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: 1

Reviewer #1 Evaluations:

This manuscript applies a relatively new machine learning model to effectively capture the temporal variability of Evaporative Fraction (EF). The model shows strong agreement with observations, demonstrating its capability to represent the dynamics of EF across different PFTs and climate zones. The authors also quantitatively assess the influence of surface hydrometeorological drivers on vegetation memory, providing valuable insights into soil-plant-atmosphere interactions.

Overall, I find this study to be of interest and with potential for publication. Several aspects of the methodological description and the presentation of the results require further clarification to ensure that readers can fully understand and evaluate the work.

Responses:

Thank you for your positive evaluation and for your great efforts to help us improve our paper. We also appreciate your recognition of our study's potential to enhance the understanding of water and carbon fluxes under changing precipitation variability in the context of climate change.

In the revised version of the manuscript, we have carefully addressed your comments, as well as those from the other reviewers.

Major Points

Comment 1:

The manuscript would benefit from a clearer and more consistent definition of "memory" throughout the text. Currently, multiple types of memory are mentioned, e.g., vegetation memory, soil memory, and meteorological memory, but the differences among these concepts are not always explicit. As a result, it can be confusing for readers to follow the main focus of the study. I recommend that the authors select one primary concept of memory and build the narrative around it, while clearly defining how other types of memory are related but distinct. This would greatly improve the conceptual clarity and help readers better understand the scope and implications of the work.

Responses:

Thank you for pointing out this.

Our primary focus is on memory effects as reflected in the observed daily EF variability. This type of memory effect, captured by the LSTM framework, implicitly integrates the influences of vegetation, soil, and atmospheric processes that contribute to the persistence of EF anomalies. We have now unified the terminology as "memory effects" throughout the manuscript for consistency.

"Current studies indicate that vegetation responses to past climate conditions exhibit lagged effects—referred to as memory effects—which can modulate ecosystem functioning, particularly after climate extremes (He et al., 2018; Hossain et al., 2022; Canarini et al., 2021; Qiu et al., 2025)."

Comment 2:

Throughout the manuscript, the authors should explicitly highlight that this is a machine learning-based approach. At present, the manuscript briefly mentions the methodology but does not clearly differentiate it from other physical approaches for estimating EF, many of which require fewer input variables. For instance, methods such as Surface Flux Equilibrium (SFE) can also calculate EF effectively with only routine weather station data. By making this distinction clearer, the authors can better justify the novelty and added value of applying a machine learning framework, and help readers understand the specific advantages and trade-offs compared to traditional physical methods.

Responses:

Thank you.

In the abstract of our original manuscript, we have already clearly state that

"We developed an explainable machine learning (ML) model based on a Long Short-Term Memory (LSTM) architecture, which explicitly incorporates memory effects, to investigate the mechanisms underlying EF dynamics."

Our study aims not simply to predict EF but to understand the temporal dynamics and memory effects that shape EF variability—an aspect that equilibrium-based physical models cannot capture, as they assume steady-state surface energy partitioning. Our ML-based approach maintains strong predictive accuracy even during dry-down periods, whereas the SFE framework is a non-predictive equilibrium theory that breaks down under non-stationary conditions such as dry-downs. Moreover, SFE does not explicitly incorporate precipitation inputs, while the ML model can extract and leverage precipitation-driven memory effects that are essential for predicting soil moisture decline and energy partitioning.

In addition, explainable machine learning allows us to uncover processes that are difficult or impossible to diagnose from the structural assumptions of physical models—for example, underrepresented meteorological memory effects. We hope the academic community will maintain an open-minded perspective toward machine learning and explainable AI techniques. These tools provide a principled way to peer inside otherwise opaque models and can help reveal new mechanisms, such as memory effects, that traditional physical frameworks often struggle to capture.

As suggested, we have revised the manuscript to more explicitly highlight the machine learning nature of our framework, its conceptual differences from physical EF estimation methods, and the complementary insights it provides.

References:

Gharari, S., Gupta, H. V., Clark, M. P., Hrachowitz, M., Fenicia, F., Matgen, P., and Savenije, H. H. G.: Understanding the Information Content in the Hierarchy of Model Development Decisions: Learning From Data, Water Resources Research, 57, e2020WR027948, https://doi.org/10.1029/2020WR027948, 2021.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resources Research, 57, e2020WR028091, https://doi.org/10.1029/2020WR028091, 2021.

Comment 3:

The manuscript would benefit from a clearer explanation of how the LSTM model improves upon previous ML approaches. At present, it is difficult for readers to fully understand why the proposed machine learning framework is needed. My own curiosity is that why earlier approaches were primarily static and, if so, why dynamic ML modeling has not been attempted before? The manuscript includes three baseline experiments, which provide a valuable opportunity to convince readers of the proposed method’s strengths. I suggest the authors provide a more detailed and systematic comparison by explaining why the baselines perform worse, which would greatly enhance the clarity of the study.

Responses:

Thank you.

First, in our original introduction, we explained the specially of LSTM structure.:

“…Long Short-Term Memory (LSTM) networks – a variant of recurrent neural networks, are particularly well suited for learning from sequential data and are increasingly used to model temporal dynamics in hydrology (Fang and Shen, 2020; Jiang et al., 2020; Li et al., 2022). LSTM, with its recurrent cells, retains previous information from input sequences akin to how meteorological data, like precipitation, and its impact is retained over long periods of time as soil moisture or snowpack (Lees et al., 2021, 2022). In the context of vegetation, previous heat stress events can cause cellular damage that impairs photosynthesis, respiration and transpiration, ultimately altering EF (Staacke et al., 2025). The hidden states in LSTM encode system memory and evolve with time, interacting with real-time climate variables to simulate time series (Xiao et al., 2024). This allows the model to account for historical vegetation-climate interactions in a data-driven way - without relying on process-based models that often misrepresent such memory effect (Kraft et al., 2022; Lees et al., 2022). Instead, it infers such information through the memory effects captured within its recurrent cells, learning functional dependencies between EF and both current and past climate drivers such as precipitation and temperature anomalies…"

Here we want to emphasis is earlier EF studies commonly used static or quasi-static machine-learning models (e.g., RF, MLP, GBM), which only map instantaneous meteorological forcing to EF. These models do not contain internal states or mechanisms for storing information across time, and therefore cannot capture ecohydrological memory processes such as soil-moisture carryover, precipitation dry-down progression, canopy conductance acclimation, or multi-day thermal effects that strongly modulate EF. These kind of mechanisms could be very complex and will be difficult to be represented by limited parameters and diagnostic model structures in physical models.

In contrast, LSTMs are explicitly designed to represent time-dependent processes through their gating architecture. The input, forget, and output gates allow the model to selectively retain or discard past information, enabling it to learn variable-length memory from the data. This makes the LSTM uniquely suited for capturing meteorological memory effects—especially the lagged influence of precipitation, soil moisture, and accumulated thermal conditions—compared with other ML models that lack recurrent dynamics. We now clarify these structural advantages in the revised text.

We also expanded the baseline comparison to clarify why the three baseline models perform worse. Static ML baselines fail because they cannot represent lag structures or propagate information across time. The diagnostic physical baseline, with its fixed functional forms and internally prescribed parameters, is unable to capture site-specific or event-level meteorological memory effects or their impacts on EF dynamics.

Overall, the revised manuscript now explicitly demonstrates: (i) why EF prediction requires a sequential machine learning model incorporating memory, (ii) how the LSTM's gating mechanisms provide structural advantages for learning meteorological memory, and (iii) why baselines are fundamentally unable to recover these lagged ecohydrological processes.

Kraft, B., Nelson, J. A., Walther, S., Gans, F., Weber, U., Duveiller, G., Reichstein, M., Zhang, W., Rußwurm, M., Tuia, D., Körner, M., Hamdi, Z. M., and Jung, M.: On the added value of sequential deep learning for upscaling evapotranspiration, https://doi.org/10.5194/egusphere-2024-2896, 10 October 2024.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Zenone, T., Vitale, L., Famulari, D., and Magliulo, V.: Application of machine learning techniques to simulate the evaporative fraction and its relationship with environmental variables in corn crops, Ecol Process, 11, 54, https://doi.org/10.1186/s13717-022-00400-1, 2022.

Han, Q., Wang, T., Kong, Z., Dai, Y., and Wang, L.: Disentangling the Impacts of Environmental Factors on Evaporative Fraction Across Climate Regimes, Journal of Geophysical Research: Atmospheres, 129, e2024JD041515, https://doi.org/10.1029/2024JD041515, 2024.

Minor Points

Comment 1:

9: What are vegetation memory effects? Please explain.

Responses:

Thank you.

As clarified in our response to Comment 1, the primary focus of this study is memory effects, defined as the lagged responses of vegetation to past climate conditions. Such memory effects can modulate ecosystem functioning, particularly following climate extremes, as demonstrated in previous studies. We have now ensured consistent use of the term “memory effects” throughout the manuscript.

In our LSTM framework, these memory effects are captured implicitly through the recurrent architecture, which integrates information from previous climate conditions. Although the model does not explicitly simulate physiological mechanisms, the resulting memory signatures reflect the combined influence of vegetation, soil, and atmospheric processes that govern the persistence of EF anomalies.

We now cite recent work demonstrating the strong contribution of antecedent climate to ecosystem functional anomalies:

Qiu, J., Zhang, Y., Cai, M., Keenan, T. F., Zhang, H., Gentine, P., Luo, X., Cattry, M., Zhou, S., and Piao, S.: Large contribution of antecedent climate to ecosystem productivity anomalies during extreme events, Nat. Geosci., 1–8, https://doi.org/10.1038/s41561-025-01856-4, 2025.

Comment 2:

10: I'm new to ML method, what's the difference between explainable ML and regular ML?

Responses:

Thank you.

We now add a brief explanation of the difference between explainable machine learning (XAI) and interpretable machine learning (IML) relative to regular machine learning (ML) at the first place where these terms are introduced. Specifically, we now clarify that regular ML focuses primarily on predictive accuracy, whereas XAI/IML aims to make model behavior interpretable through techniques such as Shapley additive explanations (SHAP; Lundberg & Lee, 2017), integrated and expected gradients (Sundararajan et al., 2017; Erion et al., 2021), and local interpretable model-agnostic explanations (LIME; Ribeiro et al., 2016). This addition is included to help readers—especially those without prior ML experience—better understand the methodological context of our study (Gunning & Aha, 2019; Murdoch et al., 2019; Rudin, 2022).

This clarification also highlights a key novelty of our work, as this study represents one of our first attempts to implement XAI/IML techniques to investigate evaporative fraction dynamics and meteorological memory effects.

References:

Gunning, D. and Aha, D. W.: DARPA's Explainable Artificial Intelligence (XAI) Program, AI Magazine, 40, 44–58, https://doi.org/10.1609/aimag.v40i2.2850, 2019.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B.: Interpretable machine learning: definitions, methods, and applications, Proc. Natl. Acad. Sci. U.S.A., 116, 22071–22080, https://doi.org/10.1073/pnas.1900654116, 2019.

Rudin, C.: Interpretable AI: Fundamental principles and 10 grand challenges, Nat. Mach. Intell., 4, 611–619, https://doi.org/10.1038/s42256-022-00534-z, 2022.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 30, 4765–4774, 2017.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic Attribution for Deep Networks, http://arxiv.org/abs/1703.01365, 2017.

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S.-I.: Improving performance of deep learning models with axiomatic attribution priors and expected gradients, Nat. Mach. Intell., 3, 620–631, https://doi.org/10.1038/s42256-021-00343-w, 2021.

Ribeiro, M. T., Singh, S., and Guestrin, C.: "Why Should I Trust You?" Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144, https://doi.org/10.1145/2939672.2939778, 2016.

Comment 3:

11: Should be "vegetation memory effects"

Responses:

Thank you. We have now unified the terminology throughout the manuscript by consistently using "memory effects" after explicitly defining this term as referring to meteorological memory effects in the revised Introduction. Please also check our responses to your minor comment 1.

Comment 4:

14: What's the advantage of this study compared to SFE (Surface Flux Equilibrium), which also only require routine weather station data for EF calculation?

Responses:

Thank you.

The Surface Flux Equilibrium (SFE) theory provides a physically grounded framework to diagnose the equilibrium relationship between evaporative fraction and meteorological states (e.g., air temperature, humidity, wind speed). However, SFE inherently assumes quasi-stationary surface flux equilibrium and does not represent temporal evolution or memory effects. Therefore, while SFE-based EF estimates are appropriate for long-term or climatological analyses (e.g., monthly or seasonal averages), they cannot capture high-frequency EF dynamics and delayed responses to meteorological forcings, especially the dry-downs. In contrast, our LSTM framework explicitly incorporates temporal dependencies, allowing it to learn non-equilibrium, memory-driven adjustments in EF that arise from soil moisture persistence, canopy heat storage, and vegetation physiological lag.

Please also check our responses to your major comment 2.

Comment 5:

17-19: Which corresponds to "water-limited" and "energy-limited" regimes?

Responses:

Thank you.

In Lines 17–19, our intention was not to explicitly classify sites into “water-limited’’ or “energy-limited’’ regimes, but rather to summarize the dominant drivers of EF dynamics within each vegetation type. When incorporating memory effects, precipitation and VPD—reflecting atmospheric dryness—emerge as primary controls in woody savanna, savanna, open shrubland, and grassland sites. In contrast, in most forest sites (which are typically less water-limited), air temperature dominates EF variability, while net radiation primarily captures instantaneous effects.

In general, savanna and shrubland ecosystems tend to operate under water-limited conditions, whereas forests are more often energy-limited. However, as shown in Fig. 7, substantial site-level variability exists within each PFT.

We added a clarifying sentence in the main text to note that these driver patterns broadly correspond to more water-limited conditions in savanna, shrubland, and many grassland sites, and to less water-limited or more energy-limited conditions in forest sites. We did not include this point in the abstract in order to keep the content length concise.

Comment 6:

24: Previously it says "vegetation memory effects", please be consistent.

Responses:

Thank you. As you suggested, we have ensured consistent use of the term throughout the manuscript.

Comment 7:

31: I think here you don't have to emphasize root-zone, since SM-EF at surface soil layer should be stronger.

Responses:

Thank you.

We have revised the text to use "soil water availability" instead of emphasizing root-zone: "Previous studies have shown that EF is mainly governed by soil water availability and vegetation structure and is intrinsically linked to precipitation-supplied water resources (Bastiaanssen et al., 1997; Gentine et al., 2007; Ichii et al., 2009)."

Comment 8:

36: I feel the cause-and-effect of this paragraph should be rephrased as how vegetation memory influence EF, rather than the other way around. The goal of this study is to predict EF, and vegetation memory is one of the key drivers. Or you should place this content after the description of EF prediction.

Responses:

Thank you.

We want to clarify that our goal is not mainly predict EF but our first aim of this study is to implement LSTM based explainable machine learning which incorporating memory effect to understand the EF dynamics. Due to the understanding task, we need an EF prediction model that can better capture the EF dynamics not only during dry-downs but also non-dry-downs.

Here we thought it's necessary to mention that this is actually a second-round review of our manuscript. In the previous review round, we have the introduction more focus the EF predictions and placing the part of memory effect following the description of EF prediction: Attach part of the previous version of introduction of EF predictions:
"…Accurate prediction of EF is essential for estimating evapotranspiration, monitoring plant water stress, and potentially inferring rooting depth(Collins and Bras, 2007; Fu et al., 2022; Liu et al., 2020; Wang et al., 2006). Among various environmental factors affecting EF, the soil moisture state dominates the day-to-day EF variations, particularly in water-limited regions(Dirmeyer et al., 2000; Dong et al., 2022; Haghighi et al., 2018; Lu et al., 2016). It serves as both a direct predictor of EF in prolonged dry spells and as a historical record, mirroring past climatic conditions and water availability — analogous to a memory effect. However, precise measurement or estimation of soil moisture in the rooting zone or plant water stress, is difficult or even impossible to acquire(O. and Orth, 2021; Skulovich and Gentine, 2023). This led to the development of various drought indices such as the Palmer Drought Severity Index, which try to infer the changes in soil moisture through a mechanistic, yet empirical, approach. These constraints underscore the importance of a more direct predictive model for EF, which could in turn be used as a more objective stress index, that can operate independently of soil moisture estimates and directly linked to surface evapotranspiration, emphasizing the incorporation of memory effects related to plant regulation within the algorithms(Kraft et al., 2019)…."

However, reviewers pointed out that simpler models can achieve EF prediction accuracies comparable to the LSTM, and therefore encouraged us to shift the introduction away from predictive performance and toward highlighting the memory effect, its underlying mechanisms, and how an LSTM-based explainable ML framework can extract these functional relationships. We revised the introduction accordingly. This restructuring better reflects the true focus of our study, as the majority of our analyses center on Expected Gradients–based driver interpretation rather than on EF predictive performance. However, we have also improved the flow of the Introduction so that the discussion of memory effects aligns more naturally with the subsequent description of the EF prediction framework.

Comment 9:

42: Please explain the terminology the first time it appears.

Responses:

Thank you. We have revised this.

"Temperature anomalies, including heat and cold stress, can impair vegetation function, alter water demand through temperature-induced soil moisture memory, that is, the persistence of soil moisture anomalies caused by past temperature effects on evapotranspiration and recharge, and deplete carbon reserves, thereby influencing transpiration and daily EF variability."

Comment 10:

74: Again, what's the difference between "explainable ML" and regular ML method?

Responses:

Thank you. We have revised this.

Please check our response to your minor comment 2.

Comment 11:

115: Did you also mask those with energy imbalance larger than a threshold (e.g., (Rn-Gs)-(LH+SH)>30W/m2)? Please explicitly indicate it in the main context.

Responses:

Thank you.

In this study, we used the energy-balance-closure–corrected fluxes provided in the FLUXNET dataset, namely LE_CORR and H_CORR, which adjust the raw latent and sensible heat fluxes to satisfy Rn - G ≈ LE + H following the standard procedure in Wilson et al., 2002. Because these corrected fluxes already account for energy imbalance, we did not apply an additional masking based on an imbalance threshold such as (Rn - G) - (LE + H) > 30 W/m2. We now explicitly clarify this in the main text and the supplementals.

In addition, as noted in the Methods section, "We assume that the errors in LE and H have comparable magnitudes, consistent with previous studies (Foken, 2008; Hollinger and Richardson, 2005; Richardson et al., 2006), and are uncorrelated. This assumption allows the mathematical elimination of errors associated with the lack of energy balance closure (Schwalm et al., 2010)."

Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., Bernhofer, C., Ceulemans, R., Dolman, H., Field, C., Grelle, A., Ibrom, A., Law, B. E., Kowalski, A., Meyers, T., Moncrieff, J., Monson, R., Oechel, W., Tenhunen, J., Valentini, R., and Verma, S.: Energy balance closure at FLUXNET sites, Agricultural and Forest Meteorology, 113, 223–243, https://doi.org/10.1016/S0168-1923(02)00109-0, 2002.

Comment 12:

121: What is "corrected" LH? Please explain.

Responses:

Thank you.

In this study, we used the energy-balance-closure–corrected fluxes provided in the FLUXNET dataset, namely LE_CORR and H_CORR, which adjust the raw latent and sensible heat fluxes to satisfy Rn - G ≈ LE + H following the standard procedure in Wilson et al., 2002.

Please also see our response in your minor comment 11.

Comment 13:

Figure 4. Please indicate the number of sites included for each PFT.

Responses:

Thank you.

Figure 4 is an example time-series plot showing selected sites and years for each plant functional type (PFT). We have revised the Figure 4 caption to explicitly indicate that this figure presents representative sites rather than all sites.

"Figure 4. Example time-series plots for representative sites and years for each plant functional type…"

In addition, the number of sites included for each PFT is clearly reported in the summarized results shown in Figure 3 and Figure 6, where the site counts are provided in parentheses following each PFT label (e.g., DBF (19), ENF (24), GRA (18), etc.).

Comment 14:

143: Section 3.2, I suggest the authors describe explicitly the difference of each baseline model from LSTM. To non-expert in ML, it looks like the settings of FNN is similar to LSTM, then why does FNN perform worse? Additionally, why do the authors want to add the SPI-based model, and why the SPI model performs so bad ($R2<0.1$)? Please add relevant discussions.

Responses:

Thank you.

We have added explanations in Section 3.2 to clarify the conceptual differences among the baseline models and to discuss why the feedforward neural network (FNN) and the SPI-based model perform substantially worse than the LSTM.

First, although the FNN uses the same meteorological time-series inputs as the LSTM and has a comparable number of parameters, it differs fundamentally in structure. The FNN treats all input features as independent and lacks any mechanism to retain information from previous time steps. In contrast, the LSTM is explicitly designed to capture temporal dependencies through gated recurrent units, allowing it to learn antecedent-condition effects that are essential for modeling EF during dry-down periods. This structural limitation explains why the FNN cannot capture memory effects and therefore underperforms relative to the LSTM.

Second, we included the SPI-based model because a reviewer in the previous round requested an additional simple baseline beyond the FNN. SPI is a widely used hydrological drought index and provides a transparent reference grounded in traditional drought assessment methods.

Attached the reviewers' comments from the previous review round:

“Reviewer #2 Comment 3:

I think the NN baseline is a fine idea to see how memory effects determine daily EF values, but only if the NN is itself convincingly good as a predictor. As a first step, I strongly suggest some simpler baseline tests:

…A site-blind soil moisture proxy. I would also not be surprised if a model such as EF = beta0 * tanh(beta1* SPI + beta2), or something similarly parameterized, might outperform an RMSE of 0.13 as well, at least in temperate conditions. This could be fit site-by-site, biome-by-biome, or globally. This would then be a function of a single variable (P), with some memory from the SPI calculation (or any other simple weighted average)…"

Finally, the poor performance of the SPI-based model ($R^2 < 0.1$) is expected because SPI summarizes only precipitation anomalies over aggregated windows (e.g., 30–365 days) and does not incorporate key EF drivers such as radiation, temperature, vegetation state, wind, or short-term soil moisture dynamics. As a result, SPI cannot capture rapid EF fluctuations following rain pulses, nor the nonlinear meteorological memory effects that the LSTM model is designed to resolve.

These clarifications have been added to the revised manuscript.

Comment 15:

145: In figure 3 it says FNN, please be consistent.

Responses:

Thank you. We have unified to FNN.

Comment 16:

176: I suggest the authors also add a brief description of EG in the main context, since it is an important component of this study.

Responses:

Thank you. We can add the brief description of EG in the main context.

Comment 17:

212: -1.21, Is it a typo? Why R2 can be negative and larger than 1?

Responses:

Thank you. This was a typo. We have corrected R^2 to the Nash–Sutcliffe Efficiency (NSE) coefficient. Unlike R^2, NSE can take values less than 0, which explains the negative value previously shown. The revision has been updated in the manuscript.

Comment 18:

216: "Figure 4" should be "Figure 3"?

Responses:

Thank you. We have revised this.

Comment 19:

293: From here to 299, can be moved to method part.

Responses:

Thank you. We will move it to the method part.

Comment 20:

302: Please indicate this is Shortwave radiation (RAD)

Responses:

Thank you. We reivsed it as suggested.

Comment 21:

307-309: Isn't air temperature correlated with radiation?

Responses:

Thank you for pointing this out. Yes, air temperature and radiation are indeed correlated (mean $R^2 = 0.35$ across our sites for the whole period), and we have revised the relevant sentences and added statements acknowledging this limitation. Nevertheless, it is important to note that most studies include both air temperature and radiation when predicting ET, despite their correlation. We note that the Ta–radiation correlation varies systematically across climate regimes: correlations tend to be lower in arid sites, where air temperature is strongly influenced by turbulent mixing and sensible-heat–dominated surface conditions, and higher in humid sites where temperature is more directly constrained by available radiation. This spatial variability further highlights that Ta contains distinct information from radiation, especially in water-limited ecosystems.

In addition, in our context, radiation primarily governs short-term EF variability through instantaneous energy input, whereas air temperature exerts a stronger influence over longer timescales when memory effects are considered. This distinction justifies their separate contributions in the model. Although future work using causal-inference frameworks could more rigorously disentangle their individual effects on EF dynamics, such an analysis is beyond the present scope. We have added this point to the limitations.

Comment 22:

351: Figure8, Can you re-arrange this figure with x-axis ranging from shallow to deep rooting-depth? Plus, how do you normalize the Contributions? Why are contributions from all variables even lower than precipitation alone? Does this indicate there could be negative feedbacks between variables?

Responses:

Thank you.

We have also revised Figure 8 in response to you and another reviewer's suggestion to improve its clarity. We further clarify that the contributions shown in Figure 8 are computed from the absolute values of Expected Gradients (EGs), so they quantify the magnitude of memory effects rather than their sign. Therefore, the fact that the precipitation-only case shows a higher long-term contribution than the "All Variables" case does not indicate negative feedbacks between variables. Instead, it mainly reflects differences in normalization and in the lag structure of the predictors: precipitation exhibits a longer effective memory, whereas several other variables included in the full predictor set have much shorter memory and increase the denominator of the normalized contribution metric. We have now added an explicit explanation of this normalization procedure and its interpretation in the revised Methods section (Section 3.6), and the updated presentation can be seen in the revised Figure 8.

Comment 23:

356: Should be "temporal EF machine learning model".

Responses:

Thank you. We have revised it as suggested.

Reviewer: 2

Reviewer #2 (Prof. Dr. Benjamin Stocker) Evaluations:

This study presents an analysis of the evaporative fraction (EF, here defined as the latent heat flux divided by the sum of the latent and the sensible heat flux) across a large set of sites with eddy covariance-based flux measurements with a focus on how EF evolved during dry-downs (consecutively dry days that lead to a progressive drying of the rooting zone of vegetation). It fits a recurrent neural network model (and two non-recurrent and simpler alternative models as a benchmark) and performs a targeted model diagnostic analysis (”explainable machine learning”) to investigate EF decay rates. These are then put in context to site characteristics, in particular the average rooting depth of the vegetation type per site, soil texture (sand fraction), etc.

Overall, I found this a very interesting analysis that yields informative insights into key properties of plant responses to water stress, inferred from ecosystem flux measurements. While the overall logic (inferring belowground properties of the vegetation from aboveground time series measurements) have been employed before, and also with a focus on rooting zone water storage capacity, the present study provides a clear added value: The application of a suitably targeted model diagnostic analysis on a recurrent neural network. This appears to yield clearly interpretable insights that comply with generally expected patterns. While generally to be expected (slow decay in tree-dominated ecosystems), the demonstration of how this method can be applied in this context, is valuable and opens the door to similar such applications. I would, however, like to encourage authors to revise the presentation of the manuscript for a clearer separation of results and interpretation in general, more detail on methods, a clarification of how inferences about rooting depths were obtained, and to make code and data publicly available for reproducibility of the results. I have, however, some general (major) and more specific points that I think should be addressed before publication.

Responses:

We sincerely thank you for the thoughtful and constructive evaluation of our manuscript. We greatly appreciate your positive assessment of the study's contribution, particularly regarding the application of interpretable machine learning to elucidate plant responses to water stress and to demonstrate its potential for broader use in ecohydrological research. Your comments have been invaluable in helping us further strengthen both the methodological clarity and the presentation of our findings.

In the revised manuscript, we have carefully addressed all points you raised. Below, we provide a detailed point-by-point response.

Major Points

Comment 1:

The model performance appears excessively strong and the analysis is not reproducible. It appears unconvincingly strong in view of other published studies that employed similar methods for ecosystem flux modelling, mostly with a focus on GPP, NEE, or ET (Nakagawa et al., 2023; Besnard et al., 2019; Kraft et al., 2024; Montero et al., 2024; Biegel et al., 2025). Here, EF is the prediction target. I expect that EF is even harder to model than GPP or ET in view of the known first-order control of solar radiation on GPP and net radiation on ET. These radiation components are reliably measurable and drive strong variations in GPP and ET, respectively. Hence, a "null-model" that is formulated as a GPP being a linear function of solar radiation explains already a large part of the variations in the data and ML models improve on this only to a limited extent (Stocker et al., 2020). Net radiation is factored out in EF and variations are accordingly smaller and should be harder to model than GPP and ET. Yet, the paper suggests that the model employed here is even better GPP and ET models. One explanation is that data from a give site is used for both training and testing. Hence, the model is not generalisable. However, I agree that it's permissible in the present case, where no spatial upscaling is performed. However, authors report an R-squared of 0.72 also for their evaluation on unseen sites, which is still extremely strong. It would be necessary to test their implementation of the model fitting and evaluation. However, code and data is not available. Therefore, in view of the unconvincingly strong model performance, I consider the lack of reproducibility a roadblock at the moment. Even with code provided, the model performance should be discussed with a view to the published literature.

Responses:

Thank you for pointing out this. In the revised manuscript, we now explicitly evaluate and compare four complementary strategies, including time-based split, time-based cross-validation, site-based cross-validation, and leave-one-site-out (LOSO). We clarify that the strongest performance is obtained under the time-based split, where training and testing are separated in time but remain within the same sites, and that this setting is used for the main analyses because our primary objective is to investigate EF dynamics and memory effects rather than spatial upscaling. We also explicitly report the lower performance under site-based cross-validation and LOSO, which better reflect the challenge of spatial transfer to unseen sites. We also now clarify that, unlike many previous ML studies on GPP, NEE, or ET, our primary objective is to understand temporal EF dynamics and memory effects, so the strongest reported performance should be interpreted as temporal generalization within sites rather than as strong transferability across unseen sites.

We also clarify the reproducibility of our workflow: the LSTM model was implemented using modules from the ealstm_regional_modeling repository, all datasets used in this study are publicly available through the sources cited in the Methods section, and the scripts for analysis and figure generation will be uploaded to a public repository and made openly accessible upon acceptance of the manuscript.

Comment 2:

This paper would be much clearer in presenting what is a result from their analysis vs. what is an interpretation, if Section 4 (now “Results and Discussions”) is separated into two sections, as commonly done, for results and discussions separately. This way, it can be made clear, e.g., how the relation between the model diagnostic analysis and rooting depth is established. I was confused for a long time when reading the paper. First I simply didn’t understand where this part was coming from. Only then, I realised that this is actually coming from an analysis (which brings me to the third point…).

Responses:

Thank you for raising this important point. In the revised manuscript, we have separated the former combined section into two standalone sections ("Results" and "Discussion") to better distinguish analytical findings from interpretation and to improve the overall clarity of the manuscript.

Comment 3:

Fig. 8 is very powerful. This could be made more central and the association with rooting depth (not even explained in the caption!) made more explicit. Could Fig. 8 be provided per site in an aggregated fashion, e.g., a line for each site in a common plot, line color distinguished by vegetation type? The relation to RD is hard to decipher. It would be more convincing if some average decay time scale measure is correlated with vegetation type-average rooting depth.

Responses:

Thank you for this valuable comment. We have revised Figure 8 to make the relationship with rooting depth more explicit and easier to interpret. Specifically, we revised the figure caption to clearly define RD (observed rooting depth) and to explain that t50 and t20 represent the characteristic decay timescales, i.e., the antecedent days required for the normalized memory contribution to decay from 100% of its initial value to 50% and 20%, respectively. Following the comparison approach of Stocker et al. (2023), we correlated these decay timescales with vegetation-type-average observed rooting depth using percentile-based aggregation, which makes the relationship between memory decay and rooting depth more explicit and quantitatively comparable across plant functional types. Our results show a positive relationship between memory decay timescales and rooting depth. At the same time, this relationship remains somewhat constrained by the limited number of vegetation types represented by the flux sites. In addition, memory decay timescales are likely co-regulated by soil properties and other ecohydrological factors, rather than rooting depth alone, which is also supported by the relationships shown in Figure 9.

Minor Points

Comment 1:

Abstract: "R2=0.82" - not clear what exactly is measured here? EF? only during dry-downs?

Responses:

Sorry for the misleading. That's the R^2 between EF ensemble mean predicitons and measurements for the whole periods, not just the dry-downs.

We have revised this in the revised manuscript.

Comment 2:

Abstract: "expected gradients" - capitalise throughout to clarify that it's a method name.

Responses:

Thank you.

We have revised it as suggested.

Comment 3:

l. 36: I wouldn't subscribe to such a definition of memory effects.

Responses:

Thank you for the comment. We agree that the original phrasing was not sufficiently precise. We have revised the sentence to use a more accurate and widely accepted definition of memory effects, following recent studies (e.g., Qiu et al., 2025), which define memory effects as lagged ecosystem responses to past climate conditions. We would be happy to further refine this wording if the reviewer has a preferred or more precise definition in mind.

"Current studies indicate that vegetation responses to past climate conditions exhibit lagged effects—referred to as memory effects—which can modulate ecosystem functioning, particularly after climate extremes (He et al., 2018; Hossain et al., 2022; Canarini et al., 2021; Qiu et al., 2025)."

Qiu, J., Zhang, Y., Cai, M., Keenan, T. F., Zhang, H., Gentine, P., Luo, X., Cattry, M., Zhou, S., and Piao, S.: Large contribution of antecedent climate to ecosystem productivity anomalies during extreme events, Nat. Geosci., 1–8, https://doi.org/10.1038/s41561-025-01856-4, 2025.

Comment 4:

l. 38 (”vegetation dynamics”) - revise term. This commonly refers to changes in the vegetation community composition

Responses:

Thank you.

We have revised the term "vegetation dynamics" to "vegetation functioning" to avoid confusion with changes in community composition.

"Vegetation functioning is shaped not only by concurrent climate conditions but also by lagged or memory-induced responses."

Comment 5:

l. 43: I find the connection of rooting depth with resistance and resilience not very convincing. The relationships illustrated in Fig. 1 make a connection to the sensitivity of EF to dry-downs. Would you define resistance as the inverse of sensitivity? Anyways, this model doesn't link to the ability of the recover (=resilience) after re-wetting.

Responses:

Thank you. We apologize for the earlier misleading connection. In the revised manuscript, we clarify our definition of memory effects as lagged ecosystem responses to antecedent climate conditions (see our response to minor comment 3). Because our analysis focuses on EF sensitivity during dry-down periods, it is related to resistance (i.e., the degree to which EF declines under sustained drying) but does not evaluate ecosystem recovery after re-wetting, which is a key aspect of resilience. We have therefore removed statements linking rooting depth to resilience and revised the text to avoid over-interpretation. Although our modeling framework could potentially be extended to analyze re-wetting responses in future work, such analyses are beyond the scope of the current study.

Comment 6:

l. 46: I guess one can debate about what 'limited' means, but some work that relied on ET or EF decay and its time scales should not go unmentioned here: Teuling et al., 2006 (this is really, as far as I am aware, the starting point of many similar analyses that followed); Giardina et al., 2023.

Responses:

Thank you for providing these references. We have revised the sentences as follows: "…Although previous studies have characterized ET or EF responses using decay-based time scales (Teuling et al., 2006) and recent ML approaches have separated drought impacts on ET (Giardina et al., 2023), these frameworks do not explicitly quantify event-level, driver-specific memory effects from observations. Consequently, direct evidence of such memory effect—arising from lagged interactions among key climate drivers—remains limited, highlighting the need for more robust analytical tools."

Comment 7:

l. 56 ("In the context of vegetation,…"): Appears a bit out of context in this paragraph that deals with ML and LSTM.

Responses:

Thank you. We have removed the sentences.

Comment 8:

l. 69 and Intro in general: Apparent rooting zone water storage capacity, not rooting depth, is more directly inferred from EF (rooting depth needs additional information about soil texture and groundwater table depth). Also, the logic is not complete. You can only infer that if you can regress EF variations against cumulative water deficits during dry-downs (Giardina et al., 2023). I guess this could be addressed by generally revising the formulations and toning it down: you can establish a relationship between EF dry-down patterns and rooting depth, but without knowing the amount of water consumed during that time and soil texture and groundwater, the association remains correlative, and not a quantitative estimate in a length unit.

Responses:

Thank you for this insightful comment. We agree that EF is more directly linked to apparent rooting-zone water storage capacity than to rooting depth itself, which additionally depends on soil texture and groundwater conditions. We also agree that a quantitative inference of rooting depth would require relating EF dry-down behaviour to cumulative water deficits during droughts, consistent with Giardina et al. (2023). In response, we have revised the manuscript to tone down our interpretation and clarify that the relationships identified in this study are correlative rather than quantitative. Specifically, we now frame the memory effects in EF as indicators of rooting-zone water storage and plant water-use strategies, rather than as direct quantitative estimates of rooting depth in length units. We further note in the manuscript that additional information, including soil texture, groundwater constraints, and water consumption during dry-down, would be needed for a more physically interpretable and quantitative inference in future work.

Comment 9:

Time scales (or cumulative water deficits) are key to making a connection to waters storage capacities or rooting zone depth (as descirbed also in your Fig. 1). The introduction should explain how functionalities of the applied ML techniques provide such insights. Traditionally, explainable ML has been used to diagnose learned functional relationships or variable importances. I think what's missing here is an explanation of the Expected Gradients method and the logic for how it can be applied in the context of this study's objectives. Can you refer to other applications of such methods?

Responses:

Thank you.

In our original introduction, we referenced some applications of attribution methods, noting that: *"…Such tools have recently led to theoretical breakthroughs in climate, ocean, and weather sciences (e.g., Tom et al., 2020; Barnes et al., 2020; Labe and Barnes, 2021), including identifying flooding drivers (Jiang et al., 2022)…".*

We agree that a clearer explanation of the Expected Gradients (EG) method and its relevance to this study's objectives will strengthen the introduction. In the revision, we could add a concise description of how EG works and why it is well suited for quantifying memory effects learned by LSTM models. Specifically, we could explain that EG estimates input–output sensitivities by integrating gradients along a data-informed path, enabling the decomposition of EF predictions into contributions from both concurrent and lagged drivers. This provides a direct way to diagnose the meteorological memory encoded in LSTM hidden states. We could also clarify its connection to hydrological applications by stating that EG can reveal how past anomalies—such as precipitaton anamolies or temperature extremes—

shape EF dynamics. This capability enables the memory effects identified by EG to be interpreted in terms of underlying hydrological processes, including soil-water storage capacity and rooting-zone water-access strategies.

In sum, to improve the manuscript structure, we could move the detailed description of Expected Gradients from the Supplement to the Methods section and add a short introductory paragraph summarizing its logic and relevance.

Comment 10:

Please cite Tumber-Davila et al. instead of Stocker et al., 2023 for the rooting depth data

Responses:

Thank you. We have updated the citation as suggested.

Comment 11:

The methods section lacks critical detail: What LE version was used (energy-balance corrected or not)? What temporal resolution of the data? Any data filtering (quality control-based cleaning) applied? Where is soil data from?

Responses:

Thank you.

First, we used the energy-balance-closure–corrected fluxes provided in the FLUXNET dataset (LE_CORR and H_CORR), which adjust the raw latent and sensible heat fluxes to satisfy $Rn - G \approx LE + H$ following Wilson et al. (2002). Because these corrected fluxes already account for energy imbalance, we did not apply an additional masking based on imbalance thresholds such as $(Rn - G) - (LE + H) > 30$ W m^{-2}. This clarification has been added to the main text.

In addition, as noted in the Methods section, "We assume that the errors in LE and H have comparable magnitudes, consistent with previous studies (Foken, 2008; Hollinger and Richardson, 2005; Richardson et al., 2006), and are uncorrelated. This assumption allows the mathematical elimination of errors associated with the lack of energy balance closure (Schwalm et al., 2010)."

Second, we now clearly describe the temporal resolution: the original half-hourly tower measurements were aggregated to daily daytime values in the main text (daily daytime mean for meteorological variables and Evaporative Fraction, and daily sum for precipitation in 24 hours).

We also move the full data-filtering description from the Supplement into the main Methods section, as follows:

"Data Preprocessing Procedures of the Eddy-Covariance Dataset

The original sampling frequency of the data is half hourly. The data filter procedure can be summarized as follows: First, to reduce the noise in nighttime measurements, the original data is filtered with sensible heat flux > 5 W/m2 and shortwave incoming radiation > 50 W/m2 to select the daytime data only. Then, the original data is averaged to daily scale value (precipitation is calculated as the daily sum in 24 hours). Secondly, we only keep days with a fraction of good quality data > 0.8. The gaps in the time series for input features were interpolated using established methods (Reichstein et al., 2005; Vuichard and Papale, 2015). We also visually checked site by site to ensure that the signal-to-noise ratio is acceptable. For latent and sensible heat fluxes, we used the energy-balance-closure–corrected variables (LE_CORR and H_CORR) provided in the FLUXNET dataset, which adjust the fluxes to satisfy $Rn – G \approx LE + H$ following Wilson et al. (2002). Due to the data limitation, only the shallowest soil moisture measurements were used for comparison with the evaporative fraction prediction dynamics during the dry-down periods."

Finally, regarding soil data, we note that this information is already included in the original Methods section. Specifically, we state that: "The soil attribute data is from SoilGrid with a resolution of 250 meters, including clay, silt and sand content (Poggio et al., 2021)."

These revisions ensure the Methods section now fully describes the flux data version, temporal aggregation, quality-control filtering, and the soil data source.

References:

Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., Bernhofer, C., Ceulemans, R., Dolman, H., Field, C., Grelle, A., Ibrom, A., Law, B. E., Kowalski, A., Meyers, T., Moncrieff, J., Monson, R., Oechel, W., Tenhunen, J., Valentini, R., and Verma, S.: Energy balance closure at FLUXNET sites, Agricultural and Forest Meteorology, 113, 223–243, https://doi.org/10.1016/S0168-1923(02)00109-0, 2002.

Comment 12:

Related to above: The analysis of outputs from the Expected Gradients analysis is entirely missing from the methods section. This is a glaring gap.

Responses:

Thank you for pointing this out. Our initial intention was to keep the main text concise by placing the full methodological description in the Supplement, but we recognize that this resulted in insufficient clarity.

In the revised manuscript, we can include a subsection in the Methods that explains:

(1) the Expected Gradient (EG) algorithmic formulation,

(2) how EG is computed for the LSTM model, and

(3) how EG outputs are summarized across time and events for interpretation.

Comment 13:

l. 168: unclear formulation.

Responses:

Thank you. We have revised the sentences:

"We define soil-moisture dry-down events as rainfall-free periods during which soil moisture first shows an immediate post-rain pulse increase and then decreases continuously for at least seven consecutive days until the next rainfall event."

Comment 14:

Fig. 8: Hard to decipher the legend. Please indicate what x axis is. The color scale tick labels are not readable. What is RD - explain in caption?

Responses:

Thank you. We have revised Fig. 8 to improve readability as suggested. Please see our response to Comment 3 and the updated manuscript.

Comment 15:

l. 421: needs a reference

Responses:

Thank you. We have added the reference.

Chen, S., Stark, S.C., Nobre, A.D., Cuartas, L.A., De Jesus Amore, D., Restrepo-Coupe, N., Smith, M.N., Chitra-Tarak, R., Ko, H., Nelson, B.W., Saleska, S.R., 2024. Amazon forest biogeography predicts resilience and vulnerability to drought. Nature 631, 111–117. https://doi.org/10.1038/s41586-02407568-w

Comment 16:

l. 436: too strong of a statement.

Responses:

Thank you. We have removed the statement.

Comment 17:

References

- Teuling, A. J., Seneviratne, S. I., Williams, C., and Troch, P. A.: Observed timescales of evapotranspiration response to soil moisture, Geophys. Res. Lett., 33, L23403, https://doi.org/10.1029/2006GL028178, 2006.
- Giardina, F., Gentine, P., Konings, A. G., Seneviratne, S. I., and Stocker, B. D.: Diagnosing evapotranspiration responses to water deficit across biomes using deep learning, New Phytologist, 240, 968–983, https://doi.org/10.1111/nph.19197, 2023.
- Tumber-Dávila, S. J., Schenk, H. J., Du, E., and Jackson, R. B.: Plant sizes and shapes above- and belowground and their interactions with climate, New Phytologist, n/a, https://doi.org/10.1111/nph.18031, 2022.
- Biegel et al. https://egusphere.copernicus.org/preprints/2025/egusphere-2025-1617/egusphere-2025-1617.pdf

- Nakagawa, R., Chau, M., Calzaretta, J., Keenan, T., Vahabi, P., Todeschini, A., ... & Kang, Y. (2023). Upscaling Global Hourly GPP with Temporal Fusion Transformer (TFT). arXiv preprint arXiv:2306.13815.
- Besnard, S., Carvalhais, N., Arain, M. A., Black, A., Brede, B., Buchmann, N., ... & Reichstein, M. (2019). Memory effects of climate and vegetation affecting net ecosystem CO2 fluxes in global forests. PloS one, 14(2), e0211510.
- Kraft, B., Nelson, J. A., Walther, S., Gans, F., Weber, U., Duveiller, G., ... & Jung, M. (2024). On the added value of sequential deep learning for upscaling evapotranspiration. EGUsphere, 2024, 1-30.
- Stocker, B. D., Wang, H., Smith, N. G., Harrison, S. P., Keenan, T. F., Sandoval, D., Davis, T., and Prentice, I. C.: P-model v1.0: an optimality-based light use efficiency model for simulating ecosystem gross primary production, Geoscientific Model Development, 13, 1545–1581, https://doi.org/10.5194/gmd-13-1545-2020, 2020.

Responses:

Thank you. We have added all the references as suggested.

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.

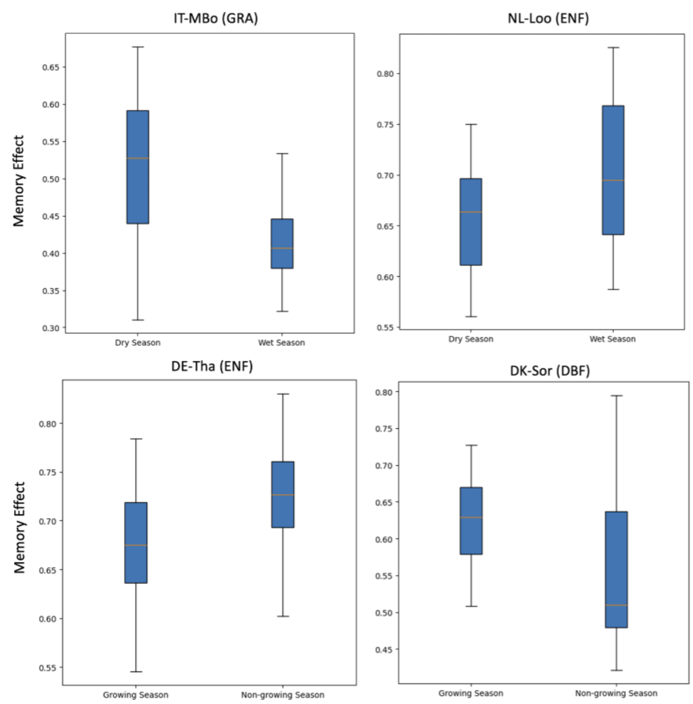

Figure R1. Memory effect comparisons between dry vs wet season and growing vs nongrowing season.

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 the temporal train–test split at each site has limitations. To address this concern, we have explicitly tested and compared several complementary training strategies, including time-based split, time-based cross-validation, site-based cross-validation, and leave-one-site-out (LOSO), and we now discuss their differences in detail in the revised manuscript. We clarify that the strongest performance is obtained under the time-based split, where training and testing are separated in time but remain within the same sites. We chose this setting for the main analyses because our primary objective is to interpret EF dynamics and memory effects at each site, rather than to emphasize spatial extrapolation or generalization to unseen sites.

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 z-score standardized using the mean and standard deviation computed from the training set, with the same statistics applied to the validation and test sets to avoid information leakage. The model follows the standard LSTM architecture (Hochreiter & Schmidhuber, 1997) as implemented in PyTorch, in which sigmoid activations are applied to the input, forget, and output gates, and tanh activations are applied to the cell-state and hidden-state transformations; the final output is produced by a single linear layer without an additional non-linearity, which is standard practice for regression targets. The model is trained using the Adam optimiser with an initial learning rate of 1×10^{-3} (decayed to 5×10^{-4} after epoch 10), a batch size of 64, a mean squared error (MSE) loss, 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.

Please check the updated Method section 3.3.

Comment 5:

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

Responses:

Thank you for pointing this out. We have clarified in the revised manuscript that separate site-years were used for testing and validation. Specifically, for each site, one entire site-year was held out as the test set, one different site-year was used as the validation set, and all remaining site-years were used for training. This strictly temporal split avoids overlap among training, validation, and testing and reduces the risk of temporal leakage. In addition, we also evaluated other complementary training strategies, including time-based cross-validation, site-based cross-validation, and leave-one-site-out (LOSO), to assess the robustness of model performance under different validation settings. We have revised Line 186 accordingly.

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 2013-10-05) 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.