Citation: https://doi.org/10.5194/egusphere-2024-3726-RC1

EcoPro-LSTM v0 : A Memory-based Machine Learning Approach to Predicting Ecosystem Dynamics across Time Scales in Mediterranean Environments

Mitra Cattry, Wenli Zhao, Juan Nathaniel, Jinghao Qiu, Yao Zhang, and Pierre Gentine

Please find our response to reviewer #1 comment. Our response is printed in blue while the comments are printed in black. The line numbers refer to the new version without track changes unless otherwise specified.

The authors introduce an initial version of an ecosystem process model using the Long Short-Term Memory approach (EcoPro-LSTM v0). This model employs a temporal multitask deep learning model to predict ecosystem respiration (RECO), gross primary productivity (GPP), evapotranspiration (ET), and surface soil water content (SWC), capturing the interdependencies of these variables across different time scales. They trained and tested the model using data from several Mediterranean sites from the FLUXNET2015 database.

My expertise lies more in general modeling and physical processes rather than deep learning. That said, I find this topic highly relevant and promising, particularly due to the challenges in predicting ecosystem responses—especially respiration and processes in dry ecosystems. My primary concerns about this manuscript are its lack of clarity, excessive text and figures, and insufficient key details. In my opinion, these issues are obscuring the full potential and effort behind this work.

I highly recommend restructuring the manuscript to improve its organization and clarity. This should involve consolidating and significantly shortening the text, ensuring all information appears in its proper section, eliminating redundant content, and reducing the number of figures—currently 12, which is excessive. Additionally, it's crucial to explicitly state all key modeling assumptions. Please see below for more details.

We thank the reviewer for his positive and constructive feedback on lack of clarity, excessive text and figures, as well as pointing where details are lacking. With implementation of this reviewer comment we indeed could enhance the quality of our work and noted more points to revise.

Major comments:

- The manuscript currently presents methodological details incrementally and intersperses them with results. For instance, key information about FLUXCOM and X-BASE is omitted from the Methods section and instead introduced in discussion. To improve readability, I recommend consolidating all methodological descriptions in the appropriate sections upfront, rather than scattering them throughout the text. Additionally, the number and nature of proposed models should be explicitly stated early in the paper (e.g., in the Introduction or Methods). After reading the entire manuscript, this remains unclear.
- The manuscript suffers from a significant gap in physical and physiological explanations. While results are presented, they lack meaningful connections to underlying ecological processes or climate drivers. For example, when describing site conditions, the authors fail to contextualize how these environmental factors relate to the observed outcomes. Strengthening these mechanistic links would greatly enhance the scientific rigor and interpretability of the findings.
- The manuscript lacks a comprehensive synthesis of model performance across sites. While site-specific conclusions are presented, they appear overgeneralized without sufficient justification. Most critically, the analysis misses an overarching assessment of when and where the model performs best—a key piece of information readers expects in the conclusions. A clear, data-

supported summary of model strengths and limitations across all studied sites would significantly strengthen the paper's impact.

Thank you for pointing out that methodological details and model variants were not explained early in the manuscript. We have revised the text as follows:

- Manuscript-wide structural review: All terminologies used in Result and Discussion are now named under Data and Method section. For instance, we now introduce all model configurations and variables in one place—SLT LSTM, and EcoPro-LSTMv0 (MLT LSTM)— and FLUXCOM and FLUXCOM-X-Base early on so readers know exactly which experiments follow.
- Improved language, equations, and references: we omitted more general and unfair comparison, added more quantitative explanations. We also integrated more literature and references to justify our finding and choices.

- Study limitations and summary of evaluation: We've relocated scattered methodological details into the Methods (Sections 2–3) and streamlined the Results/Discussion to focus on interpretation only. As our model lacks spatial generalization, and as it is beyond the scope of this article, we did not discuss how mechanistic traits such as—rooting depth, soil texture, phenology—modulate interannual flux variability, and this is actually one of the shortcomings that we would overcome in our upcoming paper. We have ensured that this limitation is clearly communicated in the last discussion section.

- Added Section to discussion "Model Evaluation Summary": In this section, we offer a data-driven summary of model performance across all sites. In addition to the detailed per-site metrics already presented in Tables 1, 2, B1, B2 and Figures 3 and 9. We now begin the Discussion with a concise, tabulated overview of model skill across all 17 sites, highlighting "when and where" EcoPro-LSTMv0 excels versus its limitations.
- Revisited all figures to adjust to reviewer comments as much as possible

We believe these additions provide a clear, high-level synthesis that directs readers immediately to the model's strengths and weaknesses, fulfilling the expectation for a comprehensive, data-supported conclusion.

Minor comments:

Abstract

- The abstract should present information more directly and specifically. Rather than using vague comparative statements like "we demonstrate our model's outperforming against state-of-the-art data," please: name the specific database used, provide quantitative performance metrics and state concrete findings.
- Please indicate how many sites from FLUXNET you used.

Done, thank you!

Introduction

- Please use “e.g.” and “see” for citations only when they are necessary (throughout the manuscript.).

Only several instances at lines 39, 45, 68 and 493 are kept, thanks.

- L22-23. Please indicate why carbon uptake in semi-arid regions depends on winter and early growing season precipitation.

It is “Mediterranean region” rather than “semi-arid region” that we claim its carbon uptake depends on early growing season precipitation, as precipitation in this climate by definition Mediterranean regions have two seasons dry and wet, as demonstrated below. Line 24 has been rewritten to better reflect our discussion here.

We changed from

“Sustained carbon uptake in these regions relies heavily on winter or early growing season precipitation (Bartsch et al., 2020).”

to

“In Mediterranean regions—characterised by wet winters and dry summers—annual rainfall is concentrated in the winter–spring growing season, and sustained carbon uptake depends on this precipitation to recharge soil moisture and fuel early season photosynthesis (Peel et al., 2007; Bartsch et al., 2020).”

- L25. Please explain why physic-based models struggled to represent legacy and lag effects. Besides, what about statistical-based models and other models using IA or machine learning?

Thanks for this wonderful point, we have now added lines 28-34 to expand this discussion (see line 32 in the version with tracked changes).

We changed from

“Physics-based models historically struggled to represent legacy and lag effects (Choat et al., 2018) or their associated interannual carbon uptake variability (Cranko Page et al., 2021; MacBean et al., 2021).”
To
“Physics-based models historically struggled to represent legacy and lag effects (Choat et al., 2018) or their associated interannual carbon uptake variability (Cranko Page et al., 2021; MacBean, et al., 2021) due to reasons such as missing mechanistic drivers, simplified dynamics, or lack of critical data. The statistical methods. Statistical approaches either rely on simple effect-size metrics that cannot capture the nonlinear nature of drought legacies, or employ autocorrelation-based methods that reduce transparency in interpreting the legacy signal (Kannenberg et al., 2020).”

About AI methods, we have expanded the literature in lines 35-49 (see lines 40 in the version with tracked changes).

"Machine learning methods are promising tools. Yet, prior applications of machine learning methods— such as tree-based, artificial neural network, support vector regression— for carbon fluxes have struggled to to capture legacy-driven variability (i.e. how past climate or vegetation states influence current fluxes), especially interannual trends and extreme-event responses (Jung et al., 2011; Bodesheim et al., 2018; Jung et al., 2020, 2009). To approximate such memory effects, studies often include proxy variables (e.g. leaf area index) rather than directly feeding lagged climate data into the model (e.g., Guo et al., 2023). By contrast, Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) are a class of recurrent neural network that use special "memory cells" and gating mechanisms to learn which information to retain or forget over time. This allows LSTMs to capture temporal dependencies—like soil-moisture carry-over or vegetation phenology—directly from the sequence of past observations. More recent architectures (e.g. temporal convolutional networks and transformer models; Lim et al. (2021); Li et al. (2023)) build on the same principle to capture even longer or more complex legacy signals. Although the use of LSTM for ecosystem dynamics is relatively recent (e.g., Besnard et al., 2019). Liu et al. (2023b) showed it can successfully reproduce interannual variability in carbon fluxes by leveraging its built-in memory. And while Besnard et al. (2019); Huang et al. (2024) have compared tree-based against LSTM and demonstrated the superior performance of LSTM in predicting ecosystem respiration with a memory of up to six months, there is still a need for further model development for capturing rare and extreme ecosystem responses (Martinuzzi et al., 2023; Kraft et al., 2024)."

- L53. What do the authors mean by "Short-Term Memory approach networks"?

The words "approach networks" is redundant, and has changed to "Short-Term Memory networks". We have also now added 68-70 (line 78 in the version with tracked changes) to explain better what do we mean by mean thanks

"Here, "Long Short-Term Memory" refers to a class of recurrent neural networks that incorporate special memory cells and gate-based mechanisms which allow the model to learn which information to retain or forget over long sequences of data. This makes LSTMs especially good at modelling processes—like ecosystem fluxes—that depend on past events occurring months or even years earlier."

- L31-32. This sentence is not clear to me. Consider rewriting it more clearly.
- L55. Consider removing "like" as they are the only variables predicted by this model.

Corrected, thanks.

Section 2

- L66. LE is not mentioned in the incorporated variables.
- L66. Please indicate how evapotranspiration is calculated.

latent heat (LE) is not incorporated directly as model input or output but rather used along with air temperature (TA) to estimate evapotranspiration as explained in line 88. We have rewritten lines 110 and 115 for further clarification (line 130 in the version with tracked changes)

- L72. Does the classification of those sites change from past to future?

To address possible ambiguity in the Köppen–Geiger (KG) scheme referenced at line 86, we have added KG_climate.zip to the Zenodo record. This archive contains the scripts KoppenClimate_FLUXNET_sites.jl and KoppenGeiger.jl, which implement both KG definitions used in our study and produce site-level classifications: present products when global version="" and future products when global version="2021". The archive also includes the resulting metadata tables (Excel) listing KG classes for semi-arid sites that were used to generate Figure 1. Our pipeline evaluates both tables and, if a site is classified as semi-arid Mediterranean under either definition, it is retained in the analysis. AU-Rig is one example where the site's classification differs between the two definitions, but we avoid further elaboration in the main text as such definitions are constantly evolving.

- L74. Please specify the criteria used to remove/consider sites.

We have now added lines 97-101 (see line 115 in the version with tracked changes)
"We retained only those FLUXNET sites that had at least two full years of high quality, continuous observations for all inputs (P, TA, PAR, VPD, SD) and outputs (RECO, GPP, SWC, ET), and that included soil water content (SWC) record. Sites failing either criterion were excluded. Accordingly, we excluded sites with only one year of data—US-LWW, IT-SR2, AU-Ync, AU-Cum, IT-Cp2, US-Me4, US-Lin, and US-Tw(1–4)—and sites lacking any SWC measurements—CA-TPD, ES-Ln2, FR-Pue, and US-Myb."

- L81. Which "established relationships"? Please be more specific.

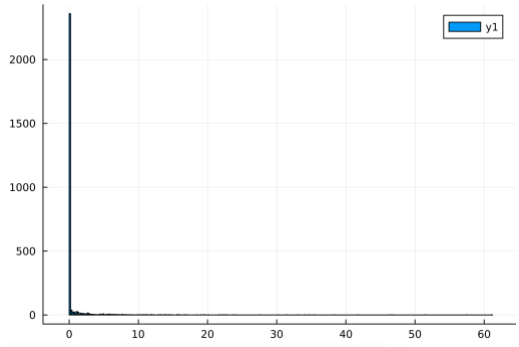
Specified, thanks

- L82. Why zero and not nan?

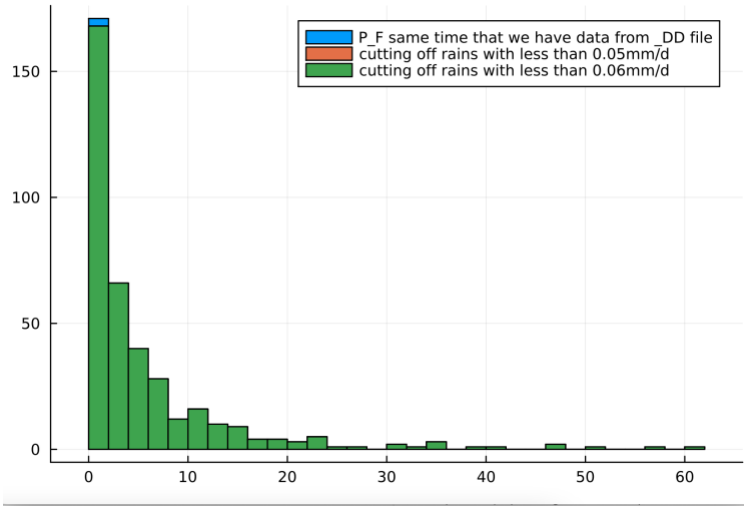
Zero PAR correctly represents "no light," whereas nan would imply "missing measurement,". To maintain continuity for input of our model, we set the negative PAR values to zero and not nan. We added a note in line 107 to reflect our discussion here

- L84. How was defined the threshold of "negligible" rainfall events? Explain how it was based on the histogram.
- L85. From where is taken the value 0.028? What does it mean?
- L86. Why cannot the SWC resulting from precipitation below this value be measured? With specific equipment?

Thanks for bringing this to our attention, so we revisited our assumptions.
The histogram of hourly data summed to daily precipitation values from FLUXNET site data is as follows, as you see it is skewed on zero, and the histogram does not resemble that of FLUXNET2015 reported at daily values (reported under _DD files)



We carried out an analysis, precipitation was first upscaled to daily values and then rainfalls less than 0.05 mm/d are removed. This gave us an exact match between histogram of precipitation reported in _DD files from FLUXNET2015 and the ones estimated from half-hourly (_HH) data.



We have rerun the code, with the 0.05 mm/d threshold instead of 1.5 mm/d threshold to ensure, the impact of small rainfalls that may accumulated in time are not ignored. In works such as Andrew Feldman et al 2021 when detecting rain pulse-response, the soil water content signal is what matters most, as there is uncertainty associated with when and where the drizzles of rain occur, to be consistent with this work, we originally applied the same filter (approximately 1.5mm/d with 0.028 mm/hr), but we decided to not consider this point. We have revisited the text, and code to ensure everything is consistent with only one filter, we didn't update the figures as results were consistent, and some level of stochasticity is always associated with each run. The only one filter is daily rainfalls less than 0.05 mm and hourly rainfall less than 0.002 mm are removed.

References:

1. Feldman, A.F., Chulakadabba, A., Short Gianotti, D.J. and Entekhabi, D., 2021. Landscape-scale plant water content and carbon flux behavior following moisture pulses: from dryland to mesic environments. *Water Resources Research*, *57*(1), p.e2020WR027592. 92. https://doi.org/10.1029/2020WR027592

- L93. Please describe how you leveraged both time scales.

We have modified line 122-124 (line 145 in the version with tracked changes) as follows to be more specific and reflect our discussion here. Further explanation is provided in the method section

"To balance the costs of using finer temporal resolution from hourly data with the reliability of daily data, we integrated both sources, leveraging hourly data to reflect short-term fluctuations and daily data to minimize errors and enhance analysis robustness."

- L94-95. This part is not necessary, especially the description of the unit symbols.

To keep our figures as polished and free of clutter as possible, we chose not to annotate every axis with units. That makes the plots easier to read at a glance, but it means we need to define our units somewhere else—hence the brief description here specifies the units of all variables that appear in figures. Thank you for bringing this point to our attention.

- L96. Please be coherent with the name of SWC, you mention both soil water content and soil moisture.

Thanks for pointing this out, soil water content (SWC), originally expressed as a percentage, was converted to a fractional scale (0–1) by dividing by 100. The variable named remained SWC for simplification in terminology in this manuscript.

- L96. SWC has units, they are volume/volume, here probably m3 m-3.

The unit of SWC from https://fluxnet.org/data/fluxnet2015-dataset/fullset-data-product/

SWC_F_MDS_#		Soil water content, gapfilled with MDS (numeric in dex "#" increases with the depth, 1 is shallowest)
HH	%	
DD	%	average from half-hourly data
WW-YY	%	average from daily data

is percentage as specified in our article. As explained in line 119, we have converted SWC to fractions (0 to 1) by dividing the original values by 100. Still, we respect reviewer point of view and modified unit from [–] to [m^3 m^{-3}]

- L98. How were defined as 4 months and 3 days?

This was partially limited by memory and computational resources, and partly justified by literature, we added lines 129-133 (see line 155 in the version with tracked changes) to reflect our discussion here. Thank you so much!

"*we structured the input data into two sequences: a daily time series spanning the past 4 months (120 days) and an hourly time series capturing the 3 days leading up to the target date. This three days window is expected to inform our model of short term pulse-response dynamics (Feldman et al., 2021) while the choice of 120 days was made due to memory limitations, and informed by previous works showing LSTM model performance doesn't improve by incorporating more than 180 days of historical data (Huang et al., 2024).*"

References:

1. Feldman, A.F., Chulakadabba, A., Short Gianotti, D.J. and Entekhabi, D., 2021. Landscape-scale plant water content and carbon flux behavior following moisture pulses: from dryland to mesic environments. Water Resources Research, 57(1), p.e2020WR027592. 92. https://doi.org/10.1029/2020WR027592
2. Huang, C., He, W., Liu, J., Nguyen, N. T., Yang, H., Lv, Y., Chen, H., and Zhao, M.: Exploring the potential of Long Short-Term Memory Networks for predicting net CO2 exchange across various ecosystems with multi-source data, Journal of Geophysical Research: Atmospheres, 129, e2023JD040 418, 2024. https://doi.org/10.1029/2023JD040418

Section 3

- L113-115. This information is repetitive.

Yes, but directly accounting for legacy effects and interannual fluctuations is the main contribution of this work. Therefore, it needs to be repeated in the method section, we have now, mentioned explicitly, that this information has been mentioned before to reflect our discussion. *"As mentioned in Section 1 and 2, we combined hourly data to capture both short-term fluctuations and daily data to directly account for legacy effects. Therefore, we use a multi-timescale LSTM framework, named EcoPro-LSTM$_{v0}$, to address short- and long- term ecosystem dynamics."*

- L116. Consider putting "see Figure 2" in parentheses.

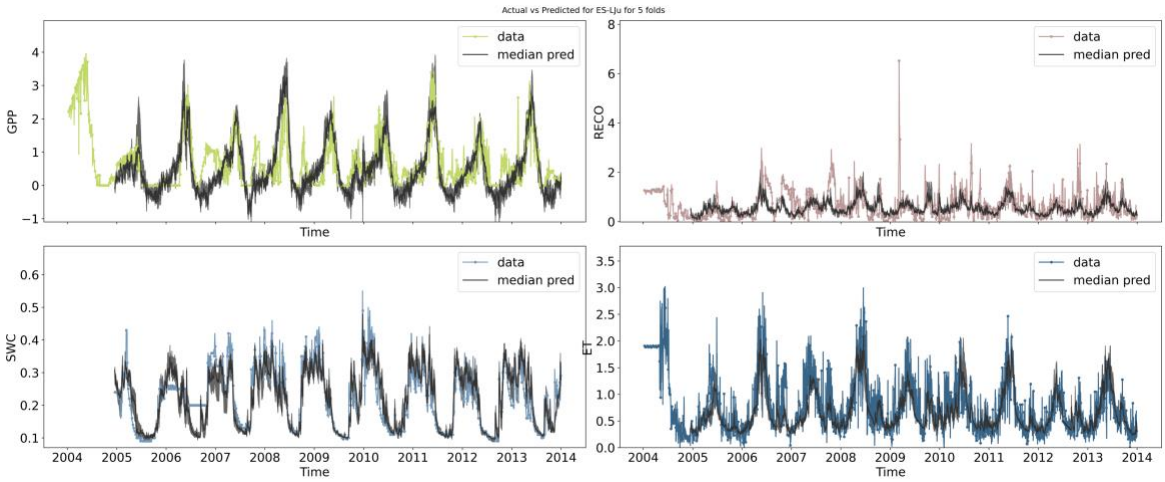
Corrected in this and everywhere else, thank you.

- L118. How were defined 110 hidden units, 0.2-0.5? What is ReLU?
- L125-126. How were define those values?
- L148. How are hyperparameters identified?

ReLU is now explained more in line 173-179 (see line 195 in the version with tracked changes). rather than relying on an automated search, we manually tuned each hyperparameter—learning rate, model architecture, hidden size, dropout rate, learning_scheduler, etc.—by testing multiple literature values.

"We use the Rectified LinearUnit (ReLU) activation (Nair and Hinton, 2010) defined by ReLU(x)=max(0,x) In words, it sets all negative inputs to zero and leaves positive inputs unchanged. Each LSTM layer contains one layer with hidden units, followed by a ReLU activation function with a dropout rate of 0.4 to prevent overfitting (Srivastava et al., 2014). ReLU activation function improves feature learning by inducing sparsity—negative inputs are set to zero, which suppresses irrelevant outputs and promotes stable gradient propagation and enhances computational efficiency as it is a simple thresholding step."

Considering we do not use GPP and RECO values directly negative values may occur at points, and use of RELU will remove these points (see figure below, where no RELU function is used and NEE observation and weekly averaged GPP is used in the error function)



Our search space was same as what commonly is applied in ML literature, as reported in the article below

https://doi.org/10.5194/egusphere-2025-1617
Preprint. Discussion started: 25 April 2025

LSTM and MLP	
Hidden dimension	32, 64, 128, 256, 512
Learning rate	$10^{-1}, 5\times10^{-2}, 10^{-2}, 10^{-3}, 10^{-4}, 3\times10^{-4}, 5\times10^{-4}, 7\times10^{-4}, 9\times10^{-4}$
Scheduler patience	5, 10, 20, 30
Scheduler factor	0.1, 0.5, 0.9
Weight decay (λ)	0.01, 0.001, 0.0001, 0.00001, 0
Batch size	16, 32, 64, 128, 256
LSTM-only	
Dropout	0, 0.1, 0.2, 0.3, 0.4, 0.5
Number of layers	1, 2, 3, 4, 5

Table 1. Hyperparameter search space for the LSTM and MLP models.

We now modified line 220-222 to explain this better (line 245 in the version with tracked changes).

"All hyperparameters were manually tuned to optimise model performance, exploring value ranges typical for LSTM architectures in the literature (see Table 1 in Biegel et al., 2025). Full search ranges and the selected values are reported in Section B1".

In SI section B1, Table B1, now we include a table of the parameters tested (a screen shot is presented here).

Table B1. Hyperparameters and tested alternatives for EcoPro-LSTM (v0).

Category	Current Value	Alternatives Used
model architecture	multi-scale LSTM with RELU	multi-scale LSTM; single-scale LSTM; GRU variants; global/multi-attention heads
hidden size	110	64, 128, 256
dropout rate	0.2	0.1, 0.3, 0.5 (0 not tested)
initial forget bias	3	1, 2, 5
Objective/Loss	WRMSE	MAE, MSE, quantile loss, spectral loss, composite losses
Optimizer	Adam	SGD
Learning rate (LR)	0.01	0.001, 0.003, 0.05, 0.1
Weight decay (WD)	0.0001	0 to 0.1 (tested up to 0.1; selected smallest among top performers)
Batch size	16	32, 64, 128, 256, 512, 1024
Epochs	80	50, 120, 250
Gradient clipping	1	None, 5, 10
LR scheduler — type	ReduceLROnPlateau	CosineAnnealingLR
ReduceLROnPlateau params	factor = 0.3; patience = 10; threshold = 0.05; min_LR = 10^{-5}	factor: 0.1/0.2/0.5; patience: 5/20; threshold: 10^{-3}/0.01; min_LR: 10^{-6}

Note. Learning rate (LR) and weight decay (WD) were tuned as a pair.

Reference:

1. Nair, V. and Hinton, G. E.: Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814, 2010.
2. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A SimpleWay to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, 15, 1929–1958, 2014.
3. Biegel, S., Schindler, K. and Stocker, B.D., 2025. Unrecognised water limitation is a main source of uncertainty for models of terrestrial photosynthesis. *EGUsphere*, *2025*, pp.1-33. https://doi.org/10.5194/egusphere-2025-1617

- Eq. 1. Very long name for a variable. Furthermore, this equation is widely known. Consider removing it.
- L137. Which scaling techniques? Why they did not yield satisfactory results?
- L141. What do the authors mean by fast time scales?
- L149-150. That time is for simulating what?
- L152-153. Any reference?

Improved, thank you.

- L152-L181. Consider to shorten this part. Leave only the relevant information to understand what you are doing.

We moved lines 152-181 in the previous version as well as its corresponding results presented formerly in lines 294-302 to Appendix A3 and Figure 5 is now A2. We only briefly mention the fact in lines 230.

Section 4

- L186-190. This information is repetitive.
- L197-198. This part is not necessary. The meaning of R2 is widely known.

Removed, thank you

- L202. Which newer metrics do the authors mean?

Explained, thanks.

- L196-. Please avoid mixing methods and results.

Moved to lines 217, thank you.

- L216-L221. This should be in the figure caption not in the main text.

Moved, thanks.

- L183. Please, use the abbreviations.

Assume, you mean, instead of using respiration, evapotranspiration, etc we use abbreviation. We tried to use abbreviation as much as possible, but still at instances, we kept full names. Thanks for bringing this to our attention.

- Table 1. It's not clear to me why, if you said you would try NEE instead of RECO, the RECO comparison appears here.

We wished to have partitioned RECO and GPP at hourly and daily timescale rather than NEE which is their residual. Yet, we wanted to remain consistent with partitioning methods. We have rewritten lines 213-5.

"*We include directly observed NEE to reduce potential sources of error and maintain consistency with the partitioning method, although NEE itself plays no major role in improving model performance."*

- L222-L224. This information is given without any context.
- L222-. Please consider describing only the key results and include information about the site when it's truly relevant. For example, when it makes the results easier to interpret.
- L268. Please avoid these very general sentences. (check the entire article)
- L276-277. Is this capability stated in the result for this specific site or is it general? If it's for the site, please specify; if it's general, please justify it.
- L284. Add space between the text and the parenthesis.

Corrected, thanks.

- L281. All labels, right? Also, consider always naming this set of variables the same way to avoid confusion.

We assume you refer to all Figures, it is corrected, thank you

- L285-286. How do the authors know that? Which specific factors?
- L288. Please indicate what is considered high data quality.

You are right, we do not know, we rewrote this sentence with less conviction

- L290. Strong correlation among them? How is that related to the challenge to predict GPP and RECO?
- L291. What is stable soil moisture? Perform the best describing which variable/s?
- L300. What do the authors mean by "matter of convenience" in this context?
- L300-301. Please fix the references.
- L301-302. Please explain what it means.
- L304. Here is the first time you use IA. Consider linking this to the introduction and methods.
- L307-314. Here you are mixing methods, legend captions, and results.

Clarified, thanks. As for legend, we are currently following this format: title, explanation, key conclusion message. If you have other suggestion to further better classify method, legend and results, please, let us know.

- L312. How is long-term defined here? Consider defining it at the beginning, arguing the consideration.

Now explained in lines 235-41 as follows (see line 295 in the version with tracked changes):

"*$\Sigma_n^m\ IG_i^X$ essentially explain how meteorological time series (X) contributed to a change in the target output relative to the baseline considering the historical input occuring over the period of n to m days prior to target date. As the variables we use in our training are scaled to be dimensionless, $\Sigma_n^m\ IG_i^X$ is also dimensionless. We separated the short-term impacts occurring during 1 week before the target date indicated by $\Sigma_1^8\ IG_i^X$ from long-term impacts occurring during the entire 3 months historical data before the target date up to 1 week before indicated by by $\Sigma_8^{120}\ IG_i^X$ where X is the meteorological time series (such as precipitation, P) used to predict the desired model output (e.g., GPP).*"

- L318. How was that date chosen?

Now explained in lines 345-7 as follows:

"*We chose this date because all four sites exhibited gross primary productivity values above their respective 75th percentiles, allowing us to explore feature importance under peak-carbon-uptake conditions.*"

- L318. Which specific environmental forcing?
- L323. In all sites? How much impact?

Clarified, thanks. The impact, is not quantitative in absolute terms as explained in Method section lines 235-41. We use also the word "often" to indicate the relationship doesn't hold necessarily for all days.

- L323. PAR is not always inversely proportional to cloudiness. By the way, what relevance does this have in this sentence? Furthermore, parentheses indicating this relationship appear at least three times in this subsection.

Very good point "inversely proportional" is indeed a strong word, corrected.

- L325. Please consider explaining their meaning instead of writing down values. Also, which panels are you referring to?

Now explained in method at lines 225-45 (lines 290-300 in the version with tracked changes)

- L328-330. What do the authors mean by "more consequential" in this context? Please explain further the relationship you are making here as it is not clear what happened in those rain events.
- L331. Why is VPD in quotes?
- L332. The term "air aridity" isn't widely used. Consider explaining it or simply describing directly what you mean.
- L331-339. What do the authors mean by words like "favourable levels", and "beneficial"? Consider replacing them with increases/reduces.
- L331-339. Please consider replacing the writing of these equations with their meaning.
- L337. Please specify what is the meaning of "beneficial influence".

Clarified, thanks.

- L339. Where can this conclusion be drawn?

This is not a conclusion, a statement, now moved to line 342 for clarity (see lines 484 and 455 in the version with tracked changes)

- L340. How can be VPD and TA confounded?

We thank the reviewer for pointing out that "confounded" implies a causal-inference issue where two variables cannot be disentangled. What we intended to convey is that VPD and TA co-vary so tightly in field observations—and that the literature has not yet reached consensus on their distinct effects on GPP—that it can be difficult to ascribe physiological responses unambiguously to one or the other.

Replaced "confounded" with "co-vary" in Discussion 5.2 (now line 425). This revision both removes the ambiguous term "confounded" and clarifies the underlying physical basis and ongoing debate regarding VPD vs. TA in controlling GPP.

- L347-350. Please be specific. What does it mean that "PAR usually exhibits reduced sunlight negatively impacting GPP"? Deviations where?

Rewritten for clarity thanks

- L359. What is the relation of that with the model?

Thanks for your note, rewrote lines 385-9 for better clarity as follows

"*There are some limitations to our method. For instance, when comparing Figures 7a and b, SWC in US-Ton are nearly double those in US-Var, primarily due to access to groundwater (Baldocchi et al., 2021). Since our model does not include groundwater as an input, the increased SWC is explained by the enhanced contribution from precipitation, temperature, and PAR on SWC.*"

- L362. 8 and 6 what?
- L362. Which environmental conditions?
- L363. I do not understand how that indicates stable water use and unstressed conditions.
- L364. Please specify which comparative analysis.

Rewritten lines 389-99 to clarify as follows:

"When comparing Figure 7 against Figure 5 (meteorological control on ET versus GPP). We observe that environmental conditions (P, SD, VPD, PAR, TA) affect ET and GPP in the same direction and order as GPP. It implies that carbon assimilation and water loss are driven by the same factors to a similar degree. Such coupling is characteristic of non-water-limited systems, where stomatal regulation and soil moisture supply allow plants to maintain a consistent water-use efficiency under unstressed conditions. As Medlyn et al. (2011) show that, when stomata operate under the "optimal" trade-off between carbon gain and water loss, the same environmental drivers (light, VPD, etc.) enter both the photosynthesis as a result, under non-water-limited (non-stressed) conditions, GPP and ET changes in VPD, PAR (light) or precipitation-driven soil moisture—maintaining a near-constant instantaneous water-use efficiency Comparing feature importance of different target variables against each other underscores the capability of interpretable AI to provide detailed insights into the ecological outcomes of climate variability on the different model outputs."

Section 5

- L368. FLUXCOM and X-BASE are not mentioned until here in any section. Again, you are mixing methods, results, and now, discussion.

A subsection is now added under data description at lines 159-165 explaining the data as follows.

- L370-373. You are comparing different metrics. Maximum R2 and then "often falls below". Please compare the same metric.
- L375. The maximum is greater than the limit range.
- L376. More balanced than what?
- L380. Performs better than what? Please specify which several sites.

Removed or corrected, thanks

- L381. FLUXCOM-X is the benchmark, you used the data of X-BASE.

Now, we use FLUXCOM-X-base.

- L379-381. Again this comparison does not seem fair. Scores up to for several sites, then ranging, then more negative. Please compare the same metric among the different products.
- L388. Please indicate better results than what. Furthermore, does that occur always? In all sites?

Removed, thanks

- L389-390. With which conclusions?
- L383. Which model? Also, do not FLUXCOM and X-BASE capture them?

Moved to discussion, thanks

- L391. What is advanced FLUXCOM?
- L393. Why KGE and R2 if Figure 9 shows only NSE.
- L394. What about the R2 and NSE of FLUXCOM and X-BASE?

Corrected thanks

- L395-397. Only in your model or also in FLUXCOM and X-BASE?

We have specific numbers in Figure 8, we have rewritten lines 432-35 to avoid confusion.

"At ES-LJu, FLUXCOM-X-base KGE of -0.82, R^2 of 0.23, and NSE of 0.15 while FLUXCOM performance metrics for this site are KGEs -1.2, R^2 0.37 and NSE -0.06, respectively. Our model, on the other hand, achieves a KGE of 0.1, an R^2 of 0.44 and NSE of 0.16, suggesting a clear advantage."

L402 - . Too much information. Please consider only writing the key points.

We removed what you noted as unfair comparison, so we hope now, there is not too much information.

- L408. Please explain what is considered "acceptable performance".
- L408-409. Please explain why you believe that.
- L414. Please directly mention the key environmental variables.

- Corrected, thanks

- L423. Why indicate precipitation infiltration? Soil moisture is much more complex than that.
- L427. Please argue that affirmation.

We agree that soil moisture is much more complex than a direct proxy for precipitation infiltration. Our intention was not to equate the two, but rather to note that changes in soil moisture partly reflect

infiltration alongside other fluxes in and out of the soil medium (e.g., redistribution, evaporation, root uptake) and that is why we include other fluxes along soil moisture. While capturing the full complexity of soil moisture dynamics remains challenging, we believe it helps to better allocate values to how precipitation contributed to carbon fluxes. We rewrote lines 423-7 as follows (now lines 468-74):

"*We hypothesise that the differences arise from the multi-task (MTL) setup, which jointly predicts SWC and ET in addition to carbon fluxes. Because SWC integrates infiltration and ET reflects plant and atmospheric responses, these auxiliary targets add constraints that help the model disentangle precipitation-related signals. In our experiments, precipitation attributions under MTL are more temporally consistent than under the single-task variant. This accords with evidence that Mediterranean ecosystems are typically water-limited, making precipitation pulses a primary driver of variability in carbon and water fluxes in Mediterranean regions (Wang et al., 2016; Bartsch et al., 2020). Given known site-scale uncertainties in precipitation (Giorgi and Lionello, 2008), we interpret this as improved attribution, rather than definitive recovery of the "true" precipitation effect."*

References:

1. Giorgi, F., & Lionello, P. (2008). Climate change projections for the Mediterranean region. *Global and planetary change*, *63*(2-3), 90-104. https://doi.org/10.1016/j.gloplacha.2007.09.005
2. Wang, J., Xiao, X., Wagle, P., Ma, S., Baldocchi, D., Carrara, A., ... & Qin, Y. (2016). Canopy and climate controls of gross primary production of Mediterranean-type deciduous and evergreen oak savannas. *Agricultural and forest meteorology*, *226*, 132-147. https://doi.org/10.1016/j.agrformet.2016.05.020
3. Bartsch, S., Stegehuis, A. I., Boissard, C., Lathière, J., Peterschmitt, J. Y., Reiter, I. M., ... & Guenet, B. (2020). Impact of precipitation, air temperature and abiotic emissions on gross primary production in Mediterranean ecosystems in Europe. *European Journal of Forest Research*, *139*(1), 111-126. https://doi.org/10.1007/s10342-019-01246-7

- L441. What do the authors mean by "adverse consequences"?

We modified line 487-91 (lines 640 in tracked changes version) as follows
"At the same time, contrary to previous studies, we find that elevated VPD negative effect on carbon uptake are limited to the late stages of the growth period; during the early stages, however, it occasionally produces a slight positive effect on GPP, although these increases remain minimal. Temperature, coupled with photosynthetically active radiation, positively supports carbon uptake during its peak phase."

Section 6

- L466-468. I do not understand this conclusion. Nor do I consider it a conclusion of this work.

Removed, thank you

Figures

Fig. 1.

- This map should be cropped, focusing only on the regions of interest and allowing for a better view of the sites. The size of the circles makes it impossible to see the other areas. For example, in the US, it's impossible to see how many sites there are.

- What ‘does diversity’ mean here?
- Consider changing the color table. Continuous color tables are for continuous values and those are discrete.
- The last sentence is not relevant in a figure caption.
- With “productivity” do you mean “GPP”? Consider using the abbreviation you defined at the beginning to avoid confusion.
- Units of productivity should be written as gC/(m$^{2} \cdot$ y) or gC(m$^{-2} \cdot$ y$^{-1}). As it is, it seems years is in the numerator.

Fig. 2.

- VPD is vapour pressure deficit, not air aridity.
- SWC is soil water content.
- What do you mean by “site-name inputs”?
- Why “including”?
- Define X and Y.
- The last sentence is not relevant in a figure caption.

All changes are implemented in addition to some requests and suggestions from reviewer #2, thank you.

Fig. 3.

- Add the units (not in the caption).
- This figure is only briefly mentioned in the main text.
- Consider changing 'panel' to ‘row’. A panel is each subplot. Besides, add the variable abbreviations in the caption.
- What is ‘set’?

Now under discussion section “model evaluation summary” this figure is mentioned. The word "set" is now explained when referring to this figure for the first time in lines 248-53 (lines 310 in the version with tracked changes). The comments are implemented. Thank you so much!

Fig. 4.

- Please put titles in y-axes.
- Consider differentiating which parts correspond to training, testing, and validation.
- The last sentence is not relevant in a figure caption.
- What are the colors?

Implemented thank you. As you can see in reviewer #2 response, the performance across all folds were similar so we averaged the data across all folds. The test, validation, and train set will be complicated to be shown as this set differs per fold.

Fig. 5.

- Please put the units in the figure.
- Consider removing the space at the beginning of each time series to enhance visibility.
- Please put the label in the color bar.

This figure is moved to Supplementary information as per suggestions of reviewer #2 and yours.

Figs. 6, 7, and 8.

- Please put the units in the figure.
- Please use variable abbreviation instead of productivity.
- The penultimate sentence is not relevant in a figure caption.
- Please correct the units of the bottom figures.

Fig. 9.

- Describe the figure in the caption.
- What is the meaning of the colors?
- The last sentence is not relevant in a figure caption.

Fig. 10.

- It is very difficult to see the lines of FLUXCOM. Please consider modifying the colors.

The colors are chosen to be colorblind friendly; it is hard to find another match.

Fig. 11.

- Please put the units in the figure.
- Please put the label in the color bar.
- The last sentence is not relevant in a figure caption.

Fig. 12.

- Please explain in the caption the figure. What are the radial figures? What do the numbers on the right mean? What are the units? The colors?
- The last sentence is not relevant in a figure caption.

We polished all the figures and their captions considering your requests. Thank you so much for taking the time to point out these improvements. In addition, we realized we have referred to FLUXNET2015 data as observation while it is solely an estimation. We have retained the final takeaway sentence in some captions to highlight the key result, in keeping with journal style for self-contained figures. We also avoided keeping units or y- and x- axis in all subpanels to keep the figure uncluttered and easy to read.