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: 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 initial stage of model development and experiments, we evaluated two training strategies: (i) time-split (training and testing within the same site but separated in time) and (ii) site-split (leave-one-site-out). We chose the time-split results for the main text because the primary goal of this study is to understand EF dynamics and memory effects, rather than spatial upscaling. To directly address your concern, in the revised manuscript we will discuss the performance differences explicitly in the context of previously published ML ecosystem-flux studies, and provide a public repository containing code and scripts to ensure reproducibility once the editor confirms that a revised submission should proceed.

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 now separated these into two standalone sections ("Results" and "Discussion"). Once the editor approves these revisions, we will upload the updated manuscript accordingly.

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 now revised the figure caption to explicitly define RD (effective rooting depth) and clarify how it relates to the antecedent memory patterns. We did examine the memory–antecedent patterns at the site level, and the site-by-site analysis indeed shows substantial within-PFT variability in memory contributions. However, the large number of sites and the considerable site-to-site variability made the full set of site-level figures too extensive for inclusion in the main manuscript. For this reason, we summarized the results at the PFT level in Fig. 8 to maintain clarity and keep the paper at a manageable length.

We are open to adding representative site-level analyses and discussing their implications in the revised Discussion section. Once the editor approves the structural revisions, we will upload the updated manuscript accordingly.

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 clarified 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 (e.g., Teuling et al., 2006) and recent ML approaches have separated drought impacts on ET (e.g., Giardina et al., 2023), these frameworks do not explicitly quantify

event-level, driver-specific memory effects from observations. Consequently, direct evidence of such memory—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 the apparent rooting-zone water storage capacity rather than rooting depth itself, which additionally depends on soil texture and groundwater table depth. We also appreciate the clarification that 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. For instance,

"We note that the present study highlights memory effects within EF as correlates of rooting-zone water storage capacity, rather than a quantitative estimate of rooting depth. Future work will incorporate additional factors—including soil texture, explicit accounting of water consumed during dry-down, and groundwater constraints—to ultimately enable a quantitative and physically interpretable model of effective rooting depth from EF decay patterns."

Importantly, your suggestion aligns closely with a direction we are actively pursuing in ongoing work: by combining EF dry-down patterns with precipitation at the onset of dry-down, radiation, soil hydraulic properties, and groundwater constraints, we aim to move beyond qualitative associations and toward a quantitative, physically interpretable model of effective rooting depth. We sincerely appreciate this constructive suggestion, which has helped refine both the manuscript framing and future research directions.

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. Specifically, we (i) clarified the x-axis label to indicate the antecedent time window (days) for each plant functional type, (ii) enlarged and simplified the color-scale tick labels to ensure they are readable, and (iii) explained the meaning of "RD" (effective rooting depth) directly in the figure caption.

We would like to clarify that the contributions shown in Figure 8 are based on the absolute values of Expected Gradients (EGs). EGs can be positive or negative depending on whether a variable at previous timesteps increases or decreases EF; however, our goal here is to quantify the magnitude of memory effects rather than their directional influence. Therefore, for each PFT we compute:

$$PC(S,t) = \frac{\sum_{i \in S} \sum_{t=k}^{t=365} \left| EG_{i(t)} \right|}{\sum_{i \in S} \sum_{t=0}^{t=365} \left| EG_{i(t)} \right|} \times 100\%$$

In this formula, S denotes the set of variables whose memory contributions are evaluated (precipitation only in Fig. 8a, and all dynamic predictors in Fig. 8b). The index i in S sums over the variables in this set. The term $EG_i(t)$ is the Expected Gradient attribution for variable i at lag t, and its absolute value is used to quantify the magnitude of memory. The variable t represents the time lag in days, and k defines the lower bound of the antecedent window, ranging from 7 to 175 days for different boxplots in Figure 8. The numerator accumulates memory contributions from lag k to 365 days, whereas the denominator sums contributions from lag 0 to 365 days. Thus, the ratio measures the percentage of total memory contributions attributable to lags deeper than k days for a given variable set S.

To generate Figure 8a and 8b, we evaluate this metric using two different variable sets: S = {P} for precipitation-only memory, and the full predictor set S = {P, Ta, RAD, VPD, WS,

LAI} for the “All Variables” case. No other aspect of the calculation differs between the two panels.

This normalization ensures that all values are positive and comparable across variables and prevents positive and negative EG attributions from canceling each other out. Accordingly, the fact that precipitation exhibits a higher long-term contribution than the aggregated “All Variables” case does not imply negative feedbacks. Instead, it reflects two factors. First, precipitation has a substantially longer effective memory than other drivers, consistent with ecohydrological theory. Second, when all variables are included, the memory contributions decreases (compared to the precipitation-only memory effect) because the normalization is performed over a larger set of predictors, many of which (e.g., shortwave radiation) exhibit much shorter memory.

To avoid ambiguity, we have resived Figure 8 as suggested.

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 revised the text as suggested.

Revised sentence:

“Our analysis suggests that memory effects may offer insights into rooting-zone water-access behavior, a component that is challenging to observe directly and is often simplified

in many studies and Earth system models. However, further investigation is required to fully understand the mechanisms linking memory effects and rooting-depth–related processes."

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.