RC1:

The preprint describes a machine-learning model ("Auto-ML") trained to diagnose the convective boundary layer height (CBLH) evolution over one day. Generally, I think the choices described to add physical grounding to the ML model are well-motivated, though the paper description of them as providing 'implicit physical constraints' may be a bit of a reach. The paper would be much stronger if it included a baseline method of CBLH prediction; without one it lacks context for judging the Auto-ML skill.

Reply: We appreciate the reviewers' constructive feedback on the physical basis of the ML model. We agree that the term "implicit physical constraints" may be too broad and will revise it to "implicit thermodynamic physical constraints." To contextualize AutoML's capabilities, we have added the Linear Regression algorithm results as a benchmark in the supplement to highlight AutoML's performance and discussed in the revised manuscript.

Specific comments:

1. Simply including LTS and surface fluxes as inputs and using the full day of CBLH as targets does not guarantee that the ML model will learn the correct physical constraints. It is fair to say that these choices introduce more physical grounding into the ML problem setup, but I think that describing these as "Implicit physical constraints" in the title and section 2.5 is too far-reaching.

Reply: We agree that including LTS and surface fluxes as inputs with full-day CBLH targets does not ensure the ML model captures all physical constraints. Morning boundary layer growth is thermodynamically driven, while afternoon CBLH peaks involve entrainment, typically parameterized (~0.2) but lacking direct physical representation. We revised it as "implicit thermodynamic physical constraints" in the title and further discussed in Section 2.4.

2. My reading of the multisite analyses in 3.3 and 3.6 is that generalizing the model to different sites is limited by the flux input differences and site-specific differences and that training the ML model on the site it is to be used on is needed to achieve the best skill. This seems to contradict the abstract ("transferability across ARM Southern Great Plains sites… confirm the model's robustness").

Reply: We appreciate the reviewer's suggestion. The core finding of our study is that current ML models for CBLH prediction exhibit limited transferability across sites due to site-specific factors. By using on thermodynamics as a primary driver, our model achieves improved transferability across ARM Southern Great Plains sites, as stated in the abstract. However, model performance (R^2) declines with distance, which is physically reasonable due to variations in non-thermodynamic factors such as terrain and vegetation, beyond just flux differences. To address your concern, we changed the abstract to "The model's generalizability across multiple sites at the ARM SGP site demonstrates its potential for transfer to greater distances, offering a scalable approach for enhancing boundary layer parameterization in atmospheric models", consistent with findings in Sections 3.3 and 3.6.

3. There is no baseline for comparison to assess how much skill the ML models are adding. I suggest including a simple baseline R2 and MAE as calculated using the training set mean CBLH target (over full time range, and also seasonal for that analysis) and including this baseline on the skill figures and tables. This would add context for how much of an improvement the AutoML model is providing.

Reply: Thank you for this highly constructive suggestion, which significantly enhances the article's readability. We have incorporated a Linear Regression algorithm as a baseline to demonstrate the extent of improvement provided by AutoML. For details, see Text S1 and S3 and Figures S1 and S3 in the Supporting Information.

4. In the description of input and output data in Sec 2.6, I would add a sentence explicitly stating the dimensionality of the input and output data. Related to this point, it sounds like aside from sunrise and sunset times, each input has (n_timestamps_in_day) values in the full input vector. However, later in the interpretability section, single importance scores are given for each input, which confused me. Are SHAP values calculated for each timestamp of an input and averaged together? Please clarify in the text.

Reply: We appreciate the reviewer's feedback. We have added a sentence in Section 2.5 (section 2.6 in the first manuscript) explicitly stating the dimensionality of the input and output data. Indeed, aside from sunrise and sunset times, each input consists of n_timestamps_in_day values in the full input vector. For the interpretability section, we clarify that SHAP values are calculated to reflect the relative importance of each input variable across the entire day, not at individual timestamps. We have revised the text to make this clear and avoid confusion in section 3.4.1.

5. What was the best model out of the set in table 2 chosen by the AutoML? This should be added to the text. Was it one of the two models in section 3.4? Did any other models in Table 2 also have comparably good skill, or were some significantly worse? Some discussion of the best performing architecture is warranted as could relate to the model's ability to generalize. e.g. one would expect a tree-based model to have difficulty generalizing as the output distribution is bounded by its training set.

Reply: Thank you for the reviewer's insightful suggestions. The best model selected by AutoML varied across different sites and even with different training data splits for the same site, making it inconsistent to highlight a single model. To avoid misleading readers, we treated the AutoML process as a black box and did not specify a single best model in the initial draft. We have now included the best model, which is the "ExtraTreesRegressor" from Section 3.4, as clarified in Section 3.1. In Table S1 (Original manuscript Table 2), many models exhibited comparable performance, with R^2 differences within 0.01. We have added a discussion in the revised manuscript noting that, with limited training data, tree-based models generally outperformed neural network architectures. AutoML underperforms in winter may relate to generalization challenges, as tree-based models' outputs are constrained by their training set.

6. The methods section should include some information about the computational resources used in training. This affects the space of model hyperparameters that can be explored by the Auto-ML algorithm. In particular the tree depth in the tree-based methods is directly related to the distribution of possible model outputs.

Reply: Thank you for the suggestion. We have added Section 2.3 to describe the computational resources used: "Windows 11 OS, Intel® Core™ i9-10900 CPU @ 2.8 GHz, 32 GB RAM." We agree that hyperparameters, such as tree depth in tree-based methods, influence the AutoML model search space and output distribution. However, a comprehensive exploration of hyperparameter tuning for each model would be computationally intensive and impractical. Therefore, this study employs default hyperparameters to facilitate model comparisons. Our related work (DOI:10.3390/rs17081399, 2025a) provides a detailed discussion on the effects of tree depth, learning rate, and number of estimators. These findings do not impact the generality of the results presented here.

7. In the interpretability section, there should be some discussion of whether the results were surprising or expected given prior knowledge of boundary layer processes. E.g. "In spring and autumn, while a comparable pattern exists, the differences between predicted and observed values are smaller, suggesting lower variability (or complexity) in meteorological conditions compared to summer." and "Potential reasons include:... distinct entrainment processes in summer compared to other seasons". I am not familiar with boundary layer processes, so for readers like me: Is it implied that it is already known that summer has lower variability in conditions and distinct entrainment processes, or are those the authors' hypotheses to explain their findings?

Reply: Thank you for your insightful and valuable feedback. We address this below and have incorporated a detailed discussion in Section 3.5.2 of the revised manuscript.

We fully concur that explicitly addressing whether observed patterns align with prior knowledge—or represent interpretive hypotheses—will aid readers unfamiliar with these processes, fostering a more accessible and rigorous discussion. The peak convective boundary layer height (CBLH) in summer (~2 km; Fig. 8b2) exceeds that in winter (~1 km), consistent with established literature. However, no prior studies have employed thermodynamic parameters to predict CBLH, rendering this approach novel. At the same time, we delineate our interpretations of summer-specific discrepancies—e.g., the pronounced widening of the interquartile range (IQR) in JJA, potentially driven by unmodeled wind-driven advection and enhanced entrainment from intense convective activity—as hypotheses grounded in process-based reasoning, rather than established consensus, to underscore the contributions of this work.

These revisions are complemented by the addition of a new panel (Figure 8c), which visually contrasts absolute and relative differences diurnally and seasonally, enabling clearer discernment of scale-dependent patterns and their implications—for instance, highlighting how relative discrepancies exceed 0.5 in autumn and winter mornings/evenings, while remaining below 0.1 during midday across seasons. We now emphasize that the winter-summer contrasts in CBLH scale and IQR are consistent with known seasonal forcings on boundary layer development, whereas the diurnal sensitivities and summer-specific variabilities represent novel insights, which we attribute to unresolved complex interactions like advection and entrainment.

To further guide future refinements, we propose incorporating parameters such as entrainment rates,tempuature and wind profiles to mitigate these gaps. We believe these enhancements not only directly address your query by distinguishing expected patterns from our proposed explanations but also elevate the manuscript's scientific depth.

We believe these clarifications and revisions strengthen the interpretability of our results and address the reviewer's concerns comprehensively. Thank you again for your constructive feedback, which has helped refine the manuscript.

8. I appreciate the breakdown of the results into the seasonal comparisons in section 3.5.2 and discussion of the physical processes affecting the CBLH and its variability. Here and in other sections, I think the writers did a good job of explaining how the physical processes involved in boundary layer changes might explain their findings.

Reply: Thank you for the positive feedback on Section 3.5.2 and our discussion of physical processes linking boundary layer dynamics to CBLH variability.

9. The readability would be greatly improved if the main text section related to importance/interpretablity just focused on the main takeaway (LTS dominates) and left the rest to an appendix. Similarly for the section about ECOR vs EBBR flux results; I did not feel the findings were salient to the main points of the paper.

Reply: We appreciate the reviewer's feedback on readability. The main focus of the paper is not solely to highlight LTS dominance but to demonstrate the accuracy and multi-site applicability of thermodynamic implicit constraints for full-day CBLH predictions, including seasonal comparisons. The comparison of ECOR and EBBR fluxes addresses a key challenge in atmospheric science regarding data assimilation, a potential further goal of using ML in this study. To improve readability and emphasize the main theme, we have moved Previous article manuscript Sections 2.3，Sections 2.4， Sections 3.1.2 and 3.2.2, along with Table2, Figures 5 and 7, to the supplementary materials.

Other comments:

10. Please define the variables in equation 4.

Reply: Added.

11. Hyperparameters for the ExtraTreesRegressor in Sec 3.4 should be provided.

Reply: Added.

12. Why is only JJA used in the comparison of the different ML methods in 3.4? Is it because the authors specifically wanted to study the season with higher DL-derived CBLH variability? Please clarify in the text.

Reply: We thank the reviewer for the comment. The choice of JJA in Section 3.4 was driven by its higher DL-derived CBLH variability and the greater availability of data, ensuring more robust results. We have added a clarification in the first paragraph of Section 3.4: "JJA was selected due to its higher DL-derived CBLH variability and larger data volume, enhancing result reliability."

13. Table 4: What is being shown in the rows labeled by the inputs? Feature importance? Please clarify in the caption.

Reply: Fixed. Yes, it is "Feature importance".

14. In the conclusion, L849 states the ML model "significantly improves the accuracy and generalizability of CBLH predictions across diverse sites and seasons." This ought to be edited as without a baseline for comparison, it is unclear this improvement is relative to.

Reply: We added the baseline. The statement on L849 has been revised to: "This implicit thermodynamic physically constrained Auto-ML approach selects the best-performing machine learning model based on the dataset, improving the accuracy and generalizability of CBLH predictions across diverse sites."