Dear Thomas Oberleitner,

We appreciate your time and effort to provide comments on our preprint. Please see below our replies (in blue) to each of your points.

Thank you for this interesting preprint. Here are some notes and suggestions regarding the presentation of the results.

1. The preprint compares the predictive performance and parameter/response relationships of random forests and neural networks. The benefit of comparing two high-capacity/complexity models on tabular data is not clear, as it is self-evident that both can achieve good results given proper handling. In fact, because regularization in NNs can be more difficult, they are generally outperformed by easier to use off-the-shelf models such as random forest and gradient-boosting [1].

Furthermore, we don't know of any literature in ML research supporting the claim that NNs would be inherently better at intra -or extrapolation unless they incorporate domain specific properties [2], for example the network architectures used in physics-informed NNs. Nor do the results suggest systematically better generalization performance. In table 3, RF models show higher R^2 than NNs for all RCPs and the slightly better performance in table 4 could be due to the choice of hyperparameters, random seeds, etc.

We therefore suggest removing the model comparison and focusing on the RF model.

Thank you for this valuable feedback. We agree that RFs are often strong models for tabular data, and that both RFs and NNs demonstrated comparable predictive accuracy in our study. However, our objective here was to go beyond model accuracy and investigate explanatory power or model interpretability. In other words, we wanted to examine how different model architectures - with their different inductive biases - capture the true relationships in complex forest carbon dynamics. In this respect, we believe that our model comparison is a valuable contribution, especially given the scarcity of studies examining the trade-offs between model accuracy and interpretability in statistical models and emulators for ecological applications (Hu et al., 2023). We will clarify our rationale for comparing these models in a revised version, as this aspect could be better articulated in the objectives and discussion of the manuscript.

We also acknowledge your concerns regarding the claim that NNs are better at extrapolation, as this statement is not widely supported in the ML literature, and model performance can be highly dependent on the domain/data, and training process quality. Additionally, since our study does not include a systematic analysis to substantiate this claim, we will revise the manuscript to remove such statements.

2. The NRMSE in the model summary seems redundant as another scale-free metric in the form of the R^2 is provided. Additionally, NRMSE is highly sensitive to outliers, whereas R^2 is much less affected.

We included the NRMSE to provide an additional perspective on model performance, particularly for readers who may find error magnitudes more intuitive. Additionally, NRMSE facilitates comparisons with other machine learning emulators in the field, such as the work by Sun et al. (2022). However, if other reviewers strongly recommend its removal from the main text, we are open to relocating it to the supplementary materials

3. The point that the emulator reproduces LPJ-GUESS outputs well is made rather strongly in section 4. For example, line 215: "… emulators were able to generalize to LPJ-GUESS outputs produced with climate projections not included in the training data without a significant decline in performance". It is not clear what "significant" refers to, nor is this evident from table 3 and 4, which show low R^2 values for most responses. The next sentence highlights the low NRMSE values, which could be deflated due to outliers (see remark 2).

We suggest more careful wording regarding emulator performance and to put it into a more applied context, i.e., by highlighting its efficiency in a specific task. While it reproduces average outputs of LPJ-GUESS well (figures 2 and 4), it most probably cannot reproduce extreme outputs of the process model. An analysis of residuals can help in verifying that. The potential inaccuracies in predicting non-average responses should then be noted somewhere, as the emulator seems to be intended as a highly efficient proxy for the process model.

The statement in question (line 215) was intended to highlight the emulator's ability to generalize to unseen RCP scenarios. Specifically, when examining the R^2 values in Tables 3 and 4, we observe no decline or clear differences between the R^2 values for RCPs used during training (RCP2.6 and RCP8.5) and those for the unseen test scenarios (RCP4.5 and RCP7.0). We noticed that the proximity of this statement to the sentence on the emulator's overall ability to reproduce LPJ-GUESS outputs may lead to confusion, and we will revise the passage to improve clarity. We also appreciate your suggestion to analyse extreme model outputs, as this would provide important information into the emulator's performance. We plan to incorporate this analysis into a revised version of the manuscript.

Additionally, we will refine our wording to avoid overstating the emulators' ability to fully reproduce LPJ-GUESS outputs and to better contextualize its performance in their intended use.

4. The attribution of importance to features using Shapley values in the way it is presented could be misleading in the presence of correlations. This is a property of all data-driven models trained on correlated data, which is why all measures of importance are affected by this to varying degrees (e.g., total information gain in random forests, coefficients in linear models, etc.). In our experience, climate and other data used to train process model emulators are highly correlated and have a major effect on explanations. This can scramble the importance ranking of correlated features and even flip their Shapley value sign [3].
Furthermore, the text does not mention the ranking method for the features, which makes it hard to compare with the SHAP plots. Provided the authors stick to Shapley values, having the rank number included in the feature names in the plot would help to understand the conclusions drawn in the text.
We recommend supplementing correlation analysis, remove correlated features and/or weakening the language and inferences made about them. In many cases, feature selection algorithms can help in removing correlated features.
Alternatively, global explanations of feature importance could be used to rank features or supplement the Shapley results, such as contribution to loss function, global information gain, permutation importance, etc. As mentioned above, such measures are also not robust against correlations, but they might warp the results in a less drastic way. For some ML models, they are directly incorporated into feature selection algorithms [4].

Thank you for your comment and the references on this issue. We are currently studying the best way to address this limitation of the method. So far, we have come across and tested an approach described by Au et al. (2022) and Molnar (2022) for dealing with correlated features in this type of analysis. We plan to conduct a feature correlation analysis, group highly correlated features together, and then calculate and interpret Shapley values at the group level rather than for individual correlated features. While this approach is not perfect, as the interpretations are coarser at the group level, we believe it will make our feature importance interpretation more robust in light of the SHAP method limitations. We will also use Shapley-based global feature importance rankings and clustering analysis (as implemented in the Python SHAP library) to visually demonstrate feature redundancy in our importance plots. Any changes will be documented in the revised methods section.

Minor remarks
a. In the NRMSE equation (2), the term under the square root in the numerator should be divided by n.

We thank you for catching this error. We will correct Equation (2) in the revised manuscript.

References
[1] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 6, pp. 7499–7519, Jun. 2024, doi: 10.1109/TNNLS.2022.3229161.
[2] V. N. Vapnik, "An overview of statistical learning theory," IEEE Trans. Neural Netw., vol. 10, no. 5, pp. 988–999, Sep. 1999, doi: 10.1109/72.788640.
[3] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values," Artif. Intell., vol. 298, p. 103502, Sep. 2021, doi: 10.1016/j.artint.2021.103502.
[4] "CatBoost Feature Selection." Accessed: Feb. 26, 2025. [Online]. Available: https://catboost.ai/docs/en/concepts/python-reference_catboost_select_features
---

Thank you once again for your constructive feedback. We look forward to implementing these points in a revised version of the manuscript.

Best regards,
Carolina Natel, on behalf of all co-authors

Hu, T., Zhang, X., Bohrer, G., Liu, Y., Zhou, Y., Martin, J., ... & Zhao, K. (2023). Crop yield prediction via explainable AI and interpretable machine learning: Dangers of black box models for evaluating climate change impacts on crop yield. *Agricultural and Forest Meteorology*, *336*, 109458.
Sun, Y., Goll, D. S., Huang, Y., Ciais, P., Wang, Y. P., Bastrikov, V., & Wang, Y. (2023). Machine learning for accelerating process-based computation of land biogeochemical cycles. *Global Change Biology*, *29*(11), 3221-3234.
Au, Quay, Julia Herbinger, Clemens Stachl, Bernd Bischl, and Giuseppe Casalicchio. "Grouped feature importance and combined features effect plot." *Data Mining and Knowledge Discovery* 36, no. 4 (2022): 1401-1450.
Molnar, Christoph. *Correlation Can Ruin Interpretability*. *Mindful Modeler*. Accessed 25 March 2025. https://open.substack.com/pub/mindfulmodeler/p/correlation-can-ruin-interpretability?utm_campaign=post&utm_medium=web