https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

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

Mitra Cattry [1], Wenli Zhao[1,2,3], Juan Nathaniel[1], Jinghao Qiu[4], Yao Zhang[4], and Pierre Gentine[1]

[1]Earth & Environmental Engineering Department, Columbia University in the city of New York, NY, USA
[2]Max Planck Institute for Biogeochemistry, Jena, Germany
[3]ELLIS Jena Unit, Jena, Germany
[4]Institute of Carbon Neutrality, Sino-French Institute for Earth System Science, College of Urban and Environmental Sciences, Peking University, Beijing, China

Correspondence: Mitra Cattry (ma4451@columbia.edu)

Abstract. Climate change is anticipated to alter the global water and carbon cycles, but the spatiotemporal effects of these climate-induced shifts remain poorly understood. Of particular relevance are the variations in rainfall intensity and frequency affecting the carbon and water cycles from daily to interannual time scales. Yet, the current models fail to reproduce these processes as capturing the complex interactions and interrelated dependencies at different timescales (daily to seasonal) requires the simultaneous estimation of multiple interconnected ecological processes. To address this challenge, here, we introduce initial version of our ecosystem process modelling using Long Short-Term Memory approach (EcoPro-LSTM$_{v0}$) which uses a temporal multitask deep learning model designed to predict ecosystem responses, focusing on critical terrestrial variables, including ecosystem respiration (RECO), gross primary productivity (GPP), evapotranspiration (ET), and surface soil water content (SWC). Our approach leverages the capabilities of LSTM networks to capture the interdependencies of those processes across time scales. LSTMs excel at time-series prediction because they can learn long-term relationships and patterns in data. We trained and tested our model using long-term data from FLUXNET2015 Mediterranean sites (at hourly and daily time-steps), mainly in the USA and Europe, known for their ecological diversity and significance. We demonstrate our model's outperforming against state-of-the-art data products and test the robustness of our model and findings through k-fold cross-validation. We also showcase the model's interpretability in revealing how short- and long-term atmospheric drivers, like precipitation, influence GPP in Mediterranean climates. This model and accompanying insights can help better understand and manage ecosystems under climate change, especially in response to changing extreme events.

1 Introduction

Changes in atmospheric drivers of the terrestrial ecosystems, especially in rainfall intensity and frequency, are expected to affect the terrestrial water and carbon cycles across all ecosystems on inter- and intra-annual time scales (see, e.g., Poulter et al., 2014; Friedlingstein et al., 2022). In this context, semi-arid Mediterranean ecosystems, which form significant areas of carbon sinks and display pronounced interannual variability, are susceptible to these changes (see, e.g., Giorgi and Lionello, 2008; Lionello

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

and Scarascia, 2018; Cos et al., 2022). Sustained carbon uptake in these regions relies heavily on winter or early growing season precipitation (e.g., Bartsch et al., 2020). The first needed step in estimating the consequences of anticipated shifts in precipitation is to create a modelling framework that properly accounts for lag and legacy effects (e.g., Katul et al., 2001; Kannenberg et al., 2020). Physics-based models historically struggled to represent legacy and lag effects (e.g., Choat et al., 2018) or their associated interannual carbon uptake variability (e.g., Cranko Page et al., 2021; MacBean et al., 2021). Thus, there is a need for a modelling framework that more readily represents the temporal variability in responses to atmospheric forcing (e.g., Yang et al., 2023).

Machine learning methods provide a promising solution to these challenges (see, e.g., Guo et al., 2023), though prior applications for carbon fluxes have struggled to represent variability over time, particularly interannual changes and responses to extreme events (see, e.g., Jung et al., 2011; Bodesheim et al., 2018; Jung et al., 2020). Liu et al. (2023b) demonstrated the ability of the Long Short-Term Memory (LSTM) modelling framework to capture interannual variabilities in carbon fluxes due to the ability of LSTM to account for temporal dependencies. Huang et al. (2024) also confirmed this finding, who compared Random Forest against LSTM and demonstrated the superior performance of LSTM in predicting ecosystem respiration with a memory of up to 6 months.

Yet, data-driven approaches, especially LSTM, may grapple with issues such as data sparsity and equifinality where multiple models can explain the observed data equally well, leading to uncertainty in model selection and interpretation (Abdar et al., 2021). In ecological modelling, the issue of equifinality is particularly problematic because different model configurations could suggest contrasting ecological responses to the same atmospheric forcing, thereby complicating efforts to understand ecosystem dynamics and forecast future conditions. Multitask Learning (MTL) is a powerful approach for addressing these equifinality challenges (see, e.g., Caruana, 1997; Ruder, 2017) while allowing for the simultaneous estimation of multiple interrelated processes. However, the application of MTL to increase the interpretability of the current machine learning models predicting carbon fluxes has been relatively limited (see, e.g., Cohrs et al., 2024; ElGhawi et al., 2023; Yan et al., 2024) and often models either solely focused on net ecosystem exchange and its derived carbon fluxes (see, e.g., Shangguan et al., 2023; Nathaniel et al., 2023) or other relevant components of carbon pool (see, e.g., Liu et al., 2024) while neglecting highly coupled ecological responses related to water and energy.

Our study focuses on Mediterranean regions using the FLUXNET2015 database to enhance the interannual variability of modeled carbon fluxes and the reliability of machine learning interpretations. These tower sites provide high-quality data from diverse conditions, enabling studying ecosystem processes like gross primary productivity (GPP) and ecosystem respiration (RECO), as well as related biophysical variables such as evapotranspiration (ET) and soil water content (SWC). To characterise these relationships at different temporal scales, we introduce our initial version of ecosystem process modelling using Long Short-Term Memory approach (EcoPro-LSTM$_{v0}$) networks, capable of preserving information in time and accounting for the temporal dependencies inherent in ecological systems.

In developing EcoPro-LSTM$_{v0}$, our objectives are: (1) to construct a robust, multitask, multi-timescale machine learning model for predicting processes like RECO, GPP, ET, and SWC; (2) to validate the model's performance using comprehensive data from Mediterranean FLUXNET2015 sites; and (3) to enhance the interpretability of the model's predictions, offering more

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

transparent insights into the impacts of different drivers on ecosystem dynamics. Our results demonstrate that our proposed modelling framework offers accurate and interpretable predictions for carbon uptake and other critical processes compared to former studies. These findings, in turn, can enhance our understanding of how atmospheric conditions shape ecosystem dynamics, particularly within the framework of climate change.

2 Data Description and Processing

Our model incorporated precipitation (P), air temperature (TA), photosynthetically active radiation (PAR), vapour pressure deficit (VPD), and snow depth (SD) as inputs. We retrieved all data except snow depth from FLUXNET2015 at half-hourly and daily timesteps. Snow depth data was retrieved hourly from the publicly available Copernicus platform, with daily estimates derived from the hourly data. Our model predicts ecosystem respiration, gross primary productivity, soil water content, and evapotranspiration (estimated from latent heat (LE) and TA in the FLUXNET2015 data repository). Additionally, we used the normalised snow cover (NDSI) data as input variables for the model; however, we excluded them from our final models due to insufficient signal coherence in the studied sites. For more details on the choice of environmental variables, see Appendix A. The data covered diverse plant functional types and environmental conditions: evergreen needle-leaf forests (US-Blo, IT-SRo), grasslands (US-Var, US-AR1, AU-Rig), woody savannas (US-Ton), croplands (US-ARM, IT-BCi, IT-CA2), shrublands (ES-LJu, IT-Noe), evergreen broad-leaf forests (IT-Cpz), and deciduous broad-leaf forests (IT-PT1, IT-CA1, IT-CA3, IT-Ro1, IT-Ro2). These sites span Mediterranean hot summer (Csb) or warm summer (Csa) climates based on the current or future Köppen climate classification (Beck et al., 2023), with data records ranging from 3 to 14 years. Despite our efforts to maximise coverage, we had to exclude some sites from our modelling due to data limitations, such as US-LWW, IT-SR2, AU-Ync, AU-Cum, IT-Cp2, and US-Me4, US-Lin, US-Tw to US-Tw4 which only had one year of high-quality data available for both input and output variables. Additionally, CA-TPD, ES-Ln2, FR-Pue, and US-Myb lacked soil moisture data, which is essential for our analysis. Figure 1 illustrates the geographic locations of the included sites, with circle size representing long-term annual GPP records and colour indicating plant functional type, highlighting the broad range and diversity of environmental conditions represented.

Before wrapping and splitting the data, we did some processing and refinement as follows: When PAR data was unavailable, it was estimated from shortwave radiation (SW) using established relationships (Jacovides et al., 2003; Meek et al., 1984). During the data cleaning, we set all negative PAR (Photosynthetically Active Radiation) to zero, as they were physically unrealistic. Additionally, we adjusted precipitation measurements below a small threshold (less than 1.5 mm/d) to zero to eliminate negligible rainfall events. We set this threshold based on the histogram of upscaled daily estimates from the half-hourly FLUXNET2015 repository, aligning with the daily data after excluding rain events below 0.028 mm (1.34 mm/day). This threshold was slightly adjusted to 1.5 mm/day, as rainfalls below this value triggered no measurable SWC response. As for the model output, we calculated ET by dividing the latent heat (LE) by the air-temperature-dependent latent heat of vapourisation (Allen et al., 1998). We converted SWC to fractions (0 to 1) by dividing the original values by 100. We focused on daytime partitioning estimates of GPP and RECO based on net ecosystem exchange (NEE). Although hourly GPP data

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

could introduce inaccuracies, mainly due to the inclusion of gap-filled data (see, Vekuri et al., 2023), daily GPP records are of higher reliability (see, e.g., Reichstein et al., 2005; Zhang et al., 2018). To balance the costs of using finer temporal resolution from hourly data with the reliability of daily data, we integrated both sources, leveraging each to minimize errors and enhance analysis robustness. We unified units in both hourly and daily timescale, as follows: water fluxes (ET, SD and P) were measured in mm water per square meter per day (mm/m^2·d), and carbon fluxes (GPP, RECO, NEE) in grams of carbon per square meter per day (gC/m^2·d). VPD is expressed in kilo pascal (kPa), TA unit is celsius (°C), and PAR is in watts per square meter (W/m^2) while soil moisture has no unit (-).

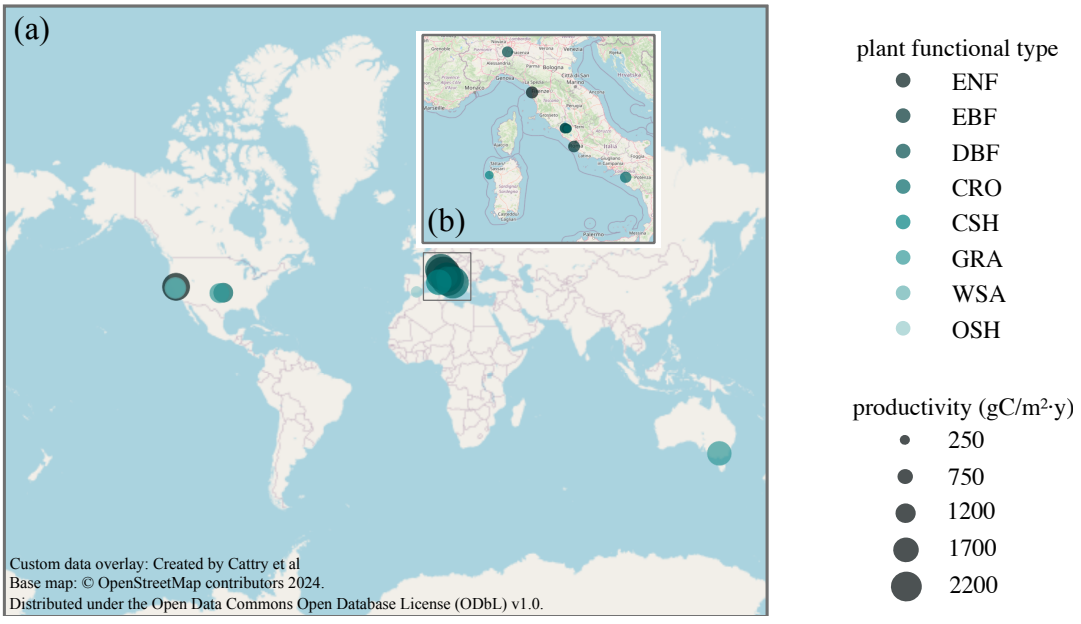

Figure 1. (a) Geographical distribution and diversity of FLUXNET2015 sites used in our study (b) a closer view at tower sites in Italy. The circles' sizes represent the long-term annual GPP, with circle size indicating values, while the colours highlight various plant functional types (PFTs). This diversity in GPP range and PFTs ensures robust modeling under varied ecological conditions.

For any target data containing all target variables (GPP, RECO, ET, and SWC), 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. The dual-resolution framework and time steps enable the LSTM to account for both long- and short-term variability. Since modelling productivity is a primary goal, we excluded target dates where GPP records were zero for over 120 consecutive days to avoid dormancy phases, ensuring the model focuses on relevant productivity signals. We scaled the daily and hourly data using a single mean and standard deviation estimated from daily records from all sites. We normalised the data before wrapping, splitting, and training to maintain consistency. To assess the model's robustness, we applied K-fold cross-validation. For K-fold, the data was split into 5 folds (20% test). After separating the test set, the remaining data was split into training and validation sets with a 90-10% ratio. All splitting were carried out without shuffling to maintain the autocorrelation

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

in the time series. Any training dates (in target date or wrapped dates) overlapping with test or validation dates were removed to prevent data leakage. Also, validation dates overlapping with the test set were removed to ensure the model's performance evaluation reflected true generalization capability. In 50 Monte Carlo (MC) simulations, the dropout layers randomly drop different neurons in each pass, which effectively perturbed the model's predictions for uncertainty evaluation. Each fold was run in parallel with a unique random seed for reproducibility.

3 Model Design and Explainability

3.1 Model Description and Evaluation

Atmospheric conditions drive ecosystem responses on interannual and intraannual levels, exhibiting lag and legacy effects. In EcoPro-LSTM$_{v0}$, we use a multi-timescale LSTM framework to address these dynamics, combining hourly and daily data to refine time-dependent variability. The architecture of our EcoPro-LSTM$_{v0}$ model consists of two LSTM layers designed to predict hourly and daily data for GPP, RECO, ET, and SWC, see Figure 2. Our model used time series of VPD, PAR, SD, TA, and P as inputs, along with one-hot encoding to differentiate between sites in an embedding format. Each LSTM layer contains one layer with 110 hidden units, followed by a ReLU activation function with a dropout rate ranging from 0.2 to 0.4 to prevent overfitting (we tuned these hyperparameters to enhance performance). The hidden and cell states for the daily LSTM layer are initialised to zeros, providing a neutral starting point for each sequence. For hourly predictions, the hidden and cell states from the daily predictions are used to initialise the hourly LSTM layer with 110 hidden units. We used an initial forget gate bias of 3 to encourage the model to retain most of the past information early in training at both time scales. The model is trained for a maximum of 150 epochs with early stopping, where training halts if the validation error does not decrease by 0.01 for 30 consecutive epochs (examples of task-specific and dataset losses are in Appendix B2, and kernel density functions per dataset in Appendix B3). The learning rate starts at 0.01 and is reduced by a factor of 0.3 if the change in validation error is less than 0.05 for ten consecutive epochs, with a minimum learning rate of 0.00001 (an example of the model's learning rate adjustment is in Appendix B1).

To ensure stable learning, we use a batch size of 16. We employ an Adam optimiser with L2 regularization of 10^{-4} (usually defined in the range of 10^{-3}-10^{-5}). The loss function used in our study is a weighted root mean square error (weighted RMSE) as illustrated in Equation 1:

$$\textit{Total Weighted RMSE} = \sum_{\text{freq}} \sqrt{\frac{\sum_{i=1}^{n} w_i \cdot (\overline{y}_{\text{true},i}^{\text{freq}} - \overline{y}_{\text{pred},i}^{\text{freq}})^2}{\sum_{i=1}^{n} w_i}} \tag{1}$$

Where *Total Weighted RMSE* is the sum of error at hourly and daily timescale, y represents model output such as GPP, $\overline{y}_{\text{pred}}$ is the scaled predictions, $\overline{y}_{\text{true}}$ is the scaled observations, the w_i are the weights here defined as $|\overline{y}_{\text{true},i}|$ which is the magnitude of the standardised observed target variable. After training the model, we compare quantile errors at the 0.65 and 0.85 percentiles to assess if the model performs similarly in different quantiles, ensuring no bias is introduced toward one or the other (refer to

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Appendix B4 for detailed comparison). This strategy advances the model's handling of extremes that are rare in the data. We also explored alternatives like data augmentation and different scaling techniques, but they did not yield the same satisfactory results as weighted RMSE.

For GPP, instead of directly estimating the error, we use a moving weekly window to compute the error. This technique aligns with the GPP daytime partitioning method, which also utilises a moving window for improved accuracy (see, e.g., Reichstein et al., 2005; Lasslop et al., 2010). We used observed net ecosystem exchange (NEE) data at fast time scales to constrain the model predictions of GPP and RECO: instead of directly comparing RECO estimates at the site level, we compare observed RECO-GPP (NEE) against the model's prediction of RECO - GPP. This approach reduces potential sources of error by incorporating directly observed data such as NEE.

Our multitask learning approach uses a single LSTM layer to simultaneously predict all sites' target variables (GPP, RECO, SWC, and ET). This design mitigates equifinality issues and clarifies the complex relationships between atmospheric forcings and ecological responses. We also evaluate the model's performance using R^2, RMSE, MAE, NSE, and KGE metrics for each fold and site. All hyperparameters are manually tuned to optimise performance. The main downside of this model and architecture is the high computational cost in terms of time (2-3 hours) and memory (on our machine: 18 GB RAM and 16 cores workstation), which limits the inclusion of more data, especially when running on a CPU.

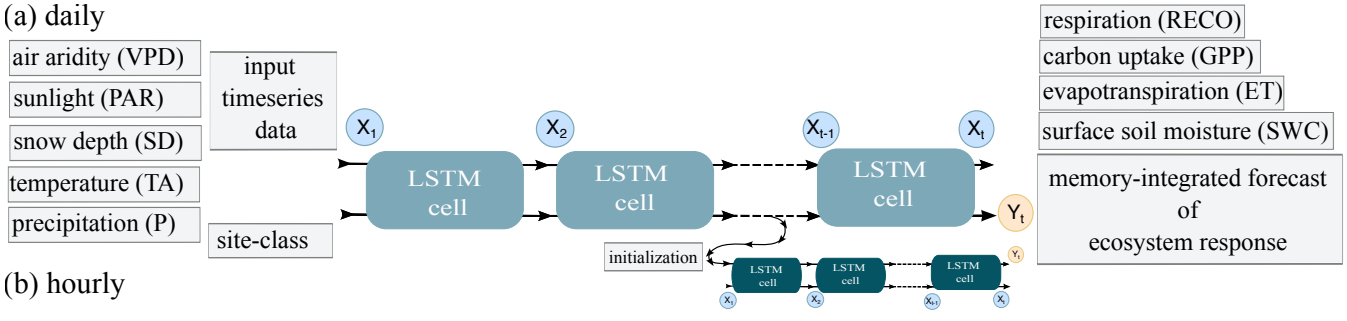

Figure 2. This figure displays the multitask multi-timescale LSTM architecture used in our EcoPro-LSTM$_{v0}$ to predict ecosystem responses. The model predicts hourly and daily values of RECO, GPP, ET, and SWC based on atmospheric and site-name inputs, including VPD, PAR, SD, TA, and precipitation. The model incorporates two LSTM layers—light blue cells handling daily time scale data and dark blue cells handling hourly data—followed by dropout mechanisms and ReLU activation for robustness against overfitting. The diagram highlights the initialization strategy for daily and hourly predictions to reflect atmospheric forcings' legacy and lag effects on ecosystem responses.

3.2 Interpretability and Transparency

After training, we further use interpretability to validate our findings through a feature importance analysis derived from the Integrated Gradients (IG) method against GradientExplainer, from SHapley Additive exPlanations (SHAP) method. These tools can be combined with a machine learning model to increase its interpretability and transparency. However, SHAP is often constrained by the assumption that the features are mutually independent leading to incorrect results when the correlation

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

between features exists (see, e.g., Hu et al., 2024). Given the feature correlation in our study, we simply compare SHAP and IG to determine a suitable baseline for the IG approach.

The SHAP framework, grounded in Shapley values from cooperative game theory (Aumann and Shapley, 2015), assigns an importance to each feature by considering all possible combinations of features. Mathematically, the GradientExplainer approximates SHAP using a weighted integral over feature inputs and their gradients, defined as:

$$\phi_i = \int_{\alpha=0}^{1} \frac{\partial f\big(h(\alpha)\big)}{\partial h_i} \cdot \big(h_i - h_i^{(0)}\big)\, d\alpha$$

where $\phi_i(f)$ is the SHAP for feature i, $h(\alpha)$ is the interpolated feature vector, and $f(S)$ is the model's prediction given the feature subset S and $h^{(0)}$ is the baseline feature vector. The GradientExplainer, a variant of SHAP, approximates these values by leveraging the gradients of the model's output to its input, making it particularly suitable for differentiable models such as neural networks (Lundberg, 2017). SHAP is robust for capturing complex feature interactions, but often considers mutual feature independence and includes unrealistic instances of data in its estimations when features are correlated (see, e.g., Aas et al., 2021), which makes the interpretation misleading. Moreover, this approach is computationally intensive because it needs to compute combinatory shuffling of the feature importance (Chen et al., 2018).

IG, proposed by Sundararajan et al. (2017), is also inspired by work from cooperative game theory Aumann and Shapley (2015). In this technique, feature attributions are computed by accumulating the gradients of the model's output its inputs along a straight path from a baseline (a point of neutral outcome) to the actual input. Mathematically, IG for feature i is defined as:

$$IG_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i}\, d\alpha$$

where x' is the baseline input, x is the actual input, and F is the model function. The choice of baseline as a point of neutrality can significantly affect the resulting attributions, and a poor choice can lead to misleading attributions, making it crucial to select a meaningful baseline for accurate explanations (Sturmfels et al., 2020). Moreover, as the path from baseline to the input value is a straight path, unnecessary noise may accumulate in the integrated gradient (see, e.g., Kapishnikov et al., 2021).

In summary, while SHAP (GradientExplainer) and IG provide valuable insights into model interpretability, they differ in handling feature dependencies and their computational cost. SHAP considers all possible subsets of features, making it robust for understanding feature interactions but less accurate in the presence of feature correlation. This method is computationally expensive, by comparison, as IG offers computationally efficient path-based attribution, excelling at local explanations. Still, it produces results comparable to SHAP only when the baseline is well-suited and noise during gradient integration is minimal. Hence, the SHAP approach served as a reference for determining the baseline in the IG technique (see Appendix 4.2 and Figure 5).

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

4 Outcomes

4.1 Model Performance

This article focuses on daily data, though our model also simulates hourly data (see Appendix C, Figure C1). Emphasizing daily data aligns with our research objectives of understanding day-to-day variations and their implications for ecological modelling. The inclusion of hourly data is intended to improve capturing the lag response at a daily resolution, which is critical for understanding ecological processes.

Figure 3 summarises the cross-fold averaged model performance in mean absolute error (MAE), root mean square error (RMSE), and weighted RMSE for each site, separated by training, validation, and test sets. Here, we display statistics for daily predictions of productivity, respiration, soil water content, and evapotranspiration, distinguished by set and site. The error is typically lower in training set than the validation or test sets. However, errors in all sets are generally similar for all variables at each site, indicating that the model is not overfitting and demonstrates consistent performance.

The overall model evaluation in terms of NSE, R^2 and KGE metrics for GPP, RECO, SWC, and ET on the test set are summarised in Table 1. The coefficient of determination (R^2) ranges from 0 to 1, with values near 1 indicating better fit, while a value of 0 suggests no predictive power, akin to a random model. The table includes additional metrics for a detailed evaluation. RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) reflect the degree of variation between prediction and observation. However, the disparity between RMSE and MAE may highlight the effect of outliers. With potential outlier presence (RMSE>MAE), it is recommended to use the coefficient of efficiency (Nash-Sutcliffe Efficiency, NSE) and other newer metrics. Note that R^2 may be inflated in these situations (see, e.g., Legates and McCabe Jr, 1999). The NSE (Nash-Sutcliffe Efficiency) can range from $-\infty$ to 1, with 1 representing a perfect fit and 0 indicating that the model performs similarly to the mean of the observed data (essentially a random model). The KGE (Kling-Gupta Efficiency) spans from $-\infty$ to 1, with 1 indicating a perfect model. NSE and KGE estimates larger than 0.5 generally indicate a good model accuracy, with one reflecting stronger predictive capability. It is important to note that due to noise in the data, achieving a perfect score of 1 for NSE, KGE, and R^2 is not realistic, even for an optimal model. In addition, we estimate NSE_β (measures the bias) and NSE_α (handles the variability), which provide a similar range to NSE. NSE and NSE_β estimates are approximately the same. It means that the model has low bias (accurately reflects the data mean). However, the NSE_α is about 10-20% lower than NSE, indicating that the model may under- or over-represent the variability of the observational data, suggesting an issue with capturing the full spread of data over time. Spectral residual and spectral divergence (range from zero to infinity) quantify the signal residual and divergence in the frequency domain, with a lower value indicating higher performance. Spectral residual and spectral divergence are zero and below 0.5, respectively, for all output variables, indicating the model's representation of periodic patterns.

We compare the model's estimations with observed data for four selected sites (Figure 4). The data presented is a mix of training, test, and validation sets averaged over all folds. The shaded area indicates the MC simulations. Notably, the EcoPro-$LSTM_{v0}$ captures seasonal patterns for all output variables (Figure 4). These sites were chosen to demonstrate our approach's robustness under various environmental conditions and at different levels of biological outputs. In panel (a), productivity is

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Figure 3. Model performance error metrics for productivity (panel 1), respiration (panel 2), soil water content (panel 3), and evapotranspiration (panel 4): Averaged key metrics (MAE, RMSE, weighted) for all cross-validation folds, categorized by site and set.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Table 1. Model performance metrics for all sites: Averaged key metrics over test sets in all cross-validation folds.

	R^2	MAE	RMSE	weighted RMSE	KGE	NSE	NSE_β	NSE_α	spectral divergence	spectral residual
GPP	0.61	1.54	2.22	2.69	0.57	0.57	0.52	0.46	0.28	0
RECO	0.45	1.34	1.79	2.22	0.43	0.35	0.37	0.21	0.37	0
ET	0.64	0.49	0.68	0.85	0.62	0.59	0.60	0.49	0.25	0
SWC	0.71	0.05	0.06	0.07	0.77	0.70	0.71	0.59	0.07	0

depicted, followed by respiration in (b), soil water content in (c), and evapotranspiration in (d). The coloured lines represent the observations, the grey patch displays the MC simulations, and the solid black line demonstrates the median of the MC simulations. The annual peaks align closely with observations, which fall within the range of uncertainty.

Panel (1) shows ES-LJu, an open shrubland (OSH) characterised by the highest variability in mean annual NEE, which ranges from -63 to 29 $gC \cdot m^{-2}$, with evapotranspiration strongly linked to both respiration and productivity (Serrano-Ortiz et al., 2009). The site features sparse and diverse vegetation, with approximately 50% of the land cover being bare soil. ES-LJu has the lowest productivity and highest interannual variability, yet our model reproduces year-to-year carbon flux fluctuations (panels a and b). For soil water content, in panel (c), the model performed particularly well in 2006, even though the observations recorded an inconsistent signal. Additionally, similar to the carbon fluxes, evapotranspiration at this site also exhibits high interannual variability, as expected in such an ecosystem. The R^2, KGE, and NSE for this site for productivity is 0.44, 0.12, and 0.39, in respect.

On panel (2), US-Ton is illustrated, a mix of Blue Oak trees, pines and grasses typical of a savanna ecosystem. The vegetation is relatively homogeneous, dominated by Blue Oaks with an understory of grasses with about 65% of tree land cover. The soil is silt loam to rocky silt loam, with a good capacity for retaining moisture. The high interannual variation in precipitation in this site doesn't translate to particularly significant evaporation and carbon fluxes interannual variability. Though ET and GPP depend on seasonal rainfall, interannual changes remain relatively stable due to the ecosystem's resilience and access to deeper water sources (Baldocchi et al., 2021). Even though in this site, soil moisture variations are expected to be less correlated with GPP, RECO and ET, all variables are modelled equally well.

IT-Ro1 (panel (3)) is a deciduous coppice oak forest in Central Italy, where soil respiration is notably high for this forest type (Rey et al., 2002). Each year exhibits two productivity peaks (panel a), which the model reproduces, except for 2004-2005. For soil moisture (panel c), the model is accurate, except during the 2004-2005 dry periods, and in 2006 despite inconsistent data. In panel (d), depicting evapotranspiration, the model excels, even in 2007, where the observed data seems erroneous.

In panel (4), we examine a pine coniferous forest with diverse vegetation in the coastal area of San Rossore, Central Italy, characterised by sandy soil. The GPP remains relatively stable, with no significant interannual variations, though carbon fluxes increase as the forest matures each year (Chiesi et al., 2005). Our model generally captures the peaks well for carbon fluxes and evapotranspiration (panels a, b, and d) from 2001 to 2008. The uncertainty band tends to overestimate these peaks. From 2007 onward, the uncertainty band consistently falls below the peaks, and by 2008, it underestimates the evapotranspiration

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

and respiration signals. In panel (c), despite the suboptimal quality of the SWC data, our model consistently predicts a coherent signal.

For GPP, US-Ton exhibits the strongest performance, with the highest NSE (0.73) and NSE_β (0.59), along with a moderate NSE_α (0.69). It also has a relatively low RMSE (1.04) and quantile loss (0.3 at the 85th quantile), indicating better overall accuracy. ES-LJu has the lowest NSE (0.21) and NSE_β (0.08), suggesting poor predictive skill. Still, the estimations maintain a relatively low RMSE (0.57) and quantile loss (0.12 at the 85th quantile), indicating some reliability in capturing parts of the distribution. IT-SRo struggles the most with capturing the extremes suggested by the highest RMSE (2.04) and quantile loss (0.63 at the 85th quantile). However, it achieves a moderate NSE (0.56) and NSE_β and NSE_α (~ 0.5), indicating that the model captures the mean and variability reasonably well despite the higher errors. IT-Ro1 has high errors, with RMSE (1.77) and quantile loss (0.54 at the 85th quantile), but its NSE (0.77) and NSE_β and NSE_α (~ 0.65) points to good performance in capturing the variability of the data, similar to US-Ton. Overall, US-Ton stands out as the best-performing site, while IT-SRo requires improvement in error reduction.

There is notable variation in the model's performance metrics and goodness-of-fit indicators for respiration. US-Ton demonstrates moderate performance with a relatively low R^2 (0.32), indicating a weak fit, and an NSE (0.21) slightly above zero, suggesting marginal predictive skill. Yet, it has a large RMSE (1.68) and negative KGE (-0.15), reflecting poor model accuracy. Our model exhibits the weakest performance in ES-LJu with negative NSE (-0.55) and KGE (-0.34), meaning the model performs worse than the mean of the data. The spectral divergence (0.68) and R^2 (0.17) are also the largest for this site. Model fitted IT-SRo better than ES-LJu, with R^2 (0.44) and NSE (0.39), but errors are still large, such as RMSE (2.05) and substantial quantile loss (0.59 at the 85th quantile), indicating challenges in reproducing high respiration. IT-Ro1 presents relatively better performance with the highest R^2 (0.49), a moderate NSE (0.43), and better error metrics than the other sites, though the quantile loss is still substantial at 0.47 (the 85th quantile). Overall, IT-Ro1 exhibits the best model performance among the sites, while ES-LJu struggles the most in predicting respiration accurately.

The model performance in estimating evapotranspiration for the selected sites varies according to statistics. US-Ton has the strongest performance with the greatest R^2 (0.86), indicating a strong fit between the model and observations, and a high NSE (0.84) and KGE (0.74), along with low quantile loss (0.08 at the 85th quantile). ES-LJu also performs relatively well, with a moderate R^2 (0.64), NSE (0.62), and KGE (0.49), though the spectral divergence (0.21) suggests some model limitations in capturing spectral patterns. A balanced performance in IT-Ro1 with a good R^2 (0.74) and NSE (0.71) with slightly higher quantile loss (0.13 at the 85th quantile) than US-Ton, suggests the model's robust performance with minor variations. In contrast, our model struggles the most in IT-SRo with a low R^2 (0.44), negative KGE (-0.17), and NSE (0.35), as well as the largest RMSE (0.84) and spectral divergence (0.6), indicating a poorer match between predictions and observations.

The model estimates SWC accurately. In US-Ton, the greatest NSE (0.95), NSE_α (0.92), and NSE_β (0.95), indicate model's ability to capture variability and minimize bias. RMSE (0.03) and quantile loss (0.01 at the 85th quantile) are minimal. Estimations in ES-LJu are also good, with NSE (0.8) and NSE_α (0.74). Still, the errors are slightly higher than in US-Ton, with RMSE (0.04) and quantile loss (0.02 at the 85th quantile). In IT-SRo, we have the lowest NSE (0.45), NSE_β (0.44), and higher

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

spectral divergence (0.19), partially due to the low quality of recorded SWC data. Moderate performance is also recorded in IT-Ro1, with NSE (0.62) and NSE_{α} (0.64).

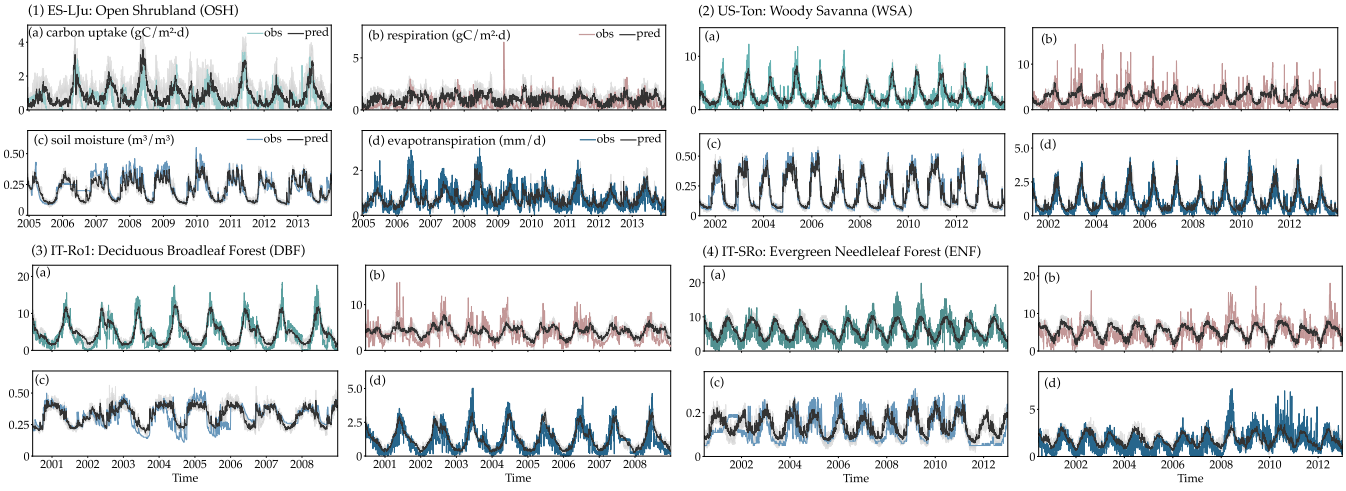

Figure 4. The median of model estimates (train, test, validation sets) in solid black compared with observed data for GPP, RECO, SWC, and ET at different sites, demonstrating performance at seasonal scales. The shaded area indicates the MC simulations. The sites represent different PFTs with varying productivity levels (250 to 2200 $gC/m^2.y$), highlighting the model's robustness in diverse ecological settings.

Table 2 summarises for one of the folds, Heidke Skill Score (HSS) at different sites for various labels (ET, GPP, RECO, SWC) at three quantile thresholds: 0.5, 0.65, and 0.85 quantiles. In general, sites with higher NSE tend to have higher HSS, indicating better model performance(See Table B1 for values in all sites).

The model's varying performance at thresholds (0.5, 0.65, 0.85) suggests accuracy is influenced by data quality and site-specific factors, not the thresholds. Table B1 illustrates that while the model captures some quantiles better, these variations do not indicate a bias toward any specific threshold. These variations could be attributed to data quality or underlying variability in the modelled environmental processes. Just as US-Ton, which has high data quality, we observe comparable performance in terms of HSS at all thresholds. Inconsistent data quality might explain these differences in performance (further statistics and data detailed in Appendix B4). Overall, SWC and ET showed strong correlations, while GPP and RECO were challenging to predict. Sites with no vegetation loss or growth and stable soil moisture (such as US-Ton and IT-Ro1) performed the best.

4.2 Optimizing Baseline Selection for Integrated Gradients: A Comparative Analysis with SHAP (GradientExplainer)

We assess the feature importance derived from these two interpretability approaches by comparing SHAP and IG techniques. SHAP provide a unified measure of feature contribution by assessing the contribution of each feature for all input combinations, offering a theoretically sound approach. In contrast, the IG attribute feature importance based on the gradient of the output to inputs, integrated along a path from a baseline to the actual input. Both techniques can highlight different aspects of feature

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Table 2. The Heidke Skill Score (HSS) for selected sites and variables (ET, GPP, RECO, SWC) at quantile thresholds of 0.5, 0.65, and 0.85 for a single fold. Performance varies by threshold, likely due to data quality and site-specific conditions.

site name	SWC			GPP			RECO			ET		
quantiles	**0.5**	**0.65**	**0.85**	**0.5**	**0.65**	**0.85**	**0.5**	**0.65**	**0.85**	**0.5**	**0.65**	**0.85**
ES-LJu	0.63	0.60	-0.02	-0.14	0.08	0.58	0.09	0.11	-0.00	0.65	0.64	0.61
IT-Ro1	0.71	0.64	0.39	0.43	0.47	0.71	0.28	0.33	0.06	0.70	0.78	0.65
IT-SRo	0.39	-0.09	0.35	0.54	0.46	0.31	0.24	0.21	0.04	0.23	0.34	0.36
US-Ton	0.87	0.89	0.42	0.61	0.65	0.76	0.17	0.34	0.34	0.63	0.84	0.75

importance. At the same time, SHAP is robust in capturing feature interactions and global importance; IG may provide detailed insights into individual predictions in deep learning models. However, the differences between the two models should be minimal and a matter of convenience Feng et al. (2022) and given IG technique's sensitivity to baseline definition, SHAP technique was used to select the appropriate baseline. Feng et al. (2022); Liu et al. (2023a). After testing multiple baselines (zero, mean, random, median), the closest match is achieved by setting the reference to the site-level median (Figure 5).

4.3 Understanding Climatic Drivers with Explainable Artificial Intelligence: xAI

We now explore the interpretability of the AI models in identifying and understanding the climatic drivers affecting our ecological systems of interest using integrated method with site median serving as a baseline. The results have been averaged over all k-folds to reduce noise. Given the similar performance in each fold, this averaging approach provides a robust depiction of feature contributions without significant loss of detail. In Figure 6, there are six rows on each panel a-d; the first five rows belong to the meteorological data used to predict the observation point on the target date. The target date is confined by the dashed line on the six rows (see the row for GPP). In these five rows, the grey line represents the meteorological records, with the colour bar illustrating each point's contribution to the GPP estimate on the target date (highlighted by the dashed line in the final row). We seperated the lag (short-term) impacts occurring during 1 week before the target date indicated by $\Sigma_1^8 \overline{IG_i}^X$ from legacy (long-term) impacts occurring during 3 months before the target date up to 1 week before indicated by $\Sigma_8^{120} \overline{IG_i}^X$ where X is the meteorological time series (such as precipitation, P) used to predict the desired model output (e.g., GPP). In the last row, the red dashed line indicates the median of estimations, and the green line is for the observations.

For each estimation, the model produces a time series of explanations, depicting how various environmental factors influenced its outputs at distinct moments in the past. The illustration in Figure 6 details the environmental parallels and contrasts on a chosen date between three sites in the United States: US-Ton and US-Var (closely located sites) and US-ARM and and one in Italy IT-Cpz for the same date (4th of May 2007). We observe similar environmental forcing on GPP for closely located sites, US-Ton and US-Var, and different impacts on the other sites. In each time series, we can focus on a specific interval and assess each input's effect on productivity(target date is May 4th 2007). In this way, we can distinguish between short-term impacts

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

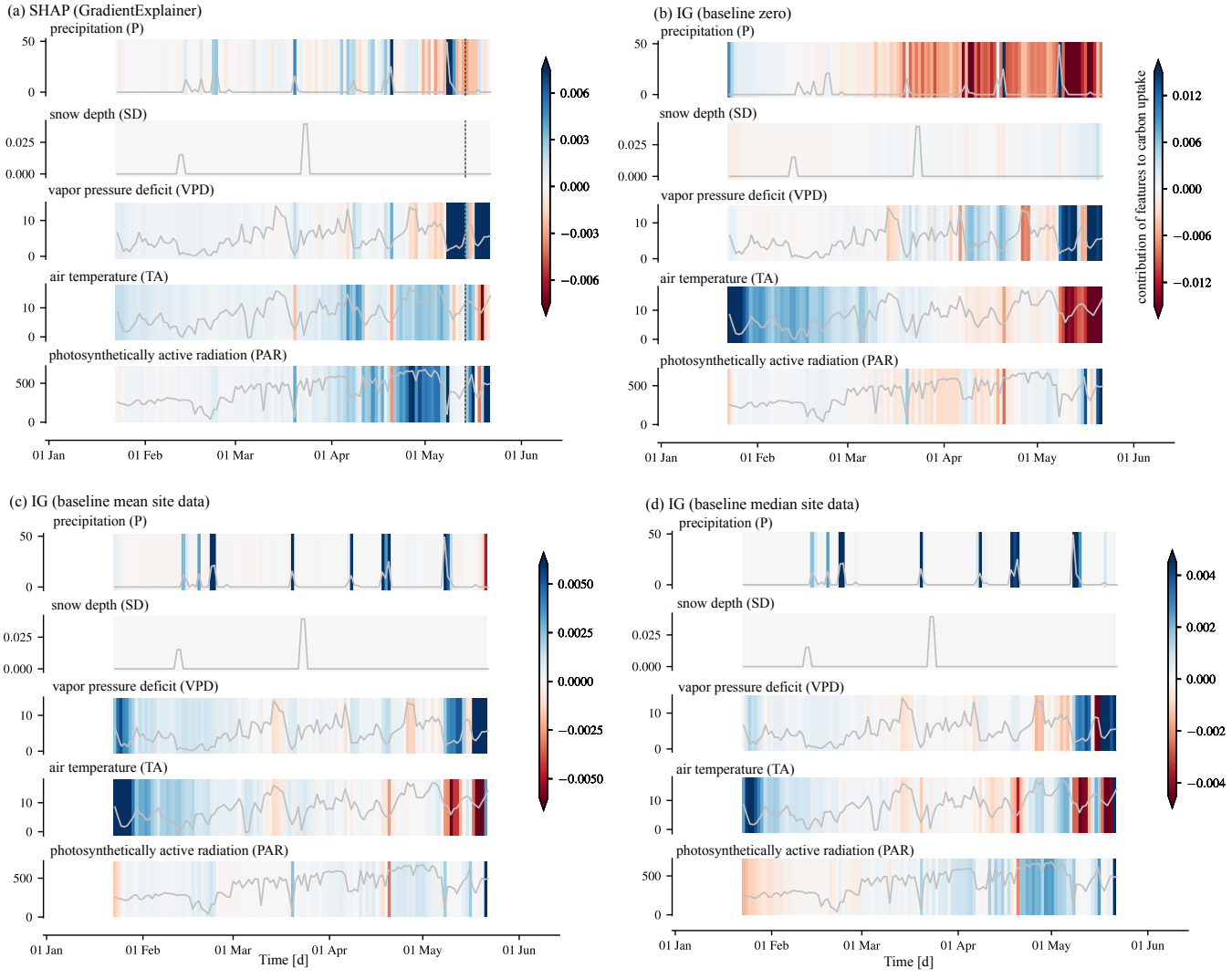

Figure 5. Feature importance in estimating gross primary production on May 25, 2008, at site ES-LJu. The feature importance is averaged over all k-folds (k=5). The baseline in the IG method, panels (b)-(d), is tuned to resemble feature importance estimates by SHAP(GradientExplainer) in panel (a). The grey line illustrate measured features in time (y-axis units are mm/d for P and SD, kPa for VPD, °C for TA and W/m^2 for PAR). The colour bar shows the feature importance with colour bar on the right hand side of each figure presenting its magnitude. When the localised median serves as the baseline in IG, we achieve the closest match to SHAP's feature importance estimates.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

(occurring over the week before), named lagged effects in this study, versus the impacts accumulated from the previous events, named legacy effects in this study.

In agreement with Marshall et al. (2018), we find that PAR (inversely proportional to cloudiness) has a short-term impact often closely related to productivity. In contrast, precipitation often has a lasting effect carried over time.

Previous precipitation events play a key role in productivity ($\Sigma_8^{120}\overline{IG}_i^P$ is 0.51 in panel a, 0.62 in panel b, 0.87 in panel c, and 0.13 in panel d). At the same time, no significant influence from snow depth is observed (shown on the second row with $\Sigma_1^8\overline{IG}_i^{SD}$ and $\Sigma_8^{120}\overline{IG}_i^{SD}$ equal to zero). The impact of each antecedent rainfall event on productivity varies in strength. Some events are marked in darker blue, indicating greater significance, while others appear lighter. Heavier rains tend to be more consequential, though the relationship is highly non-linear and transient. For example, two early April rain events contribute as much to carbon uptake on May 4th as the February event.

In Figure 6a-b, the "vapour pressure deficit" panel illustrates a significant negative effect from VPD ($\Sigma_1^8\overline{IG}_i^{VPD}$=-0.07) on productivity over a brief period (about one week) before May 4th due to increased air aridity. Prior to this, from 120 days to 8 days before the target date, VPD remained at favourable levels, positively influencing productivity ($\Sigma_8^{120}\overline{IG}_i^{VPD}$=0.29). In Figure 6c, a similar pattern is observed. Rainfall shortly before May 4th sharply reduces VPD at the US-ARM site, leading to a positive lag effect ($\Sigma_1^8\overline{IG}_i^{VPD}$=0.09) as VPD quickly returns to favourable levels. Figure 6d focuses on an evergreen broad-leaf forest, where precipitation is no longer the primary driver of productivity ($\Sigma_1^{120}\overline{IG}_i^P$=0.02+0.01). Instead, PAR ($\Sigma_1^{120}\overline{IG}_i^{PAR}$=0.07+0.06) and TA ($\Sigma_1^{120}\overline{IG}_i^{TA}$=0.08+0.08) have a stronger, beneficial influence over both short and long timescales in driving carbon uptake. The memory effects are notably transient, sensitive to noise, yet exhibit a general temporal consistency.

Although high VPD and temperature are often correlated or confounded, their relationship is more complex than it may initially appear. Let's examine the lagged (short-term) impacts from VPD ($\Sigma_1^8\overline{IG}_i^{VPD}$) and temperature ($\Sigma_1^8\overline{IG}_i^{TA}$) in panels a-d. Temperature and VPD both influence GPP in the same direction, showing simultaneous rises and falls. Still, while a rise in VPD directly leads to a decline in GPP, it does not coincide with an immediate adverse effect from air temperature (note the blue peak for air temperature and the red peak for VPD). The legacy effects of VPD ($\Sigma_8^{120}\overline{IG}_i^{VPD}$) and air temperature ($\Sigma_8^{120}\overline{IG}_i^{TA}$) show greater correlation, as both remain favourably low during much of this period and exert a comparable contribution to GPP.

While photosynthetically active radiation (a sunlight proxy inversely associated with cloudiness) usually exhibit reduced sunlight negatively impacting GPP, occasional deviations suggest noise in the IG method or nuanced underlying mechanisms. Consistently, when examining the short-term effects of PAR ($\Sigma_1^8\overline{IG}_i^{PAR}$), the direction of changes in PAR aligns with that of GPP (an increase in PAR enhances GPP, while a decrease suppresses it).

Figures 7 and 8 now illustrate SWC and ET instead of GPP. Each panel (a-d) shows meteorological inputs (first five rows) and their contributions to SWC and ET observed on the target date (4th of May on the last row). Antecedent conditions from the prior week ($\Sigma_1^8\overline{IG}_i$) are separated from long-term memory effects spanning three months up to one week before ($\Sigma_8^{120}\overline{IG}_i$).

When comparing Figure 6 and Figure 7 (environmental outcome on productivity versus soil moisture), the consequences of meteorological inputs like precipitation, snow depth and VPD remain consistent. Still, for PAR (indicating sunlight and lack of

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

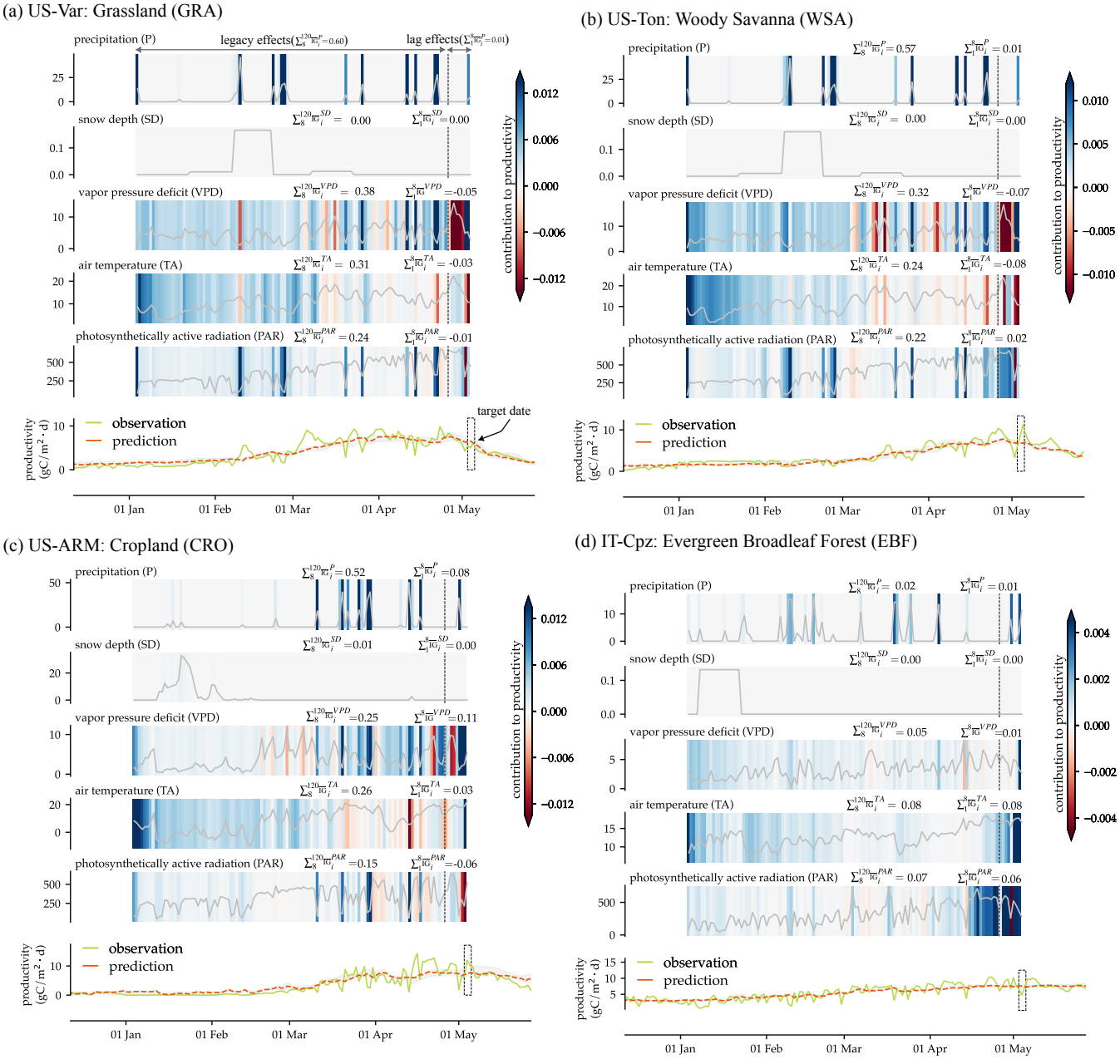

Figure 6. Assessing feature contributions to GPP on May 4th, 2007, in four sites: US-Var (GRA), US-Ton (WSA), US-ARM (CRO), and IT-Cpz (EBF). Each subplot presents the contributions of environmental factors to the model's estimations on May 4th, 2007, highlighting the commonalities and divergences in in the role of atmospheric drivers on carbon exchange. The results point to comparable environmental outcomes at nearby sites (US-Var and US-Ton) and varied outcomes at distant sites (US-ARM and IT-Cpz). The short- and long-term effects of these factors highlight the model's ability to offer detailed insights into site-specific climatic influences. P and SD units are mm/d, VPD is in kPa, TA unit is °C, and PAR in W/m^2.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

cloud cover), the direction of impact on SWC and GPP is reversed, as expected. Sunlight promotes ecosystem water use, which negatively affects SWC while benefiting GPP. Temperature also, at times, contributes differently in terms of direction (negative or positive) to GPP (shown in Figure 6) and SWC (Figure 7). Still, it remains unclear if these differences reflect a consistent signal or noise. Additionally, when comparing Figures 7a and b, soil moisture levels 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 soil moisture is explained by the enhanced contribution from precipitation, temperature, and PAR on soil moisture. When comparing 8 against 6 (meteorological control on ET versus GPP). We observe that environmental conditions affect ET in the same direction and order as GPP, indicating stable water use and unstressed conditions. This comparative analysis underscores the capability of interpretable AI to provide detailed insights into the ecological outcomes of climate variability on the different model outputs.

5 Discussions

5.1 Improved Interannual Variability in Prediction of Carbon Fluxes

We benchmark our model against established frameworks, including FLUXCOM (Jung et al., 2020) and FLUXCOM-X (X-base) (Nelson et al., 2024), with a focus on daily GPP estimates, as shown in Figures 9 and 10.

Figure 9a shows our EcoPro-LSTM$_{v0}$ robust performance, with higher R^2 at multiple sites, indicating strong correlation between observed and predicted data. For example, EcoPro-LSTM$_{v0}$ achieves R^2 above 0.6 for most sites, with a maximum of 0.89. In contrast, FLUXCOM-X reaches up to 0.78, while FLUXCOM often falls below 0.5, indicating challenges in modelling local variations in semi-arid ecosystems.

The Kling-Gupta Efficiency (KGE), shown in Figure 9b, further underscores variability in model performance. EcoPro-LSTM$_{v0}$ renders superior results in most locations, with positive KGE often around 0.6 to 0.8 and a maximum of 0.79, indicating a more balanced representation of both variability and bias. Conversely, FLUXCOM shows some negative KGE, with a minimum of -0.75, suggesting challenges in reliably modelling seasonal fluctuations. FLUXCOM-X outperforms FLUXCOM with a KGE up to 0.78 but lacks the precision of EcoPro-LSTM$_{v0}$ predictions.

Nash-Sutcliffe Efficiency (NSE), Figure 9c, also indicate that EcoPro-LSTM$_{v0}$ performs generally better, achieving positive NSE scores of up to approximately 0.8 for several sites. FLUXCOM, in comparison, usually delivers NSE ranging from 0.3 to 0.6, whereas FLUXCOM-X often demonstrates negative NSE, dropping to -1.1, reflecting suboptimal model reliability. Lower NSE scores point to challenges in replicating field data while showcasing Multi-Timescale LSTM framework's adaptive memory in capturing transient dynamics vital for semi-arid and Mediterranean systems. The model reliably captures inter-annual variability and extremes (Figure 10). This enhanced performance is achieved through three fundamental mechanisms: (1) LSTM's inherent ability to account for time dependencies, (2) the use of a weighted Root Mean Square Error (weighted RMSE) that addresses potential biases in the data distribution, and (3) the integration of multiple timescales, each encompassing distinct processes and information connected to certain temporal trends. The multiple timescales included in our model structure contains different information; and thus better results (see Appendix A2). Improved efficiency through integrating

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

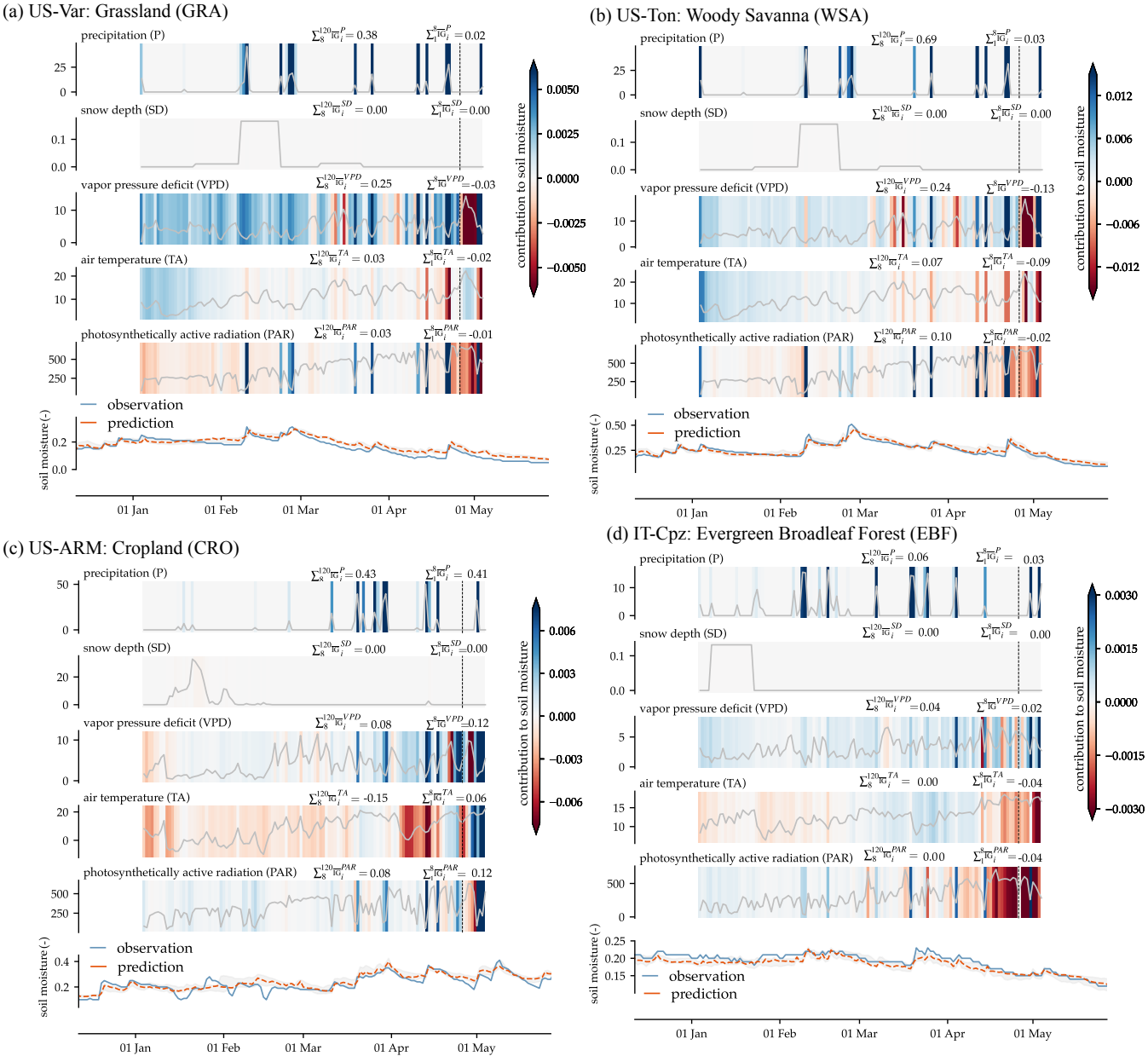

Figure 7. Impact of climatic drivers on soil water content on May 4th, 2007 in four sites: US-Var, US-Ton, US-ARM, and IT-Cpz. Each subplot shows how environmental factors influence soil moisture recorded on May 4th, 2007. The higher soil water content at US-Ton compared to US-Var is attributed to precipitation, PAR, and TA, as groundwater data is unavailable. P and SD units are mm/d, VPD is in kPa, TA unit is °C, and PAR in W/m^2.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

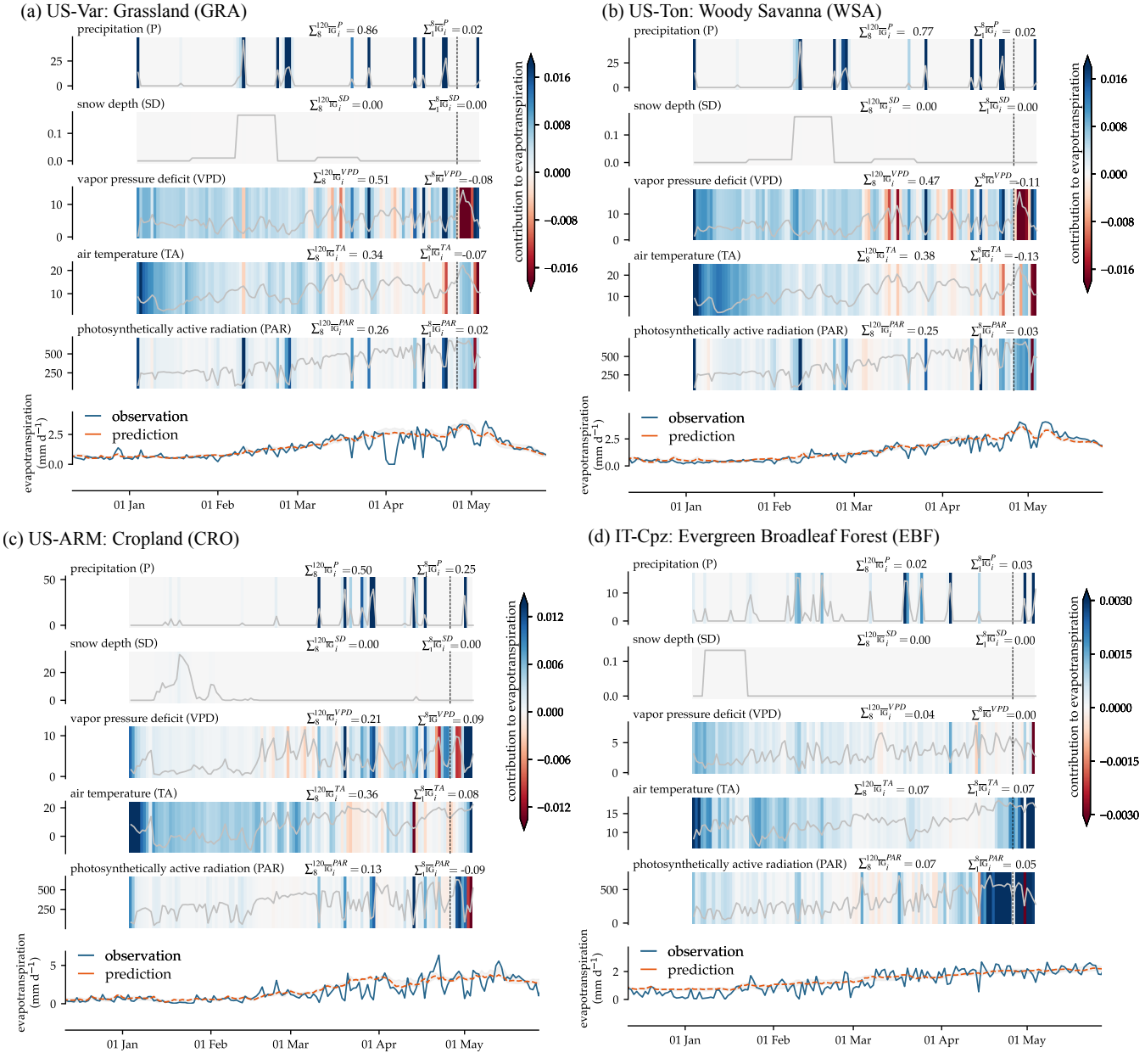

Figure 8. Feature importance analysis for ET on May 4th, 2007, in sites: US-Var, US-Ton, US-ARM, and IT-Cpz. Each subplot presents the contributions of environmental factors to the model's estimations on May 4th, 2007, highlighting the similarities and differences in the impact of climatic drivers on evapotranspiration. P and SD units are mm/d, VPD is in kPa, TA unit is °C, and PAR in W/m^2.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

data from multiple temporal dimensions aligns with conclusions in recent studies employing convolutional GRU frameworks for imagery data processing (Nguyen et al., 2024).

Figure 9 presents a comparison against the Advanced FLUXCOM versions for four sites in terms of NSE. When comparing the performance of IT-SRo, IT-Ro1, ES-LJu, and US-Ton for FLUXCOM-X, FLUXCOM against EcoPro-LSTM$_{v0}$, our model consistently outperforms in key metrics like KGE, R^2, and NSE. At ES-LJu, FLUXCOM-X and FLUXCOM display negative KGEs of -0.82 and -1.2, respectively, while our model achieves a KGE of 0.1, an R^2 of 0.44 and NSE of 0.16, suggesting a clear advantage. The observed data quality at ES-LJu appears poorer than other sites, likely due to minimal productivity—the least among all sites—and sparse vegetation (open shrubland and bare soil), which may contribute to greater noise in the measurements, see Figure 10(1). At US-Ton, NSE score of our model is 0.73, larger than both FLUXCOM-X's 0.56 and FLUXCOM's 0.52, indicating better error reduction. EcoPro-LSTM$_{v0}$ predictions lead with an R^2 of 0.75, outperforming FLUXCOM-X (0.6) and FLUXCOM (0.56) in explaining variance. However, its KGE is slightly lower at 0.64 compared to FLUXCOM-X (0.78) and FLUXCOM (0.7). This higher KGE for FLUXCOM-X may be due to its better capture of peak magnitude, although it does not match EcoPro-LSTM$_{v0}$'s accuracy in timing the peak, see Figure 10(2).

This discrepancy suggests that while FLUXCOM-X achieves a balanced fit in magnitude, EcoPro-LSTM$_{v0}$'s strength lies in capturing the precise seasonal timing of the peak, contributing to its higher NSE and R^2. In IT-Ro1, EcoPro-LSTM$_{v0}$ forcasts' KGE (0.6) and R^2 (0.79) scores are better than FLUXCOM-X (KGE is 0.43 and R^2 is 0.23) and FLUXCOM (a KGE of 0.28 and R^2 of 0.55). Additionally, EcoPro-LSTM$_{v0}$ predisctions' NSE score, in IT-Ro1, is 0.77, larger than NSE for FLUXCOM-X (0.07) and FLUXCOM (0.35). In IT-SRo, EcoPro-LSTM$_{v0}$ results outperform FLUXCOM-X by achieving a KGE of 0.46, R^2 of 0.6, and NSE of 0.58. This superior performance at IT-Ro1 and IT-SRo is also visually visible in Figure 10(3-4).

Although our model demonstrates acceptable performance statistics, it underperforms at the AU-Rig site (Figure 9). We believe that the lower performance metrics at this site are due to suboptimal observation data quality and may not necessarily indicate an underperforming model; further details in Appendix D.

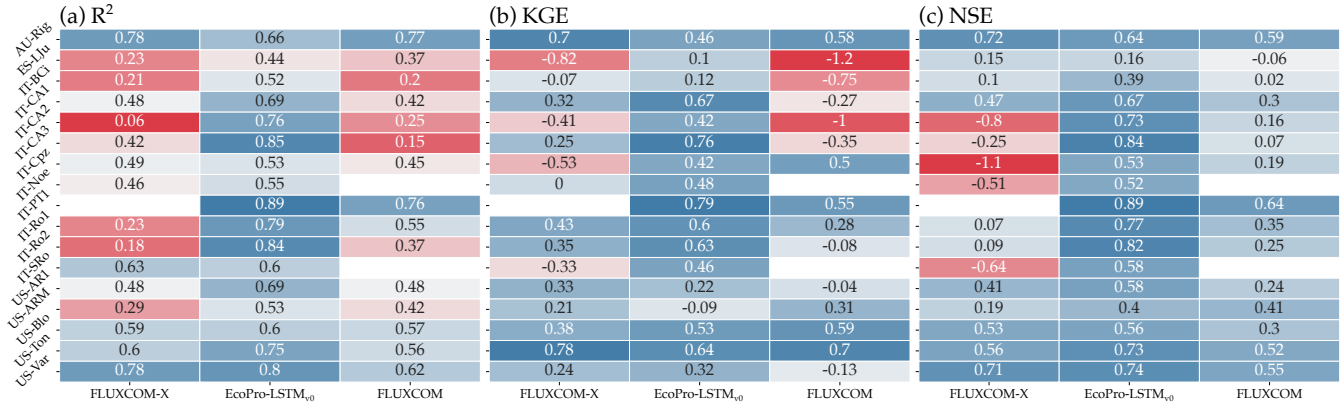

Figure 9. Evaluation of our EcoPro-LSTM$_{v0}$ against benchmark datasets using R^2, KGE, and NSE metrics on a per-site basis. The reduced performance metrics of EcoPro-LSTM$_{v0}$ at certain sites do not indicate overall inferiority to the benchmark datasets.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

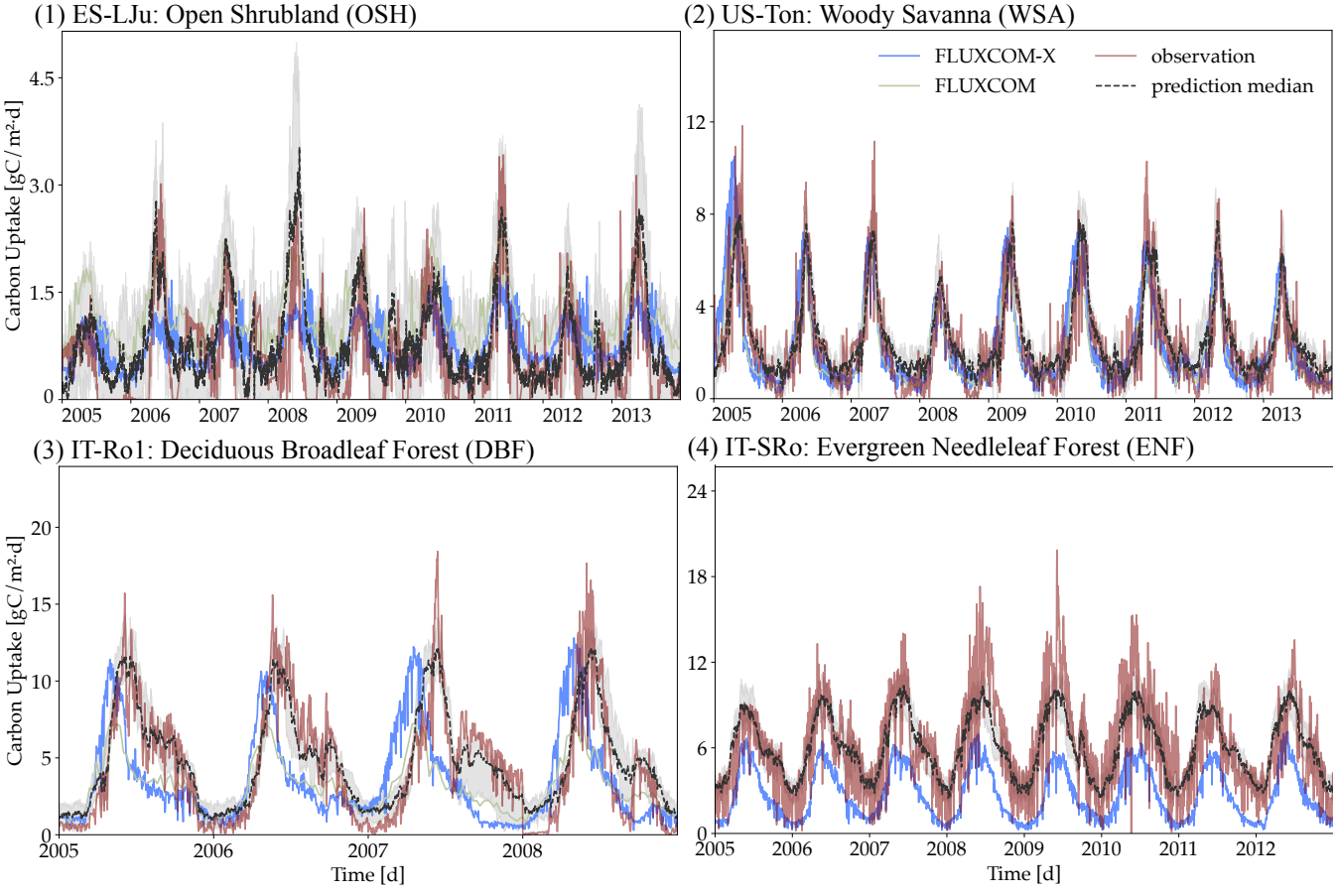

Figure 10. EcoPro-LSTM$_{v0}$ predictions benchmarked against the widely recognized FLUXCOM and FLUXCOM-X datasets. The grey area represents 50 Monte Carlo simulations, with the solid line showing the median.

5.2 Interpretablity: Single- vs Multi- Task Learning

In this study, we evaluate feature importance explanations derived from multitask learning (MTL) and single-task learning (STL) models, with both approaches employing the Multi-Timescale LSTM (MT-LSTM) architecture used in our EcoPro-LSTM$_{v0}$. The MTL model simultaneously predicts productivity with other key environmental variables (soil moisture and evapotranspiration). In contrast, the STL model is limited to productivity and respiration alone (comparing productivity and respiration with net ecosystem exchange using a rolling window for productivity error, similar to the principal framework). We consider this a single target since net ecosystem exchange is the sole directly sampled data at tower sites. To ensure comparability between the models, we use the same experimental setup as described in Section 3.1 and set the batch sizes to 64.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Our analysis using integrated gradients shows that feature importance magnitudes are similar between the two models for most input variables. Differences arise in driver attributions, particularly precipitation (see Figure 11). The MTL model assigns greater importance to certain rain events than the STL model.

We attribute these differences to the additional predictive tasks in the MTL framework, which includes variables such as soil moisture—indicating precipitation infiltration—and evapotranspiration, which reflects plant and soil responses to atmospheric conditions. These factors provide a holistic understanding of rainfall dynamics through the soil-plant-atmosphere continuum; the MTL model refines the nuanced interplay between precipitation pulses and productivity (see, e.g., Giorgi and Lionello, 2008), likely mitigates the inherent challenges in measuring and interpreting precipitation accurately.

5.3 Spatiotemporal Variation of Climate Forcing

To understand the spatiotemporal influence of environmental factors on carbon uptake in Mediterranean sites, we analysed 3 to 15 data points per month from the 75th to 99th quantiles of monthly data. By aggregating feature importance over months, years, and k-folds, as demonstrated in Figure 6, we can understand the principal climatic drivers that both positively and negatively influence carbon uptake (see Figure 12). Our results reveal a comparable pattern of environmental drivers across all sites, with precipitation playing a critical role, especially at the early and late phases of the productive season. We observed a lack of a direct relationship between snow depth and carbon uptake, suggesting the potential involvement of other intermediary factors. Temperature and photosynthetically active radiation predominantly regulate peak annual carbon uptake, though they can sometimes exert adverse influences during the early stages of the growing season. The mounting adverse impacts of air aridity (indicated by VPD) toward the end of the growing season further illustrate the complex interactions among these climatic factors, shaping ecological productivity in Mediterranean regions. Like other previous studies in Mediterranian sites (Wang et al., 2016; Markos et al., 2024), we observe that carbon uptake is primarily driven by precipitation, especially during the transitional phases of productivity, and noted key outcomes of temperature (both positive and negative). At the same time, contrary to previous studies, we find the adverse consequences of VPD on carbon uptake are limited to the late stages of the growth period, with occasional favourable outcomes during its early stages (often remaining minimal). Temperature, coupled with photosynthetically active radiation, positively supports carbon uptake during its peak phase.

5.4 Next Steps

The development of our data-driven MLT MT-LSTM framework within EcoPro-LSTM$_{v0}$ model represents a leap forward in predicting the outcomes of climatic variability at semi-arid sites, considering the multiscale and multivariable nature of the coupled water and carbon processes. Leveraging the capabilities of this memory-enhanced model, we are poised to explore a broader spectrum of environments beyond the well-studied Mediterranean landscapes. An intriguing question arises concerning the nature of episodic rainfall events (often termed rain pulse) and the role of pulses intensity and frequency on ecosystem functions (see, e.g., Seager et al., 2019). Yet, unpacking these dynamics poses several challenges. The transient and short-term (few hours) nature of rainfall events makes them inherently difficult to characterise with precision, and the variability in different semi-arid regions introduces additional complexity (see, e.g., Haverd et al., 2017; Arca et al., 2021; Hao et al., 2020; Bachman

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

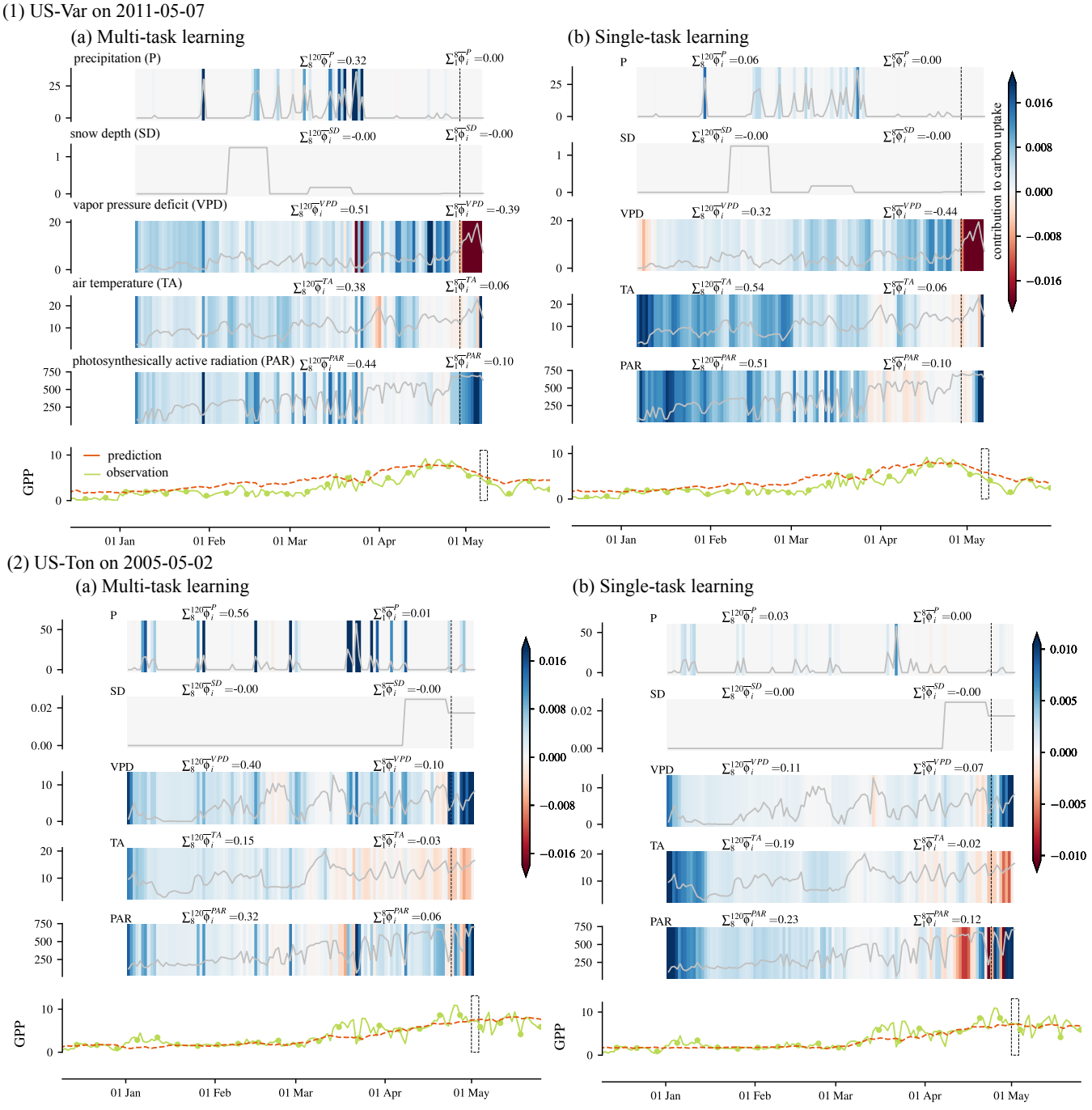

Figure 11. Comparison of explanatory insights into variable contributions between multitask learning (MTL) and single-task learning (STL) models utilising the MT-LSTM framework. The MTL model incorporates productivity with other key environmental variables (soil moisture and evapotranspiration), whereas the STL model focuses on productivity and respiration alone. IG analysis reveals that the MTL model attributes more weight to rain events than the STL model.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

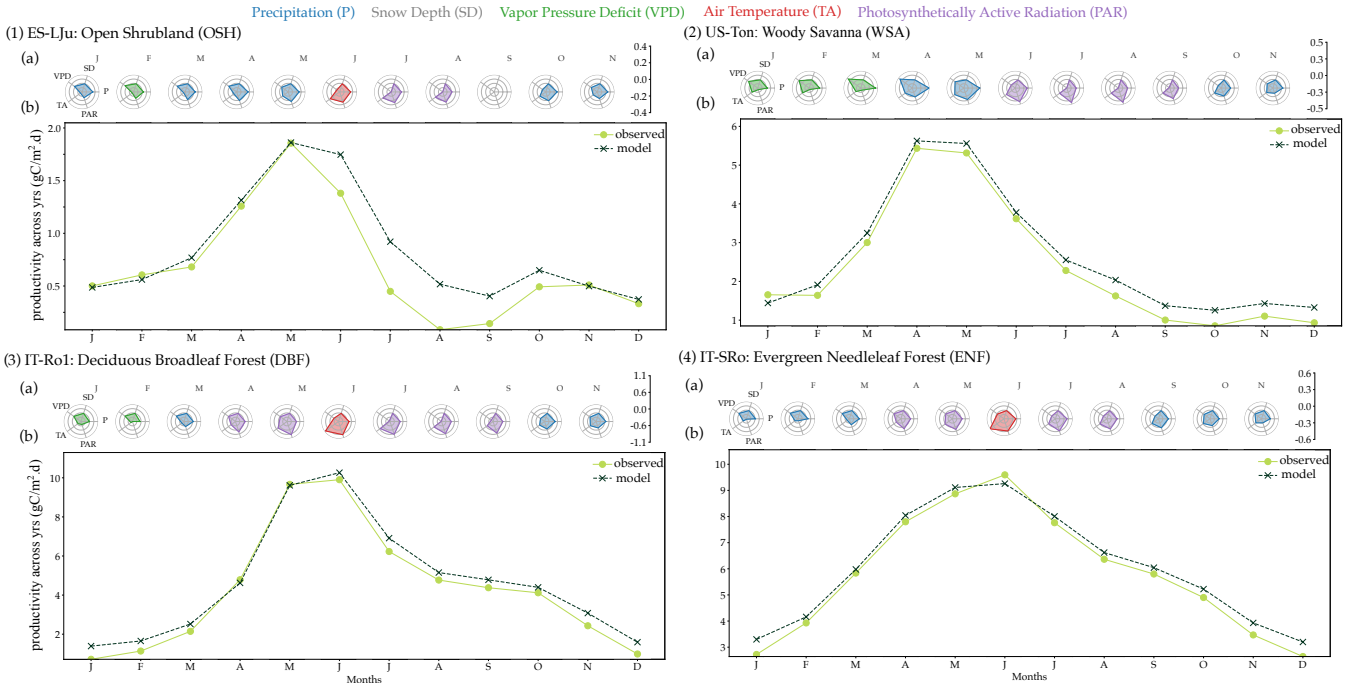

Figure 12. Feature importance analysis for productivity at Mediterranean sites, focusing on the 75th to 99th percentiles of monthly data. Each subplot represents a site with a distinct PFT: ES-Lju (Open Shrubland), US-Ton (Woody Savanna), IT-Ro1 (Deciduous Broad-leaf Forest), and IT-SRo (Evergreen Needle-leaf Forest). The analysis highlights the importance of precipitation at the start and end of the growth period, with temperature and radiation driving annual peaks.

et al., 2010). An ideal setting to explore this question is within semi-arid ecosystems (see, e.g., Hu et al., 2022), characterised by pulse-driven ecological processes, as further demonstrated here. These regions exhibit a layered interplay where the imme-
455 diate consequences of rainfall pulses—such as alterations in carbon uptake, nutrient cycling, and plant growth—are intricately linked with the broader narrative of long-term ecological impacts. Future endeavours aim to investigate the translatability of pulse-driven ecosystem dynamics in semi-arid landscapes across diverse plant functional classifications, climatic regimes, and precipitation regimes.

6 Conclusions

460 In this study, we developed and leveraged a memory-based machine learning model to test and advance our understanding of ecosystem dynamics in semi-arid regions, especially in Mediterranean environments. By embedding long-range correlations and temporal structures at multiple scales, our multi-timescale LSTM network has addressed critical ecological processes, particularly in response to extreme events, and interannual variability. We demonstrated how our model explains the influence of environmental controls on ecosystem responses. Consistent with previous research, our model indicates the crucial

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

role of precipitation in the Mediterranean region in determining the carbon uptake across these diverse and ecologically significant sites. Meanwhile, the absence of a strong direct relationship between snow depth and carbon uptake, alongside the notable role of temperature and photosynthetically active radiation on annual carbon uptake peaks, underscores the complexity of these ecosystems. Our model accurately explains extremes in the terrestrial water and carbon cycles. It also offers interpretable insights into atmospheric forcings' short- and long-term outcomes on these processes with enhanced representation of precipitation dynamics and their cascading effects.

This study highlights the necessity of advanced modelling techniques to capture such dynamics, pointing to the potential for further refinement and application of our model across broader ecological landscapes. As we face increasing climatic variability and changes, the insights gained here will be crucial for developing robust conservation strategies and sustainable management practices, ensuring the resilience and sustainability of the Mediterranean and similar ecosystems worldwide.

Acknowledgements. MA and PG thank Swiss National Foundation, Award #P500PN_206603 as well as National Science Foundation (NSF) Science and Technology Center (STC) Learning the Earth with Artificial Intelligence and Physics (LEAP), Award #2019625-STC for their financial support of this project. We thank FLUXNET2015 and Copernicus for their open-source datasets.

Author contributions. MA: Conceptualization, Methodology, Data curation, Visualization, Formal analysis, Funding acquisition, Writing - original draft. PG Conceptualization, Supervision , Funding acquisition, Writing - review & editing. WZ, JN, JQ, YZ Writing - review & editing.

Competing interests. The authors declare no conflicts of interest regarding the publication of this article.

Code and data availability. The code will be available publicly after acceptance and is available under https://zenodo.org/records/13963773 and the tutorial to run the model is available at https://ecoclimate-at-g-lab-columbia-university.github.io/MT-LSTM/. The FLUXNET2015 data used in this article are publicly available under fluxnet.org, and the snow depth and NDSI tested during this study were downloaded from Copernicus reanalysis datasets.

Disclaimer. The authors are not responsible for any inaccuracies in the publicly available datasets used in this study, which may be subject to updates. The findings, interpretations, and conclusions are those of the authors and do not necessarily reflect the views of the funding agencies or institutions. The authors are not liable for any damages arising from the use of this research. The data and findings are intended for scientific research and educational purposes only, and any unauthorised use is discouraged.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Appendix A: Model Design

A1 Feature Selection

Feature selection is a fundamental process in building models, aimed at boosting accuracy, simplifying structure, and enhance interpretability. Here, we followed a simplified forward selection and backward elimination technique to select the atmospheric variables used as input in our model as supported by James et al. (2013)

- **Forward Selection** involves starting with a null model and adding variables one at a time. Here, We evaluated each variable's explanatory power in a single-parameter model; the variable that results in the most significant improvement in the model's performance (based on the NSE metric) is added. Forward Selection helps in identifying the most important predictors for the model incrementally (Table A1 second column).

- **Backward Elimination** starts with a complete model containing all candidate variables (shown by a checkmark in Table A1). At each step, the least significant variable (based on NSE score) is removed (numbers in the second column of Table A1 indicates the order of removal), and the model is re-evaluated. Backward Elimination is useful for identifying and removing redundant features that do not contribute meaningfully to the model's predictive power (Table A1 last column). Though it greatly affects GPP, TS (temperature sensitivity) was not selected as an input variable, as it is not readily measurable and may not be available at most sites. We chose TA to substitute TS as the two are highly correlated.

Table A1. NSE for atmospheric variables in feature selection for the FLUXNET Mediterranean networks. The columns indicate the significance of individual variables in a single predictor model, the feature selection order, and the backward elimination sequence.

Features	NSE (one param model)	Included features	NSE after feature elimination
P (Precip.)	0.10-0.27	✓	-
SD (Snow Depth)	-	✓	-
TS (Soil Temp.)	0.34-0.5	-	-
TA (Air Temp.)	0.29-0.37	3-✓	0.15-0.3
Rn (Radiation)	0.17-0.29	-	-
VPD	-0.09-0.3	1-✓	0.45-0.56
RH (Rel. Hum.)	0.02-0.15	-	-
PAR	0.20-0.27	2-✓	0.51-0.58
SW_{IN}	0.18-0.24	-	-
PA (Atm. Pr.)	-	-	-
WS (Wind Speed)	0.04-0.30	-	-

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

A2 Input Variables and Colinearities

Understanding the co-linearity of input data helps build robust and interpretable models. Co-linearity, or multicollinearity, occurs when highly correlated predictor variables lead to redundancy and potential instability in model estimates. High co-linearity can inflate the variance of coefficient estimates and reduce the reliability of statistical tests of predictors. Here, in Figure A1, we present the co-linearity of the data in site ES-LJU (CSH), using 500 random samples - the lower triangle displays the correlation in scatter plot format , while the upper triangle presents the correlation coefficient in numerical format. The correlation coefficients between input features and output variables were consistent across all sites. Despite a strong correlation between air aridity (VPD) and air temperature, the variance inflation factors (VIF) are regularly below five among input features (except for US-Var, US-Ton, and US-Blo sites). Moreover, we had Lasso (Least Absolute Shrinkage and Selection Operator) regression with a weight decay of 0.0001(often in range of 0.001-0.00001) included in our training. This technique offers an alternative to traditional feature selection by penalizing less essential features, shrinking some coefficients to zero for variable selection. It helps address multicollinearity, prevents overfitting, and creates interpretable models focused on the most influential predictors. Figure A1a and Figure A1b highlight differing input-output relationships, suggesting that incorporating diverse temporal scales into the LSTM framework enhances insights and performance.

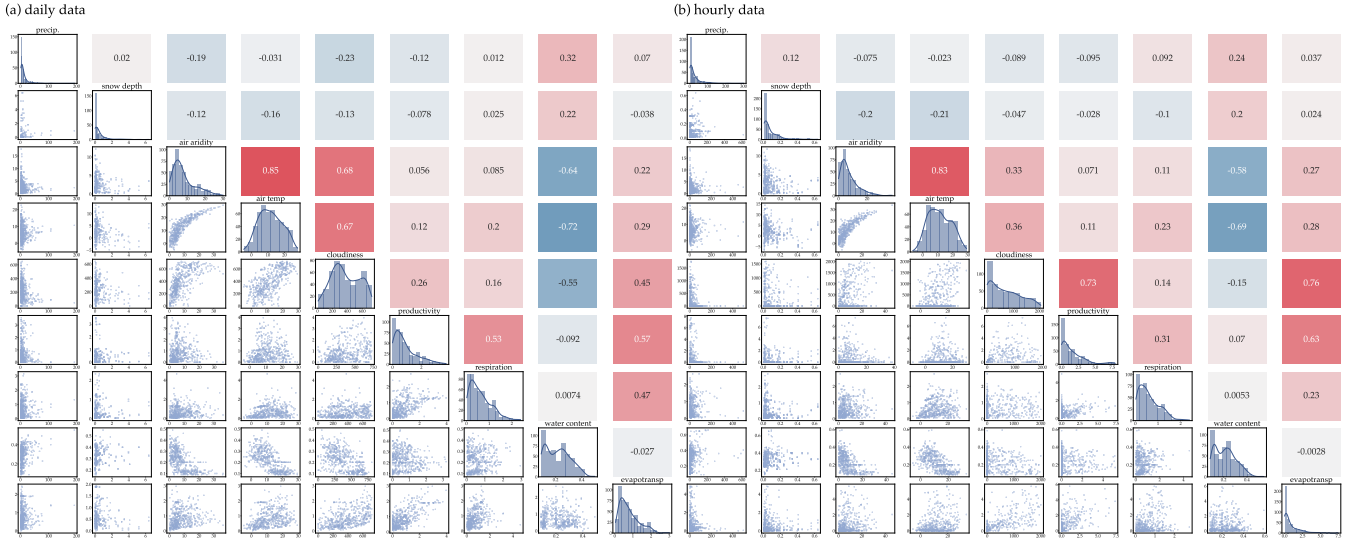

Figure A1. The diagonal elements show the variable distributions, while the lower triangle presents scatter plots of correlations, and the upper triangle displays the correlation coefficients. Stronger positive correlations are indicated by darker red, and negative correlations by darker blue. Panels (a) and (b) show collinearities among input and output variables at daily and half-hourly scales, respectively. Collinearities and correlations vary across scales. This figure, based on 500 random samples, reflects general trends rather than precise measurements.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Appendix B: Training Performance

B1 Learning Rate Adjustment

Figure B1 depicts the learning rate adjustments during training, controlled by the ReduceLROnPlateau scheduler. This method dynamically reduces the learning rate when a monitored metric (in this case, the loss) stops improving, helping to refine model convergence. We also evaluated model efficiency using the CosineAnnealingLR scheduler, which gradually reduces the learning rate following a cosine curve. Both approaches aim to optimise training efficiency and prevent the model from getting stuck in local minima or overfitting.

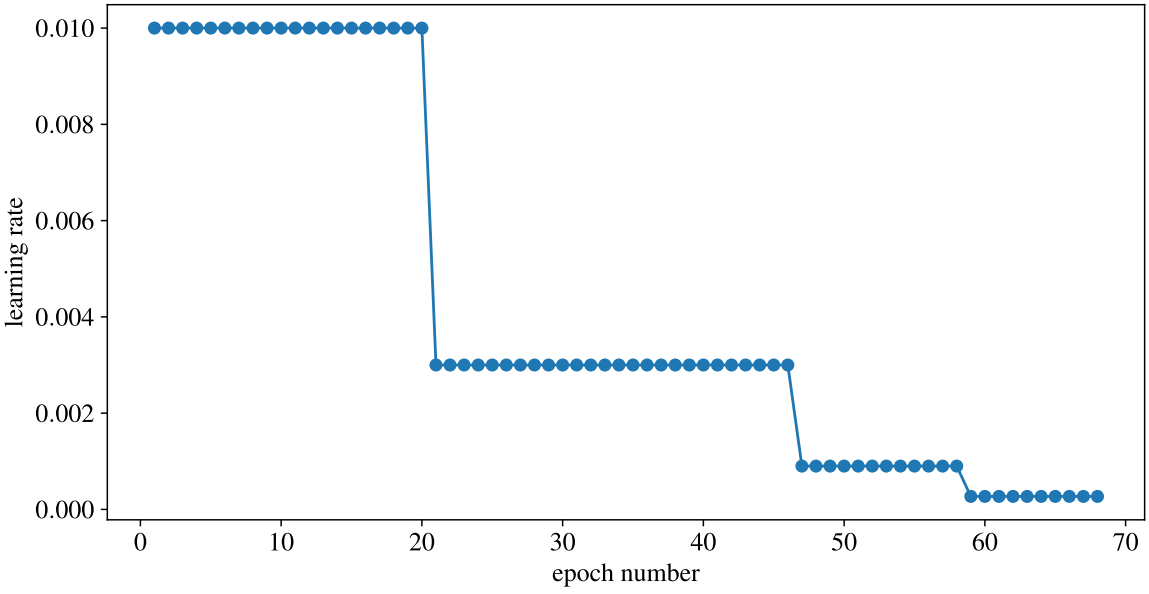

Figure B1. Learning rate adjustments during the training process using the 'ReduceLROnPlateau' scheduler. This method monitors validation loss and reduces the learning rate by 30 per cent when the loss has stopped improving by a minimum threshold of 0.05 for ten number of epochs, allowing finer adjustments to the training process.

B2 Loss versus Epochs

Figure B2 displays the error progression over epochs for a single fold from the K-fold cross-validation process. Since the patterns observed are comparable for all k-folds, we present results for one fold to avoid redundancy. Figure B2a illustrates the error trajectories for the training and validation sets, illustrating the model's performance across epochs during the training phase. These trajectories allow for analyzing convergence behavior and identifying signs of potential overfitting or underfitting. Ideally, we anticipate a monotonous decline in the training error, while the validation error should initially decrease and then

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

stabilise, indicating effective learning without overfitting. If the validation error rises after a certain number of epochs, it may suggest overfitting, necessitating the implementation of regularisation methods such as dropout or early stopping. Figure B2b depicts the reduction in training error for each output variable. The patterns in this figure suggest that the gradient descent optimisation during backpropagation efficiently minimises the error for all tasks, indicating that the gradients are not in conflict. Instead, they converge towards a shared optimal solution, suggesting that the learning process is well-aligned for multitask objectives.

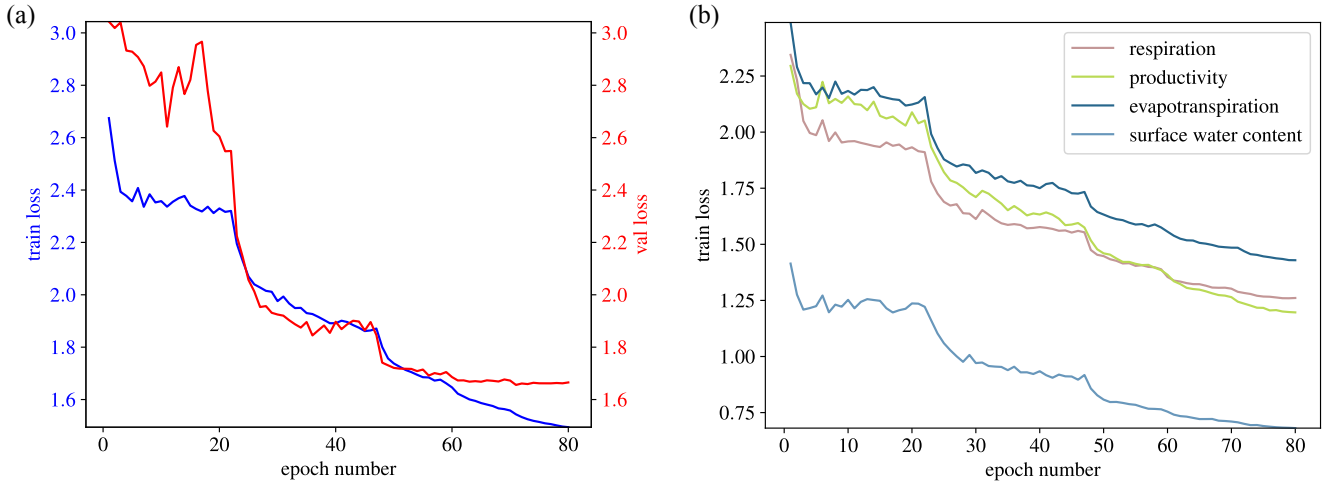

Figure B2. Error progression during model training across epochs for one fold in the K-fold cross-validation process. (a) Training and validation loss reduction versus epochs illustrate the convergence behaviour of the model. (b) Reduction in training loss per epoch for each output variable points to a comparable learning process for all tasks.

B3 Kernel Density Estimates

Figure B3 shows the Kernel Density Estimate (KDE) plots for IT-SRo for training, test, and validation sets in K-fold (k=5) cross-validation. The KDE plots provide a visual representation of the distribution of data points in different sets, allowing us to examine the consistency and overlap of features between these subsets. These KDE plots help assess whether the splitting strategy has maintained a similar distribution for all folds, which is crucial for ensuring robust and generalizable predictive performance. KDE for all sites has a similar trend, but at sites with 3 years of data, the validation and test set showed a narrower distribution than the training set.

B4 Statistics: Error across Quantiles

In Table B2, we present 65, and 85 quantile losses to exhibit the predictive efficiency for all sites at different quantiles. Also, these results imply our modified loss function has not caused a bias in favouring the higher quantiles for gross primary production.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

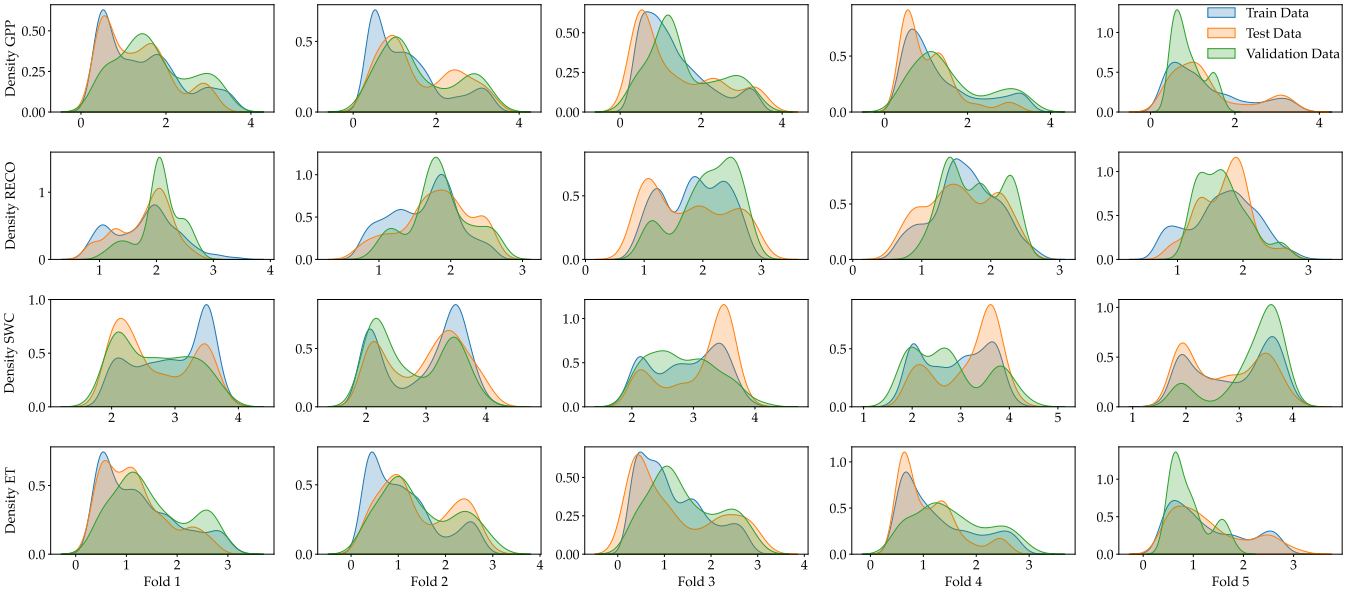

Figure B3. KDE for each site helps assess whether the splitting strategy has maintained a similar distribution across the folds and sets, ensuring robust and generalizable predictive efficiency. Here, we present the KDE of the IT-Ro1 site.

Table B1 summarises for one of the folds, Heidke Skill Score (HSS) for different sites and variables (ET, GPP, RECO, SWC) at three different thresholds: 0.5, 0.65, and 0.85 quantiles. In general, sites with lower MAE and quantile loss (e.g., US-Var) tend to have higher HSS. However, high MAE (e.g., IT-BCi) reflects poor predictive skill, especially for RECO and GPP, suggesting that the model struggles more with carbon flux predictions than with water-related variables such as SWC and ET. lower HSS could indicate that the model needs refinement for capturing complex ecological processes related to carbon dynamics, but it can also simply indicate the data is noisier in these sites.

B5 Statistics: Error per Site and Set

The data in Table B3 is used to plot Figure 3 in the manuscript and presents fold average error distinguished per site and train, test, validation set.

Appendix C: Predictions at hourly time-step

Figure C1, we present our model prediction at hourly time-step compared against hourly observational data.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Table B1. HSS for variables ET, GPP, RECO, SWC at quantiles of 0.5, 0.65, and 0.85 quantile for one fold.

site name	SWC			GPP			RECO			ET		
quantiles	**0.5**	**0.65**	**0.85**	**0.5**	**0.65**	**0.85**	**0.5**	**0.65**	**0.85**	**0.5**	**0.65**	**0.85**
AU-Rig	0.30	0.61	0.56	0.00	0.05	0.27	0.02	0.07	0.15	0.79	0.64	0.33
ES-LJu	0.63	0.60	-0.02	-0.14	0.08	0.58	0.09	0.11	-0.00	0.65	0.64	0.61
IT-BCi	0.13	0.08	0.03	0.11	0.57	0.76	0.58	0.47	0.40	0.53	0.68	0.55
IT-CA1	0.16	0.00	0.00	0.41	0.85	0.12	0.00	0.48	0.12	0.59	0.74	0.43
IT-CA2	0.07	0.07	0.00	0.16	0.10	0.00	0.26	0.31	0.08	0.31	0.40	0.38
IT-CA3	0.07	0.58	-0.07	0.44	0.79	NaN	0.41	0.00	0.00	0.42	0.06	NaN
IT-Cpz	0.92	0.79	0.58	0.59	0.58	0.00	0.27	0.15	0.40	0.68	0.65	-0.00
IT-Noe	0.65	0.83	0.38	0.54	0.61	0.53	0.52	0.31	0.07	0.44	0.46	0.30
IT-PT1	0.11	-0.00	0.00	0.26	0.62	0.60	-0.01	0.31	0.04	0.67	0.64	0.46
IT-Ro1	0.71	0.64	0.39	0.43	0.47	0.71	0.28	0.33	0.06	0.70	0.78	0.65
IT-Ro2	0.07	0.05	0.00	0.76	0.68	0.53	0.59	0.46	0.00	0.75	0.76	0.53
IT-SRo	0.39	-0.09	0.35	0.54	0.46	0.31	0.24	0.21	0.04	0.23	0.34	0.36
US-AR1	0.37	0.34	0.68	0.35	0.54	0.35	0.46	0.59	0.41	0.64	0.71	0.60
US-ARM	0.15	0.17	0.07	0.07	0.17	0.41	0.35	0.49	0.29	0.56	0.57	0.46
US-Blo	0.73	0.52	0.24	0.33	0.21	0.04	0.07	0.10	0.09	0.79	0.75	0.41
US-Ton	0.87	0.89	0.42	0.61	0.65	0.76	0.17	0.34	0.34	0.63	0.84	0.75
US-Var	0.83	0.84	0.66	0.24	0.72	0.60	0.15	0.48	0.26	0.57	0.70	0.73

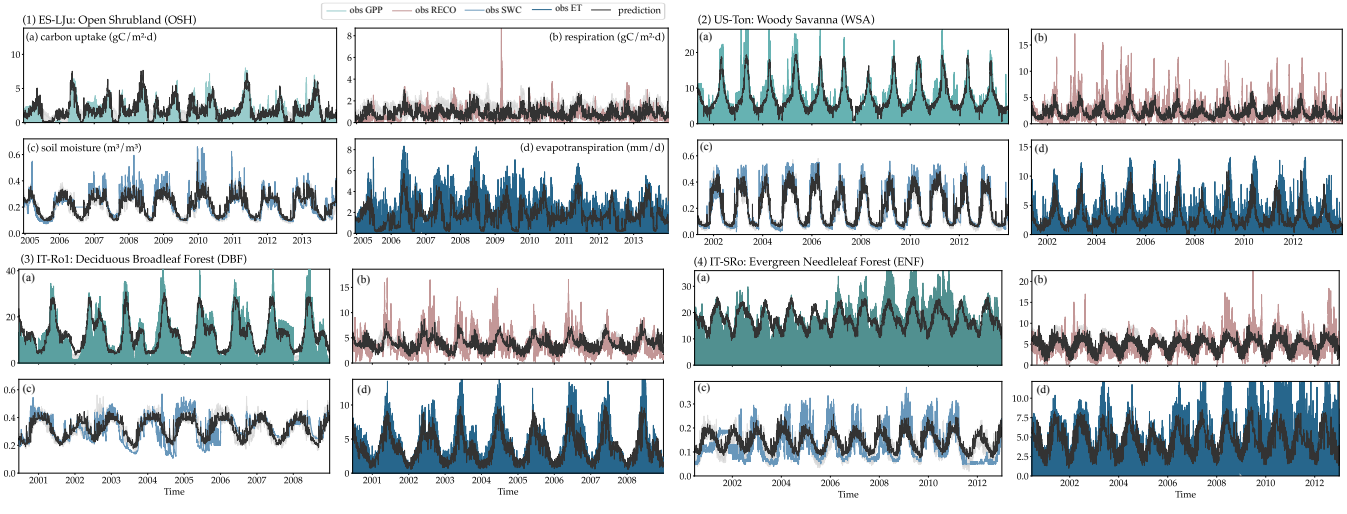

Figure C1. Comparison of our model estimations at hourly time-step against observational data

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Table B2. Performance metrics for different sites in k-folding (k=5) averaged across folds for GPP.

site name	mean absolute error	65 quantile loss	85 quantile loss
IT-CA2	1.3	0.55	0.43
IT-CA3	0.81	0.38	0.33
AU-Rig	1.21	0.55	0.47
US-Blo	1.05	0.47	0.40
US-Var	1.03	0.44	0.33
US-Ton	0.80	0.36	0.30
US-ARM	1.44	0.60	0.44
US-AR1	1.31	0.54	0.37
ES-LJu	0.45	0.18	0.13
IT-Noe	0.85	0.38	0.33
IT-SRo	1.59	0.74	0.66
IT-BCi	3.04	1.27	0.93
IT-Cpz	1.28	0.64	0.65
IT-PT1	1.23	0.58	0.53
IT-CA1	1.34	0.59	0.49
IT-Ro1	1.39	0.63	0.54
IT-Ro2	1.56	0.66	0.51

Appendix D: Model comparison for extended sites

Figure D1, we present our model prediction compared against FLUXCOM(-X). The KGE assesses predictive efficiency by combining correlation, variability, and bias, offering a broader evaluation than R^2 and NSE. For AU-Rig, EcoPro-LSTM$_{v0}$ predictions achieve a KGE of 0.36, R^2 of 0.67, and NSE of 0.63, while FLUXCOM-X performs with a KGE of 0.7, R^2 of 0.78, and NSE of 0.72. This increased KGE suggests a balanced fit between variability and bias in FLUXCOM-X data.

The EcoPro-LSTM$_{v0}$ in this site consistently identifies double-peaked growth periods (during the second and third years), while observations manifest a double peak only in the second season and no clear peak in the third. FLUXCOM and FLUXCOM-X adapt more flexibly to these interannual changes, contributing to their larger KGE. To ensure KGE performance is valid and not distorted by data noise, we plan to incorporate extended records from OzFlux in future work.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Table B3. Performance metrics for productivity averaged for all cross-validation folds, distinct by site and dataset.

	mean_ae			rmse			weighted_rmse_torch		
Site	train set	val test	test set	train set	val test	test set	train set	val test	test set
AU-Rig	1.14	1.8	1.75	1.5	1.99	2.21	1.8	1.91	2.69
IT-CA2	1.28	1.64	1.94	1.61	1.94	2.57	1.93	2.03	2.78
IT-CA3	0.88	0.64	1.37	1.24	0.76	1.8	1.42	0.81	1.86
ES-LJu	0.48	0.45	0.54	0.61	0.59	0.68	0.6	0.63	0.67
IT-Noe	0.83	1.23	0.99	1.09	1.45	1.28	1.18	1.19	1.31
IT-SRo	1.58	1.67	1.9	2.02	2.07	2.34	2.1	2.17	2.42
IT-BCi	3.00	3.27	3.47	3.97	4.12	4.58	3.45	4.16	4.24
IT-Cpz	1.21	1.9	1.42	1.54	2.43	1.81	1.59	2.92	1.98
IT-PT1	1.26	1.51	1.65	1.78	1.8	2.34	2.14	2.59	2.88
IT-CA1	1.37	1.78	2.05	2.03	2.13	2.91	2.48	1.66	3.88
IT-Ro1	1.36	1.89	1.67	1.75	2.33	2.07	2.1	2.68	2.48
IT-Ro2	1.50	2.22	1.80	1.95	2.8	2.34	2.25	2.62	2.67
US-Blo	1.06	1.19	1.34	1.35	1.48	1.66	1.39	1.5	1.77
US-Var	1.05	1.00	1.12	1.32	1.16	1.42	1.83	1.26	1.95
US-Ton	0.81	0.91	0.88	1.05	1.14	1.17	1.27	1.27	1.37
US-ARM	1.46	1.53	1.69	1.90	1.95	2.16	2.57	2.95	2.84
US-AR1	1.35	0.71	1.81	1.73	0.91	2.42	1.96	1.09	2.86

References

Aas, K., Jullum, M., and Løland, A.: Explaining individual predictions when features are dependent: More accurate approximations to Shapley values, Artificial Intelligence, 298, 103 502, 2021.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information fusion, 76, 243–297, 2021.

Allen, R. G., Pereira, L. S., Raes, D., Smith, M., et al.: Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56, Fao, Rome, 300, D05 109, 1998.

Arca, V., Power, S. A., Delgado-Baquerizo, M., Pendall, E., and Ochoa-Hueso, R.: Seasonal effects of altered precipitation regimes on ecosystem-level CO 2 fluxes and their drivers in a grassland from Eastern Australia, Plant and Soil, 460, 435–451, 2021.

Aumann, R. J. and Shapley, L. S.: Values of non-atomic games, Princeton University Press, 2015.

Bachman, S., Heisler-White, J. L., Pendall, E., Williams, D. G., Morgan, J. A., and Newcomb, J.: Elevated carbon dioxide alters impacts of precipitation pulses on ecosystem photosynthesis and respiration in a semi-arid grassland, Oecologia, 162, 791–802, 2010.

Baldocchi, D., Ma, S., and Verfaillie, J.: On the inter-and intra-annual variability of ecosystem evapotranspiration and water use efficiency of an oak savanna and annual grassland subjected to booms and busts in rainfall, Global Change Biology, 27, 359–375, 2021.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

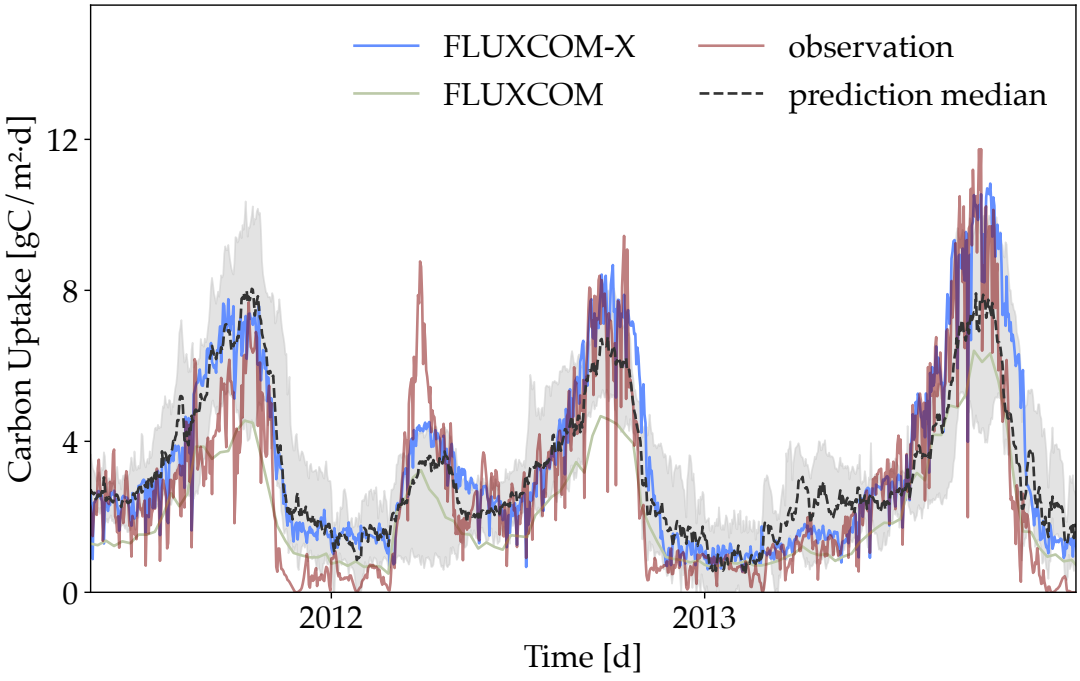

Figure D1. Comparison of our model estimations against FLUXCOM(-X) for the AU-Rig site. The results indicate that our model exhibits relatively poorer performance R^2, KGE, and NSE metrics.

Bartsch, S., Stegehuis, A., Boissard, C., Lathière, J., Peterschmitt, J.-Y., Reiter, I., Gauquelin, T., Baldy, V., Genesio, L., Matteucci, G., et al.: Impact of precipitation, air temperature and abiotic emissions on gross primary production in Mediterranean ecosystems in Europe, European Journal of Forest Research, 139, 111–126, 2020.

Beck, H. E., McVicar, T. R., Vergopolan, N., Berg, A., Lutsko, N. J., Dufour, A., Zeng, Z., Jiang, X., van Dijk, A. I., and Miralles, D. G.: High-resolution (1 km) Köppen-Geiger maps for 1901–2099 based on constrained CMIP6 projections, Scientific data, 10, 724, 2023.

Bodesheim, P., Jung, M., Gans, F., Mahecha, M. D., and Reichstein, M.: Upscaled diurnal cycles of land–atmosphere fluxes: a new global half-hourly data product, Earth System Science Data, 10, 1327–1365, 2018.

Caruana, R.: Multitask learning, Machine learning, 28, 41–75, 1997.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I.: L-shapley and c-shapley: Efficient model interpretation for structured data, arXiv preprint arXiv:1808.02610, 2018.

Chiesi, M., Maselli, F., Bindi, M., Fibbi, L., Cherubini, P., Arlotta, E., Tirone, G., Matteucci, G., and Seufert, G.: Modelling carbon budget of Mediterranean forests using ground and remote sensing measurements, Agricultural and Forest Meteorology, 135, 22–34, 2005.

Choat, B., Brodribb, T. J., Brodersen, C. R., Duursma, R. A., López, R., and Medlyn, B. E.: Triggers of tree mortality under drought, Nature, 558, 531–539, 2018.

Cohrs, K.-H., Varando, G., Carvalhais, N., Reichstein, M., and Camps-Valls, G.: Causal hybrid modeling with double machine learning—applications in carbon flux modeling, Machine Learning: Science and Technology, 5, 035 021, 2024.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Cos, J., Doblas-Reyes, F., Jury, M., Marcos, R., Bretonnière, P.-A., and Samsó, M.: The Mediterranean climate change hotspot in the CMIP5 and CMIP6 projections, Earth System Dynamics, 13, 321–340, 2022.

Cranko Page, J., De Kauwe, M. G., Abramowitz, G., Cleverly, J., Hinko-Najera, N., Hovenden, M. J., Liu, Y., Pitman, A. J., and Ogle, K.: Examining the role of environmental memory in the predictability of carbon and water fluxes across Australian ecosystems, Biogeosciences Discussions, 2021, 1–29, 2021.

ElGhawi, R., Kraft, B., Reimers, C., Reichstein, M., Körner, M., Gentine, P., and Winkler, A. J.: Hybrid modeling of evapotranspiration: inferring stomatal and aerodynamic resistances using combined physics-based and machine learning, Environmental Research Letters, 18, 034 039, 2023.

Feng, T., Zhou, Z., Tarun, J., and Nair, V. N.: Comparing Baseline Shapley and Integrated Gradients for Local Explanation: Some Additional Insights, arXiv preprint arXiv:2208.06096, 2022.

Friedlingstein, P., O'sullivan, M., Jones, M. W., Andrew, R. M., Gregor, L., Hauck, J., Le Quéré, C., Luijkx, I. T., Olsen, A., Peters, G. P., et al.: Global carbon budget 2022, Earth System Science Data, 14, 4811–4900, 2022.

Giorgi, F. and Lionello, P.: Climate change projections for the Mediterranean region, Global and planetary change, 63, 90–104, 2008.

Guo, R., Chen, T., Chen, X., Yuan, W., Liu, S., He, B., Li, L., Wang, S., Hu, T., Yan, Q., et al.: Estimating global GPP from the plant functional type perspective using a machine learning approach, Journal of Geophysical Research: Biogeosciences, 128, e2022JG007 100, 2023.

Hao, G., Hu, Z., Di, K., and Li, S.: Rainfall pulse response of carbon exchange to the timing of natural intra-annual rainfall in a temperate grass ecosystem, Ecological Indicators, 118, 106 730, 2020.

Haverd, V., Ahlström, A., Smith, B., and Canadell, J. G.: Carbon cycle responses of semi-arid ecosystems to positive asymmetry in rainfall, Global Change Biology, 23, 793–800, 2017.

Hu, X., Zhu, M., Feng, Z., and Stanković, L.: Manifold-based Shapley explanations for high dimensional correlated features, Neural Networks, 180, 106 634, 2024.

Hu, Z., Piao, S., Knapp, A. K., Wang, X., Peng, S., Yuan, W., Running, S., Mao, J., Shi, X., Ciais, P., et al.: Decoupling of greenness and gross primary productivity as aridity decreases, Remote Sensing of Environment, 279, 113 120, 2022.

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.

Jacovides, C., Tymvios, F., Asimakopoulos, D., Theofilou, K., and Pashiardes, S.: Global photosynthetically active radiation and its relationship with global solar radiation in the Eastern Mediterranean basin, Theoretical and Applied Climatology, 74, 227–233, 2003.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al.: An introduction to statistical learning, vol. 112, Springer, 2013.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., et al.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, Journal of Geophysical Research: Biogeosciences, 116, 2011.

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S., Bodesheim, P., Carvalhais, N., et al.: Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the FLUXCOM approach, Biogeosciences, 2020.

Kannenberg, S. A., Schwalm, C. R., and Anderegg, W. R.: Ghosts of the past: how drought legacy effects shape forest functioning and carbon cycling, Ecology letters, 23, 891–901, 2020.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Kapishnikov, A., Venugopalan, S., Avci, B., Wedin, B., Terry, M., and Bolukbasi, T.: Guided integrated gradients: An adaptive path method for removing noise, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5050–5058, 2021.

Katul, G., Lai, C.-T., Schäfer, K., Vidakovic, B., Albertson, J., Ellsworth, D., and Oren, R.: Multiscale analysis of vegetation surface fluxes: from seconds to years, Advances in Water Resources, 24, 1119–1132, 2001.

Lasslop, G., Reichstein, M., Papale, D., Richardson, A. D., Arneth, A., Barr, A., Stoy, P., and Wohlfahrt, G.: Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation, Global change biology, 16, 187–208, 2010.

Legates, D. R. and McCabe Jr, G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water resources research, 35, 233–241, 1999.

Lionello, P. and Scarascia, L.: The relation between climate change in the Mediterranean region and global warming, Regional Environmental Change, 18, 1481–1493, 2018.

Liu, L., Zhou, W., Guan, K., Peng, B., Xu, S., Tang, J., Zhu, Q., Till, J., Jia, X., Jiang, C., et al.: Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems, Nature communications, 15, 357, 2024.

Liu, S., Chen, Z., Ge, S., Wang, J., Fan, C., Xiong, Y., Hu, R. W. Y., Ji, Z., and Gao, Y.: Rethink Baseline of Integrated Gradients from the Perspective of Shapley Value, arXiv preprint arXiv:2310.04821, 2023a.

Liu, W., He, H., Wu, X., Ren, X., Zhang, L., Feng, L., Wang, Y., Lv, Y., et al.: Importance of the memory effect for assessing interannual variation in net ecosystem exchange, Agricultural and Forest Meteorology, 341, 109 691, 2023b.

Lundberg, S.: A unified approach to interpreting model predictions, arXiv preprint arXiv:1705.07874, 2017.

MacBean, N., Scott, R. L., Biederman, J. A., Peylin, P., Kolb, T., Litvak, M. E., Krishnan, P., Meyers, T. P., Arora, V. K., Bastrikov, V., et al.: Dynamic global vegetation models underestimate net CO2 flux mean and inter-annual variability in dryland ecosystems, Environmental Research Letters, 16, 094 023, 2021.

Markos, N., Preisler, Y., Radoglou, K., Rotenberg, E., and Yakir, D.: Physiological and phenological adjustments in water and carbon fluxes of Aleppo pine forests under contrasting climates in the Eastern Mediterranean, Tree Physiology, 44, tpad125, 2024.

Marshall, M., Tu, K., and Brown, J.: Optimizing a remote sensing production efficiency model for macro-scale GPP and yield estimation in agroecosystems, Remote sensing of environment, 217, 258–271, 2018.

Meek, D. W., Hatfield, J. L., Howell, T. A., Idso, S. B., and Reginato, R. J.: A generalized relationship between photosynthetically active radiation and solar radiation 1, Agronomy journal, 76, 939–945, 1984.

Nathaniel, J., Liu, J., and Gentine, P.: MetaFlux: Meta-learning global carbon fluxes from sparse spatiotemporal observations, Scientific Data, 10, 440, 2023.

Nelson, J. A. et al.: X-BASE: the first terrestrial carbon and water flux products from an extended data-driven scaling framework, FLUXCOM-X, EGUsphere, 2024.

Nguyen, T.-A., Rußwurm, M., Lenczner, G., and Tuia, D.: Multi-temporal forest monitoring in the Swiss Alps with knowledge-guided deep learning, Remote Sensing of Environment, 305, 114 109, 2024.

Poulter, B., Frank, D., Ciais, P., Myneni, R. B., Andela, N., Bi, J., Broquet, G., Canadell, J. G., Chevallier, F., Liu, Y. Y., et al.: Contribution of semi-arid ecosystems to interannual variability of the global carbon cycle, Nature, 509, 600–603, 2014.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., et al.: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm, Global change biology, 11, 1424–1439, 2005.

https://doi.org/10.5194/egusphere-2024-3726
Preprint. Discussion started: 13 February 2025

Rey, A., Pegoraro, E., Tedeschi, V., De Parri, I., Jarvis, P. G., and Valentini, R.: Annual variation in soil respiration and its components in a coppice oak forest in Central Italy, Global Change Biology, 8, 851–866, 2002.

Ruder, S.: An Overview of Multi-Task Learning in Deep Neural Networks, arXiv preprint arXiv:1706.05098, 2017.

Seager, R., Osborn, T. J., Kushnir, Y., Simpson, I. R., Nakamura, J., and Liu, H.: Climate variability and change of Mediterranean-type climates, Journal of Climate, 32, 2887–2915, 2019.

Serrano-Ortiz, P., Domingo, F., Cazorla, A., Were, A., Cuezva, S., Villagarcía, L., Alados-Arboledas, L., and Kowalski, A.: Interannual CO2 exchange of a sparse Mediterranean shrubland on a carbonaceous substrate, Journal of Geophysical Research: Biogeosciences, 114, 2009.

Shangguan, W., Xiong, Z., Nourani, V., Li, Q., Lu, X., Li, L., Huang, F., Zhang, Y., Sun, W., and Dai, Y.: A 1 km global carbon flux dataset using in situ measurements and deep learning, Forests, 14, 913, 2023.

Sturmfels, P., Lundberg, S., and Lee, S.-I.: Visualizing the impact of feature attribution baselines, Distill, 5, e22, 2020.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic attribution for deep networks, in: International conference on machine learning, pp. 3319–3328, PMLR, 2017.

Vekuri, H., Tuovinen, J.-P., Kulmala, L., Papale, D., Kolari, P., Aurela, M., Laurila, T., Liski, J., and Lohila, A.: A widely-used eddy covariance gap-filling method creates systematic bias in carbon balance estimates, Scientific Reports, 13, 1720, 2023.

Wang, J., Xiao, X., Wagle, P., Ma, S., Baldocchi, D., Carrara, A., Zhang, Y., Dong, J., and Qin, Y.: Canopy and climate controls of gross primary production of Mediterranean-type deciduous and evergreen oak savannas, Agricultural and forest meteorology, 226, 132–147, 2016.

Yan, Y., Li, G., Li, Q., and Zhu, J.: Enhancing Hydrological Variable Prediction through Multitask LSTM Models, Water, 16, 2156, 2024.

Yang, Y., Roderick, M. L., Guo, H., Miralles, D. G., Zhang, L., Fatichi, S., Luo, X., Zhang, Y., McVicar, T. R., Tu, Z., et al.: Evapotranspiration on a greening Earth, Nature Reviews Earth & Environment, 4, 626–641, 2023.

Zhang, Y., Xiao, X., Zhang, Y., Wolf, S., Zhou, S., Joiner, J., Guanter, L., Verma, M., Sun, Y., Yang, X., et al.: On the relationship between sub-daily instantaneous and daily total gross primary production: Implications for interpreting satellite-based SIF retrievals, Remote Sensing of Environment, 205, 276–289, 2018.