

Earth Syst. Sci. Data Discuss., referee comment RC2  
<https://doi.org/10.5194/essd-2022-150-RC2>, 2022  
© Author(s) 2022. This work is distributed under  
the Creative Commons Attribution 4.0 License.

## Comment on **essd-2022-150**

Anonymous Referee #2

---

Referee comment on "A merged continental planetary boundary layer height dataset based on high-resolution radiosonde measurements, ERA5 reanalysis, and GLDAS" by Jianping Guo et al., Earth Syst. Sci. Data Discuss., <https://doi.org/10.5194/essd-2022-150-RC2>, 2022

---

The authors have compared radiosonde-based PBL height estimates with PBL heights derived from the ERA5 reanalysis for over 300 available land stations, showing a significant bias in ERA5 PBL heights. A machine learning routine is developed to predict the ERA5 PBL height bias based on numerous input parameters, and this bias is subtracted from the ERA5 PBL height to produce a corrected dataset. This produces an immediately useful and relevant dataset that can be applied in many future studies. The work is novel, well-constructed, and succinctly explained in the paper. There are a few non-structural fixes that could improve the manuscript, but no major issues with the work, so I would only call these minor revisions.

Notes:

Line 203: can you add some detail on what you mean by the 'second level'?

In equation 2, it appears that PBLH-M and PBLH-E are mis-formatted as PBLH – M and PBLH – E (where M and E are variables being subtracted) This is probably a formatting error, but is initially very confusing.

Line 260: I worry that randomly dividing the data can cause an issue if certain geographic regions are underrepresented in the training data. I would recommend dividing your stations into specific regions (for example: valleys, mountains, coastal, continental, tropical, polar...) and ensuring that a subset for each region is then randomly drawn for each training/validation pool. An easier solution may be just to show that the randomly selected data already chosen for training represents these differing types of regions using a map and/or histogram.

Figure 2: Dividing by calendar season for stations on both sides of the equator is not recommended, since you are lumping winter with summer, autumn with spring, etc... It would be better to combine similar seasons, so that southern hemisphere DJF is combined with northern hemisphere JJA, etc... This would better illustrate seasonal biases.

Figure 8: the panels are too small for a meaningful comparison between PBLH-R and PBLH-M (comparing the dots to the shading). I recommend making larger maps available as supporting material, or showing this comparison some other way