Dear Reviewer,

We appreciate your time and effort in reviewing our work. Below, we provide our replies (in blue) to your comments. For easier reference, we've highlighted the revised text in green.

Carolina Natel et al. developed both a Random Forest model and a neural network model to emulate the dynamics of ecosystem carbon fluxes and carbon pool changes. These machine learning models have been widely applied to land models or components in the past and have consistently demonstrated effectiveness. Similarly, this study shows reasonable performance in emulating the target variables. The paper is well-written, with well-documented data and code. Overall, this is a solid modeling paper. Below, I have a few specific comments:

1. Interpretability vs. Physical Consistency

My primary concern is the balance between model interpretability and physical consistency. While SHAP has been used to interpret the ML models, it does not ensure that the emulators capture established physical knowledge embedded within the land model (in this case, LPJ-GUESS). I encourage the authors to further explore this aspect, as it is fundamentally important to understand the functional relationship emerging from ML emulators.

For example, in land models, atmospheric CO_2 concentration is a key driver of vegetation productivity, while temperature (T) strongly influences soil carbon stocks. Ideally, such first-order relationships should also be reflected in the trained ML emulators. One way to test this would be to leverage factorial LPJ-GUESS simulations with:

(a) Future SSP climate scenarios + historical atmospheric CO2 levels.

(b) Future SSP CO2 levels + historical climate conditions (e.g., repeated climate from 2010–2020).

If the trained ML emulators can reproduce the results of these factorial runs, it would provide strong evidence that the emulators have captured critical relationships between environmental drivers and carbon dynamics.

We appreciate this comment and recognize the importance of evaluating whether the emulators capture the physical relationships embedded in LPJ-GUESS, such as those mentioned by the reviewer. In response to the first comment from another referee (RC1), we have revised our discussion section on ML explainability, specifically interpreting the SHAP values in the context of previously published LPJ-GUESS sensitivity studies. We invite the reviewer to refer to the updated section, which can be found in our reply to RC1 (https://doi.org/10.5194/egusphere-2024-4064-AC3).

Regarding the suggestion to evaluate the emulation using factorial LPJ-GUESS simulations, we offer the following clarifications. A similar approach was partially explored during the early stages of emulator development, as outlined in the manuscript (Lines 355–360). Our initial hypothesis was that training the emulators on a factorial experiment, with independent variations in temperature, precipitation, and CO_2, would allow them to generalize to future RCP scenarios. However, we encountered a key challenge the factorial experiments generated feature spaces with physically implausible combinations in the feature and target space (e.g., low CO_2 paired with high temperatures, which only arise in extreme future scenarios with elevated CO_2, or extreme unrealistic ecosystem carbon in regions in which bioclimatic or soil constraints would not allow). This mismatch between training data and the intended application (RCP-based projections) led to poor generalization, as ML models rely on statistical patterns within the training distribution.

To address this issue, we tested a two-step training approach using neural networks. We pre-trained the NN on factorial experiments and fine-tuned it on RCP scenarios. However, the extreme imbalance between these two datasets caused one of two problems: (i) models became biased toward the factorial scenarios and performed poorly on realistic RCP trajectories, or (ii) when the weights were allowed to adapt flexibly to the new training data, fine-tuning effectively overwrote the pre-trained weights. This rendered pretraining ineffective and substantially increased training time.

It is important to note that purely ML-based emulators do not explicitly enforce process-based relationships in the same way as process-based models. As such, they differ fundamentally from process-based models and cannot generalize across different data distributions. We could have included in our training process some sort of pre-processing technique to balance the training data, in which we for example, oversample the examples similar to realistic RCP scenarios to balance out the training data, however, having such a flexible emulation approach was out of the scope of this work. Despite this limitation, our ML emulators remain highly effective for their intended application. Our training dataset consisted of physically plausible RCP scenarios at each grid cell, ensuring that the emulators learned realistic covariances among drivers. We do, however, acknowledge that further development of the LPJ-GUESS emulator for LandSyMM could benefit from constraining the emulator with LPJ-GUESS 'physics', which would make it more useful in a broader range of applications. We believe the revisions we made in response to RC1 (Main Comment #1) partially address this concern, especially the newly added paragraph in Section 5.4, "Emulator Application", which reads:

New passage to address RC1 and RC2:

It's important to note that while the emulators were generally able to reproduce LPJ-GUESS's outputs related to forest carbon dynamics for the employed RCP scenarios, they should not be expected to capture all original model sensitivities, including both parameter sensitivities (e.g. parameters governing vegetation dynamics), and the original model' physical responses (e.g. the response of carbon dynamics to atmospheric CO_2 outside the bounds of training data).

Additionally, we plan to expand the explanation of our preliminary tests in Lines 355–360 to further clarify the issues encountered when testing the emulator with factorial analysis scenarios, as discussed in this reply.

2. Justification for Annual Time Step

Further justification is needed regarding the choice of an annual time step. Land models typically operate at much finer temporal resolutions (e.g., daily, hourly, or at least monthly). It would be helpful to explain why the annual scale was selected and how potential loss of information at shorter timescales may affect the emulator's performance.

Yes, indeed, some LPJ-GUESS processes are updated daily (e.g., phenology and soil organic matter dynamics), while carbon allocation occurs annually. Our decision to use an annual time step for the input features was based on practical considerations. On one hand, we aimed to develop an emulator that could replicate LPJ-GUESS's interannual variability under RCP scenarios with sufficient accuracy (Fig. 2 and Fig. 4). On the other hand, we wanted to avoid including variables that might introduce unnecessary complexity without improving model performance. While finer temporal resolutions capture more information, they also introduce significant noise and variability, which may not enhance emulator performance and could even degrade it, in addition to increasing model complexity and training time.

We recognize that certain intra-annual climate information is essential for modeling the impact of climate change on carbon dynamics, such as the effect of rising mean annual temperatures on extending the growing season and increasing GPP, particularly in boreal regions (Piao et al., 2012). Therefore, we included seasonality-related variables, such as the highest mean monthly temperature (m_temp_max) and total annual growing degree days (gdd0), to capture these climate-driven processes while maintaining model simplicity.

The following clarification has been added to the methods section:

New passage 1:

To emulate the LPJ-GUESS model, we used input features aggregated to an annual time step. While some LPJ-GUESS processes operate at finer temporal resolutions (e.g., daily updates for phenology and soil dynamics), key carbon fluxes, such as allocation, are computed annually. Our goal was to develop an emulator that accurately captures interannual variability in carbon dynamics under future climate scenarios while avoiding the complexity and noise associated with higher-frequency inputs. To retain essential intra-annual climate signals relevant to carbon responses, such as the effects of seasonality on productivity, we included variables such as the total annual growing degree days above 0°C (gdd0). This approach balances model simplicity with the need to represent critical climate-driven processes affecting carbon dynamics.

Additionally, we have included the following statement in the discussion of model explainability to highlight the potential loss of information at shorter timescales. This sentence will be included after the discussion of the SHAP values alongside LPJ-GUESS climate sensitivity.

New passage 2:

It is important to note that, due to the use of an annual time step for input features, some LPJ-GUESS sensitivities to intra-annual climate cannot be fully captured in our simulations. However, we believe the selected input features are well-suited for capturing the interannual variability in carbon dynamics for our intended application.

3. Capturing Inter-Annual Variability

Given the focus on annual time steps, evaluating the emulator's ability to capture inter-annual variability in carbon fluxes (in addition to long-term trends) would be an important validation metric. Although the training is performed at the grid cell level (random sampling), it may also be valuable to include spatially aggregated fluxes (e.g., global/regional totals) as part of the loss function. This could improve the model's ability to represent inter-annual variability at regional or global scales.

As we have demonstrated, the emulator not only captures long-term carbon dynamics trends (e.g., long-term carbon uptake or losses), but also the year-to-year variability. Therefore, we are uncertain about the reviewer's comment. It's possible the reviewer is referring to intra-annual (monthly or seasonal) variability in carbon dynamics; however, this does not align with the intended use of the emulator, nor is it a standard output in LPJ-GUESS climate assessment studies.

The suggestion to include spatially aggregated fluxes and more physics in the loss function is indeed appreciated, and we might consider improving the realism of the emulators along with further applications we might need in the future, for example, representing a diverse range of forest management options (thinning, planted species etc).

4. Treatment of Disturbance Intensity

The results show disturbance as one of the most important features. However, it remains unclear how disturbance intensity (e.g., fractional area burned by wildfire, or land-use/land-cover change) is handled. How does the ML model represent partially disturbed grid cells? Additional clarification on this point would be needed.

To clarify, the "disturbance" feature refers to the time since the last stand-replacing disturbance, such as a forest clearcut event in the LandSyMM framework. This feature is not designed to capture detailed disturbance intensity, but rather to track forest recovery. To avoid any confusion, we are adding a sentence in our Methods section to clarify what type of disturbances are being represented in the emulation and what this feature means.

New passage:

In our LPJ-GUESS setup, we simulate only Potential Natural Vegetation and do not account for land-use changes. Additionally, since the emulator is designed to predict forest carbon potentials, we have disabled the fire module. The only disturbance incorporated in the emulation is the LPJ-GUESS representation of external disturbances (e.g., windstorms, plant diseases) through a generic patch-destroying regime with a stochastic probability based on the expected return time. Disturbance return times vary significantly across global forest areas (Pugh et al., 2019), and the interval chosen in this study (100 years) is a simplification commonly adopted in previous studies using LPJ-GUESS and other vegetation models, as reported by Zaehle et al. (2005).

As our emulation application is designed to model forest regrowth following a clearcut event within LandSyMM, we have included a feature called "time since the last disturbance" (in years) to track forest recovery. Within LandSyMM, this feature is reset to zero whenever a clearcut is performed.

Additionally, we believe there was a misunderstanding regarding the emulator's application, and we apologize for not making this clearer earlier. The LPJ-GUESS emulator will serve as a fast proxy for carbon dynamics when coupled with land-use models (e.g., PLUM within LandSyMM) to optimize future decisions. However, it is not intended to fully replace LPJ-GUESS within the LandSyMM framework. For example, the full LPJ-GUESS model will still be used to simulate crop productivity and other land uses that do not involve forests. Once the forest use scenarios are optimized for a certain simulation step, the resulting maps will be passed back to the original LPJ-GUESS model to predict actual carbon dynamics, including full disturbance effects (e.g., we might activate the fire module depending on the scenario). This final simulation step will ensure that we fully utilize the original model's sensitivities, but it's not computationally intensive.

To solve this misunderstanding, we are replacing the old passage:

Within LandSyMM, the emulators will replace LPJ-GUESS simulations in an online coupling to optimize global forestry decisions over multi-decadal timescales.

New passage:

In the context of the LandSyMM application, where we couple a land-use predicting model (e.g., PLUM) with LPJ-GUESS, the emulator will replace the original model at each coupling time step (e.g., every 5 years) to quickly estimate forest carbon potentials. This eliminates the need for full 100-year forest potential simulations for several scenarios. However, the emulator will not be used to predict actual vegetation carbon under the predicted land use for the next 5 years. Instead, a real LPJ-GUESS simulation will estimate vegetation carbon based on the predicted land use, and a new coupling step will then occur.

5. Recommendations for future work: Since land models simulate continuous, time-dependent changes in carbon fluxes and pools, it may be worthwhile to explore time-series ML models (e.g., RNNs, LSTMs, or Transformers) in future work. Such models could potentially outperform static models like Random Forests and ANNs by better capturing temporal dependencies and dynamics.

Thank you for this suggestion. We agree that exploring time-series ML models, such as RNNs, LSTMs, or Transformers, could be highly beneficial as we expand the emulator's capabilities to include more complex dynamics, such as multi-year recovery after disturbances and memory effects from land-use changes in soil carbon.

Thank you once again for your valuable feedback. We hope that we have addressed your concerns and we are positive that your comments will help strengthen our manuscript.

Best regards,

Carolina Natel, on behalf of all coauthors