Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., referee comment RC2
https://doi.org/10.5194/amt-2022-235-RC2, 2022

Comment on amt-2022-235

Anonymous Referee #1

Referee comment on "Estimates of the spatially complete, observational-data-driven planetary boundary layer height over the contiguous United States" by Zolal Ayazpour et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2022-235-RC2, 2022

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.

Major comments:

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.

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.

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.

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.

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.

Minor

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?

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.

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.

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.

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.