Reviewer #2:

The paper proposes an innovative method that combines Bayesian interpolation (BI) and basic MultiLayer Perception (MLP) to map refractivity from Global Navigation Satellite Systems (GNSS) radio occultation (RO) data, i.e., COSMIC-2. While the study suggests that the BI&ML model outperforms individual BI and MLP models, there are some concerns regarding the presented conclusions.

Dear reviewer, thank you very much for taking the time to review our work, we appreciate your comments. Below, you can find our answers to all your points in an extended version.

Review Comments:

1. Training Consistency: The paper suggests that BI&ML outperforms MLP alone, but upon examining tables 1 and 2, the difference does not appear significant. It is common knowledge that the performance of the MLP model is tied to its initial state and training quality. To address this concern, it is recommended that the author trains multiple models for all methods to provide a clearer understanding of uncertainty associated with each model.

Thank you for your comment. To address this comment, we performed further experiments where we trained several models.

In Table 1, we summarize the results on an ensemble of 10 neural networks for ML and BI&ML. We display these results for a network of 5 layers and for a network of 8 layers (related to your second comment). Similarly to the results in the paper, there is a smaller standard deviation for the BI&ML compared to ML, for a network with 5 layers. For a network with 8 layers, the improvement is smaller, however the uncertainty of the standard deviation is also improved.

Note that Table 1 has not been included in the paper, however, we have added the following paragraph: 'Note that the results displayed in this section (Table 1 and Table 2), are a result of one single trained network, and not of an ensemble of networks. To further validate our results, we performed additional experiments for the refractivity at 2 km iso-height, where we trained multiple (10) models for ML and BI&ML. For our architecture, similar results were achieved on the results of the ensemble of the models, with ~0.2 N-unit worse standard deviation. '

Table 1: Statistics (in N-unit) over 10 trained models on ML and BI&ML models, (not in manuscript).

	5 Layers				8 Layers			
	Av. STD	**Std. St.dev**	**Av. Bias**	**Std. Bias**	**Av. STD**	**Std. St.dev**	**Av. Bias**	**Std. Bias**
ML	9.16	0.08	0.19	0.31	8.90	0.13	0.41	0.17
BI&ML	**8.96**	**0.08**	**0.17**	**0.45**	**8.85**	**0.09**	**0.28**	**0.35**

2. Hyperparameter Tuning: Figure 2 illustrates hyperparameter tuning on wandb, which is commendable. However, it would be beneficial to explore more critical hyperparameters, such as the optimization method (e.g., Adam, AdamW, RMSprop), number of layers, and weight decay. These parameters are likely to have a more substantial impact than batch size, epochs, and learning rate.

Thank you for this input. Following your comment we have trained also the other hyperparameters that you suggested. In Table 2 we have summarized all the results. We point out that all the statistics are computed over 10 different models for each tunned parameter.

Firstly, the main hyperparameter that further impacts the results is the number of layers. Smaller number of layers results in worse results, especially in terms of standard deviation and larger number of layers further improves the results (again, mainly in terms of standard deviation). We also point out that when we trained networks with 7 (or 8) layers and 9 (or 10) layers the time to train the model was about 1.4 and 1.5 time longer than for 5 layers.

Secondly, we point out that there are no significant changes if we use different optimizers or weight decay. In some cases, the same results are obtained, for example, if we use weight decay 0, 1e-5 or 1e-4, and if we use Adam or AdamW optimizers.

In the paper, we have added the following (section 4.1.1): 'We also trained the number of layers, the optimizer, and the weight decay. However, we did not notice any significant differences when using different optimizers or weight decay.'

And at the end of section 4.2:

'In addition, for the ensemble of models, we noticed further improvements (mainly on the standard deviation) when we added more hidden layers. On an ensemble of 10 trained models, ~0.3 N-unit improvement can be achieved when using 10 hidden layers, compared to 5 layers for the ML model. However, this further increases the training time, which is especially important for the ML method, given that we use a total of 30000 epochs. A higher number of layers is more suitable for BI&ML, where the number of epochs is much smaller.

We point out that the scope of this study is to present ML as an alternative method to grid RO observations. The results obtained herein indicate better performance compared to BI. Considering our results and the additional tests with multiple models, BI&ML brings additional (small) improvement compared to ML methods, with an obvious advantage in terms of a much shorter time to train the model. For future work related to continuous long-term RO-gridded products, an ensemble of models will be trained to also provide the uncertainty related to the models.'

Table 2: Tuning of number of layers, optimizer, weight decay. Statistics in N-units, (not in manuscript).

Nb. of Layers	1	2	3	4	5	6	7	8	9	10
Av. St.dev	13.05	10.64	9.65	9.39	9.16	9.13	9.07	8.90	8.91	8.87
Std. St.dev	0.24	0.16	0.13	0.14	0.08	0.17	0.11	0.13	0.11	0.07
Av. Bias	0.17	0.26	0.30	0.26	0.19	0.31	0.24	0.41	0.09	0.29
Std. Bias	0.10	0.20	0.43	0.40	0.31	0.32	0.40	0.17	0.35	0.24

Optimizer	Adam	AdamW	RMSprop
Av. St.dev	9.16	9.16	9.12
Std. St.dev	0.08	0.08	0.08
Av. Bias	0.19	0.19	0.22
Std. Bias	0.31	0.31	0.39

Weight Decay	0	1e-5	1e-4	1e-3
Av. St.dev	9.16	9.16	9.16	9.18
Std. St.dev	0.08	0.08	0.08	0.10
Av. Bias	0.19	0.19	0.19	0.38
Std. Bias	0.31	0.31	0.31	0.17

3. Data Preprocessing: The paper does not explicitly mention any preprocessing steps for COSMIC-2 data. It would be insightful to provide details on any preprocessing carried out, as this could significantly influence the model's performance.

Thank you for this comment. For our application, not much pre-processing is required. The only preprocessing we do is mentioned in the paper, as follows:

i. COSMIC-2 refractivities are interpolated at isohypsic heights, 2 km, 3 km, 5 km, 8km, 15 km and 20 km. The vertical resolution of RO is relatively high and therefore we do not expect this interpolation to change the distribution of the data. Since the resolution is high, we used a simple linear interpolation. The resulting interpolation error is much smaller than the uncertainty from training different models.

   From Figure 1 in the paper, you can see the resulting refractivities at these heights, which appear appropriate to our expectations.

ii. The input data latitude, longitude and time are standardized.

4. Data Splitting Strategy: Randomly splitting data into train, valid, and test sets might not be optimal, as these sets could be interrelated (so called 'data leaking'). The suggestion is to have an independent validation set and test set for a more robust evaluation of the proposed models.

We appreciate this comment and have made revisions to make this clearer in the paper. Although the data are randomly chosen, we point out that the test, validation, and train datasets do not have the same refractivity. By random we mean that we do not split them according to time or geolocation or refractivity values, however, these datasets do not contain the same points.

In addition, we would like to emphasize that our problem is an interpolation problem and not a prediction problem. We aim to improve the resolution of RO refractivities and produce gridded products. Therefore, it is fine for our application if the training/validation and testing data are not separated epoch-wise, for example,days 1-8 for training and validation and days 9-10 for testing.

In the paper, in Section 4.1 we changed the following sentence:
'As is customary in ML, we randomly split the nature dataset into two segments: 80% of the data for training, 20% for testing.'
to
'As is customary in ML, we randomly split the nature dataset into three segments: 72% of the data for training, 8% for validation and 20% for testing. We point out that the training, validation, and testing datasets do not overlap. By random choice, we mean that we do not split them according to a specific parameter (such as time , geolocation, or refractivity values).'

5. Code and Data Availability: To enhance the reproducibility and validation of the research, it is recommended that the authors provide complete code and data. This would enable other researchers to replicate the experiments more easily and validate the results effectively.

Thank you for this suggestion.

- We provide the Matlab routines we have used to read COSMIC-2 refractivity at 2 km height. The data can directly be downloaded from the UCAR website.
- We provide the ML implementation in Python to train and evaluate the mapping of refractivity using ML.
- We provide the mapped refractivities using BI for train and test dataset.
- We provide the COSMIC-2 data (as CSV files), used as input to ML routines.

The codes and data will be available for the scenario at 2 km height when we train COSMIC-2 refractivities. Using the scripts researchers can reproduce the ML and BI&ML results for COSMIC-2 at 2 km directly and can easily adapt the routines to reproduce the results at the other evaluated altitudes.

The section 'Code and data availability' has changed as follows:
'We provide sample routines for the readers to be able to reproduce the results for COSMIC-2 observations at 2 km, https://doi.org/10.3929/ethz-b-000670139. These include sample data (train and test data for refractivity at 2 km) and code implementation (Matlab code to read the COSMIC-2 data at 2 km and Python codes to train and evaluate the ML model). Additional codes associated with this study are available from the corresponding author upon reasonable request. Additional datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request. The COSMIC-2/FORMOSAT-7 data is freely available from UCAR. The ECMWF data are a product of the European Centre for Medium-Range Weather Forecasts (ECMWF) (© ECMWF).'