level 2 SNR-ML retrieval product for the prediction period (see Table 1) has been published on zenodo (Gong et al. (2024)). We welcome use and feedbacks.

COSMIC-1 and COSMIC-2 Level 1 and Level 2 data are downloaded from https://data.cosmic.ucar.edu/gnss-ro/. Metop-A and Metop-B data are downloaded from https://gpsmet.umd.edu/gnssro/download.php. ATOMIC data are downloaded from https://psl.noaa.gov/atomic/data/. EUREC4A data are downloaded from https://doi.org/10.25326/137. SOCRATES data are downloaded from https://data.eol.ucar.edu/master_lists/generated/socrates/. MARCUS data are downloaded from ARM data request portal. MAGIC data are downloaded from ARM data request portal. ARRecon data are downloaded from https://ARRecon.ucsd.edu/arrecon_data/ specially processed to fit the needs of this research. Interested users are encouraged to contact the last author for assistance of post-processed data.

Appendix A: A

Table A1. Summary of GNSS-RO instrument noise (σ) used in this work, separated by rising and setting modes.

Instrument Name	Orbit	Noise (σ)
COSMIC-1/C1	Rising	10.1
	Setting	10.9
COSMIC-1/C2	Rising	10.2
	Setting	10.9
COSMIC-1/C3	Rising	9.6
	Setting	10.4
COSMIC-1/C4	Rising	10.6
	Setting	11.2
COSMIC-1/C5	Rising	10.1
	Setting	11.1
COSMIC-1/C6	Rising	9.2
	Setting	10.7
COSMIC-2/E1	Rising	17.0
	Setting	17.5
COSMIC-2/E2	Rising	17.5
	Setting	17.8
COSMIC-2/E3	Rising	17.2
	Setting	17.9
COSMIC-2/E4	Rising	17.5
	Setting	17.7
COSMIC-2/E5	Rising	17.4
	Setting	17.8
COSMIC-2/E6	Rising	17.5
	Setting	17.8

Table B1. Excess Phase L1 grid for this work

Parmeter	Grid values
$Log_{10}(\phi_{L1})$	1.26245, 1.33846, 1.41162, 1.48144, 1.54777, 1.62428, 1.69679, 1.76530, 1.82995, 1.89098, 1.94866, 1.97000, 2.00325, 2.02000, 2.05500, 2.08000, 2.10415, 2.13000, 2.15091, 2.17000, 2.19548, 2.23805, 2.27875, 2.30103, 2.32222, 2.37000, 2.41497, 2.44000, 2.55630, 2.59000, 2.63000, 2.69020, 2.75000, 2.81291, 2.86000, 2.92428, 2.95000, 3.02531, 3.10000, 3.11727, 3.15000, 3.20140, 3.22000, 3.25000, 3.27875, 3.30000, 3.32000, 3.35025, 3.41664, 3.47857, 3.53656, 3.59106
Rough corresponding H_{SL} [km]	-150, -140, -130, -120, -110, -108, -106, -104, -102, -100, -98, -96, -94, -92, -90, -80, -70, -60, -50, -40, -37, -33, -30, -26, -23, -20, -19, -17, -15, -13, -11, -9, -7, -5, -3, -2, -1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 15, 17, 19

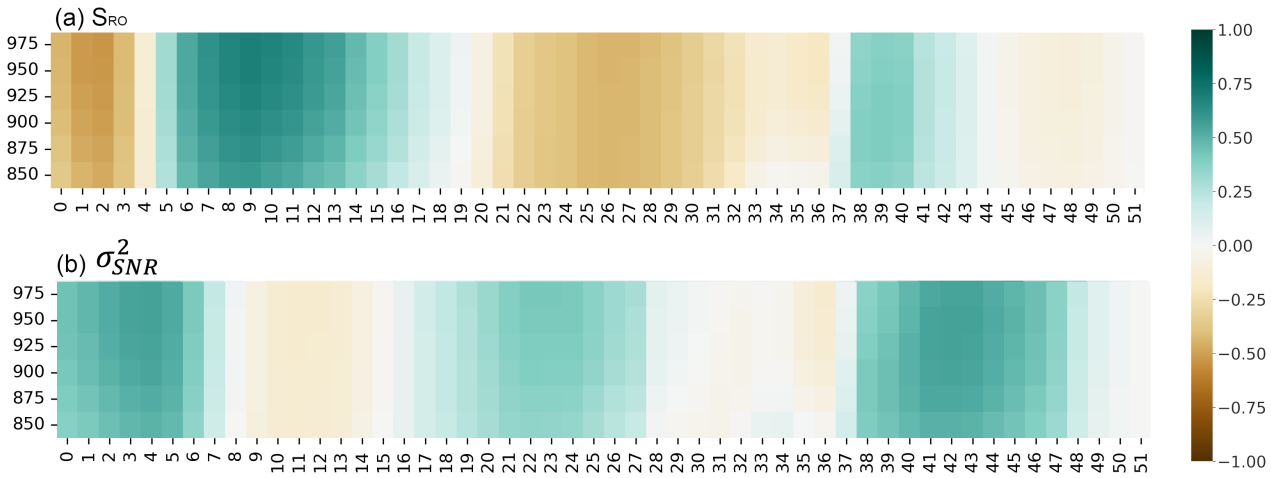


Figure A1. Same with Fig. 2, except for Metop-A training dataset.

Author contributions. TEXT

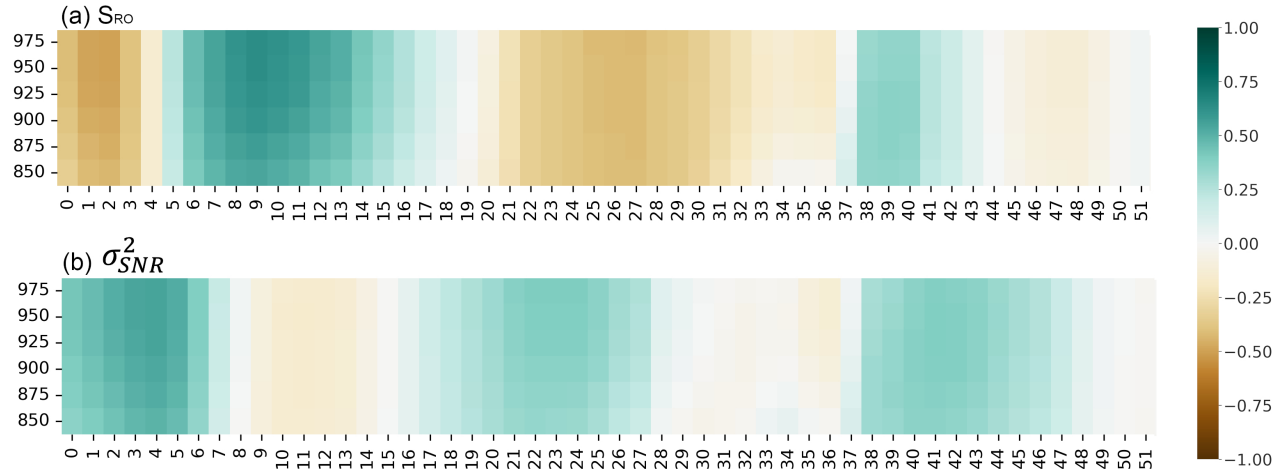


Figure A2. Same with Fig. 2, except for Metop-B training dataset.

D.L.W. came up with the initial idea. J. G. designed the methodology, executed the plan, built the model, and conducted the validation and data analysis. M.B. helped conducted the hyperparameter tuning. M.G. provided Fig.1. M. Z. provided the high vertical resolution AR-Recon data. All authors participated in result discussion and interpretation.

Competing interests. TEXT

The authors have no competing interests.

Acknowledgements. J.G. is grateful to Dr. Mariel Friburg at NASA Goddard in providing financial support of M.B. We thank the editor and two reviewers for their thoroughly and insightful comments, which helped greatly in improving the readability and clarity of this paper.

References

John J. D'Alessandro, Greg M. McFarquhar, Wei Wu, Jeff L. Stith, Jorgen B. Jensen, Robert M. Rauber: Characterizing the Occurrence and Spatial Heterogeneity of Liquid, Ice, and Mixed Phase Low-Level Clouds Over the Southern Ocean Using in Situ Observations Acquired During SOCRATES, J. Geophys. Res., https://doi.org/10.1029/2020JD034482.

Boisvert, L. N., Wu, D. L., Vihma,T., Susskind, J.: Verification of air/surface humidity differences from AIRS and ERA-Interim in support of turbulent flux estimation in the Arctic, J. Geophys. Res., doi:10.1002/2014JD021666, 2015.

Chang, H., Lee, J., Yoon, H., Morton, Y. J., and Saltman, A.: Performance assessment of radio occultation data from GeoOptics by comparing with COSMIC data, Earth, Planets and Space, doi:10.1186/s40623-022-01667-6, 2022.

Chepfer, H., H. Brogniez, and V. Noel: Diurnal variations of cloud and relative humidity profiles across the tropics, Sci. Rep., https://doi.org/10.1038/s41598-019-52437-6, 2019.

Cobb, A., A. Michaelis, S. Iacobellis, F. M. Ralph, and L. Delle Monache: Atmospheric River Sectors: definition and characteristics observed using dropsondes from 2014-20 CalWater and AR Recon, Mon. Wea. Rev., doi:10.1175/MWR-D-20-0177.1, 2021.

Keeler, Evan, Burk, Ken, Kyrouac, Jenni: ARM - Balloon-borne sounding system (BBSS) WNPN output data, sondewnpn.b1, https://doi.org/10.5439/1595321, 2022.

Feng, X., F. Xie, C. Ao and R. Anthes: Ducting and Biases of GPS Radio Occultation Bending Angle and Refractivity in the Moist Lower Troposphere, J. Atm. Ocn. Tech., https://doi.org/10.1175/JTECH-D-19-0206.1, 2020.

Yarin Gal and Zoubin Ghahramani: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, Proceedings of the 33 rd International Conference on Machine Learning, https://proceedings.mlr.press/v48/gal16.pdf

Ganeshan, M., Y. Yang: Evaluation of the Antarctic Boundary Layer Thermodynamic Structure in MERRA2 Using Dropsonde Observations from the Concordiasi Campaign, Earth and Space Sci., https://doi.org/10.1029/2019EA000890, 2019.

Ganeshan, M., Wu, D. L., Santanello, J. A., Gong, J., Ao, C. O., Vergados, P., and Nelson, K.: Exploring commercial GNSS RO products for Planetary Boundary Layer studies in the Arctic Region, doi:10.5194/amt-2024-83, in review for AMT, 2024.

George, G., Stevens, B., Bony, S., Pincus, R., Fairall, C., Schulz, H., Kölling, T., Kalen, Q. T., Klingebiel, M., Konow, H., Lundry, A., Prange, M., and Radtke, J.: JOANNE: Joint dropsonde Observations of the Atmosphere in tropical North atlaNtic meso-scale Environments, Earth Syst. Sci. Data, 13, 5253–5272, doi:10.5194/essd-13-5253-2021, 2021.

Gong, J., Wu, D. L.: GNSS deep SNR retrievals of marine atmosphere boundary layer (MABL) specific humidity, zenodo, https://doi.org/10.5281/zenodo.13946112, 2024.

Johnston, B.R.; Randel, W.J.; Sjoberg, J.P.: Evaluation of Tropospheric Moisture Characteristics Among COSMIC-2, ERA5 and MERRA-2 in the Tropics and Subtropics. Remote Sens., doi:10.3390/rs13050880, 2021.

Krüger, K., Schäfler, A., Wirth, M., Weissmann, M., and Craig, G. C.: Vertical structure of the lower-stratospheric moist bias in the ERA5 reanalysis and its connection to mixing processes, Atmos. Chem. Phys., 22, 15559–15577, https://doi.org/10.5194/acp-22-15559-2022, 2022.

Kuo, Y., Sokolovskiy,S., Anthes,R.A., and Vandenberghe, F.: Assimilation of GPS Radio Occultation Data for Numerical Weather Prediction, Space Sci., doi:10.3319/TAO.2000.11.1.157(COSMIC), 2000.

LeCun, Y., Bengio, Y. and Hinton, G., : Deep learning. Nature, 521(7553), pp.436-444, https://doi.org/10.1038/nature14539, 2015.

Liu, S.,Liang, X.:Observed Diurnal Cycle Climatology of Planetary Boundary Layer Height,J. Clim., doi:10.1175/2010JCLI3552.1, 2010.

Millán, L. F., Lebsock, M. D., and Teixeira, J.: Variability of bulk water vapor content in the marine cloudy boundary layers from microwave and near-infrared imagery, Atmos. Chem. Phys., 19, 8491–8502, doi:10.5194/acp-19-8491-2019, 2019.

Milestein, A., Blackwell, B.: Neural network temperature and moisture retrieval algorithm validation for AIRS/AMSU and CrIS/ATMS, J. Geophys. Res.-Atm., doi:10.1002/2015JD024008, 2016.

Milestein, A.: Planetary Boundary layer (PBL) Final Report, Lincoln Lab, MIT, https://apps.dtic.mil/sti/pdfs/AD1166514.pdf

Milstein, A., Santanello, J. A., Blackwell, B.: Detail Enhancement of AIRS/AMSU Temperature and Moisture Profiles Using a 3D Deep Neural Network, Artificial Intelligence in Earth Science, doi:10.1175/AIES-D-22-0037.1, 2023.

Seethala, C., Zuidema, P., Edson, J., Brunke, M., Chen, G., Li, X.-Y., et al.: On assessing ERA5 and MERRA2 representations of cold-air outbreaks across the Gulf Stream. Geophysical Research Letters, https://doi.org/10.1029/2021GL094364, 2021.

Sokolovskiy, S.; Schreiner, W.; Zeng, Z.; Hunt, D.; Lin, Y.-C.; Kuo, Y.-H. Observation, analysis, and modeling of deep radio occultation signals: Effects of tropospheric ducts and interfering signals. Radio Sci. 2014, 49, 954–970. https://doi.org/10.1002/2014RS005436, 2014.

Sokolovskiy, S., Z. Zheng, D. C. Hunt, J.-P. Weiss, J. J. Braun, W. S. Schreiner, R. A. Anthes, Y.-H. Kuo, H. Zhang, D. Lenchow and T. Vanhove: Detection of Superrefraction at the Top of the Atmospheric Boundary Layer from COSMIC-2 Radio Occultations, J. Atms. Ocn. Tech., doi:10.1175/JTECH-D-22-0100.1, 2024.

Stephan, C. C., Schnitt, S., Schulz, H., Bellenger, H., de Szoeke, S. P., Acquistapace, C., Baier, K., Dauhut, T., Laxenaire, R., Morfa-Avalos, Y., Person, R., Quiñones Meléndez, E., Bagheri, G., Böck, T., Daley, A., Güttler, J., Helfer, K. C., Los, S. A., Neuberger, A., Röttenbacher, J., Raeke, A., Ringel, M., Ritschel, M., Sadoulet, P., Schirmacher, I., Stolla, M. K., Wright, E., Charpentier, B., Doerenbecher, A., Wilson, R., Jansen, F., Kinne, S., Reverdin, G., Speich, S., Bony, S., and Stevens, B.: Ship- and island-based atmospheric soundings from the 2020 EUREC4A field campaign, Earth Syst. Sci. Data, 13, 491–514, doi:10.5194/essd-13-491-2021, 2021.

Stevens, B., and co-authors: EUREC4A, Earth Syst. Sci. Data, doi:10.5194/essd-13-4067-2021, 2021.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, http://jmlr.org/papers/v15/srivastava14a.html, 2014.

Teixeira, J., Piepmeier, J. R., Nehrir, A. R., Ao, C. O., Chen, S. S., Clayson, C. A., Fridlind, A. M., Lebsock, M., McCarty, W., Salmun, M., Santanello, J. A., Turner, D. A., Wang, Z., Zeng, X.: TOWARD A GLOBAL PLANETARY BOUNDARY LAYER OBSERVING SYSTEM: THE NASA PBL INCUBATION STUDY TEAM REPORT, NASA, https://ntrs.nasa.gov/api/citations/20230001633/downloads/AFridlindPBLTowardsReport.pdf, 2021.

Tian, B., B. Soden and X. Wu: Diurnal cycle of convection, clouds, and water vapor in the tropical upper troposphere: Satellites versus a general circulation model, J. Geophs. Res., https://doi.org/10.1029/2003JD004117.

Virman, M., M. Bister, J. Räisänen, V. A. Sinclair, H. Järvinen: Radiosonde comparison of ERA5 and ERA-Interim reanalysis datasets over tropical oceans, Tellus A: dynamic meteorology and oceanography, https://doi.org/10.1080/16000870.2021.1929752, 2021.

Vömel, H., Brown, W.: SOCRATES -2018 Radiosonde Data Quality Report, UCAR/NCAR - Earth Observing Laboratory, University Corporation for Atmospheric Research, doi:10.5065/D69P30HG, 2018.

Wee, T.-K.; R.A. Anthes, D.C. Hunt, W.S. Schreiner, and Y.-H. Kuo: Atmospheric GNSS RO 1D-Var in Use at UCAR: Description and Validation. Remote Sens., 14, 5614. https://doi.org/10.3390/rs14215614, 2022.

Wu, D.L.,Gong, J., and Ganeshan, M.: GNSS-RO Deep Refraction Signals from Moist Marine Atmospheric Boundary Layer (MABL), Atmosphere, doi:10.3390/atmos13060953,2022.

Xu, X., Zou, X.: COSMIC-2 RO Profile Ending at PBL Top with Strong Vertical Gradient of Refractivity, doi:10.3390/rs14092189, 2022.

Ye, J., Liu, L., Wang, Q., Hu, S. and Li, S.: A Novel Machine Learning Algorithm for Planetary Boundary Layer Height Estimation Using AERI Measurement Data, IEEE Geosci., doi:10.1109/LGRS.2021.3073048, 2021.

Yin, J., A. Porporato: Diurnal cloud cycle biases in climate models, Nature Comm., https://doi.org/10.1038/s41467-017-02369-4, 2017.

Zhran, M.: An evaluation of GNSS radio occultation atmospheric profiles from Sentinel-6, The Egypsian J. of Remote Sensing and Space Sci., doi:10.1016/j.ejrs.2023.07.004, 2023.

Zheng, M., Torn, R., Delle Monache, L., Doyle, J., Ralph, F. M., Tallapragada, V., Davis, C., Steinhoff, D., Wu, X., Wilson, A. M., Papadopoulos, C., and Mulrooney, P. (2024). An Assessment of Dropsonde Sampling Strategies for Atmospheric River Reconnaissance. Monthly Weather Review, doi:10.1175/MWR-D-23-0111.1, 2024.