Reply on RC2:

The manuscript presents a deep learning framework using an attention-augmented ResNet with transfer learning to estimate diurnal variations of near-global planetary boundary layer height (PBLH) from CATS lidar, explicitly addressing multi-layer structures in spaceborne backscatter profiles. The topic is timely, and the approach is interesting and potentially impactful. Please see the detailed comments below.

Specific comments:

Line 129: Please spell out the date as “January 10” rather than using an abbreviation, for consistency with the rest of the manuscript.
Response: The abbreviation has been expanded as “January 10” in the revised manuscript.

Lines 205–210: The accuracy metric is defined as the fraction of predictions within 500 m of radiosonde PBLH. While 500 m is a reasonable tolerance for some regimes, it can be relatively large for diurnal PBLH over land. To demonstrate robustness, please justify the choice of the 500 m threshold or provide a sensitivity analysis showing how key conclusions change with the tolerance chosen.
Response: We are grateful to the reviewer for pointing out this deficiency. Using a fixed threshold to determine the accuracy of the model is indeed too arbitrary, especially for this PBLH with obvious diurnal variations. Therefore, we conducted some sensitivity tests on the selection of the threshold, ranging from 300 m to 700 m. As shown in Table S2, it can be observed that as the threshold increases, the hit rates of all five peaks increase linearly. The conclusion in lines 205–210 of our manuscript is mainly to reveal that using multiple WCT peaks can better capture the true PBLH, providing theoretical support for the subsequent model construction; rather than to explain the deviation between these peaks and the true PBLH. Therefore, it can be said that the selection of the threshold does not affect the main conclusion of this work. In fact, choosing 500 m as the threshold is basically consistent with the dilation factor of 480 m for calculating the WCT profiles in this work, and it is almost the uncertainty range of the WCT retrieval algorithm. In the revised manuscript, we provided a demonstration of the results of these sensitivity tests.

Table S2. Hit rate (%) of the first five peaks when setting different threshold for calculating accuracy.

Threshold (m)	300	400	500	600	700
First Peak	30.39	33.59	35.77	36.99	38.11
Second Peak	13.31	15.8	16.97	17.78	18.19
Third Peak	11.38	12.65	13.62	14.23	14.63
Fourth Peak	10.26	11.38	12.55	12.75	13.06
Fifth Peak	8.74	9.6	10.16	10.57	10.87

Line 269: Please specify the interpolation method used to map MERRA2 meteorological profiles onto the 84 CATS bins.
Response: We sincerely apologize for the omissions in processing the MERRA-2 reanalysis vertical profiles in the manuscript. The meteorological profiles include temperature, relative humidity, and wind speeds obtained from MERRA2 reanalysis, which were first matched with each CATS orbit, following inverse distance weight for spatial matchup and most proximity for temporal matchup. And then the

spatio-temporally matched MERRA2 profiles were vertically aligned to 84 CATS bins using a linear interpolation method.

Lines 293–296: In the transfer-learning stage, the transfer-training set comprises 4,000 samples and the remaining 662 samples serve as a common test set. Please describe how you minimized spatial and temporal leakage between training and test sets. For example, indicate whether you used station-wise or region-wise splits, any temporal separation, and provide summaries/maps of the train/test distributions to verify independence.

Response: We fully agree with the reviewer's concerns regarding data leakage. In the process of choosing testing samples, we indeed overlooked this aspect and only selected 662 samples through random sampling, which obviously posed a risk of data leakage. Therefore, in the revised manuscript, we re-screened the testing samples. Due to the strict temporal and spatial isolation when matching the CATS profiles of different orbits with the sounding stations, we believe that the risk of data leakage is extremely low. In the revised version, we only processed the samples from the same orbit that were relatively close to each other. Specifically, if the distance between the sounding sites matched to the same CATS orbit is less than 300 km, we consider there is a risk of data leakage in the spatial dimension. This type of sample will not be included in the testing set but will all be placed into the training set. Additionally, in the revised manuscript, the radiosonde sites in the ocean-island were removed. To ensure that there were still sufficient samples available for training the transfer model after eliminating the ocean-based radiosondes, we increased the matching distance between CATS orbits and the sounding sites from less than 150 km to less than 200 km. Eventually, we obtained 5008 matching samples, of which 750 (15%) were used for the testing. Based on this new test set, we retrained the transfer model. The results showed that the accuracies of the base model, the pre-trained model, and the new transfer model on the new test set were 60.0%, 59.2%, and 68.3% respectively. This indicates that adopting a transfer learning strategy in this work is still appropriate. The new transfer model we trained did not change the key findings, but some of the results are different from those in our original manuscript. Therefore, we updated most figures in revised manuscript (from Fig.3 to Fig.10) when adopting newly trained transfer model and modified some of the conclusions.

Line 348: You indicate ocean profiles were removed during pre-training due to limited radiosonde matchups, yet Fig. 5 shows results over oceanic areas. Please clarify whether the model was trained only on land but applied over oceans at inference.

Response: Our pre-train model was only trained on land and excluded the grid points in the ocean. In Fig. 5 of manuscript, there are still some results on the ocean. Actually, these ocean-based results are not from the ocean surface but are the results of matching with the sounding sites on islands. To enable our model to have more possible samples for training, we took these island radiosondes into our work. However, we must admit that our work overlooked this point raised by the reviewer. The CATS data matched to the island-based radiosonde could potentially come from the ocean. Therefore, we re-examined the samples that matched these sounding sites and found that some of the samples had a surface type of water body, which contradicted the settings of the pre-trained model. In the revised manuscript, we removed all the samples with a land type of water body, but this would further reduce the already insufficient training dataset. To ensure a sufficient sample size as much as possible, we therefore increased the spatial range of CATS matching radiosondes from 150 km to 200 km (as the model's prediction biases show less dependence on matchup distance). Ultimately, we generated 5008 matching samples, and we chosen 750 samples (~15%) from them as the common testing set for both the pre-train and transfer train model.

Lines 372–376: Please elaborate on why the model performs more poorly from April to September. If available, add supporting analyses or references.

Response: Thanks the reviewer for pointing this out. The poorer performance of the model in the months of April to September mainly has two reasons. As the poorer performances are primarily sourced from the Northern Hemisphere, it can be concluded that the model's representation in spring and summer seasons were somewhat weaker than that in autumn and winter. For the spring and summer seasons, the atmosphere is vigorous, accompanied by more convective activities. This leads to more complex aerosol structures (more noised CATS signal), but also limits the representation ability of MERRA2. In contrast, the atmosphere is more stagnant, and the aerosol structure is simpler (Li et al., 2025). Additionally, our assessment is mainly based on absolute deviation. The higher PBLH magnitude in the spring and summer seasons will make the assessment worse. When considering relative deviation (NMAE, Fig. 3a), the performance improves somewhat, but it is still slightly poorer than that in the autumn and winter seasons.

References: Li, Y., He, J., Ren, Y., and Wang, H. (2025). Aerosol-PBL relationship under diverse meteorological conditions: Insights from satellite/radiosonde measurements in North China. Atmospheric Research, 321, 108125.

Lines 380–381: The permutation importance approach is appropriate, but shuffling individual features across samples in a sequence task can yield unrealistic feature combinations when predictors are correlated (e.g., temperature and local time).

Response: The reviewer's comments are very insightful. We carefully considered the issue pointed out by the reviewers. It is objectively true that by randomly shuffling individual features, some false correlations between features can be generated. In our manuscript, we calculate the importance of features by shuffling individual features across samples. However, we do not randomly change the order of a single feature completely; instead, we perform block-level shuffling using 84 bins (one sample contains 84 vertical layers). Each shuffling is only performed between blocks, and the order of this feature within a block is not shuffled. That is to say, these features retain their values as the height changes. And shuffling between blocks, each sample is independent (from different times and locations), and we believe this approach can effectively alleviate the problem pointed out by the reviewer. In this work, what we have shuffled is the sequence between blocks, rather than the internal structure of the blocks. In each block of samples, the local time, longitude, latitude, and altitude have the same value for 84 bins. As the reviewer pointed out, there is indeed a correlation between temperature and local time, but when the temperature feature is shuffled at the block-level, the temperature layers of the 84 vertical layers are replaced by the values of other samples. It can be understood that the sample is from other different locations or different seasons, and the physical correlation between temperature and local time is still reasonable. Of course, using permutation importance inevitably leads to some false feature combinations. We just minimize this impact as much as possible. We hope to hear further feedback from the reviewer. If necessary, we may replace it with another more suitable feature importance estimation method.

Lines 455–457: Given that absolute PBLH magnitudes vary substantially across regions and seasons, please report relative bias metrics in addition to absolute errors. This will better reflect performance where PBLH is small or large.

Response: Thank to the reviewer for your suggestion. Using an absolute PBLH deviation can only indicate a certain aspect of the model's performance and cannot represent the complete performance of the actual reaction model. Following to the reviewer's suggestion, we have added curves of relative deviations in Figs. 3a-b, Fig. 3f, and Fig. 5a-d. Compared to the absolute deviation, the changes in relative deviation are more gradual, which also reflects that the model's performance has higher temporal and spatial consistency. It is smaller in both diurnal variations and spatial distributions compared to the absolute deviation. Additionally, we have calculated the relative deviation indicators spatially. Although the spatial distribution is more uniform, there are still large deviations in desert areas and high-altitude regions. This reflects that the poor performance of the model in these areas.

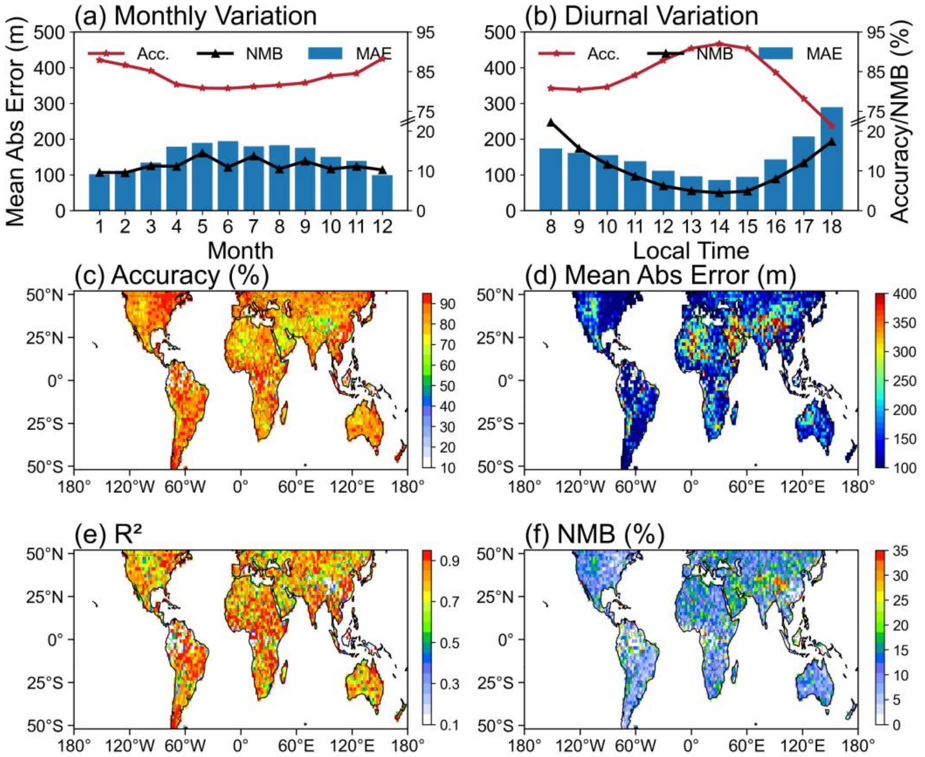

Fig. 3. Assessment of the pre-train model. (a-b) give the accuracy (column), MAE (black solid line) and NMAE (red solid line) at monthly and hourly scale, respectively; (c-f) denote the spatial distributions of accuracy,MAE, R2, and NMAE, respectively.

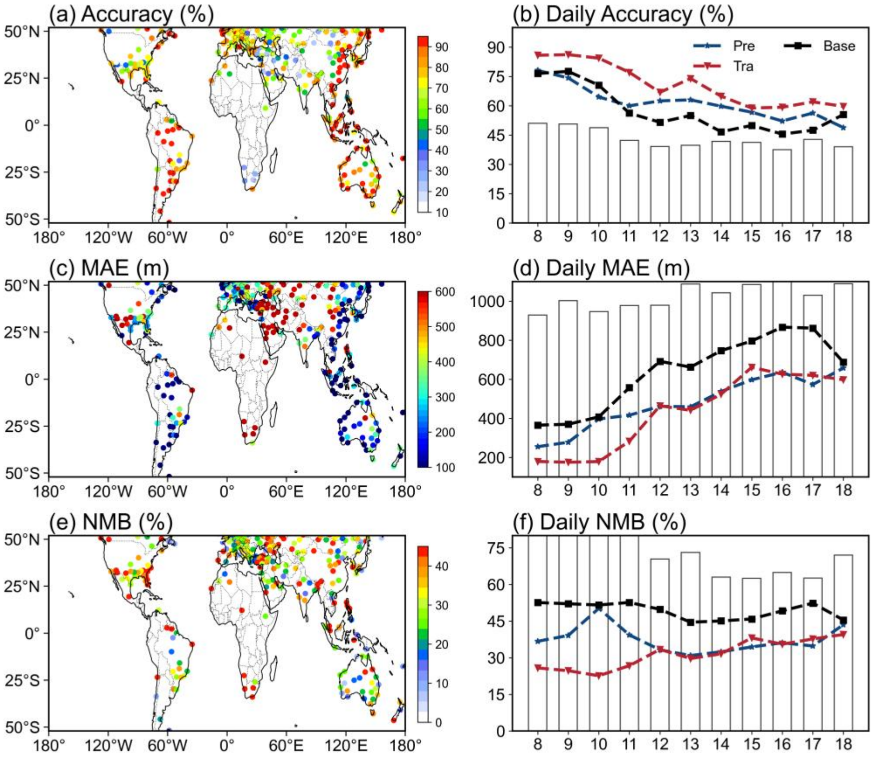

Fig. 5. Performance comparisons of the WCT, base, pre-train and transfer model against radiosonde constrained target labels. (a, c, e) show the spatial distributions of accuracy, MAE, and NMAE for transfer model, (b, d, f) display the diurnal variations of these metrics for WCT (column), base (back dash), pre-train (blue dash), and transfer (red dash) models.

Line 584: Please specify the source of the land surface type categories used.

Response: We appreciate the question raised by the reviewer. In fact, the land surface type data used in this work is all derived from the CATS dataset, which integrates the surface categories from MODIS. These data are one-to-one corresponding to each CATS profile and no additional processing is required. We provide further illustration of the surface type categories in the Section 3.1. In addition, the surface classifications shown in Figs. 9 and 10 in the manuscript are also based on them.

Please review the manuscript for tense consistency.

Response: We have carefully reviewed the entire manuscript and revised all the inconsistent tenses to ensure tense consistency throughout the paper. We have followed academic writing conventions for tense usage: the past tense is used to describe our experimental procedures and results, the present tense is used to state general scientific facts and the purpose of this study, and the present perfect tense is used to summarize the previous research progress. All the revisions are marked in red in the revised manuscript.