We would like to thank the referee for the useful comments and constructive suggestions. In the following, we address the referee's comments and describe corresponding changes we have made to the manuscript. The referee's comments are listed in *italics*, followed by our response in blue. New/modified text in the manuscript is in **bold**.

Getting the best estimate of the PBLH is a worthwhile endeavor, which can help with our basic understanding of processes associated with the PBL and can be used to help modelling. The manuscript is suitably organized, is generally well written, and includes appropriate figures. As with any empirical technique, the interpretation needs to be done in a way that clearly points out the limitations and bias of choices. In addition, the case must be made that these results are robust. In particular, are the results actually generalizable or is it too dependent on what is or is not included in both the method (e.g., parameter selection) and training data (e.g. which airports or years used). This is a common issue in applying any sort of empirical method, but it is really important to be clear about this. Otherwise, one could repeat these steps, but alter a couple of choices, and come up with different results.

We thank the referee for this important comment. In the revised manuscript, we point out limitations of our study and provide all the key details of our model such as parameters and training data (see our responses to several comments below). We revise the sentences at lines 69-75 to the following:

"The existing AMDAR PBLH dataset is available hourly from 2005 to 2019 at 54 airport locations. However, many applications require PBLH in other locations or completely covering a region. The objective of this study is to produce observation-based, spatially-complete PBLH fields over the CONUS. We develop a data-driven predictive model using various meteorological and geographical predictors to match the AMDAR PBLH observations. Since the predictors all have complete spatial coverage over the CONUS, running the model forwardly can yield PBLH prediction at arbitrary locations in the domain. We cross-validate the model in space by randomly splitting the airports into training and testing sets, and the model is selected based on averaged metrics on the testing sets. The predicted PBLH is then compared to PBLHs from three widely-used reanalysis products (ERA5, MERRA-2, and NARR), and all of them are further compared to independently diagnosed PBLH from observations (e.g., from research aircraft profiles and HSRL airborne lidar) and the CALIPSO PBLH product."

We also discuss the generalizability of our work by adding the following sentence to the first paragraph of the conclusions:

"The model configurations have been specifically optimized and evaluated to generate the spatially complete PBLH dataset over the CONUS during the period when AMDAR PBLH data are available (2005–2019). Further generalization of the work will require additional tuning and model evaluation."

Input data: Data is preferentially excluded. For instance, if the AMDAR PBLH is too far away from the ERA5 estimate it is thrown out, so as stated around line 115, "...which accounted for about one third of all AMDAR data, half of data under stable condition, and only 10% of data under convective condition." This has many implications for all subsequent analysis. Importance of the ERA5 PBLH (at times 0, -1, and -2) for permutation and SHAP feature is extremely large (Fig. 5). Throwing out all "bad" AMDAR data contributes to that importance, and basically implies overfitting to the ERA5 PBLH. The method itself accounts for overfitting, but if the input data is already filtered to get rid of 'bad' data before the method is applied, it will artificially create a 'better' fit.

Thanks for this important comment. It is not uncommon to detect and correct corrupt or inaccurate records from the dataset before any valid statistical model can be built. In this case, the data cleaning is a crucial preprocessing step. As seen from Fig. 2a-b in the manuscript, a portion of data points are characterized by very low ERA5 PBLH but much higher AMDAR PBLH. We have tried to explain this "troubled group" using a wide range of variables but did not find any meaningful correlation. The most likely cause of the troubled group is the ambiguity of PBL top diagnosed from ERA5 and AMDAR. Although both ERA5 and AMDAR use the critical bulk Richardson number method, the footprint of their profiles is very different. ERA5 profiles supposedly represent averages over a 0.25° grid cell and AMDAR profiles (at least below a few km) are in-situ measurements at points.

These challenging cases happen much more often in stable and neutral conditions, suggesting that they are likely due to ERA5 and AMDAR identifying different vertical structures as the PBL top. Including this troubled group will derail the whole model fitting as the algorithm would have struggled to explain the large differences between AMDAR and ERA5 PBLH, which is the dominant feature. To clarify this point, we revise the sentences at lines 106–108 to the following:

"We did not find any meteorological or geographical factors that would explain the occurrence of AMDAR vs. ERA5 PBLH data pairs in these clusters. The most likely cause of these clusters of data is the ambiguity of identifying the PBL top. Under challenging conditions, the critical bulk Richardson number algorithm that is used in AMDAR and ERA5 data to diagnose PBLH may identify different vertical structures as PBL top, leading to very different and uncorrelated PBLH values. Including these uncorrelated clusters of points will strongly bias the model training results."

We do understand the referee's concern and to further address the this concern, we explicitly acknowledge this data preprocessing step in the conclusions as a limitation of our study after line 337:

"Significant challenges still exist due to the lack of PBLH observations and the uncertainty of existing datasets. We observe clusters of AMDAR observations that are uncorrelated with collocated ERA5 PBLH, mostly under stable conditions, and no meteorological or geographical factors could explain this discrep-

ancy. A preprocessing step had to be implemented to mitigate its impacts on the model training. The satellite-based CALIPSO dataset is the most spatiotemporally complete for model evaluation, but it is subject to large uncertainties, which gives essentially no correlations with reanalysis datasets and the prediction from this work when PBLH is lower than 1 km. Future spaceborne PBLH observations with higher fidelity and more routine suborbital measurements, especially under stable conditions, will be beneficial."

Comparisons: Fig. 3 gives distributions of PBLH from various datasets at various locations, times, and sample sizes. If the point of the figure is to show how different places and times have different distributions of PBLH, is this really necessary? If the point of this figure is to compare distributions of PBLH obtained from different data sets, then the data sets must use the same locations and times for a fair comparison. Otherwise, the differences seen in the plot have no meaning since the differences could just be a result of when it was sampled. As it is now in Fig. 3a, CALIPSO and AMDAR have extremely different distributions, so the results of essentially no relationship in Fig. 7 is not surprising.

We deem that the point of this figure is neither "to show how different places and times have different distributions of PBLH" nor "to compare distributions of PBLH obtained from different data set". We would like to think of this figure as a necessary overview of PBLH from various datasets, which sets the stage for the following sections.

We completely understand the reviewer's concern about comparison with CALIPSO. Besides a few more revision to be brought up in the following responses, we add a note that the distributions of CALIPSO and AMDAR should not be directly compared given their different sampling. The sentence at lines 157–158 is revised to:

"One should note that the distributions of CALIPSO and AMDAR should not be directly compared given their different sampling and the large uncertainty from CALIPSO."

Nonetheless, one could compare the spirals (in-situ profiles) and HSRL (airborne lidar), which did happen during the same campaigns, in Fig. 3.

Mountain West: Given the high average PBLH in the mountain west compared to the rest of the country, the variance is likely to be much larger as well. This has a couple major implications. First, any differences between data sets are likely dominated by differences in the mountain west. Has this been assessed with this data set? This could be done fairly easily in two ways. Either use only the eastern or western half of CONUS and repeat the analysis, or normalize by PBLH. Again, because this is an empirical method, the results could be much different by sector.

Figure 2 in the manuscript does not seem to suggest that high PBLH values are associated with high variances. Moreover, the RMSEs of XGB vs. HSRL and spiral (Fig. 9c and Fig. 10c) do not show outstandingly high variances in Colorado than other regions. We have tested normalization by fitting the log of PBLH early on in this study, but the results

were not as good as using the PBLH. In the revised manuscript, we do acknowledge that a further step of our work can be separating the CONUS into different geographic regions, as the reviewer suggested. The following is added after line 344:

"Future improvements of model performance may be achieved by focusing on smaller geographic regions and fine tuning region-specific predictors."

PBLH Reference: With PBLH, as we are all aware, there is no 'gold standard' that is a reliable reference for comparison given limitations in spatial or temporal resolution, retrieval method, etc. When comparing XGB with the reanalysis and CALIOP, it is not clear if the same time periods are used. For instance, AMDAR used 2005 to 2019 AMDAR (Line 189), but CALIOP from 2006 to 2013 (Line 150). So do all these comparisons use a consistent period of time? If not, this may lead to biases from using different times.

After the XGB model is trained using AMDAR data (from 2005 to 2019 as the reviewer correctly pointed out), it can be used to predict PBLH at any other locations and times within the domain, so the comparisons in sections 4.1-4.3 are consistent in space and time. The following sentence is added to line 151 in the original manuscript to clarify this point:

"During the evaluation, we first obtain the model prediction at the same location and time of each CALIPSO sounding and then compare the predicted PBLH with the CALIPSO PBLH."

For a more thorough discussion about the PBLH references, we group the descriptions of the three evaluation datasets (CALIPSO, HSRL, and spirals) into a single subsection ("**2.3 Observational datasets used for evaluation**") and add a new subsubsection to overview the pros and cons of these datasets:

"2.3.4 Comparisons of observational datasets

As summarized by Figs. 1 and 3, none of the observational datasets described above can uniformly represent the PBLH over the study domain. CALIPSO features the most homogeneous spatial coverage (Fig. 1b), but its PBLH product relies on an automatic, global algorithm that may be subject to significant uncertainties. Yet the unique benefit of including CALIPSO data is that it can indicate errors in the spatial prediction made by our model, as the availability of AMDAR airports is spatially clustered (Fig. 1a). For example, no AMDAR sites are available in the large area over the Northern Rockies and Plains and the Southeast. Because of the large differences in AMDAR and CALIPSO PBLH, we consider the intercomparison involving CALIPSO more relative than absolute and focus on correlations rather than biases.

One should also note that CALIPSO and HSRL PBLH data are based on aerosol backscatter gradients, which is quite distinct from AMDAR, DISCOVER-AQ spiral profiles, and ERA5, where PBLH values are diagnosed thermodynamically. Although systematic differences between aerosol-based and thermodynamics-based PBLH may exist, we do not observe them by comparing spatiotemporally

close spiral and HSRL measurements in the same DISCOVER-AQ campaigns (i.e., comparing d vs. g, c vs. h, d vs. i, and e vs. j in Fig. 3). Furthermore, the model prediction from this work may serve as a "traveling standard" when evaluated against HSRL and spiral datasets. As will be shown in Sections 4.2 and 4.3, the biases between HSRL data and collocated model prediction do not show significant differences from the biases between spiral data and the corresponding model prediction."

Tuning and Training: Selecting 800 trees with a depth of 8, which is a large amount, still results in a rather large IQR for the test set, even considering differences of sample size. If this were just an issue with large variance, at least the IQRs would overlap. None of the IQRs between training and testing overlap (and even the 97.5 percentiles barely overlap!), suggesting little utility of using this method outside of the training data. This really points to some large underlying flaw, which could be related to a number of factors.

The number of trees, the tree depth, and a few other hyperparameters were determined from the data by cross validation using various metrics on the testing dataset (Section 3.2). The selected hyperparameters are the best performers on the testing data on average. Since the main motivation of this study is to fill the spatial gaps between AMDAR sites, we put more weight on the model performance than the simplicity of model. The model predictions are evaluated on three other observational datasets in Section 4, and neither large biases nor large variances are observed on locations away from AMDAR sites. The following sentence is added to line 220 of the original manuscript:

"Although driven by the data, this selection yields a complicated XGB model. Since the main motivation of this study is to fill the spatial gaps between AMDAR sites over the CONUS during the AMDAR period without further extrapolating in space or time, we put more weight on the model performance than the simplicity or computational cost of model."

It is unclear to us why *"None of the IQRs between training and testing overlap"* would suggest *"little utility of using this method outside of the training data"*. If there were really little utility of using the method outside of training set, the metrics on testing should approach a null model, i.e., zero R^2 and RMSE approaching the variance of the predicted variable. That's not what we observe in Fig. 4 of the manuscript.

Line 125: A good reason to use AMDAR and ERA5 is that they can both use the bulk Richardson number to find PBLH. Even though a critical Ri of 0.5 was used in a previous study with AMDAR, why shouldn't this work use a consistent critical Ri?

The AMDAR and ERA5 profiles are fundamentally different as one is measured in-situ, while the other is model-based and on a 0.25° grid. AMDAR profiles contain structures that cannot be resolved by ERA5. Hence the optimal parameters for both profiles are unlikely the same. In addition to a critical Ri of 0.25, ERA5 also used another parameter $b = 0$, whereas AMDAR used $b = 100$ (Zhang et al., 2020). Zhang et al. (2020) compared AMDAR

PBLH estimated with the same parameters as ERA5, but the results gave larger differences from ERA5 and larger biases relative to other observation datasets. We revised the sentence at line 125 accordingly:

"The PBLH from the ERA5 product is identified using the bulk Richardson number method but with slightly different parameters (ECMWF, 2017). Zhang et al. (2020) compared AMDAR PBLH estimated with the same parameters as ERA5, but the results gave larger biases relative to ERA5 and other observations."

Line 252: Using year as a factor in the final model is a surprising feature since there is no physical basis for this. This suggests that if extending to a new year of 2022, it is not possible to use relationships developed in this model, so it calls the generality or robustness of the model into question.

The year as predictor can capture any interannual variation that cannot be explained by the features. Similarly we tested day of year for any remaining potential seasonal variation, but day of year is ranked low in the feature important tests and not included. See lines 236–238 of the original manuscript. The main motivation is to fill the spatial gap when the AMDAR data are available, so we will not simply extrapolate to future years (revisions have been made in the previous responses). However, when the AMDAR data in 2022 become available, the model can be developed in the same way. We add the following sentence to address this point:

"The significance of year as a predictor indicates interannual variations that cannot be explained by other physics-based predictors. Therefore, this model should be used during the years when AMDAR data are available."

Fig. 5: Because the BL height at time 0, -1, and -2 is so important in this model, do you think that a linear trend would work just as well to get the BL height? If so, the simpler model is better.

We tested linear model using the same predictors, the RMSE on testing sets are about 5% higher than XGB and 3% higher than random forest. We add the following to line 204:

"The linear regression model, although the simplest and fastest, was not used because of its slightly lower performance than XGB (testing RMSE higher by $\sim 5\%$) and random forest (testing RMSE higher by $\sim 3\%$) and the fact that the computing cost of XGB is not of concern."

Section 4.1: Using CALIPSO as a benchmark seems problematic; there are many issues with PBLH retrievals from CALIPSO, and Fig. 7 shows that there is really no agreement at all with any data set to CALIPSO.

We thank the referee for this comment, which was also raised by the other referee. We were not necessarily treating CALIPSO as a benchmark, but an independent, observation-based sanity check at locations without AMDAR sites (e.g., Fig. 8 shows no excessive biases at

locations without AMDAR sites). This was perhaps not clear in the previous submission but is now clarified in multiple places in the revised manuscript, including abstract, introduction, and the newly added section 2.3.4.

Line 340: Yes, a natural next step is to extend it to other times, but the above issues would be much worse given the added difficulty of defining the nocturnal boundary layer.

Yes, we completely agree with the referee on the difficulty in defining the height of nocturnal boundary layer, which is exactly why we need future studies on this topic. We revise this sentence to:

"Since the AMDAR PBLH data are available hourly, it is possible to extend this work to other daytime hours and even nighttime hours with the caution that it will be more challenging due to the increase of stable conditions and less observational datasets available for evaluation."

References

ECMWF: Part IV: Physical processes, in: IFS Documentation CY43R3, IFS Documentation, ECMWF, URL `https://www.ecmwf.int/node/17736`, 2017.

Zhang, Y., Sun, K., Gao, Z., Pan, Z., Shook, M. A., and Li, D.: Diurnal Climatology of Planetary Boundary Layer Height Over the Contiguous United States Derived From AMDAR and Reanalysis Data, Journal of Geophysical Research: Atmospheres, 125, e2020JD032 803, https://doi.org/https://doi.org/10.1029/2020JD032803, e2020JD032803 2020JD032803, 2020.