Revision Notes

Manuscript title: Learning Evaporative Fraction with Memory

We sincerely appreciate the opportunity to revise our manuscript. We are deeply grateful to the editor and reviewers for their time, thoughtful insights, and constructive suggestions, which have been invaluable in improving the quality and clarity of our work. In response to the comments received, we have undertaken comprehensive revisions to address all concerns raised. A detailed point-by-point response to each reviewer's comment is provided below, with reviewers' comments shown in black and our responses in blue.
The comments from the reviewers are shown in black followed and our responses in blue.

Reviewer: 3

Reviewer #3 Evaluations:

This study investigated the role of environmental variables in the prediction of evaporative fraction (EF). Vegetation memory effects on evaporative fraction were studied through the use of a temporal deep neural network, which explicitly incorporates effects from historical conditions. The experiments explored how the network predictions behave during soil moisture dry-down periods, how different features contribute to the predictions at different points in history, and how average contributions from past feature values (at least 7 days in the past) vary across plant functional types and other ecosystem features (e.g. seasonality, rooting depth, etc.). Notably, variations in memory length are detected across different ecosystem types with variations in rooting depth.

Overall, the study brings very interesting insights. The analysis with a temporal network is an interesting approach that leads to substantial new insights into the behaviour of EF over time and across ecosystem types. I would recommend improving the manuscript in a few areas before publication.

Responses:

Thank you for your positive evaluation and for your great efforts to help us improve our paper. In the revised version of the manuscript, we have carefully addressed your comments, as well as those from the other reviewers.

Comments

Comment 1:

First, the connection between EF and its learned memory effects and dynamic rooting depth could be explained further. It is a large component of the motivation (Figure 1), which describes how during a dry-down period, EF decays at different rates depending on the plants' water use strategies. Figure 9 then shows that there is a link between the learned memory effects and rooting depth. While it appears that an increased rooting depth is associated with stronger memory effects, the distributions of the memory effects overlap significantly between rooting depth "bins". As a major goal of the study is to show the potential of using the memory effect estimates for inferring rooting depth and even using them as a proxy, the work should more clearly state or show how this could be done. As the (previously estimated) rooting depth used here is a static estimate, one suggestion would be to further compare the memory effects during dry-down periods and non dry-down periods, a distinction that was also used in the overall model evaluation. In particular, variations in memory effects during dry-down periods may reveal additional insights on water use strategies.

Responses:

Thank you for this insightful comment. We agree that the link between EF, its learned memory effects, and rooting depth is a central motivation of this study. In the revised manuscript, we strengthened the introduction to more clearly explain the conceptual pathway through which LSTM-derived memory effects relate to root-zone water storage and effective rooting depth. Importantly, we now emphasize that quantitatively inferring rooting depth from EF decay alone would require additional information—including soil texture, hydraulic properties, and groundwater constraints—which is beyond the scope of the current analysis but represents an important future direction.

Regarding the reviewer's suggestion to compare memory effects during dry-down versus non-dry-down periods: we conducted similar exploratory analyses during earlier revisions in the last round review. However, memory effects are jointly shaped by multiple factors—such as rooting depth, soil hydraulic properties, and meterological anomalies—making it difficult

to attribute seasonal (dry or wet season) or event-scale differences to rooting depth alone. Because a full mechanistic decomposition would substantially expand the manuscript, we decided not to include these additional analyses here. Instead, we have added clarifying text in the Discussion to position this as a promising avenue for future research. Specifically, in future work we aim to integrate EF dry-down behavior with precipitation inputs, radiation, and soil/groundwater properties to enable more quantitative inference of effective rooting depth and additional insights on water use strategies.

Attached our analysis for dry and wet season from the previous review round:
"In our extended analysis, we compared memory contributions between dry and wet seasons as well as growing and non-growing seasons. However, we did not observe a consistent pattern across sites. That is, memory contributions were not always higher in the dry season than in the wet season, nor consistently higher during the growing season compared to the non-growing season. We believe this inconsistency reflects the complex interactions among climatic drivers. For instance, during dry seasons, plants may rely more heavily on water stored from previous wet seasons, aligning with findings such as those by Miguez-Macho and Fan (2021), who show that transpiration in dry seasons is often sustained by deep soil moisture. However, EF is co-regulated by multiple drivers—particularly precipitation and temperature—and the dominant controls can vary depending on climate regimes (e.g., energy-limited vs. waterlimited systems). In some cases, temperature anomalies (e.g., heat stress) during the wet season can induce prolonged memory effects that rival or exceed those in the dry season. Similar observations were made when comparing growing and non-growing seasons.

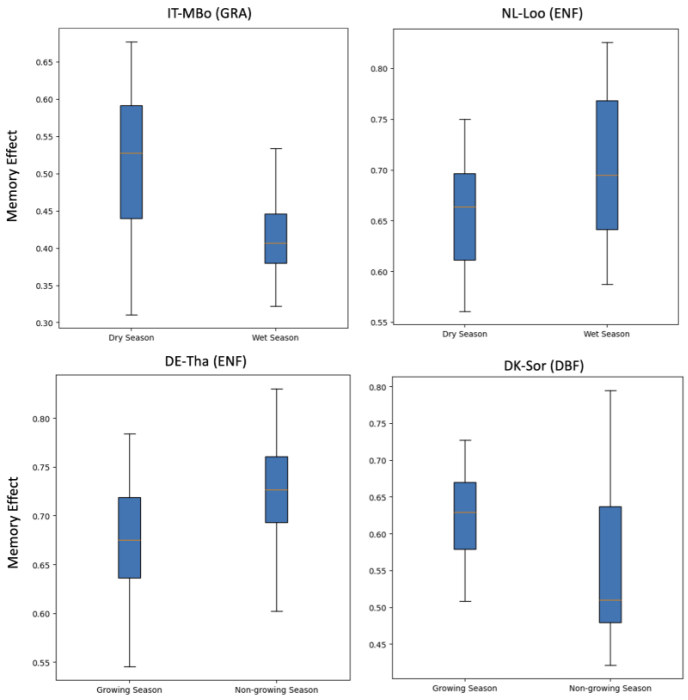

Figure R1. Memory effect comparisons between dry vs wet season and growing vs nongrowing season.

This complexity suggests that memory effects are not governed by a single seasonal pattern, but rather emerge from compound effects and interactions among drivers. Our daily-scale explainable machine learning framework enables event-level analyses that can capture such nuanced behaviors. In future work, we plan to explore clustering or other data-driven classification techniques to systematically identify dominant memory regimes and their seasonality.

Reference:

Miguez-Macho, G. and Fan, Y.: Spatiotemporal origin of soil water taken up by vegetation, Nature, 598, 624–628, https://doi.org/10.1038/s41586-021-03958-6, 2021."

Comment 2:

Additionally, the model training setup brings some limitations that should be addressed. As each site appears in the training set with 2 site-years left out for testing, there is potential for overfitting on the patterns in the training data. While I think the motivation given for this setup is reasonable, it would be good to discuss these limitations. In particular, a good predictive

performance does not automatically mean that the model represents physical processes. The use of interpretability methods, as done in this work, is suitable for tackling this. However, L221-223 should be revised or discussed later on in the manuscript as this conclusion should not be made based on overall performance.

Responses:

Thank you for this helpful suggestion. We agree that a temporal train–test split at each site may introduce limitations. While this setup is reasonable for our primary goal—to use the LSTM as an interpretive tool for understanding EF dynamics rather than for upscaling or out-of-sample generalization—we acknowledge that strong predictive performance alone may lead to some misleadings to readers.

In the revised manuscript, we explicitly highlight this limitation in the Methods section. We clarify that our intent is not to emphasize the absolute predictive accuracy of EF, but to leverage the LSTM's ability to encode memory effects, which we subsequently analyze using interpretability methods. We also note that achieving robust EF prediction for extrapolation or large-scale upscaling remains a challenging task that would require alternative training strategies and dedicated evaluation of generalization performance.

Regarding the conclusion originally stated in Lines 221–223, we have relocated this statement to the later discussion accompanying Figures 4 and 5, where it can be interpreted in the appropriate context of model interpretability rather than overall performance.

Comment 3:

Some details on the data and model are missing. Is the input data standardised? Which activation function is used for the model (L137)? Other model training details such as optimiser (settings), number of training iterations, data batch size, learning rate, etc. should also be included.

Responses:

Thank you for pointing out the missing methodological details. We have now updated the Methods section to include a full description of the data preprocessing and model training

procedures. Specifically, all input variables were standardized using their mean and standard deviation. The LSTM model uses a Leaky-ReLU activation function, and is trained using the Adam optimiser with a learning rate of 0.001, a batch size of 64, and 50 training epochs. These details have been added to the revised manuscript for clarity and reproducibility.

Comment 4:

L160-161: this description is confusing as it suggests the 10-model ensemble consists of models with different sequence lengths, whereas it becomes clear later on that the final model is an ensemble of models each trained with 365-day sequences.

Responses:

Thank you for pointing out this ambiguity. We apologize for the unclear wording. In our workflow, each memory length (10–360 days, plus 365 days) was evaluated using an ensemble of 10 models, each differing only in the random initialization of weights and biases. We then compared the prediction skills across all memory lengths and found that the 365-day sequence length consistently produced the best performance. The final model used in the analysis is therefore an ensemble of 10 models, all trained with a 365-day sequence length, not a mixture of different memory lengths.

"For each candidate memory length (10–360 days in 30-day intervals, plus 365 days), we trained an ensemble of 10 models that differed only in the random initialization of weights and biases. This ensemble approach reduces stochastic variability and provides an estimate of prediction uncertainty. After comparing performance across memory lengths, we selected the 365-day configuration, which yielded the highest prediction skill. All subsequent analyses therefore use the 10-member ensemble trained with a 365-day sequence length."

Comment 5:

L186: clarify if separate years are used for testing and validation.

Responses:

Thank you for pointing this out. We have clarified the data splitting strategy in the revised manuscript. For each site, we used a strictly temporal split: one entire site-year was held out as the test set (never seen during training or validation), one additional site-year from the

remaining years was randomly selected as the validation set, and all other site-years were used for training. This ensures that training, validation, and testing are conducted on non-overlapping years, avoiding temporal leakage and allowing us to evaluate generalization to unseen periods at the same site. We have revised Line 186 to explicitly describe this procedure.

Comment 6:

L183-192: several remarks are repeated and should only be part of the methods section (on the ensembles in particular).

Responses:

Thank you for pointing this out. In the revised manuscript, we have removed these redundant explanations from Lines 183–192 and kept such descriptions only in the Methods section.

Comment 7:

L194: improved compared to what? What was the difference? If an ensemble is recommended, it would be useful to show more details of the comparison.

Responses:

Thank you for pointing this out. In the original text, "improved performance" referred to the fact that the ensemble mean consistently achieved higher predictive skill and reduced variance compared to individual LSTM model runs with different random initializations. To clarify this, we have revised the manuscript to explicitly state the comparison.

Specifically, across all sites, the ensemble mean improved the median R^2 by approximately 0.02–0.05 relative to individual model realizations and substantially reduced run-to-run variability. These improvements arise because the ensemble mean helps stabilize nonlinear dynamics and mitigate stochastic noise from parameter initialization.

We have updated the text to reflect this comparison and removed ambiguous wording.

Comment 8:

Figure 3 is not mentioned in the text.

Responses:

Thank you for pointing out this. We have now added explicit references to Figure 3 in the Results section to ensure that it is clearly introduced and discussed in the main text.

Comment 9:

L249: typo in the dates

Responses:

Thank you. We have revised it as suggested.

"…uses the previous 365 days of input data (from 2012-10-05 to 2013-10-15) to predict EF on 2013-10-15…"

Comment 10:

Figure 5: the caption appears to contain some interpretations. It would be more suitable to discuss these in the text.

Responses:

Thank you for pointing this out. We have revised the figure caption by removing interpretative statements and retaining only essential descriptive information about the figure. All interpretative content (e.g., explanations of driver influence, interpretation of delayed responses) has now been moved to the main text in the Results section for clearer and more appropriate discussion.

Comment 11:

Figure 10: there is a typo in the top left. The seasonality index is not defined.

Responses:

Thank you for pointing this out. We have corrected the typo in the top-left label of Figure 10. We also added a definition of the seasonality index in the Methods section to clarify how it is calculated and applied in our analysis.