August 1, 2024

Nonlinear Processes in Geophysics
Manuscript ID: EGUSPHERE-2023-2368

Dear Dr. Zoltan Toth, dear Reviewer 2,

We thank the editor, Dr. Zoltan Toth, for reading through our paper, the replies, and the reviews and for the additional feedback provided. Additionally, we thank Reviewer 2 for taking the time to address our replies to previous reviews and for providing additional comments. This letter details the changes made to the manuscript to address the points raised.

We highlight the referee comments in *blue italics*. We provide our replies to the comments and the changes we made to the manuscript in **bold font**. In the manuscript, changes are highlighted in red.

1. *Title. Does the "extremes" refer to climate drivers or vegetation response? I assume it should be vegetation responses if my understanding is correct.*

   That is correct, extremes in the title refers to the vegetation response.

2. *Neural network training. I am concerned about the training method and do not agree with the authors' claim about the "universal model". See these relevant studies: [1] and [2] It has been proven that a diverse dataset would increase the model performance.*

   We thank the reviewer for providing us with interesting and relevant literature. Further reading and discussion into the topic helped us realize that, although 20 locations are fewer compared to other studies, their impact on the overall performance should still be noticeable, as noted in [1], Figure 5. Therefore, we believe that a global model with additional features would demonstrate increased performance, as shown in previous research [3], even when trained with just the 20 locations used in our study.

   Our goal was to investigate the immediate response of vegetation to climate data on a site level with a focus on extremes. We choose to train on specific locations in order to highlight the difference of the models in simple settings, which can be increased in complexity in follow up studies. We now see that this point is not stressed in the manuscript.

   Lastly, similar work to the one kindly provided by the reviewer is not yet available in the vegetation community. It would be interesting to investigate the differences using a global approach versus a local one for general accuracy and prediction of extremes in vegetation responses, especially given the importance that memory effects play in biosphere dynamics [4] and the effects extremes have on its resilience [5]. We agree with the editor in the assessment that our study represents a first step in this direction. To this end, we have added a paragraph in the discussion section (Lines 420-431):

   Most existing studies center on creating a single global model, which is key to understanding performances beyond the training sample [3, 6]. However, we opted to train individual models for each site to understand how, under optimal conditions, each architecture could capture the specific dynamics of an ecosystem without the need for extrapolation. In this way, we could focus on the models' inherent capability to understand each ecosystem's dynamics, and we are sure that model differences do not arise due to different data demands for generalizations. Increasing the locations, features and expanding the models are necessary future steps for the continued investigation of modeling extremes in vegetation with ML. It has been shown that including additional features and multiple locations improves the ML model performance in hydrological applications [1, 2]. While similar studies do exist in the context of biosphere dynamics [3], more investigation is needed. For example,

including more locations, as highlighted in [1] represents the easiest way to improve performance. Investigating the effects of more location on the performance of ML for vegetation extremes would be an important contribution for the continued adoption of ML models in predicting biosphere dynamics.

3. *In addition, for the current training setup. How did the authors tune the hyperparameters? Do all the sites share the same set of hyperparameters? Line 139-141 mentioned 2000 to 2013 is used for training and 2014 to 2020 for testing, suggesting there is no standalone validation dataset for hyperparameter tuning. Section B2 mentioned "Split temporal cross-validation". How many folds are used here? Please clarify the hyperparameter tuning method.*

   Hyperparameters are selected through temporal cross-validation for each site. Sites do not share the same set of hyperparameters. For the temporal cross-validation, we used three folds, and for each fold, 20% of the training data was reserved for validation. We appreciate the comment pointing out that our explanation in the text lacked details. We added the following text to provide more context for the training of our models (Lines 501-502):

   We used three folds for cross-validation, with 20% of the training dataset left for validation in each fold.

4. *In Section B1, 50 out of 100 runs are selected based on performance. Do these 100 runs share the same set of hyperparameters? If so, is this process a selection of initial weights? An ensemble prediction is reasonable but cannot be used for models/runs selections since the testing set should be treated unseen.*

   We thank the reviewer for this observation. The 100 runs share the same set of hyperparameters obtained with the tuning selection described above. The different runs are initialized with different weights. We now include the results of all 100 runs and have updated Table 1, Figures 5, 6, 7, A1, and section B1. This does not change the results qualitatively, as can be seen from the updated table and figures.

5. *Input features. Line 153 mentioned T is the time window after which only the input variables are available. What's the value of T in this study? Is it only concurrent variables as input (e.g., temperature at day 10 to predict NVDI at day 10), or T is specified to construct a time window (e.g., temperature from day1 to day T to predict NVDI at day T)? Is it a sequence-to-sequence model or sequence-to-one model? Please clarify it.*

   It only uses concurrent variables as inputs, and T represents the length of the training dataset. We claryfied some notation in the technical description of the setup. We thank the reviewer this this observation.

6. *Line 155. What's the purpose of introducing "observer"? It seems irrelevant.*

   Thanks for poiting this out, we removed the observer explanation in the revised manuscript.

7. *Metrics for extremes. I would suggest considering F1-score (or area under ROC curve) since these metrics consider true positives, false negatives, and false positives at the same time.*

   Thanks for pointing out alternative metrics to use for studying the extremes. We added the F1 scores for the single locations in the appendices. This further showcases the different performances of the models for each site.

8. *Figure 6, 7. Are these figures for all the sites? Is there any performance difference among the sites?*

   These figures include results from all the sites. This is now clarified in the figure captions. Additionally, we now provide a figure in Appendix C (Fig B1), which shows the F1 scores for all the unique locations for all different quantiles studied. Figure 1 shows the added figure to Appendix C.

We look forward to hearing from you.
Sincerely and on behalf of all authors,

Francesco Martinuzzi

Leipzig University
Center for Scalable Data Analytics and Artificial Intelligence,
Institute for Earth System Science and Remote Sensing, and
Remote Sensing Centre for Earth System Research

References

[1] F. Kratzert, M. Gauch, D. Klotz, and G. Nearing. Hess opinions: Never train an lstm on a single basin. *Hydrology and Earth System Sciences Discussions*, 2024:1–19, 2024.

[2] Kuai Fang, Daniel Kifer, Kathryn Lawson, Dapeng Feng, and Chaopeng Shen. The data synergy effects of time-series deep learning models in hydrology. *Water Resources Research*, 58(4):e2021WR029583, 2022.

[3] Basil Kraft, Martin Jung, Marco Körner, Christian Requena Mesa, José Cortés, and Markus Reichstein. Identifying dynamic memory effects on vegetation state using recurrent neural networks. *Frontiers in big Data*, 2:31, 2019.

[4] Kiona Ogle, Jarrett J Barber, Greg A Barron-Gafford, Lisa Patrick Bentley, Jessica M Young, Travis E Huxman, Michael E Loik, and David T Tissue. Quantifying ecological memory in plant and ecosystem processes. *Ecology letters*, 18(3):221–235, 2015.

[5] Jaboury Ghazoul, Zuzana Burivalova, John Garcia-Ulloa, and Lisa A King. Conceptualizing forest degradation. *Trends in ecology & evolution*, 30(10):622–632, 2015.

[6] Codruț-Andrei Diaconu, Sudipan Saha, Stephan Günnemann, and Xiao Xiang Zhu. Understanding the role of weather data for earth surface forecasting using a convlstm-based model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1362–1371, 2022.

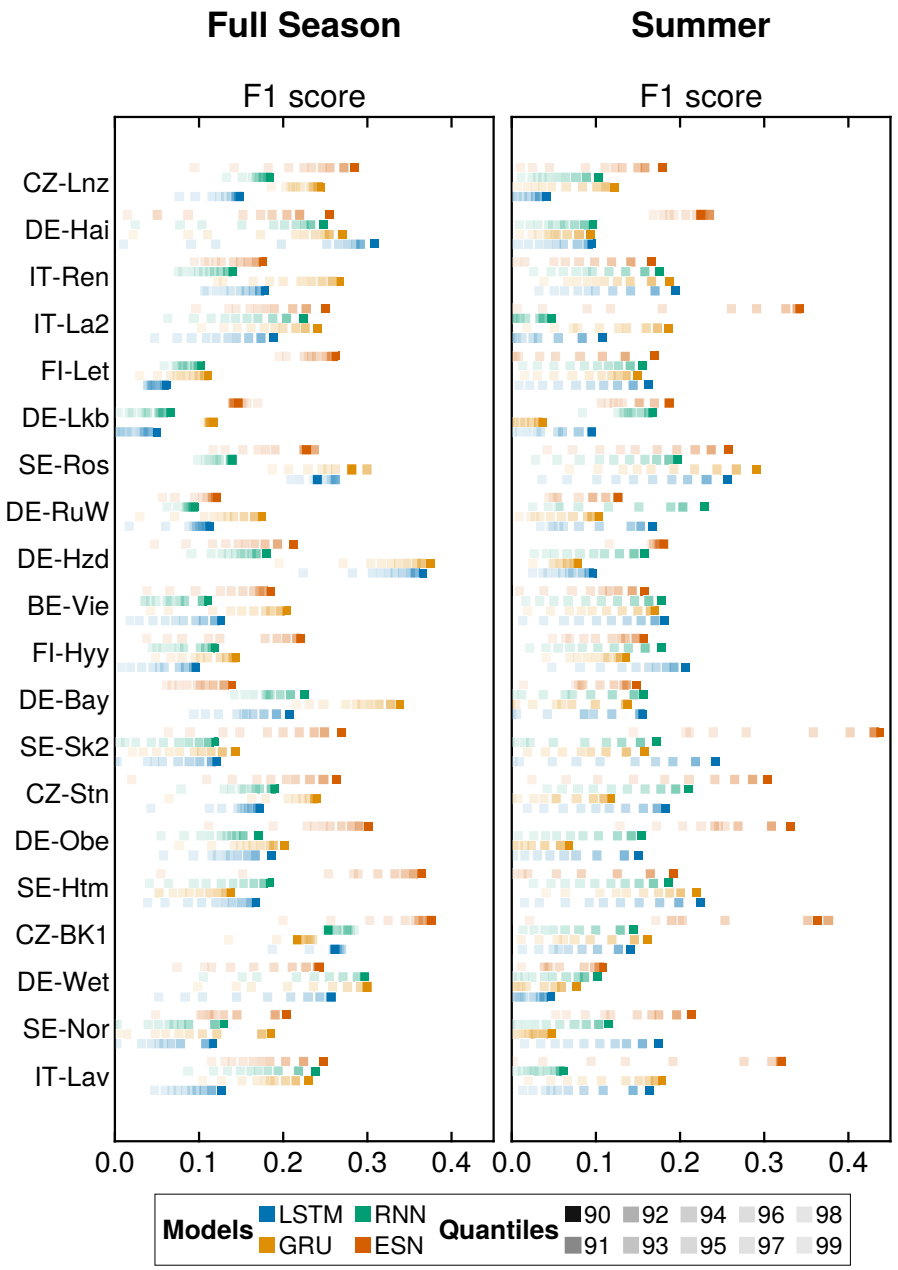

Figure 1: Additional figure in Appendix C.