Response to Reviewer 1

Legend
Black text: reviewer comments
Blue text: responses to comments

https://doi.org/10.5194/egusphere-2025-1617-RC1

General comments:

To predict ecosystem gross primary productivity (GPP), this manuscript utilizes ecosystem flux data, meteorological measurements from 109 globally distributed sites, and remotely sensed vegetation indices to train three models: a mechanistic, theory-based photosynthesis model, a memoryless multilayer perceptron (MLP) and a recurrent neural network (Long Short-Term Memory, LSTM). The authors found that both deep learning models outperform the P-model, and the LSTM performs best. Particularly, model skill is consistently good across moist sites with strong seasonality. Model error tends to increase with increasing potential cumulative water deficits. The LSTM adapts better to arid environments affected by water stress, yet there is still a large variability in model skill across relatively arid sites.

This is an interesting analysis and the topic is pretty important. Overall, I find the paper compelling and fit for publication after revision. I include my comments below, which I hope help the authors to further strengthen the paper.

1. Some important details should be provided in this paper. Firstly, the methods and dataset for determining the optimal hyper-parameter of the proposed model are not provided.

   Fair point. Our aim is certainly full reproducibility of all experiments, and we will make sure to clarify all remaining details, e.g., the hyper-parameter tuning approach, in the revised manuscript. We will add details on the number of data folds used, the stratification approach, and evaluation metric. Moreover, the code to reproduce the experiments has been published open-source.

   Additionally, the rationale behind selecting these three specific models, especially the combination of machine learning models (LSTM and MLP) with the process-based P-model, should be better justified. It is also recommended to explore and compare other state-of-the-art models, such as gated recurrent units (GRU), convolutional neural networks (CNN), and other sequence modeling approaches, to provide a more comprehensive evaluation.

   OK. While we do not expect big differences with alternative sequence models, we will add further comparisons in the appendix. We will include this in our

justification of the temporal model choice. We will also further clarify the choice of the process-based model as a benchmark with a known treatment of temporal effects.

2. In order to prove the superiority of the proposed models, this paper has carried out several forecasting experiments. Three models simulate GPP dynamics across a range of environmental conditions and vegetation types, please give numbers of samples for each classification.

   You are right, we overlooked this. We will annotate the figures with the number of sites in each category.

   Furthermore, figure descriptions must be more precise. For example, in Section 3.3 (Lines 240-255), the discussion of Figure 6 should clearly indicate that it comprises two sub-figures and describe each accordingly.

   Thank you, we will add more comprehensive descriptions to the affected figures. In general, we prefer not to overload figure captions, we hope it is ok not to repeat information that is already described in the text.

3. Please make sure all figures are clear. The parameters and other details of the proposed model and methods should be organized in some tables.

   We will double-check the clarity of the figures and will add tables with model details in the appendix.

Minor Suggestions:

1. Please provide a description of the machine learning model construction, including the procedures for training and test dataset selection, normalization or preprocessing methods, and any other relevant implementation details necessary for reproducibility.

   We will revise the experiments section to further clarify the details of the model construction and architecture: we divided all sites into 5 folds with stratified random sampling, then each fold in turn serves as the test set, with the four others as training set. A small subset of the training sites (20%) was set aside as validation set to enable early stopping of the training process and hyperparameter tuning.

2. In Figure 1, it is recommended to include the number of observation sites corresponding to each aridity type.

   Thanks, we will add the number of sites in each category to the legend.

3. Methods requires citing references, please check.
4. The explanation of part “3.3 Spatial patterns in model performance” (corresponding to Figure 6) is somewhat unclear. “Across sites with moisture index P/PET ≥ 0.75 the R_2 is 0.76, whereas it was only 0.57 for more arid sites

(MI <0.75). The (normalised) RMSE follows a similar pattern, with a value of 0.88 for sites with MI <0.75, compared to 0.57 for moist sites." However, these values are not directly visible in the two subplots on the left panel of Figure 6.

These values are indeed not directly visible in the referenced figures: they are derived from aggregations of the sites belonging to the two different aridity levels, whereas the current figures only display the performances at each individual site. We will make this clearer by adding panels for the specific categorisation used here.