On the added value of sequential deep learning for upscaling of evapotranspiration

Basil Kraft[1,2,3], Jacob A. Nelson[1], Sophia Walther[1], Fabian Gans[1], Ulrich Weber[1], Gregory Duveiller[1], Markus Reichstein[1], Weijie Zhang[1], Marc Rußwurm[5], Devis Tuia[4], Marco Körner[3], and Martin Jung[1]

[1]Max Planck Institute for Biogeochemistry (MPI BGC)
[2]Swiss Federal Institute of Technology Zurich (ETH)
[3]Technical University of Munich (TUM)
[4]École Polytechnique Fédérale de Lausanne (EPFL)
[5]Wageningen University (WU)

Correspondence: Basil Kraft (basil.kraft@protonmail.com)

Abstract.

Estimating ecosystem-atmosphere fluxes such as evapotranspiration (ET) in a robust manner and at global scale remains a challenge. Machine learning–based methods have shown promising results to achieve such upscaling, providing a complementary methodology that is independent from process-based and semi-empirical approaches. However, a systematic evaluation of the skill and robustness of different ~~ML~~ machine learning (ML) approaches is an active field of research that requires more investigations. Concretely, deep learning approaches in the time domain have not been explored systematically for this task.

In this study, we compared instantaneous (i.e., non-sequential) models—extreme gradient boosting (XGBoost) and a fully-connected neural network (FCN)—with sequential models—a long short-term memory (LSTM) model and a temporal convolutional network (TCN)~~, for~~ —for the modeling and upscaling of ET. We compared different types of covariates (meteorological ~~,~~ without precipitation, precipitation, remote sensing, and plant functional types) and their impact on model performance at the site level in a cross-validation setup.

When using only meteorological covariates, we found that the sequential models (LSTM and TCN) performed better—each with a Nash-Sutcliffe modeling efficiency (NSE) of 0.73—than the instantaneous models (FCN and XGBoost)—both with an NSE of 0.70—in site level cross-validation at the hourly scale. The advantage of the sequential models diminished with the inclusion of remote-sensing-based predictors (NSE of 0.75 to 0.76 versus 0.74). On the anomaly scale, the sequential models consistently outperformed the non-sequential models across covariate setups, with an NSE of 0.36 (LSTM) and 0.38 (TCN) versus 0.33 (FCN) and 0.32 (XGBoost) when using all covariates.

For the upscaling from site to global coverage, we input the two best-performing ~~combination of covariates—which was meteorological~~ combinations of covariates—meteorological and remote sensing ~~observations—with~~ observations, and with precipitation and plant functional types in addition—with globally available gridded data. To evaluate and compare the robustness of the modeling approaches, we generated a cross-validation-based ensemble of upscaled ET, compared the ensemble mean and variance among models, and contrasted it with independent global ET data.

We found that the sequential models performed better than the instantaneous models (FCN and XGBoost) in cross-validation, while the advantage of In particular, we investigate three questions regarding the performance of sequential models compared to the non-sequential models in the sequential models diminished with the inclusion of remote-sensing-based predictors. When including remote sensing covariates, context of spatial upscaling: a) whether they lead to more realistic and robust global and regional ET; b) whether they are able to capture the temporal dynamics of ET better; and c) how robust they are to the covariate setup and training data subsets.

The generated patterns of global ET variability were highly consistent across all relatively consistent across the ML models overall. However, the temporal models yielded 6-9% lower globally integrated ET compared to the non-temporal counterparts and estimates from independent land surface models , which was likely due to their enhanced vulnerability to changes in the predictor distributions from site-level training data to global prediction data. The neural network ensembles showed a sizable spread due to training data subsets, which exceeds differences among neural network variants. but in regions with low data support via EC stations, we observed substantial biases across models and covariate setups, and large ensemble uncertainties. The sequential models better capture temporal dynamics of ET when upscaled to global coverage, especially when using precipitation as additional input, and they seem to be more robust to covariate setups, particularly the LSTM model. However, they exhibited, together with the non-temporal FCN model, larger ensemble spread than XGBoost, and they yielded lower global ET estimates than what is currently understood. XGBoost showed smaller ensemble spread compared to neural networks in particular when conditions were poorly represented in the training data-, but it was more sensitive to the covariate setup. Plant functional types were useful at site level for improved representation of spatial patterns, but had a large leverage on upscaling results—i.e., having disproportionate impact on the spatial patterns—especially for XGBoost, but less for the LSTM model.

Our findings highlight non-linear model responses to biases in the training data and underscore the need for improved upscaling methodologies, which could be achieved by increasing the amount and quality of training data or by the extraction of more targeted features representing spatial variability. The neural networks seem to yield more realistic ensemble uncertainty compared to XGBoost. Approaches such as transfer learning, knowledge-guided ML, or hybrid modeling, which encourage physically consistent results while harnessing the efficiency of ML, or transfer learning, should be should be further investigated. Deep learning for flux upscaling holds large promise, while remedies for its vulnerability to training data distribution changes , especially of sequential models, still need consideration by the community.

. TEXT

1

Measurements of land-atmosphere fluxes of gases, such as water vapor or carbon dioxide, are crucial for understanding the interactions between climate and ecosystems. Instruments at eddy covariance (EC) stations measure such fluxes integrated over

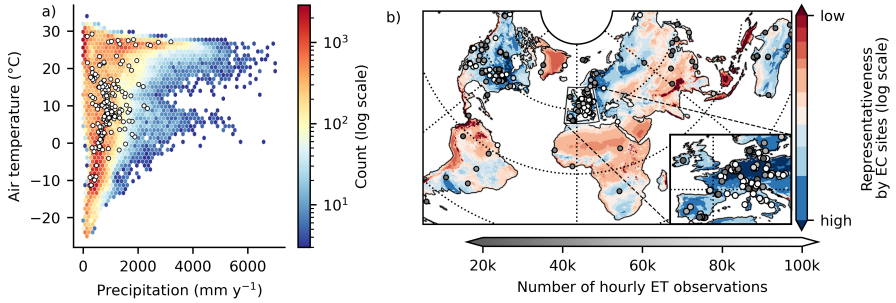

Figure 1. Overview of eddy covariance (EC) sites used in this work: **a)** Distribution of EC sites (white points) and map grid-cells (background color) within the global climate in terms of mean temperature and annual precipitation. **b)** Geographic EC site locations in different gray scales according to the number of hourly observations of evapotranspiration. The map color corresponds to the representativeness of a geographic location by the EC station sites. It is the average Euclidean distance in climate space (mean and standard deviation of normalized 15-daily temperature, precipitation, and radiation) to the ten closest stations. A lower representativeness (red) means a given location is further away from EC sites in climate space.

a time span of 30 or 60 minutes and a small spatial footprint, spanning a couple of hundred meters to over a kilometer, depending on the ~~station~~ instrument height, terrain roughness, and wind conditions. The measurement is performed at *ecosystem level*, as it represents the integral of biotic and abiotic processes across scales (Baldocchi et al., 2001). While EC stations provide a crucial source of data to measure these fluxes, they come with challenges. For instance, their representativeness and applicability for regional to global analysis may be restricted due to the sparsity of EC sites in geographic and climate space (Fig. 1a and b).

Evapotranspiration (ET) is the combined flux of water vapor via evaporation from bare surfaces and plant transpiration. The ET flux is of high relevance for modeling and understanding the Earth system because it links water, carbon, and energy cycles (Jung et al., 2010; Nelson et al., 2018). However, the modeling of ET is challenging due to the highly dynamic nature ~~of~~ and modulation by ecosystems. Their behavior depends on past system exposure via so-called dynamic memory effects (Ogle et al., 2015; Besnard et al., 2019; Kraft et al., 2019, 2021). Among other factors, ET depends on soil moisture, which is primarily driven by the recent past rather than by instantaneous weather conditions. Other processes impacting ET that depend on past meteorology are related to vegetation states, such as the leaf area or phenology (Migliavacca et al., 2012).

To consider such complex memory effects, a model must ~~either~~ incorporate past system exposure, such as temperature or precipitation. Alternatively, the model can be fed with states that represent past exposure, such as leaf area index (LAI) and soil moisture observations, or aggregations of past meteorology like temperature or precipitation sums. However, the observation of ecosystem states is challenging and often not possible. *In-situ* measurements, e.g., of soil moisture, are not consistently ~~measured~~ available at all EC ~~station sites~~ stations and may not always precisely coincide with the eddy covariance measurements in space or time, limiting the applicability for across-site modeling. As an alternative, remotely sensed observations can serve as proxies of ecosystem states, like vegetation indices for foliage or phenology. These observations alone can only partially explain EC measurements, as they represent structural or optical properties of the canopy rather than plant physiology

or subsurface water states, and especially optical observations tend to saturate with dense vegetation (Huete et al., 2002). Therefore, it may be beneficial to *learn* the non-observable states for the modeling of land-atmosphere fluxes as non-linear functions of available covariates. Here, sequential machine learning (ML) models may offer a unique opportunity, as they are able to extract dynamic proxies from temporal data (Rußwurm and Körner, 2017; Kraft et al., 2019).

ET can be quantified at large scales employing process-based paradigms, i.e., land surface models, or semi-empirical approaches, based on inputs from remote sensing observations and predefined empirical relationships (e.g., the Global Land Evaporation Amsterdam Model (GLEAM), Martens et al., 2017). As a complementary approach, the data-driven upscaling, i.e., the generalization from the irregularly distributed EC stations to a regular spatio-temporal field, ~~provides~~ can provide independent insights into ecosystem processes (Jung et al., 2017). The upscaling is achieved by training an ML model at the EC sites with covariates that are also available as spatio-temporal fields (Jung et al., 2009). The optimized model is then fed with the contiguous covariates to generate regional to global scale products.

Due to the availability of long-term records of both eddy covariance data and remote sensing products, increased computational capacities, and a higher acceptance of ML approaches in the geosciences (Camps-Valls et al., 2021), data-driven approaches to model ecosystem-atmosphere fluxes have gained momentum in the past decade (Tramontana et al., 2016; Jung et al., 2011; Nelson & Walther et al., 2024; Zhu et al., 2024). Today, ML is widely used to model and upscale EC data, but the field is still dominated by non-sequential modeling (i.e., an instantaneous model that does not learn memory effects), such as decision trees or fully-connected neural networks.

An ensemble of global, harmonized products of upscaled EC fluxes from different ML algorithms (tree, kernel, regression splines, and neural network-based methods) was released by the FLUXCOM initiative (FLUXCOM, 2017), founded on previous work by Beer et al. (2010), Jung et al. (2010, 2011), and Tramontana et al. (2016). These products are ~~build~~ built upon non-sequential models, and they account for memory via manually designed features, such as seasonal amplitudes or water availability indices, and remote sensing-based ecosystem state proxies, like vegetation indices (Huete et al., 2002). The FLUXCOM products of energy (Jung et al., 2019) and carbon (Jung et al., 2020) are utilized in contemporary ~~land–atmosphere~~ land-atmosphere interaction studies and function as benchmarks for Earth system models. To improve the temporal resolution and resolve the diurnal cycle, Bodesheim et al. (2018) upscaled 30-minute fluxes of carbon and energy using randomized decision forests (Breiman, 2001), with a non-sequential modeling approach. Xiao et al. (2014) upscaled daily carbon and water fluxes in North America using moderate imaging spectroradiometer (MODIS) data with non-sequential ML approaches. Xu et al. (2018) evaluated different non-sequential ML methods to upscale ET with high-resolution features available regionally in China. Zhao et al. (2019) and ElGhawi et al. (2023) both used a non-sequential physics-constrained neural networks approach to model ET, which has the potential to yield physically consistent and partially interpretable models. Recently, Nelson & Walther et al. (2024) published an hourly upscaling product of carbon and energy fluxes (X-BASE), built upon a novel framework (FLUXCOM-X), which enables the testing and application of different data streams and ML methods for upscaling in a flexible manner. They use a non-sequential model based on boosted regression trees (XGBoost; Chen and Guestrin, 2016) and account for memory effects via remote sensing state proxies. ~~This framework and a similar data setup are also used within this study.~~

Non-sequential ML approaches, however, cannot represent temporal variable interactions beyond the observable state proxies in contrast to, for instance, recurrent neural networks (RNNs; Lipton et al., 2015). For time series regression, the long short-term Memory network (LSTM; Hochreiter and Schmidhuber, 1997) is a widely used architecture based on the RNN paradigm (Van Houdt et al., 2020). Such sequential approaches have been evaluated for EC flux modeling at the site level. Reichstein et al. (2018) applied RNNs to model weekly net ecosystem exchange of carbon (NEE) from 9 European flux stations with meteorological forcing and showed the relevance of temporal information via a permutation test. Besnard et al. (2019) employed an LSTM architecture to model monthly NEE at EC sites and achieve better performance ~~as~~ than with a non-sequential random forest. But still, they reported poor representation of temporal dynamics both in terms of interannual variability and anomalies, the deviations from the mean seasonal cycle.

In the domain of deep learning, different model architectures are capable of processing sequential data. In the Earth sciences, the LSTM has become the *de facto* standard, even though other architectures have been developed, such as the temporal convolutional network (TCN; Oord et al., 2016; Bai et al., 2018). The TCNs use sparse convolution along the temporal dimension to consider long-term effects more efficiently. More recently, models employing self-attention (Vaswani et al., 2017) have shown noteworthy performance in many domains. These sequential models could also hold potential for EC flux ~~model~~modeling, as has been shown by Armstrong et al. (2022) and Nakagawa et al. (2023). While conceptually apparent, there is little systematic evidence of whether such sequential deep learning methods provide an advantage over non-sequential approaches ~~with hand-designed features and state proxies~~ for the upscaling of EC fluxes, and about how these models respond to other issues ~~with~~ inherent in upscaling, such as limited and unevenly sampled training data and distribution shift from the local point data to gridded fields.

In this study, we provide a systematic comparison of different machine-learning approaches to the modeling of site-level ET fluxes and ~~upscaling to a global scale~~ subsequently their upscaling to the global scale using the FLUXCOM-X modelling framework (Nelson & Walther et al., 2024). A simple linear model, XGBoost, and a feed-forward fully connected neural network serve as baselines for non-sequential models. Two sequential models, one based on the LSTM architecture, and another based on a TCN, account for temporal effects. We chose these models as they are conceptually different but both commonly used for time-series simulation and forecasting tasks, and we acknowledge that other architectures could be used as well. We compare the model performances at the site level in a cross-validation setup and assess the relevance of dynamical memory effects for ~~land-atmosphere flux modeling, with specific attention to ET~~ ET modeling. For each model, we conduct a feature ablation experiment, where we drop feature groups. The groups considered in addition to meteorology are precipitation (which we obtained from reanalysis data and not from site-level observations due to large gaps), dynamic state representations, based on remotely sensed observations, and plant functional types (PFTs), which are static descriptors of site characteristics. We provide and investigate cross-validation–based upscaling ensembles from the independent cross-validation models to test for robustness. To assess the impact of the model architecture on upscaling, we contrast our products globally to a set of land surface model simulations and to a semi-empirical approach (GLEAM). We investigate, in the context of upscaling, a) whether the sequential models lead to more realistic and robust global and regional ET compared to independent estimates, b) whether

they are able to capture the temporal dynamics of ET better, and c) how robust they are to the covariate setup and training data subsets, compared to the non-sequential models.

The key contributions of this study are:

- A systematic comparison of the effectiveness of different ML methods for site-level ~~land–atmosphere~~ land-atmosphere ET flux modeling.
- An assessment and discussion of the relevance of different covariates in the context of ecological memory effects for ET.
- A characterization and comparison of an ensemble of upscaled ET estimates generated with different ML models.

2 Data sources and processing

We used hourly EC data from 2001 to 2020 processed by the ONEFLUX pipeline (Pastorello et al., 2020). Only sites available under the CC BY 4.0 license were included in this analysis, i.e., FLUXNET 2015 (Pastorello et al., 2020), ICOS Drought 2018 (Drought 2018 Team and ICOS Ecosystem Thematic Centre, 2020), ICOS Warm Winter 2020 (Warm Winter 2020 Team and ICOS Ecosystem Thematic Centre, 2022), or more recent ICOS or Ameriflux releases when present. In total, we used 287 sites with approximately 19 million hourly observations of ET and meteorological conditions~~distributed across 7.7 years per site, on average~~. The approach by Jung et al. (2023) was used for quality flagging. We used latent heat energy as target flux and convert it to ET assuming a constant latent heat of vaporization of 2.45 $\mathrm{MJ\,mm^{-1}}$. The following meteorological covariates were considered: near-surface air temperature (T_{air}), vapor pressure deficit (Δe), shortwave irradiation (R_{in}), potential shortwave irradiation ($R_{\mathrm{in,\,pot}}$), and time-derivative of potential shortwave irradiation ($\Delta R_{\mathrm{in,\,pot}}$). Note that the time-derivative, which is the difference between potential shortwave irradiation values for two consecutive hours, are intended to help the non-sequential models discern the diurnal cycle. As precipitation (P) observation at site level is often missing, we used the hourly ERA5 reanalysis instead (Hersbach et al., 2020), extracted from the nearest pixel to the site. In addition, we used remote sensing observations from the moderate imaging spectroradiometer (MODIS) sensor on board both Terra and Aqua satellite platforms, collection ~~v006.~~ 006. These include the enhanced vegetation index (EVI, Huete et al., 2002) ~~, the near infrared reflectance of vegetation (NIRv, Badgley et al., 2017),~~ and the normalized difference water index (NDWI, Gao, 1996), ~~all~~ both retrieved at site level from the MCD43A4 product ~~(Schaaf and Wang, 2015b, spatial resolution of 500 m)~~and quality filtered based on the MCD43A2 product (spatial resolution of 500 m, Schaaf and Wang, 2015a, b), and from MCD43C4 for the global data runs ~~(Schaaf and Wang, 2015c, spatial resolution of 0.05°)~~(spatial resolution of 0.05 °, Schaaf and Wang, 2015c). Additionally, the land surface temperature (LST) was obtained from MOD11A1 at site level (Wan et al., 2015a, spatial resolution of 1km), and from MOD11C1 globally (Wan et al., 2015b, spatial resolution of 0.05°). ~~Each remote sensing product was interpolated to daily resolution~~Although the MCD43A4 product for the reflectances uses observations from a period of 16 days to characterize and invert the bidirectional reflectance distribution function of a given pixel for the day at the center of the period, this operation is done over a temporally moving window at daily timesteps, resulting in output data with daily frequency. Processing of the datasets, cutouts at the sites, and quality control correspond to the set-up used in the FLUXCOM-X-BASE data set (Nelson &

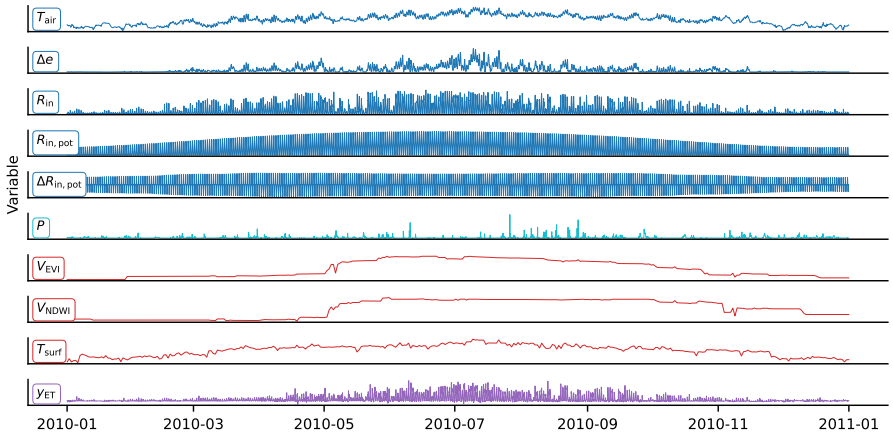

Figure 2. ~~Two-year~~ One-year time series from the Hainich site (DE-Hai) in Germany. Meteorological covariates (hourly): near-surface air temperature (T_{air}), vapor pressure deficit (Δe), shortwave irradiation (R_{in}), potential shortwave irradiation ($R_{\text{in, pot}}$), ~~and~~ time-derivative of potential shortwave irradiation ($\Delta R_{\text{in, pot}}$), and precipitation ($P$). Remote sensing ~~vegetation indices (interpolated to~~ daily): enhanced vegetation index (V_{EVI}), normalized difference water index (V_{NDWI}), and land surface temperature (T_{surf}). Land-atmosphere target flux (hourly): evapotranspiration (y_{ET}).

Walther et al., 2024; Walther & Besnard et al., 2022; Jung et al., 2023). As an optional covariate, we use the plant functional type (PFT), available for all EC station sites. The nine PFTs were one-hot-encoded and repeated in time to match the hourly time series. One-hot encoding represents categorical variables as binary values, assigning a unique binary digit to each category. Sample time series of the covariates and ET are shown in Fig. 2.

For upscaling, we used global meteorological data from the ERA5 reanalysis (Hersbach et al., 2020) corresponding to the site level variables. ~~For the remote sensing data, the same products were used for upscaling and for site level modeling.~~ The hourly data was spatially resampled to a resolution of 0.05° ~~spatial resolution~~ using bi-linear interpolation. This data was also used to fill gaps in site-level meteorological observations.

For the evaluation of the upscaling results, due to the lack of direct and spatially contiguous observations of ET, we used the Global Land Evaporation Amsterdam Model (GLEAM) v3 (Martens et al., 2017) and global sums of yearly ET from 14 land surface modes (TRENDY v6, values extracted from Pan et al., 2020) ~~)~~ as reference. Note that these reference data sources do not represent the ground truth, but are estimates derived using different approaches, independent from the data-driven upscaling performed here.

3 Methods

3.1 Experimental setup

We evaluate a set of sequential and non-sequential ML models at the site level in a ~~spatial~~ leave-sites-out cross-validation setup. The models are trained with different types of covariates: meteorological site-level observations without precipitation (`met`), precipitation from ERA5 (`prec`), remote sensing (`rs`), and PFTs (`pft`). ~~This~~ These experiments with different sets of variables as model inputs ~~,~~ are summarized in Tab. 1~~, gives insights into the relevance of the types of covariates~~. In total, ~~four~~ six covariate setups were tested and combined with five machine-learning models, i.e., ~~twenty~~ thirty models were trained and evaluated at the site level. For the evaluation, we use the Nash-Sutcliffe modeling efficiency (Nash and Sutcliffe, 1970)

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} \quad , \tag{1}$$

where y_t is the observed and $\hat{y}_t$ the predicted ET at time t, and $\bar{y}$ represents the mean of the observations. The NSE is calculated per site and can take values from $-\infty$ to 1 and reflects model performance relative to the mean of the observations. Values above 0 indicate better prediction than using the mean observations, and 1 is a perfect prediction. Note that for the evaluation of spatial patterns, the NSE was not computed per site but across sites, which corresponds to the R^2.

3.2 Modeling approach

With the goal of evaluating model performance at EC station locations and afterwards upscaling to the global scale, we tested a number of ML algorithms in a site-level cross-validation setup. We denote the modeling problem as

$$\hat{y}_{s,t} = f_\theta(\boldsymbol{X}_{s,t-K:t}, \boldsymbol{c}_s) \quad . \tag{2}$$

Here, $\boldsymbol{X}_{s,t-K:t} \in \mathbb{R}^{(K+1)\times D}$ are the D dynamic input covariates with up to K antecedent time steps, and $\boldsymbol{c}_s \in \mathbb{R}^M$ are the M static (constant) input features. The target flux of ET is represented as $\hat{y}_{s,t} \in \mathbb{R}$ at site s and time step t. Note that $K = 0$ with only instantaneous covariates $\boldsymbol{X}_{s,t}$ is a special case where no antecedent time steps are considered (i.e., a non-sequential model). We aim to find the parameters

$$\theta^* = \arg\min_\theta \mathcal{L}(f_\theta(\boldsymbol{X}_{s,t-K:t}, \boldsymbol{c}_s), \boldsymbol{y}_t) \tag{3}$$

of a function f_θ that minimize the loss function $\mathcal{L}$, given by the mean ~~square~~ squared error (MSE).

As baselines, we used a linear regression (`linearreg`) as well as two non-sequential models, a fully connected feed-forward neural network (`fcn`), and extreme gradient boosting (`xgboost`). The latter was also used in the recent state-of-the-art global upscaling product `xbase` (Nelson & Walther et al., 2024). The setup for these models was ~~kept constant, *i.e.*~~ largely consistent with `xbase` , i.e., the same covariates were used here plus precipitation. The remote sensing and PFT covariates were repeated in time for every hour to obtain uniform inputs~~.~~, i.e., the former were constant over a day and the latter over the entire time series. Although the remote sensing covariates do, in theory, vary on a sub-daily basis, these variables are not

Table 1. The ablation experiment with different covariate groups: Meteorological without precipitation (met, hourly), ~~plant functional type~~ precipitation (~~pft~~prec, ~~constant~~hourly), ~~and~~ remote sensing–based (rs, daily), and plant functional type (pft, constant). Each item corresponds to a unique ~~model~~ covariate setup.

Setup	~~Covariate groups~~ Covariates
met	~~Meteorology~~ T_{air}, Δe, R_{in}, $R_{\text{in, pot}}$, $\Delta R_{\text{in, pot}}$
met+prec	met + P
met+pft	~~Meteorology, PFT~~ met + S_{PFT}
met+rs	~~Meteorology, remote sensing~~ met + V_{EVI}, V_{NDWI}, T_{surf}
met+pft+rs	~~Meteorology, remote sensing, PFT~~ met + V_{EVI}, V_{NDWI}, T_{surf} + S_{PFT}
met+prec+pft+rs (=full)	met + P + V_{EVI}, V_{NDWI}, T_{surf} + S_{PFT}

met: near surface air temperature T_{air}; vapor pressure deficit Δe; shortwave irradiation R_{in}; potential shortwave irradiation $R_{\text{in, pot}}$; time-derivative of the potential shortwave irradiation $\Delta R_{\text{in, pot}}$; prec: precipitation P; rs: enhanced vegetation index V_{EVI}; normalized difference water index V_{NDWI}; land surface temperature T_{surf}; pft: plant functional type S_{PFT}.

available at the hourly resolution, and the diurnal variations are driven primarily by hourly-varying meteorological variables, though they interact with satellite-based features that change only on a daily to weekly basis. In addition to these non-sequential models, we used two sequential models: A simple LSTM architecture, a model able to learn temporal dynamics via its built-in memory processing mechanism, and a TCN model, which applies 1D convolutions in time. Those sequential layers were stacked to achieve the extraction of complex temporal features. While the LSTM has, conceptually, an unlimited receptive field, the temporal context considered by the TCN depends on its hyperparameters. The neural network-based models use the building blocks illustrated in Fig. 3 and were implemented in PyTorch (Paszke et al., 2019) v1.13.

3.3 Model training

To identify models with the capacity to generalize well to unseen sites, we trained them following an eight-fold cross-validation scheme, for which the data splitting between sites was kept identical across different models and ~~architectures~~covariate setups. To decrease the dependency between the sets, we ensure that sites in close spatial proximity (below 0.05° distance) are part of the same set using clustering of coordinates. The site groups are provided in the Appendix (Tab. B1). For each of the eight folds, six of the cross-validation sets were used for training (75%), one for validation (12.5%), and one for testing (12.5%),

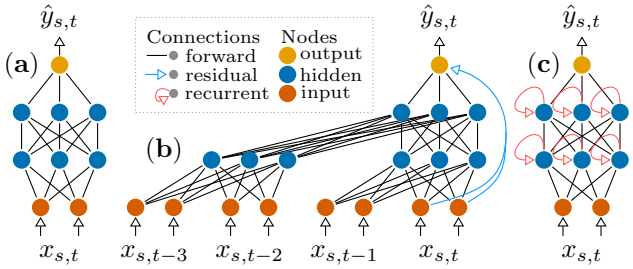

Figure 3. The three neural network layers used in this study: **a)** a feed-forward neural network, **b)** a temporal convolutional network (TCN), which applies causal (i.e., does not consider future time steps) 1D convolutions in the time dimension and **c)** a long short-term memory (LSTM) model, which uses recursion for information flow in the time dimension. The model inputs ($\boldsymbol{x}_{s,t}$) at site s and time t are mapped to the output $\hat{y}_{s,t}$.

such that each site appeared in the testing set once. The training and validation sets were used for model tuning with the early stopping algorithm: The model parameters were optimized on the training set, while the validation set was used to evaluate the generalizability regularly (ten times in each training epoch). Once the validation loss converged over a given number of validation steps (the "patience"), model training was halted, and the best parameters were restored. With these parameters, the model was applied to the independent test set. This approach yielded independent predictions for each site, which we then used to evaluate the model's performance on a site-level basis. For a speedup of the training, the model was iteratively fed with randomly selected sequences of two years. The first year was used for providing temporal context ~~(~~similar to the "spinup" in dynamic process models~~)~~, while the second was used for tuning. Note that the two years were randomly sampled in every epoch, ensuring that all observations were, potentially, used for training with high likelihood.

We used a random search over a predefined set of hyperparameters. For each model, 20 parameter sets were sampled uniformly with replacement. The sets are reported in Table A1 in the Appendix. Note that we selected hyperparameter ranges based on prior experiments, i.e., we excluded values that performed consistently badly in order to obtain a denser sampling of the sensitive ranges. With this protocol, we tuned hyperparameters independently for each model except for `linearreg`, which has no hyperparameters. The same hyperparameter set was used throughout the cross-validation for each model setup. Thus, the cross-validation ensemble is composed of models with the same hyperparameters, but trained on different subsets of the data.

To quantify model uncertainty on the site level, we performed the cross validation for the six best-performing models (in terms of validation MSE) from hyperparameter tuning for each model and setup individually. For later analysis and upscaling, only the best-performing model was used.

3.4 Upscaling

To achieve global coverage, we fed the models with harmonized and gridded data from 2001 to 2021 with 0.05° spatial and hourly temporal resolution. Due to the high computational demands of the upscaling, we decided to use only ~~the overall best covariates setup, which was `met+rs`, for~~ two well-performing covariate setups across all models. We ~~did not use the~~ selected the `met+rs` and the `met+prec+rs+pft (full)` setups as they were among the best-performing setups. The `linearreg` model ~~for upscaling~~was excluded, as it ~~showed significantly worse performance compared to~~ performed significantly worse than the non-linear algorithms in the site-level cross-validation. For each of the four remaining ML models, we compute an ensemble of eight upscaling products. The members, ~~herein~~ hereafter referred to as "cross-validation ensemble", correspond to the models obtained from the cross-validation folds, i.e., each fold yielded one model which was trained and evaluated on an independent set of sites. Note that this differs from the X-BASE setup (Nelson & Walther et al., 2024), where the cross-validation was used exclusively for model evaluation, and the upscaling was done with a single model trained again on additional sites without holding out a test set. This method does not yield an ensemble, and is, therefore, not suited for the evaluation of within-model upscaling robustness. The upscaled products are then evaluated by a ML model inter-comparison and by contrasting global yearly sums and regional cross-validation ensemble mean and variability to independent products.

4 Results and discussion

4.1 Site-level modeling of evapotranspiration (ET)

In this section, the EC site-level prediction of ET is evaluated based on the cross-validation setup. We aim to understand the impact of different covariate types and ML approaches on performance at different temporal scales and assess the relevance of sequential model architectures on reproducing ~~observed ET~~ ET observed at EC sites.

4.1.1 Model performances across scales

~~The overall~~ Figure 4 presents site-level ~~performance of hourly ET , shown in~~ ET accuracy in terms of NSE for different ML models and covariate groups across scales. We now focus on the best-performing models (solid lines in Fig. 4~~, depended more on~~). Overall, model outcomes were more influenced by the choice of covariates ~~rather than on the choice of the ML algorithm , except for the `linearreg` model, which performed poorly. However, we observed a strong interaction between the~~ than by the ML algorithm used. Notably, a significant interaction between ML models and covariates ~~. Figure 4 shows model performance in terms of the NSE for different ML models and covariate groups by temporal scales: the raw time series (`raw`), the daily average (`daily`), the mean seasonal cycle (`seasonality`), the daily anomalies (`anom`, the deviation from the seasonality), and the interannual variability (`iav`, the year-to-year variability). Overall, linear regression was outperformed by the ML models by a large~~ was observed.

In general, the ML models outperformed linear regression by a substantial margin. On the `raw` and `daily` time ~~scale, the sequential models performed best , with a relatively stable NSE of about 0.75 and 0.64~~ scales, sequential models exhibited the

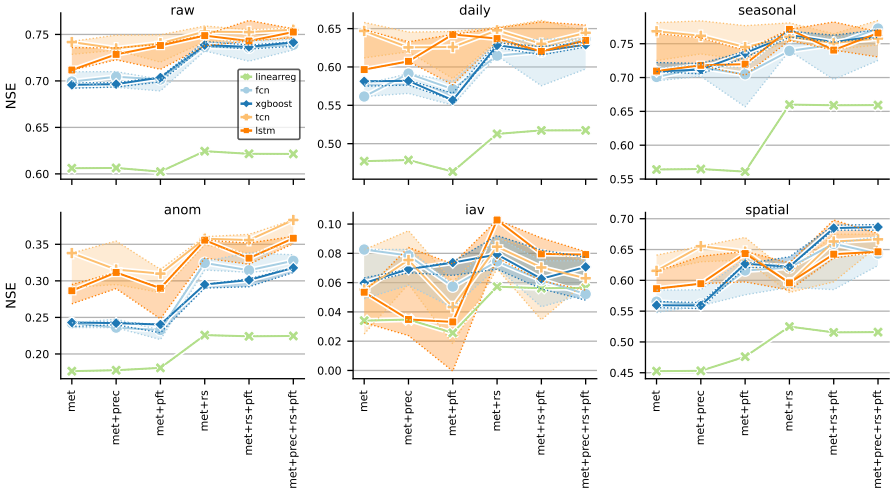

Figure 4. Site-level evaluation for modeled evapotranspiration: Median Nash–Sutcliffe model efficiency (NSE) across sites for different models (lines) and covariates (x-axis) for various scales (panels). The shaded area represents the range of the top six models from hyperparameter tuning per model setup, with the solid line indicating the best-performing model (top one). This can be interpreted as model uncertainty. Note that the model selection during hyperparameter tuning was based on the validation set and the mean squared error. Consequently, the best model is not necessarily the top performer in terms of NSE. The scales shown are `raw` for hourly, `daily` for daily aggregates, `seasonal` for daily seasonal, `anom` for daily anomalies (`daily` minus `seasonal`), and `iav` for interannual variability. The `spatial` scale reflects the NSE across site mean values, indicating the ability of the models to capture spatial variability. For certain temporal scales, some sites had to be excluded due to missing or infinite values; the number of sites used is indicated in the respective panel title.

best performance, maintaining stable NSE values of 0.70-0.75 (`raw`) and 0.60-0.65 (`daily`) across data setups~~, respectively. The non-sequential ML models showed a significant increase in performance from the `met` and `met+pft` setup to the setups including remote sensing observations, where they achieved similar performance as the~~ . Non-sequential models performed worse when using only meteorological covariates but showed significant improvement when remote sensing covariates were included, achieving similar performance to sequential models.

On the seasonal scale~~and without remote sensing covariates, the sequential models performed best. With~~ , sequential models outperformed others in the absence of remote sensing covariates. However, when remote sensing covariates were included, the performance differences between models became less pronounced. For anomalies, ~~the `tcn` model performed the best, and the `lstm` the worst, while differences among models remained relatively small (see y-axis range). For the anomalies, all models benefited from~~ adding remote sensing covariates ~~, and the sequential models performed consistently better across all data setups. For the~~ enhanced model performance across all setups, with sequential models consistently outperforming non-sequential models. Notably, this was the only scale where precipitation had a clear positive impact across all models compared to the `met+rs+pft` setup.

For interannual variability, ~~all models show very low performance and~~ model performance was generally poor across all setups, with no clear patterns, and the small y-axis range underscores this. Adding PFTs and precipitation did not improve, and on the contrary in some cases, reduced performance. On the spatial scale, model performance was similar across models, although sequential models exhibited a slight advantage with the ~~patterns are less clear; while the `lstm` improved with adding additional covariates, the other modelsshowed a decreased performance with PFT~~`met` and `met+prec` setups. Including PFT covariates notably improved performance for all models on this scale.

~~Site-level evaluation for modeled evapotranspiration: Median Nash–Sutcliffe model efficiency (NSE) across sites for different models (lines) and covariates (x-axis) for different temporal scales (panels). The scales shown are `raw` for hourly, `daily` for daily aggregates, `seasonal` for daily seasonal, `anom` for daily anomalies (`daily` minus `seasonal`), and `iav` for interannual variability. For certain temporal scales, some sites had to be removed due to NaN or Inf values; the number of sites used is indicated in the respective panel title.~~ Regarding model uncertainty (shaded areas in Fig. 4), sequential models generally showed lower robustness than non-sequential models, likely due to their added complexity. Sequential models consistently outperformed non-sequential ones on the `raw` scale. On the `daily` and `seasonal` scales, the differences between sequential models were primarily driven by model uncertainty, i.e., the uncertainty ranges displayed in Fig. 4 overlapped strongly on the spatial scale. On the anomaly scale, sequential models reliably outperformed non-sequential models across the top six setups. Given the small performance range and large model uncertainties on the `iav` scale, caution is needed when interpreting these differences. Similarly, the differences on the spatial scale were minimal, and the relative uncertainties were large.

Next, we discuss the broader implications of these findings. The linear models (`linearreg`) ~~fell short consistently~~ consistently underperformed because evapotranspiration is ~~characterized~~ governed by complex interactions and non-linear functions~~such that~~ , making the advantages of ML ~~become noteworthy. While the sequential models only marginally improved with~~ methods particularly evident. While sequential models showed only marginal improvement when adding covariates related to ecosystem state, ~~the~~ non-sequential models ~~improved more prominently. This is, on the one hand,~~ exhibited more substantial improvements. This serves as a sanity check for the sequential models: ~~They~~ they were able to extract additional information from the temporal meteorological covariates, as expected. ~~Still, adding~~ However, incorporating remote sensing covariates improved and stabilized their performance. ~~On the other hand, this shows~~ Conversely, this suggests that remote sensing covariates ~~are~~ serve as useful proxies for ecological memory: ~~The sequential models were able to~~ while sequential models could extract additional information from antecedent covariates, ~~but most of the information seems to be comprised in~~ most of this information appeared to be contained in the remote sensing covariates, ~~and thus, the~~ allowing non-sequential models ~~achieved~~ to achieve similar performance. ~~This could be interpreted as follows: Consider a drought at an EC site. By using past meteorology, the sequential models can infer a severe water deficit. In contrast, the non-sequential models do not have access to such information. The drought stress is, to a certain extent, also reflected in the vegetation indices, and therefore, all models with access to these covariates are informed about the drought event.~~

On the anomaly scale, ~~however,~~ we observed a more ~~distinct performance increase also~~ noticeable performance increase when considering remote sensing even for the sequential models~~, and the `lstm` model in particular~~. This is ~~noteworthy, as the anomalies are highly relevant to study and quantify ecosystem response~~ important, as anomalies are crucial for studying

and quantifying ecosystem responses to uncommon or extreme conditions. This ~~could be related to processes that are partially~~ performance increase could be linked to processes observable by remote sensing ~~, but cannot be derived~~ but not directly derivable from meteorology, such as ~~forest or crop management and~~ the effect of management on crops and forests and natural disturbances. Here, adding precipitation improved the performance of all models. This is not surprising, as precipitation is a key driver of ET through its influence on soil moisture (Nelson et al., 2020). However, the improvement was comparatively small, which may be due to the use of precipitation reanalysis data that may not fully represent local conditions. The low performance on the `iav` scale was also reported by Jung et al. (2019) and Nelson & Walther et al. (2024).

It is ~~notable~~ worth noting that adding PFTs as covariates did not improve~~(and sometimes even harmed) model performance .~~, and in some cases even reduced, model performance on the temporal scales. On the spatial scale, however, their inclusion was beneficial, highlighting the potential importance of spatial covariates for modeling EC fluxes and upscaling. PFTs have long been criticized for not ~~being representative of~~ accurately representing the continuous characteristics of ecosystems (Reichstein et al., 2014; Kattge et al., 2011). ~~As our experiments suggest , adding PFTs brings little to no information~~ Our experiments suggest that while adding PFTs provides additional information for representing spatial patterns, they may also harm extrapolation on other scales, arguably due to the inflated covariate space; indeed, ~~while increasing the input features space; in fact,~~ each of the nine PFTs ~~adds~~ introduces another input dimension due to ~~the~~ one-hot encoding. ~~This can—in general, and particularly here due to the data-limitedness of the modeling problem—deteriorate model performance, supposedly due to additional overfitting on the sparse information provided by the PFTs. We , therefore , suggest not to use the full stack of PFTs as covatiates for EC flux modeling, but we encourage the exploration of alternatives.~~ We therefore recommend exploring alternative spatially continuous variables, such as soil properties or plant traits~~. This finding advocates for a~~ , that could summarize ecosystem functional properties.

These findings advocate for comprehensive feature selection to identify more relevant static features~~and, therefore, to avoid inflating~~ , thereby avoiding unnecessary inflation of the input dimensionality. Alternatively, or in addition, location embeddings, such as SatCLIP (Klemmer et al., 2024), could ~~help improve model generalizeability~~ improve model generalizability by providing a condensed representation of land surface characteristics.

4.1.2 Memory effects matter

As ~~noted before, the difference in model performance~~ previously mentioned, the performance gap between the non-sequential and sequential models ~~shrank~~ decreased when remote sensing observations were ~~added~~ incorporated as covariates. We ~~investigate~~ explore these differences in Fig. 5: As ~~illustrated in the top left~~ shown in the top-left panel (Fig. 5a), which ~~shows~~ displays the absolute error difference between the ~~the~~ `lstm` and `xgboost` ~~, the~~ models, non-sequential models ~~perform worse with~~ performed worse when high incoming radiation (> 200 Wm^{-2}) was paired with either low or high observed ET. To represent these conditions, ~~there must be an implicit knowledge about the water availabilitylearned by the models. It seems~~ the models must implicitly learn about water availability. It appears that the sequential model was able to learn proxies of wetness from the meteorological time series, ~~but~~ whereas the non-sequential model was ~~not. Rather, the latter~~ unable to do so. Instead,

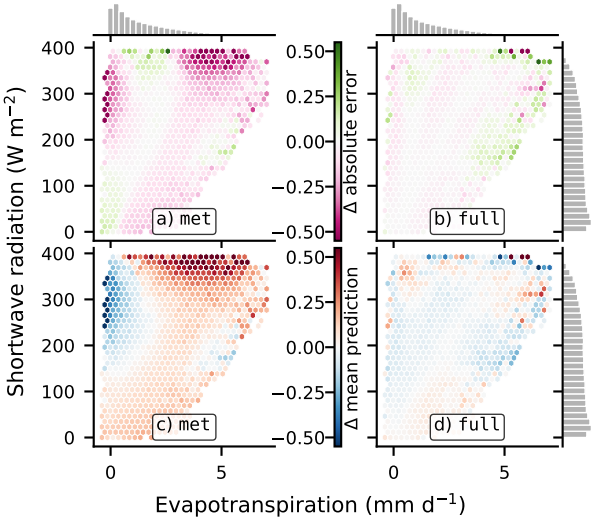

Figure 5. Comparison of a sequential (`lstm`) and a non-sequential (`xgboost`) model in terms of absolute error and mean predicted ET in the space of observed evapotranspiration × shortwave irradiation: Panel **a)** and **b)** show the *difference in absolute error* ~~of~~ between the `lstm` ~~minus the~~ and `xgboost` ~~model~~models, ~~**a)**~~ with **a)** showing the difference when only meteorological covariates ~~only~~are used, and ~~**b) when using**~~ **b)** showing the difference when precipitation, remote sensing, and PFT covariates ~~in addition~~(`full` setup) are included. Here, magenta ~~represents~~ indicates cases where the sequential model ~~performs better~~outperforms the non-sequential model, and green vice-versa. The bottom panels **c)** and **d)** show the *difference in mean predicted ET* ~~of~~ between the `lstm` ~~minus the~~ and `xgboost` ~~model~~ models for the respective covariate setups. ~~Here~~In these panels, red colors indicate an underestimation of ET by `xgboost` compared to `lstm`, and blue ~~vice-versa~~indicates the opposite. The histograms represent the marginal data distribution.

the non-sequential model seems to have learned an average behavior, which ~~worked~~ performed well in most ~~instances. As~~ situations.

When remote sensing covariates were added as well as PFTs and precipitation, the differences in performance were reduced but ~~did not entirely disappear~~ not entirely eliminated (Fig. 5b). This interpretation is ~~supported by~~ further supported by the bottom panels, Fig. 5c-d, which show the difference in mean predicted ET between the sequential and ~~the~~ non-sequential ~~model. Without access to the remote sensing~~ models. With access to only meteorological covariates (Fig. 5c), the non-sequential model overestimated ET with high incoming radiation but low observed ~~ET; these are~~ ET—representing dry conditions that the model failed to ~~identify. On the contrary, large~~ recognize. In contrast, high observed ET was underestimated by the non-sequential model~~; these are, supposedly, ~~, which likely corresponds to wet conditions. ~~When adding the remote sensing~~observations as After incorporating remote sensing, PFTs, and precipitation covariates (Fig. 5d), the ~~differences were reduced significantly~~performance differences were substantially reduced. This comparison ~~illustrates why memory effects play a role in modeling ET and~~ underscores the importance of memory effects in ET modeling and illustrates how remote sensing covariates~~are good, but not perfect~~ , while useful, are not perfect proxies for ecological memory.

~~The temporal context length considered by the lstm cannot be quantified easily. It can—in principle—, access a minimum of one year (the spinup time) and a maximum of two years (the sample sequence length) of context during training. For the ten, the context length depends on the tuned hyper-parameters: While the lstm processes the entire time series sequentially, the ten's context depends on the number of layers and the kernel size. For ten, the temporal context was 19 days for the met setup, 9 days for the met+pft setup, and 4 days for the setups including the remote sensing covariates, met+rs and met+pft+rs (also see Appendix A). This, again, indicates that the remote sensing features are well-suited proxies for ecosystem memory, as a model having access to these observations needs a shorter context of meteorological conditions (19 versus 4 days).It is, however, not clear why the met+pft setup works best with a shorter context of 9 days. This could be an artifact of the random search for the hyperparameter tuning or because PTFs contain some information about the climate, providing a shortcut to bypass the extraction of temporal features containing similar information. Overall, it is somewhat surprising that additional context did not improve model performance, as ecological memory can span across multiple months or even years (Ogle et al., 2015; Besnard et al., 2019; Kraft et al., 2021). We hypothesize that the sparse nature of extreme events (e.g., disturbances or droughts, which can have long-term effects) and biases in the observations (Jung et al., 2023) pose a challenge for the ML models to identify the more fine-grained, long-term memory effects.~~ Furthermore, including precipitation may be beneficial for the sequential models in principle, as these models could learn soil moisture dynamics. However, as noted earlier, the performance increase was not as pronounced as expected (Fig. 4), possibly due to the use of reanalysis data that may not fully capture local conditions.

4.2 Scaling evapotranspiration to global coverage

With the models optimized at the site level, we create global ensembles of ET estimates for the met+rs and the full (met+prec+rs+pft) covariate setups. The ensemble members were trained with different subsets of the training data within the cross-validation scheme. At the site level, the differences between ML models were small when considering remotely sensed observations or PFTs as covariates. However, when scaling globally, data distribution shifts can (and will) affect different model types in different ways. ~~The shifts evolve~~ These shifts arise from the different scales of the measurements (point at EC site versus grid globally), the different data products used (direct observation of meteorological variables at EC site versus reanalysis globally), and the spatial extrapolation into different ecoclimatological conditions from irregularly and sparsely sampled locations. In this section, we consider the performance of the different ML approaches and covariate setups while scaling out of the ~~flux~~ EC station locations.

4.2.1 Global patterns of evapotranspiration

~~On the global scale, the ML ensembles yielded mean annual sums of ET within less than 10 percent difference, while the neural networks showed considerably larger ensemble spread than xgboost. Global annual ET (the cross-validation ensemble mean) amounted to about $65 \cdot 10^3$ for the~~ Figure 6 shows global annual ET estimates for all ML models and the two selected covariate setups. With the met+rs setup, non-sequential models ~~xgboost ($65.3 \pm 0.9 \cdot 10^3$) and fcn ($65.6 \pm 5.0 \cdot 10^3$)~~ xgboost, ~~in agreement with~~ fcn) estimated about $65 \cdot 10^3$ $\mathrm{km^3\,y^{-1}}$ global annual ET, which is close to the land surface

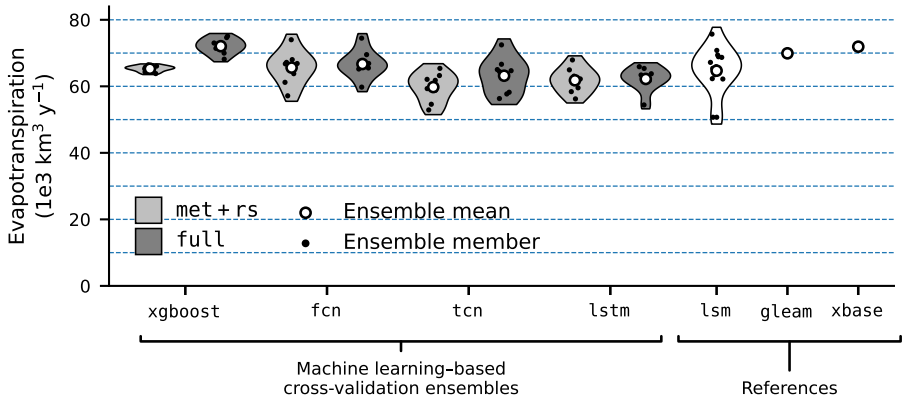

Figure 6. Global annual evapotranspiration (ET) per model. The violin plots represent the density of independent cross-validation runs (black dots), with their mean values across runs displayed as white dots. For the machine learning models, the `met+rs` setup (meteorological and remote sensing covariates) is shown in light grey, while the `full` setup (meteorological, precipitation, remote sensing, and PFT covariates) is shown in dark grey. The data from a number of land surface models (`lsm`), the GLEAM product (`gleam`), and FLUXCOM X-Base (`xbase`) are added as reference.

models ~~ensemble mean~~ (`lsm`) ~~of~~ ensemble mean ($64.7 \pm 6.9 \cdot 10^3$ km^3 y^{-1}), while sequential models (`tcn`, ~~and to about $60 \cdot 10^3$ for the sequential models `tcn` ($59.8 \pm 4.3 \cdot 10^3$) and~~ `lstm`~~($61.8 \pm 3.7 \cdot 10^3$) on average (Fig. 6). This amounts to a~~) predicted roughly 8% lower ~~annual ET estimate by the sequential models. Both `gleam` and `xbase` estimate a larger global ET around 70×10^3~~ values (59.8 and $61.8 \cdot 10^3$ km^3 y^{-1}~~. The large range of neural network-based models was encompassed by the large spread of results from an ensemble of land surface models (`lsm`), whereas the spread between~~, respectively). With the `full` setup, `xgboost` ~~members was considerably smaller~~showed the strongest increase in global ET (+10%), while `fcn`, `lstm`, and `tcn` increased modestly (1%, 0.5%, and 5%, respectively). `xgboost` was closest to `gleam` and `xbase`, whereas neural networks aligned more with `lsm` estimates.

~~Overall, the ML models showed consistent patterns of spatial mean , while systematic deviations are evident in mid to low latitudes. This is shown in~~ Figure 7 displays spatial patterns of ensemble mean ET per model and covariate setup. The models aligned well spatially in the `met+rs` setup (Fig. ~~??a, which represents the spatial model ensemble means per grid cell of `xgboost` (most left), and its difference to the means predicted by the other ML models . While the differences were low in Northern America, Europe, and Central Asia, we saw larger discrepancies elsewhere. In sub-equatorial zones of Africa, South Asia and the Himalayas, `xgboost` estimated larger ET on average than `fcn`, and vice-versa for arid to hyper-arid deserts. Compared to the sequential models (the two right-hand side panels in~~ 7a), with lower ET by sequential models especially in regions in the Southern Hemisphere which have sparse coverage by EC stations. The `full` setup (Fig. ~~??a) , `xgboost` estimated larger ET globally, but in particular in the tropics and sub-tropics, but not for rain forests and deserts. In the temperate zone, the differences were marginal. Still, the bias between models on grid-cell level was relatively small compared to the uncertainties of the site-level ET measurements (Bambach et al., 2022)~~7b) revealed stronger divergences across models,

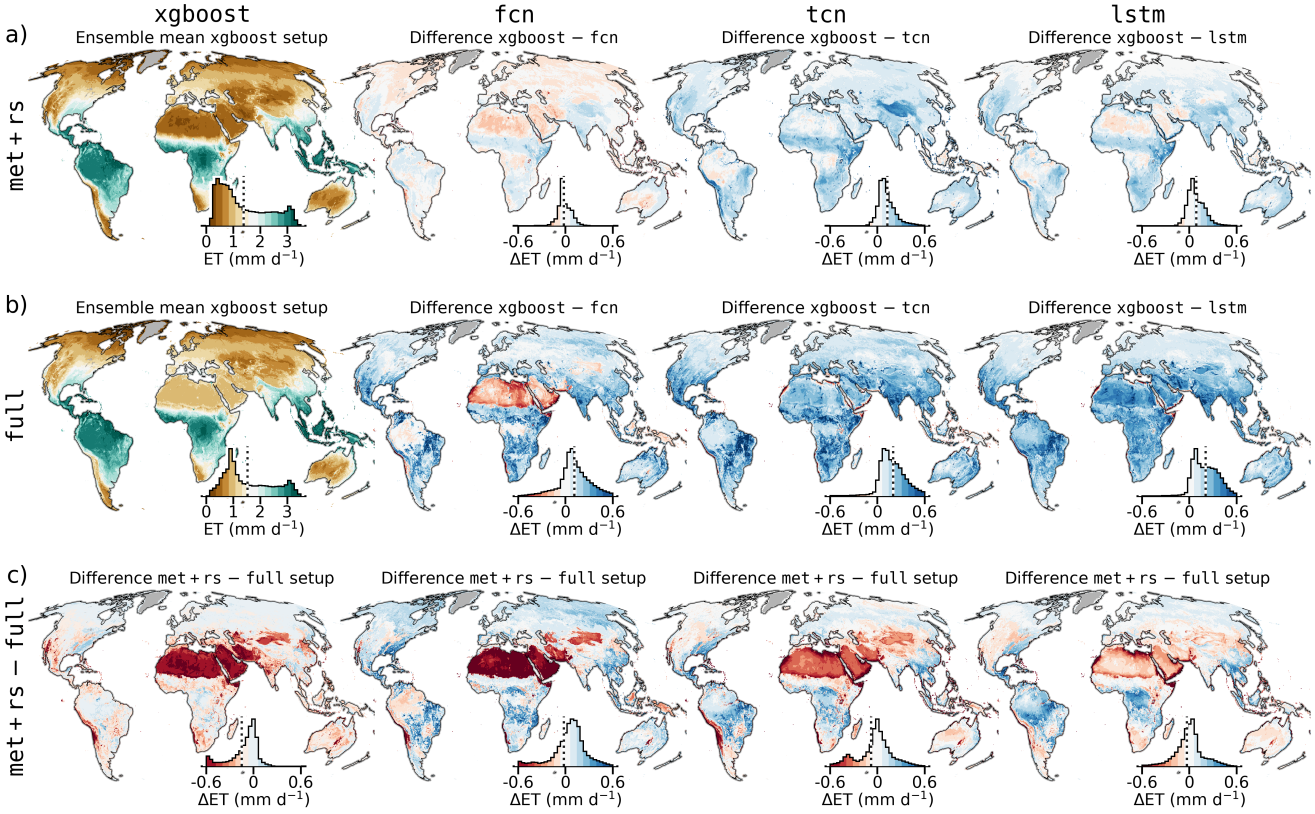

Figure 7. Spatial model evaluation and comparison for the `met+rs` and `full` (meteorological, precipitation, remote sensing, and PFT covariates) setup. ~~**a)** Grid level~~ Grid-level cross-validation ensemble mean ET for ~~the `xgboost`~~ each model is shown in the ~~leftmost map~~top row (**a**)) for `met+rs`, and the ~~difference between~~ middle row (**b**)) for `full`. The first column represents `xgboost`, and the ~~neural network–based models is shown in the~~ remaining columns show the difference to `xgboost`. ~~**b)** The grid level median absolute deviation~~ bottom row (~~MAD~~**c**)) ~~per ML model quantifies~~ shows the ~~cross-validation ensemble uncertainty~~ difference `met+rs` minus `full`. The color scale is consistent across panels and reflects mean ET in $\mathrm{mm\,d^{-1}}$. ~~The map inset~~ Inset histograms ~~represent~~ show the distribution of ~~the~~ values weighted by ~~the~~ grid cell area; the median is ~~shown as~~ indicated by a dashed black line.

particularly in arid zones such as the Sahara and Arabian Peninsula and generally in the subtropic zones. Figure 7c shows that changes in ET estimates due to covariate setup differences were low in temperate zones but more pronounced in arid and tropical regions. The `lstm` appeared relatively robust across setups.

~~In terms of spatio-temporal patters, all ML models showed a very similar agreement with `gleam`. This is shown in the~~ 430 ~~diagonal panels in Fig. **??**, which represents the monthly ET values of the ML models versus `gleam`, with pooled temporal and spatial dimension. The linear correlation was between $r = 0.86$ and $r = 0.87$ in all cases. In the upper triangular panels in~~ A temporal anomaly correlation analysis (Fig. ~~**??**, which displays the realtionship between pooled spatio-temporal anomaly values, we see that the linear correlation was between $r = 0.92$ and $r = 0.97$. Here, the strongest correlation ($r = 0.97$) was~~

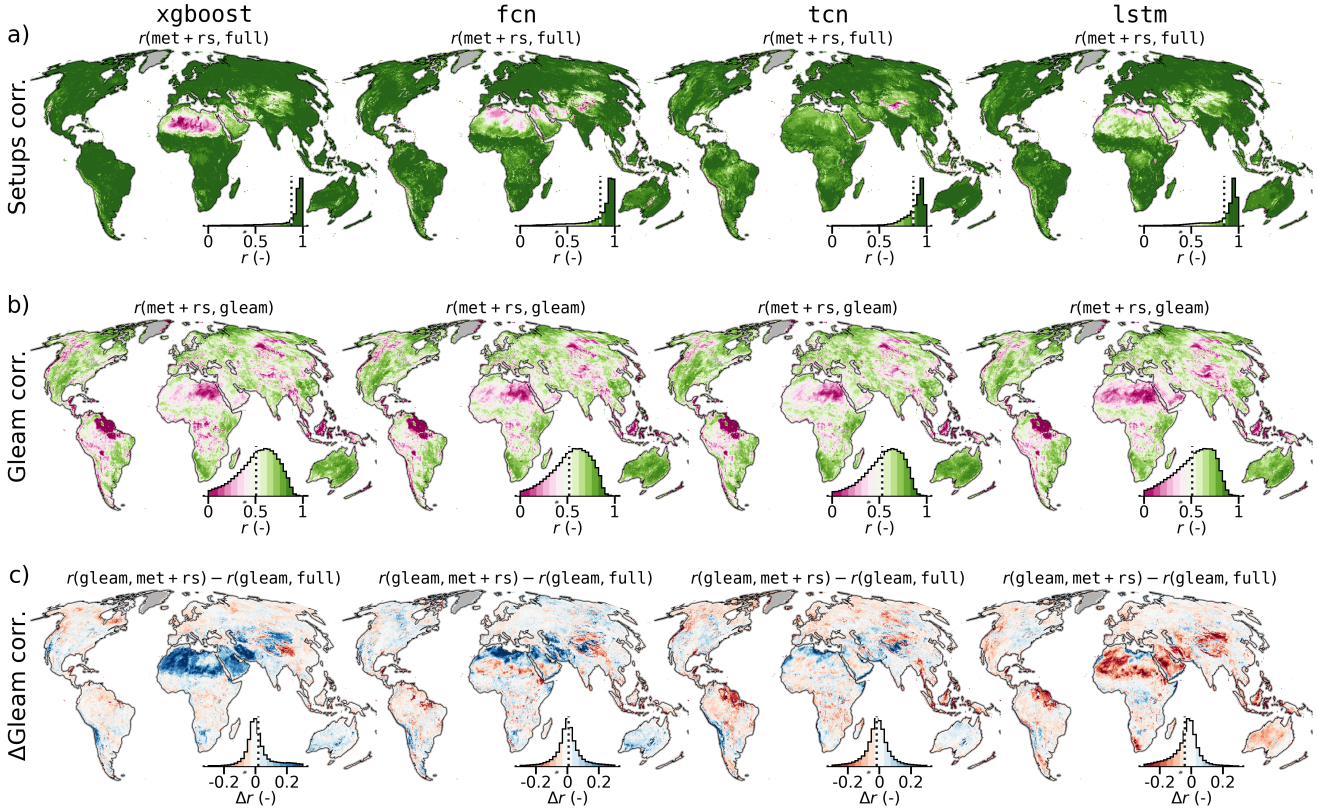

Figure 8. ~~Comparison~~ Temporal correlation of monthly ET anomalies among ~~machine learning (ML)~~ models and ~~with the GLEAM product (`gleam`)~~covariate setups. ~~All panels represent relationships (log-density) of pooled spatial~~ **a)** Correlation between monthly ET anomaly time series from `met+rs` and ~~temporal values, axes have units~~ `full` setups. ~~The strength~~ **b)** Correlation between monthly ET anomalies of `met+rs` and the ~~relationship is quantified~~ GLEAM product. **c)** Difference in temporal correlation with GLEAM between the ~~Pearson correlation (r)~~`full` and `met+rs` setups. ~~The lower triangular (magenta hues) shows~~ Inset histograms show area-weighted value distributions with dashed black lines indicating the ~~relationship of the~~ median~~absolute deviation (MAD), i.e., the cross-validation ensemble spread, between ML models~~. ~~The diagonal (blue hues) shows the association of the ensemble mean per ML model (y-axis) with `gleam` (x-axis). The upper triangular (orange hues) displays the relationship between ensemble mean ET anomalies among the ML models.~~

~~found between the sequential models.The weakest relationships were found between `xgboost` and the sequential models ($r = 0.92$ and $r = 0.93$).~~ 8) reveals high agreement ($r > 0.8$) between covariate setups in most areas, except for arid zones, and only the `tcn` maintained high consistency globally. A comparison with `gleam` (Fig. 8b) showed moderate correlations (mean $r \approx 0.5$), with discrepancies mainly in tropical and arid regions. Figure 8c highlights where models improved alignment with GLEAM when using additional covariates: `tcn` improved slightly, while `lstm` improved most consistently across dry regions.

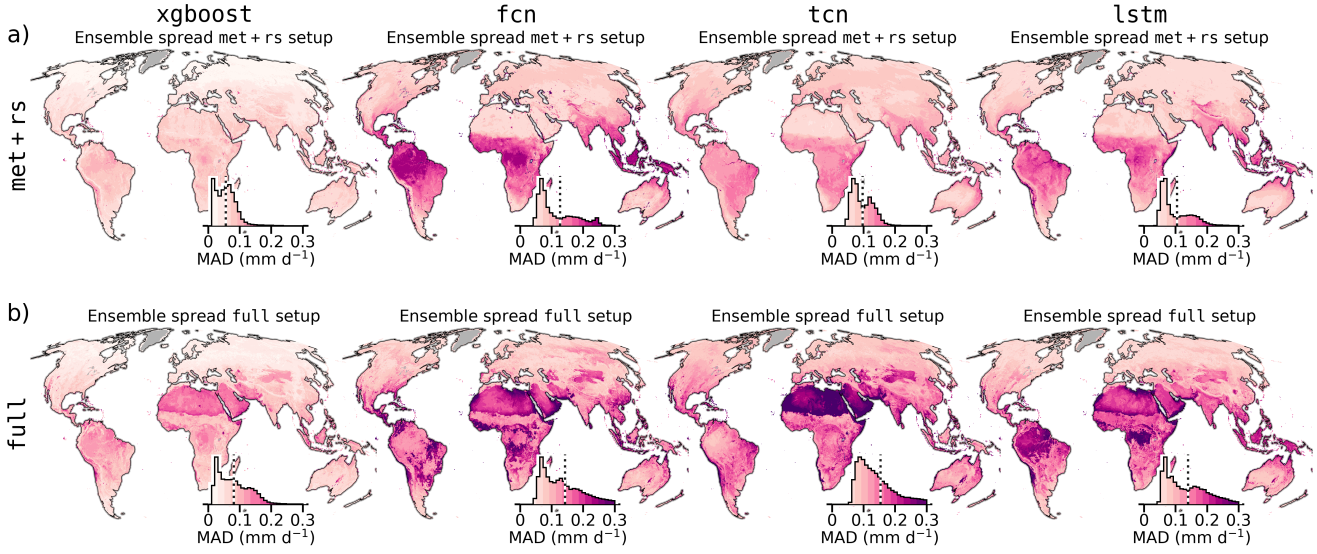

Figure 9. Median absolute deviation (MAD) of the cross-validation ensemble for each model setup, representing spatial uncertainty in mm d^{-1}. **a)** MAD for the `met+rs` setup. **b)** MAD for the `full` setup. Inset histograms show the distribution of values weighted by grid cell area; the dashed black line marks the median.

~~The neural network-based models exhibited considerably larger ensemble spread mainly in the tropics, as displayed in Fig. ??b. The figure shows the grid cell~~

Figure 9 shows uncertainty via median absolute deviation (MAD) ~~per model in , computed on a monthly scale and averaged across time afterwards. Here,~~ across ensemble members based on monthly values. `xgboost` ~~(left-hand side panel) had a low ensemble spread in general, with slightly larger values~~ had the lowest spread overall, with modest increases in tropical and ~~sub-tropical regions and moderate hot spots in rainforests. The other models showed a considerably larger ensemble spread. The `fcn` yielded the largest spread, with high values in the topical zone. The sequential models showed a lower spread than `fcn`, but the spatial distribution aligned well. The ensemble spread of the neural networks did not show a strong agreement in terms of spatio-tempoal patterns . The lower triangular panels in~~ more pronounced increase in arid zones in the `full` setup. The neural networks showed higher ensemble spread: in the tropics and subtropics for the `met+rs`, and in arid regions for the `full` setup. In the latter, uncertainty patterns aligned with ecosystem boundaries, likely reflecting PFT transitions.

We saw that models showed biases in terms of global ET sums, but apart from `xgboost`, the values were relatively robust to changes on covariate sets. From the spatial and temporal patterns (Fig. 7, Fig. ~~?? indicate that the association was weak between `tcn`~~ 8, and ~~`fcn` ($r = 0.64$), as well as between `tcn` and `lstm` ($r = 0.71$), and slightly larger between `lstm` and `fcn` ($r = 0.82$). Interestingly, the relationship between the neural network-based models and~~ Fig. 9), it becomes evident that the uncertainty originates mainly from tropical and arid to semiarid regions. Finally, we want to investigate if the differences are related to the technical setup (architecture and covariates) or to the variability in training data used in cross validation.

Table 2. Global evapotranspiration (ET) correlation across model ensemble members. The correlation is based on global ET sums across cross-validation members trained on the same sites. The upper triangle (bold) shows Pearson correlation, the lower triangle Spearman rank correlation.

		xgboost		fcn		tcn		lstm	
		rs	full	rs	full	rs	full	rs	full
xgboost	rs	1.00	**0.90**	**0.25**	**0.23**	**0.63**	**0.45**	**0.67**	**0.20**
	full	0.76	1.00	**0.57**	**0.36**	**0.58**	**0.65**	**0.80**	**0.39**
fcn	rs	0.12	0.71	1.00	**0.58**	**0.16**	**0.43**	**0.44**	**0.78**
	full	0.26	0.62	0.52	1.00	**0.20**	**0.25**	**-0.08**	**0.45**
tcn	rs	0.33	0.38	0.29	-0.05	1.00	**0.75**	**0.39**	**0.08**
	full	0.40	0.81	0.81	0.38	0.62	1.00	**0.64**	**0.09**
lstm	rs	0.43	0.55	0.52	-0.12	0.38	0.62	1.00	**0.32**
	full	-0.10	0.36	0.76	0.31	0.10	0.33	0.48	1.00

Therefore, we investigate the alignment of members shown in Fig. 6: We compute the linear (r) and rank (ρ) correlation between global sums of ET between all models and covariate setups. In other words, we quantify if using the same data subset in cross validation leads to a consistent upscaling behavior in terms of global ET. The results are shown in Table 2. Global ET estimates from `xgboost` ~~was not much lower with values from $r = 0.59$ to $r = 0.69$~~were highly correlated between setups ($r = 0.90$, $\rho = 0.76$), suggesting consistent behavior across training data subsets. Neural networks exhibited lower correlation ($r = 0.32$–0.75), indicating greater variability. Overall, consistency in upscaling behavior was not strongly tied to model type or covariate setup.

~~We identified three noteworthy features of our upscaling results, which we discuss in more depth in the following subsections. First, we discuss the lower global integral of predicted ET of our upscaling results compared to the similar `xbase` approach.Second, we consider the lower mean of our upscaling results, with a focus on the low ET predicted by the sequential models . Third, we have a closer look at the larger ensemble spread of the neural networks when compared to `xgboost`.~~

4.2.2 Model biases across architectures and covariate setups

~~The global ET estimates, especially by the sequential neural networks, were low compared to our current understanding of global ET magnitude of~~ Most models, particularly the sequential ones, estimated global ET—averaged across members—at the lower end of current estimates from independent methods of about $70 \pm 5 \cdot 10^3$ km^3 y^{-1} ~~based on a variety of methods (Jung et al., 2019)~~(Jung et al., 2019), and compared to `gleam` and `xbase`. ~~Some underestimation in our global ET estimates would be expected due to the systematic energy balance closure gap problem across flux station sites (Stoy et al., 2013; zha, 2023) for which no correction has been applied here. The associated uncertainty for global ET is estimated to be about~~ Energy-balance

non-closure at EC sites could depress ET by up to 20% (Jung et al., 2019)~~and could explain the apparent underestimations, while multiple indications suggest only a comparatively small underestimation of global ET due to the energy balance closure gap problem (Mauder et al. in review, will add in future version). However,~~ , yet recent findings suggest that the closure gap may contribute only moderately to the global ET bias, as the closure gap very likely stems primarily from sensible rather than latent heat fluxes, i.e., ET (Zhang et al., 2024; Mauder et al., 2024). Because both `xgboost-full` and the `xbase` ~~approach suffers from the same issues, but estimates a relatively large global ET. Spatially, the differences originate mostly from mid to low latitudes. These are regions that are generally underrepresented by EC sites (see Fig. 1b and Jung et al., 2011, 2019; Nelson & Walther et a . The larger estimates of ET by `fcn` and `lstm` in the Sahara and Arabic desert are notable, as Nelson & Walther et al. (2024) already pointed out an overestimation in arid regions by their `xbase` setup, which was based on the XGBoost method~~product would share the same closure issue but still reach the benchmark range, the deficit of the other models must originate elsewhere; options are training data subsets, the ML approaches, covariate setup, or data distribution shift due to extrapolation into underconstrained regions and product types (site level versus gridded).

~~The inconsistencies between `xbase` and our results were surprising, as the same framework, and in the case of~~ Disentangling these possible causes is difficult, but we can draw some conclusions from the results. We have shown that different training subsets did not lead to consistent upscaling behavior (Table 2). In other words, taking the same training data subset for model training did not lead to similar behavior across ML models in terms of global ET. This suggests that the ensemble variance at global scale is not driven by training data, but rather by how the ML models extrapolate out of the training data distribution. It indicates that, when evaluating the global sums alone, neither neural networks versus `xgboost` ~~even the same machine-learning model , was used. A key difference between our approach and `xbase` is the cross-validation setup and the way how the upscaling was done after training. While we used each member of the 8-fold cross-validation for upscaling, `xbase` was based on 10-fold cross-validation, and a final model was trained on nine folds for training and one for early stopping. This re-training of `xbase` on additional data may have a positive impact on model quality, yet it yields only one upscalingproduct instead of an ensemble. However, it seems unlikely that this retraining caused an increase of about 10% from `xgboost` to `xbase`. Furthermore, `xbase` uses some additional input variables , namely PFT (which we did use for the site-level experiment but not for the upscaling), near-infrared reflectance of vegetation (NIRv), and nighttime land surface temperature. While all of those variables may have had an impact on the upscaling result, we posit that the PFT had the largest impact. The PFTs, or a subset thereof, may provide valuable information for site-level modeling and spatial extrapolation. Considering the relatively small number of approximately 215 EC training sites per cross-validation fold and the redundancy within the data due to spatial autocorrelation, incorporating additional covariates can present challenges. The sparseness of the one-hot-encoded PFTs is particularly problematic, as certain PFTs are represented only by a handful of EC site instances. Hence, these sparse instances can exert disproportionate leverage on the upscaling results . This hypothesis is supported by the site-level cross-validation results (~~nor sequential versus non-sequential models can be regarded as inherently more consistent—or therefore more reliable—for global upscaling.

A further complication is the change from EC site observed meteorology to reanalysis grids. At EC sites, meteorological variables are locally observed, whereas global inputs are derived from reanalysis grids (with a spatial resolution of 0.25°) with

assimilation and spatial smoothing (Hersbach et al., 2020; Parker, 2016; Grusson and Barron, 2022; Valmassoi et al., 2023). Sequential networks, which exploit fine-scale temporal structure, are theoretically more sensitive to such a distribution shift than tree-based or feed-forward models. Yet our results show that spatial patterns are robust in data-rich regions, both across architectures (ensemble means, Fig. 4) , where using PFTs did have inconsistent impact on the model performance.While it is unclear why `xbase` achieved global ET that is more consistent with other sources, we note that seemingly small methodological choices had a larger impact on the upscaling results than the ML model type. 7) and within architectures (ensemble spread, Fig. 9). Because every model and covariate set converges where observational support is strong, the station-to-grid shift cannot be the dominant source of the spatial or global-sum discrepancies. Instead, the residual differences arise from the extrapolation into data-scarse regions, with different behavior across models and covariate setups.

The systematic 8% difference in global annual ET between sequential and non-sequential models is noteworthy, since the feature sets and training data were identical. From the site-level cross-validation experiment, we learned that the sequential models represent ET slightly better The model runs without PFTs (Fig. 7a and Fig. 9a) showed a better agreement among each other (Fig. 4) . It is possible that the sequential models learned a better representation of the underlying processes, and that we should trust them more than the other ML models and prior studies. To investigate this, we analyze the similarity between the two sequential models by quantifying the alignment of the ensemble members in terms of global ET (the black dots in 7a vs. Fig. 7b) and within ensemble members (Fig. 9a vs. Fig. 9b). This suggests that excluding the PFTs from the covariates improves the model robustness. However, we saw at site level that the PFTs improved the model performance particularly at the spatial scale (Fig. 6)using the Pearson (r) and Spearman (ρ) correlation between them. The former quantifies a linear, the latter a monotonic (regardless of linear or not) relationship. As the ensemble members were trained on the same training subsets, we would see a high correlation between models if they learned similar representations from the data. This is, however, not the case, as shown in Tab. 2. The largest agreement was found between `tcn` and 4). At the same time, we observed unrealistic spatial patterns with the `full` setup: In particular, `xgboost` ($r = 0.65$, $\rho = 0.81$), followed by `lstm` and and `fcn` with the `full` setup estimated large ET in arid regions like the Sahara and Arabian Peninsula. These large ET estimates align poorly with our understanding of ET processes, and `fcn`xbase ($r = 0.45$, $\rho = 0.31$) and estimates in these areas have also been shown to be too high (Nelson & Walther et al., 2024). Thus, while `fcn` and particularly `xgboost` and with the `fcn`full ($r = 0.36$, $\rho = 0.62$)covariate setup appear close to the global mean ET benchmark, this is likely for the wrong reasons, i.e., overestimation in arid regions. The sequential models showed the lowest alignment between ensemble members in terms of the linear relationship ($r = 0.09$) and moderate alignment in terms of the rank correlation ($\rho = 0.33$). In summary, the sequential models did, when trained on the same subsets of sites, not behave similar in terms of global annual ET. Therefore, it seems unreasonable to assume that the lower global ET estimated by the sequential neural networks is due to a better (and hence more consistent) representationof the processes. also yielded higher global ET when including PFTs, likewise originating from arid regions, but the increase was less pronounced. The `lstm` showed the largest robustness across setups. For all models, we observed an increase of ensemble spread in arid and also tropical regions when adding PFTs (Fig. 9). It is possible that in the site-level cross validation, these effects were not visible because of a lack of EC stations in arid and tropical regions.

~~Global evapotranspiration (ET) correlation across model ensemble members. The correlation is calculated from the yearly ET of the cross-validation ensemble members, which are based on the same training sites. The upper triangular (in boldface) represents the Pearson correlation (the linear relationship) and the lower triangular shows the Spearman rank correlation (the monotonic relationship).~~ ~~`xgboost` `fcn` `tcn`~~ While we cannot answer which model in combination with which covariate setup yields the "best" upscaling results, we can conclude that non-temporal models yielded more realistic global ET estimates, even if this is largely due to covariate-driven artifacts in data-scarse regions rather than genuinely improved process representation. The sequential models were also sensitive to covariate setup, but the `lstm` ~~`xgboost` 1.00 **0.36** **0.65** **0.39** `fcn` 0.62 1.00 **0.25** **0.45** `tcn` 0.81 0.38 1.00 **0.09** `lstm` 0.36 0.31 0.33 1.00~~

~~The lower ET estimates may be attributed to shifts in covariate distribution for several reasons. Site-level data is measured by local meteorological stations, whereas global grid data is reanalysis-based, introducing a twofold shift. Firstly, site-level observations represent point measurements, while the reanalysis data, with a 0.05° spatial resolution, smoothes local variations. Secondly, reanalysis is based on data assimilation, harmonizing diverse observations with process models , and often exhibits biases and overly smooth solutions, contributing to data shifts in upscaling approaches (Parker, 2016; Hersbach et al., 2020; Grusson and Ba .Additionally, upscaling involves extrapolating into poorly represented ecoclimatic regions~~ was more robust compared to the other models.

4.2.3 Consistent temporal dynamics

Despite variability in absolute ET magnitudes, the models showed strong internal consistency in their temporal dynamics (Fig. 8). The high correlation between covariate setups (Fig. ~~1b) , potentially affecting sequential models more severely due to their reliance on past meteorology and remote sensing-based observations. Consequently, we hypothesize that sequential models are more impacted by distribution shifts, which calls for efforts to close this gap.~~

~~In contrastto the systematic differences in terms of mean ET, its relative patterns were robustly predicted across ML models . The consistent correlation with `gleam` in terms of ensemble means of monthly ET (diagonal panels in Fig. **??**), and the small differences in terms of monthly ET anomalies (upper triangular panels in~~ 8a) indicates that models captured similar temporal variations and responded consistently to climatic forcing.

Temporal correlation with `gleam` was moderate overall, as discussed previously. The GLEAM ET product is based on conceptual models driven by remote sensing inputs and meteorological forcing. We do not treat it as ground truth, but we assume it provides spatially consistent temporal patterns due to its incorporation of prior knowledge. In contrast, our ML models rely solely on data-driven representations, and may be sensitive to data shifts due to generalization beyond data support and data types. Given this, the overall moderate agreement with GLEAM is not surprising and does not imply poor model performance. Rather, we interpret divergences—especially in tropical and arid regions—as a consequence of low data availability and known challenges in these environments.

Temporal correlation with `gleam` improved notably for the `lstm` model in arid and some tropical regions when additional covariates were included (Fig. **??**~~), indicate that the distribution shift from site level to global gridded scale introduces stronger shifts on the mean values than on the relative spatial patterns and temporal dynamics.~~

The substantial variation in the model ensemble spread indicates a significant impact of model architectures on the upscaling process. The neural network-based models exhibited ensemble variability comparable to the spread of land surface models (lsm) on a global annual scale . The 8c). This improvement was not observed for xgboost model demonstrated considerably lower variability across cross-validation members , which showed a slight degradation, particularly in arid climates, nor for fcn, which remained largely unchanged. We attribute the improved performance of sequential models primarily to the inclusion of precipitation, consistent with their enhanced performance on anomaly scale at site level when this covariate was added (Fig. 6). A low ensemble spread is generally deemed advantageous. Nonetheless, it is unclear whether the high robustness of 4).

It is unclear why xgboosttcn signifies low uncertainty or if it indicates underfitting or a rigid extrapolation behavior, which could lead to heavily biased predictions. In terms of underfitting, no such behavior was observed at the did not show a similar improvement in arid regions. At site level (Fig. 4), where the performance of fcn 5), tcn actually outperformed lstm on the anomaly scale, albeit with higher model uncertainty. This could suggest that lstm learns a more robust representation of temporal dynamics in these regions, which is not evident from the site-level cross-validation as those regions are largely absent from the training data. In contrast, tcn may be more sensitive to distribution shifts. There is evidence that lstm architectures are better suited for capturing hydrologically relevant temporal patterns compared to convolutional models (Kraft et al., 2025). While the lstm can theoretically access the full temporal context, tcn is constrained to a fixed temporal window of approximately 8 days in both the met+rs and xgboost was nearly identical across various scales and datasetups. Rigid extrapolation behavior could be linked to the "bound truncation" behavior of regression trees during out-of-distribution extrapolation (Malistov and Trushin, 2019). This behavior constrains predictions to the range of the training data, limiting the model's ability to extrapolate beyond observed values and potentially introducing bias into the upscaling results.If this was the case, we would expect the neural networks to show better alignment among each other than with full setups (Tab. A1). Note that this temporal context is, for the xgboosttcn . However, xgboost was not an outlier among the models in terms of patterns of spatial ensemble spread (the lower triangular in Fig. **??**), alignment with GLEAM (diagonal in Fig. **??**), or spatio-temporal patterns of upscaled anomalies (the upper triangular in Fig. **??**). Thus, it seems that the xgboost ensemble spread was not necessarily a sign of a rigid extrapolation behavior but rather a sign of more robust predictions. In contrast, the neural network-based models showed an extensively large ensemble spread, which could indeed be related to the notoriously challenging out-of-distribution prediction with such flexible models (Pastore and Carnini, 2021)architecture, related to the hyperparameters. This limited context may explain its reduced ability to generalize temporal patterns in data-sparse regions.

In summary, temporal dynamics remained relatively consistent across methods. Sequential models, especially lstm, benefited from richer covariate input, particularly precipitation, improving alignment with an independent product in arid regions. This highlights the added value of using sequential models for representing temporal dynamics in ET modeling and upscaling, yet the interaction between model architectures and covariates, particularly in data-sparse regions, also suggests that a profound covariate selection is necessary to identify best-working setups.

4.3 Lessons learned and outlook

From site-level analyses, it was observed that sequential models generally outperform

4.3.1 Site-level performance: small but consistent advantages for sequential models

At the site level, sequential models consistently outperformed non-sequential models in ET flux modeling (see for ET flux prediction, especially when evaluated on the anomaly scale (Fig. 4). This finding is consistent with the results by Besnard et al. (2019), who found similar behavior for NEE flux modeling . When covariates that effectively represent ecosystem stateis in line with previous findings for NEE modeling (Besnard et al., 2019), suggesting that temporal dependencies, such as vegetation indices, were incorporated, ecosystem memory effects, are captured more effectively by models with recurrent or convolutional memory. The inclusion of remote sensing variables (e.g., vegetation indices) reduced the performance gap between sequential and non-sequential models (e.g., XGBoost and fully-connected feed-forward neural networks) and sequential models narrowed, although latter continued to exhibit superior by providing state proxies, although the advantage for sequential models persisted. Adding precipitation from reanalysis further improved performance at the anomaly scale. It is conceivable that advanced deep learning , particularly for the sequential models, confirming the value of dynamic hydrological information. However, these improvements were relatively modest, indicating that data availability and quality, rather than model architecture, remain the primary bottlenecks. More advanced architectures, such as those based on transformers, might further enhance model performance. However, results by Nakagawa et al. (2023) on modeling EC gross primary production (GPP) using a temporal fusion transformer showed temporal transformers, may offer further gains, but recent work in GPP modeling has shown only marginal improvements , aligning with our observations (Nakagawa et al., 2023), providing further evidence that EC flux modeling remains is still a data-limited challenge with limited benefits from time-domain deep learning techniques to dateproblem.

Upscaling to global coverage introduces significant covariate shifts, resulting in unexpected impacts on the global estimates. Notably, minor modifications in the experimental setup relative to `xbase` (Nelson & Walther et al., 2024), such as the exclusion of some covariates, resulted in substantial deviations in global ET estimates (see

4.3.2 Global upscaling: both model architecture and covariates drive uncertainty

Small changes in the covariate setup led to moderate differences in global ET means (Fig. 6). Sequential deep learning models tended to predict lower global ET values compared to non-sequential models and independent evaluations. This discrepancy may be attributed to the resilience of non-sequential models to covariate shifts, particularly those utilizing robust remote sensing data. Previous studies, such as Jung et al. (2019), have acknowledged these uncertainties in ML-driven upscaling, yet our study underscores the critical role of , especially for `xgboost` with the `full` setup, which showed an increase of 10% in global ET, supposedly due to strong leverage effects of PFTs in sparsely sampled regions. The moderate differences were partially due to error compensation. While we found no clear evidence that any architecture yields more realistic absolute ET estimates, our results point to systematic biases that depend on both model design and the covariates used. This highlights the importance of methodological choices, including architecture, covariate selection, and cross-validation , data handling, and ML configuration

~~in influencing these uncertainties. Notably, similar findings were reported by Zhu et al. (2024).~~ design, in driving upscaling uncertainty.

~~The XGBoost method resulted in a more robust upscaling ensemble compared to those derived from neural network models (Fig. ?? & ??). Nevertheless, all tested machine learning models showed similar agreement with the independent GLEAM product in terms of spatio-temporal patterns, and there was significant agreement among the models at monthly anomaly time scales in terms of correlation. This suggests that the upscaling was robust in terms of spatio-temporal patterns, apart from the previously mentioned biases, across the machine learning models. From this analysis, we consider XGBoost to be a well-suited tool for upscaling of EC fluxes, while the complexity and higher energy consumption of sequential approaches with their small added value renders such methods, currently, less favorable for practical applications in environmental data modeling. This finding is supported by the analysis of Zhu et al. (2024), which found LSTMs to perform only marginally better in challenging regions, i.e., the tropics. We strongly encourage the investigation of methodological aspects and their impact on upscaling beyond machine learning type. This involves the role of covariates, ML approaches, cross-validation schemes, and distribution shifts. Special emphasis should be placed on investigating the role of spatial features, either through more targeted ablation studies similar to the one performed here, via feature selection, or by considering continuous EC site data, such as plant traits (Kattge et al., 2011), soil properties (Hengl et al., 2017), or deep learning-based location embeddings (Klemmer et al., 2024) . This approach could provide deeper insights into the contribution of these features to the overall model performance and upscaling results. A tool for conducting such systematic methodological experiments ("FLUXCOM-X"), which was also used within this study, was recently introduced by Nelson & Walther et al. (2024).~~

4.3.3 Spatio-temporal patterns: robust dynamics, variable magnitudes

Despite differences in magnitude, all models showed consistent temporal ET patterns and strong alignment in anomaly correlations, both among themselves and, to a lower extent, also with the GLEAM product (Fig. 8), yet the robustness drastically decreased outside of the training domain. With the inclusion of precipitation and PFTs in the `full` setup, the sequential models—particularly `lstm`—showed improved temporal agreement with GLEAM in arid regions, supporting their value for representing soil moisture-related memory effects. However, these benefits were largely limited to the temporal dimension; biases in global means and ensemble spread remained pronounced, especially for data-sparse regions where additional covariates such as PFTs introduce high leverage. Given their lower computational complexity and more stable behavior, we find that non-sequential models like `xgboost` currently offer a pragmatic and robust solution for global-scale ET upscaling, especially when paired with carefully selected meteorological and remote sensing inputs.

4.3.4 Outlook: addressing biases and uncertainties in data-scarce regions

Moving forward, reducing model biases and uncertainty in upscaling requires both methodological advances and improved data. Rather than prioritizing complex model architectures, future efforts should focus on four complementary pathways:

i) Feature selection and additional data constraints: Deep learning presents promise for enhancing flux modeling and upscaling, and offers advanced computational techniques capable of managing complex, non-linear interactions within ecosystems.

However, to maximize the effectiveness of deep learning in such a data-limited setting, it is essential to implement additional constraints and integrate richer data sources (Reichstein et al., 2019). The accuracy of deep learning models heavily relies on the quality and diversity of the input data (Karpatne et al., 2019). Enhancing these models with additional covariates that accurately reflect ecological and atmospheric conditions can significantly improve their predictive power. Additionally, expanding the network of flux stations and sharing the data for scientific applications would enhance the data base to cover more diverse ecological conditions and climate zones, thereby enriching the training data used for model calibration and validation. Furthermore, applying constraints at a regional level, akin to the approach by Upton et al. (2024), who used an ensemble of atmospheric inversions of NEE as large-scale guidance for flux upscaling, could be used to reduce biases. For ET modeling, large-scale water balance could be used as a regional constraint, for example.

ii) Additional data sources: To further refine the performance of deep learning approaches in EC measurement upscaling, it could be beneficial to tap additional data sources. Approaches such as transfer learning can be particularly effective (Caruana, 1997; Pan and Yang, 2010). By applying knowledge gained from one region to another, or from related and richer datasets, models can achieve better generalization, especially in data-sparse areas. To deal with shifts in covariates due to the various reasons discussed previously, domain adaptation (He et al., 2023) could provide a useful toolbox to reduce upscaling biases.

iii) Physical constraints: As a complementary pathway, incorporating prior scientific knowledge into deep learning models could help address challenges associated with data extrapolation and distribution shifts encountered in upscaling (Reichstein et al., 2019; Kraft et al., 2022). Such integration aids in aligning model outputs with established physical laws and ecological principles, thereby improving the reliability of the predictions (Reichstein et al., 2022). Physics-informed and hybrid physics/ML approaches represent a cutting-edge direction in the field of flux modeling, as they merge the empirical strengths of deep learning with the deterministic nature of physical models. For upscaling into undersampled regions, such constraints can nudge the model outputs towards physically more plausible solutions. As an example, encoding simple relationships between precipitation and evaporation, or vegetation and transpiration, could help reducing ET estimates in arid regions, where EC stations are lacking. Although challenging, more comprehensive physical process parameterizations, such as the Penman-Monteith equations, can be combined with machine learning to estimate ET (Zhao et al., 2019; ElGhawi et al., 2023). This could, in principle, reduce the widely reported regional biases in upscaling EC fluxes with machine learning, which we identified ~~to be~~ as currently the main challenge in flux upscaling.

iv) Systematic benchmarking: Finally, systematic benchmarking frameworks, such as FLUXCOM-X (Nelson & Walther et al., 2024), are essential for disentangling the complex interactions between models, covariates, and evaluation setups. Such frameworks enable controlled ablation studies and targeted diagnostics, helping to build a more transparent understanding of uncertainty sources in ML-based flux upscaling.

In conclusion, while deep learning provides valuable tools for modeling land-atmosphere interactions, the key to better global ET estimates lies in more comprehensive observational data, thoughtful covariate selection, targeted physical constraints, and methods to reduce extrapolation bias—rather than in model complexity alone.

5

In this study, we assessed different ~~data setups and ML~~ covariate setups and machine learning approaches for modeling ET fluxes at ~~EC sites in a~~ eddy covariance sites through cross-validation~~setup and assessed~~ , and evaluated the robustness and quality of ~~upscaled ET at a global scale~~globally upscaled ET estimates. From our ~~analysis at the site level~~site-level analysis, we conclude that sequential deep learning ~~approaches~~ models can outperform non-sequential models for ET flux modeling~~.~~ , particularly on the anomaly scale. The sequential models ~~learned~~ captured memory effects related to water availability, which ~~led to a better representation of arid and wet conditions~~improved their ability to represent temporal dynamics of ET. However, ~~when adding remote sensing observations, the advantage of using sequential models shrank as these covariates provided appropriate proxies for ecological memory. Using PFTs did not increase model performance overall and even decreased performance in some cases. Thus, we suggest to further investigate the role of PFTs for the modeling of EC fluxes and we encourage exploring other static variables instead, or, alternatively, to perform~~ this advantage diminished when remote sensing covariates were included, as these effectively act as proxies for ecosystem memory. Adding precipitation as a covariate led to small performance improvements, especially for sequential models and on the anomaly scale, confirming its value in representing hydrological constraints. The inclusion of PFTs increased the ability of the models to capture spatial variability across sites, yet their discrete nature raises concerns about their broader utility. We, therefore, recommend further exploration of alternative static variables or targeted feature selection to ~~keep the number of covariates low~~maintain a parsimonious covariate set.

~~The~~ Globally, sequential and non-sequential models ~~yielded~~ produced small but systematic differences ~~of global mean ET. Therefore, it seems that the models learned different representations and behaved not the same when fed with new, potentially differently distributed, gridded data . As long as we do not understand the sources of those uncertainties, it is beneficial to use structurally different ML modelsto get a plausible estimate of robustness .~~ in mean ET, suggesting that different model types learn distinct representations and respond differently and randomly when extrapolating out of the training distribution. When using PFTs as additional covariate, both divergence of spatial means of ET among models and ensemble spread within models increased, particularly in arid regions where data support is limited. This highlights the large leverage of the PFTs on upscaling results when extrapolating out of the covariate space, and underscores the need for better understanding of the role of covariates in upscaling. While the sequential models suffered from similar biases as their non-sequential counterparts, they achieved better temporal alignment with the GLEAM product, particularly when precipitation was included.

~~Given the additional complexity of the~~ In the context of our experiments, we can answer the research questions posed in the introduction as follows:

a) *Do the sequential models lead to more realistic and robust global and regional ET estimates?*
No, sequential models yielded lower global ET estimates than expected from independent estimates. However, the larger estimates of the non-sequential models were likely due to an overestimation of ET in arid regions. The sequential models, especially the LSTM, were more robust to covariate shifts, though, supposedly due to their lower dependency on PFT due to their ability to link spatial variability to temporal features in the covariates.

b) *Do the sequential models capture the temporal dynamics of ET better?*

Yes, the sequential models captured temporal dynamics better than non-sequential models, at both site and upscaling level, particularly when using precipitation as additional input and with the LSTM architecture.

c) *How stable are the sequential models to the covariate setup and training data subsets, compared to the non-sequential models?*

The sequential models showed a similar ensemble spread than the non sequential neural network, indicating that robustness towards training data subsets is rather linked to the model type (decision tree versus neural network) than to the sequential nature of the models.

Considering the added complexity and computational cost of sequential neural networks~~and the sequential data handling~~, the relatively ~~small performance increase~~ modest performance gain at site level, and ~~considering the underestimation of ET globally and the large ensemble spread~~their underestimation of global ET alongside larger ensemble spreads, we conclude that ~~using such advanced modeling techniques is not a strict requirement for modeling ET globally~~these models are not strictly required for global ET upscaling. Non-sequential machine learning approaches, such as XGBoost, can ~~provide comparably robust predictions~~ deliver similarly robust estimates across scales when ~~paired with good-quality~~ supported by high-quality meteorological and remote sensing covariates. Yet, XGBoost overestimated ET in arid regions, which raises some concerns, but these are also not exclusive to this model type. Particularly the LSTM architecture was robust towards covariate setups and able to learn subtle temporal dynamics, which could be beneficial in specific applications. These findings highlight the importance of structural model diversity in ensemble setups to assess the robustness and uncertainty of upscaling results.

The potential ~~of~~ for upscaling ET to a global scale ~~via modern ML approaches seems to be limited~~ using modern ML methods is constrained by the information content ~~in the EC site-level data, and hence,~~ and representativeness of the EC training data. As such, seemingly small changes in ~~the setup might have a big leverage on the poorly constrained upscaling problem. Thus, other pathways need to be explored. The integration of richer data sources, such as additional covariatesor additional EC stations,~~ covariate setups can exert large influence over upscaling results, particularly in data-sparse regions. To enhance the robustness and physical consistency of global ET products, we recommend pursuing complementary strategies: integrating richer covariates, increasing EC site coverage, applying regional constraints, ~~related data via~~ leveraging transfer learning, and ~~the incorporation of~~ embedding prior scientific knowledge ~~could increase both the robustness and physical consistency of the global upscaling products. By embracing these strategies~~into ML architectures. By following these pathways, deep learning has the potential to ~~deliver more precise~~produce more accurate, robust, and physically grounded predictions of terrestrial evapotranspiration across diverse environmental ~~scenarios~~settings.

. TEXT

. TEXT

The upscaled ET fields generated for this study are available from the corresponding author upon reasonable request.

. TEXT

. TEXT

. TEXT

Appendix A

Table A1 provides an overview of the hyperparameter search space and the best-performing combination of settings for each model and covariate setup. We sampled 20 random hyperparameter configurations for each model, evaluating them via early stopping on a validation set to determine the final selections. For the `tcn` model, the temporal context length considered (`temp. contex`) depends on the hyperparameters of the model, leading to a range from about 4 days (91 hours) to about 19 days (451 hours). The number of parameters in each model (`# parameters`) was derived from the chosen architecture, highlighting that the complexity increases substantially with deeper layers or larger hidden sizes.

Appendix B

Table B1 displays the eight cross-validation groups, each column representing one group, with its corresponding eddy covariance site IDs. We used this grouping to ensure that training and validation sets differ systematically across folds, minimizing spatial autocorrelation effects at the site level. The sites within each group span a range of climatic and ecophysiological conditions, allowing for more robust evaluation of model generalization.

. BK implemented the model architectures and data pipeline for the neural networks and performed the analysis. He took the lead in writing the manuscript. JN, SW, FG, MJ, BK, and ZH contributed to the FLUXCOM-X framework, on which this study builds upon. SW, JN, FG, UW, and MJ provided, processed, and cleaned the datasets used. BK, JN, SW, MJ, GD, MRe, and WZ contributed to the scientific evaluation of the results. MK, MRu, and DT contributed with their expertise to the data-scientific and machine learning aspects of the study. All co-authors contributed to the manuscript.

Table A1. Hyperparameter search space and the best performing hyperparameter per model and setup. The best combination was found by evaluating 20 random samples based on early stopping validation loss. The `xgboost` and `fcn` models are non-sequential, the `lstm` has minimum of one year (9k hourly time steps) and maximum of two years of theoretical context (18k hourly time steps), and the `tcn` model's temporal context depends on the hyperparameters (reported in the row `temp. context`), ranging from 90 to 450 hours, *i.e.*, about 4 to 19 days. The number of parameters per model is reported in the row `# parameters`.

Model	Hyperparameter	Search space	Selected hyperparameter per setup				
			met	met+pft met+prec	met+rs met+pft	met+pft+rs met+rs	
xgboost	max_depth	{6, 8, 10, 12}	10	10	12	8	
	learning_rate	$\{10^{-2}, 10^{-1}, 2 \times 10^{-1}\}$	10^{-1}	10^{-1}	10^{-1}	10^{-1}	
	min_child_weight	{1, 5, 10}	1	1	5	5	
	max_delta_step	{1, 5, 10}	10	10	10	5	
	# parameters	derived	~~208~~209K	212K	640K	~~132~~128K	~~1~~
fcn	num_hidden	{128, 256}	128	128	128	256	
	num_layers	{3, 4}	~~3~~ 1	~~3~~ 1	~~4~~ 1	~~3~~ 2	
	dropout	{0.0, 0.2}	0.2	0.2	0.2	0.2	
	learning_rate	$\{10^{-6}, 10^{-5}, 10^{-4}\}$	10^{-6}	10^{-4}	10^{-6}	10^{-5}	
	weight_decay	$\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$	10^{-3}	10^{-1}	10^{-3}	10^{-3}	
	# parameters	derived	17K	17K	19K	134K	
tcn	num_hidden	{64, 128, 256}	64	256	256	256	
	num_layers	{2, 3, 4}	4	4	3	4	
	kernel_size	{4, 8, 16}	16	4	16	4	
	dropout	{0.0, 0.2}	0.2	0.2	0.2	0.2	
	learning_rate	$\{10^{-6}, 10^{-5}, 10^{-4}\}$	10^{-6}	10^{-5}	10^{-6}	10^{-5}	
	weight_decay	$\{10^{-3}, 10^{-2}, 10^{-1}\}$	10^{-3}	~~10^{-1}~~ 10^{-2}	10^{-2}	10^{-2}	
	temp. context	derived	~~450~~ 451	~~210~~ 91	~~90~~ 211	~~90~~ 91	
	# parameters	derived	473K	~~5 600K~~ 2.0M	~~2 000K~~ 5.6M	~~2 200K~~ 2.0M	
lstm	num_hidden	{64, 128, 256}	128	256	256	128	
	num_layers	{1, 2}	2	1	1	2	
	dropout	{0.0, 0.2}	~~0~~ 0.0	0.2	0.2	0.2	
	learning_rate	$\{10^{-6}, 10^{-5}, 10^{-4}\}$	10^{-4}	10^{-4}	10^{-6}	10^{-6}	
	weight_decay	$\{10^{-3}, 10^{-2}, 10^{-1}\}$	10^{-1}	~~10^{-3}~~ 10^{-1}	10^{-3}	10^{3}	
	# parameters	derived	217K	336K	344K	219K	

Table B1. The eddy covariance site groups used for cross validation. The eight groups correspond to columns, the items correspond to site IDs.

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
AR-Vir	AR-TF1	AU-Fog	AU-DaP	AU-ASM	AU-Cum	AR-SLu	AU-Ade
AT-Neu	AU-Wac	BE-Lcr	AU-Whr	AU-Dry	AU-RDF	AU-Cpr	AU-DaS
AU-Tum	BR-Sa3	CA-ER1	CH-Cha	AU-Emr	AU-TTE	AU-Gin	BE-Maa
BE-Dor	CA-NS2	CA-SF1	CH-Fru	BE-Bra	AU-Wom	AU-Rob	CA-Qfo
CA-NS3	CA-NS5	CA-SF3	CN-Dan	BE-Lon	BE-Vie	AU-Ync	CH-Dav
CA-NS4	CA-TP1	CN-Qia	CN-Du2	BR-Sa1	CA-Cbo	BR-Npw	CZ-KrP
CN-HaM	CA-TPD	DK-Sor	CN-Du3	CA-Obs	CA-LP1	CA-DB2	CZ-Lnz
CZ-BK1	DE-RuR	FR-Aur	CN-Sw2	CA-SF2	CA-NS6	CA-DBB	DE-Hte
CZ-BK2	DE-RuS	FR-Bil	CZ-Stn	CA-TP3	CH-Lae	CA-Gro	DK-Gds
CZ-RAJ	DE-RuW	FR-Fon	DE-Kli	CA-TP4	CN-Cha	CA-NS7	ES-Agu
DE-Akm	DE-Seh	FR-Tou	DE-Lnf	DE-Hai	CN-Cng	CA-Oas	ES-LJu
DE-HoH	DE-Spw	GF-Guy	ES-Abr	DE-Tha	CN-Din	CA-TP2	ES-LM1
DE-Lkb	DE-Zrk	GL-Dsk	ES-LgS	FI-Ken	DE-Geb	CH-Aws	ES-LM2
DE-Obe	ES-Amo	IE-Cra	FI-Hyy	FI-Lom	FI-Sii	CH-Oe1	ES-Ln2
DE-SfN	FI-Var	IT-BFt	FI-Let	FR-Lam	GH-Ank	CH-Oe2	FI-Qvd
DK-Eng	FR-Hes	IT-Cp2	FR-EM2	GL-NuF	PA-SPn	CZ-wet	FR-FBn
ES-Cnd	IL-Yat	IT-Cpz	IT-Tor	IT-Col	PA-SPs	DE-Gri	FR-LBr
FI-Jok	IT-BCi	JP-SMF	MX-Tes	IT-PT1	SE-Lnn	DE-Hzd	FR-Pue
GL-ZaF	IT-Ro1	US-Atq	NL-Hor	IT-SR2	SE-Ros	DK-Fou	IT-Isp
GL-ZaH	IT-Ro2	US-CS1	PE-QFR	IT-SRo	US-A32	FI-Sod	IT-La2
IT-Lsn	NL-Loo	US-CS2	RU-Fy2	JP-MBF	US-Cop	FR-Gri	IT-Lav
IT-MBo	RU-Ha1	US-CS3	RU-Fyo	MY-PSO	US-EDN	FR-LGt	IT-Noe
IT-Ren	SD-Dem	US-CS4	SE-Htm	RU-Cok	US-Ho2	IT-CA1	RU-Che
SE-Nor	US-IB2	US-GBT	US-ARb	SJ-Blv	US-Me1	IT-CA2	SE-Svb
SJ-Adv	US-KS1	US-GLE	US-ARc	US-CRT	US-Me3	IT-CA3	US-Blo
US-AR2	US-Los	US-Ha1	US-CF1	US-ORv	US-Me6	SE-Deg	US-Ivo
US-ARM	US-MOz	US-KS2	US-CF2	US-Prr	US-Myb	SN-Dhr	US-KFS
US-BZB	US-NR1	US-KS3	US-CF3	US-Tw1	US-Ne1	US-AR1	US-UMB
US-BZF	US-ONA	US-Lin	US-CF4	US-Tw2	US-Ne2	US-Bi1	US-UMd
US-BZS	US-Ro1	US-Me2	US-Jo2	US-Tw3	US-Ne3	US-Bi2	US-Wi0
US-BZo	US-Ro4	US-Me5	US-NGB	US-Tw4	US-Sne	US-Goo	US-Wi1
US-ICs	US-Ro5	US-SRG	US-OWC	US-Tw5	US-Ton	US-Hn3	US-Wi3
US-ICt	US-Ro6	US-SRM	US-Oho	US-Twt	US-Var	US-MMS	US-Wi5
US-KLS	US-Rwf	US-Wjs	US-Rms	US-WPT	US-WCr	US-PFa	US-Wi7
US-Mpj	US-Rws	US-Wkg	US-Rwe	US-Wi6	US-Wi2	US-Sta	US-Wi8
US-Syv	US-SRC	ZM-Mon	US-Whs	US-xBR	US-Wi4		US-Wi9

. The contact author has declared that none of the authors has any competing interests.

. This research was funded by the European Research Council (ERC) Synergy Grant *Understanding and modeling the Earth System with Machine Learning (USMILE)* under the Horizon 2020 research and innovation program (Grant Agreement No. 855187). We thank the Max

Planck Institute for Biogeochemistry for providing the computational and data infrastructure for conducting this research. We appreciate the financial support by the Max Planck Society to publish this manuscript open-access.

References

The effect of relative humidity on eddy covariance latent heat flux measurements and its implication for partitioning into transpiration and evaporation, Agricultural and Forest Meteorology, 330, 109 305, https://doi.org/https://doi.org/10.1016/j.agrformet.2022.109305, 2023.

Armstrong, S., Khandelwal, P., Padalia, D., Senay, G., Schulte, D., Andales, A., Breidt, F. J., Pallickara, S., and Pallickara, S. L.: Attention-based convolutional capsules for evapotranspiration estimation at scale, Environmental Modelling & Software, 152, 105 366, https://doi.org/10.1016/j.envsoft.2022.105366, 2022.

Badgley, G., Field, C. B., and Berry, J. A.: Canopy Near-Infrared Reflectance and Terrestrial Photosynthesis, Science Advances, 3, e1602 244, https://doi.org/10.1126/sciadv.1602244, 2017.

Bai, S., Kolter, J. Z., and Koltun, V.: An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, https://arxiv.org/abs/1803.01271, 2018.

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., U, K. T. P., Pilegaard, K., Schmid, H. P., Valentini, R., Verma, S., Vesala, T., Wilson, K., and Wofsy, S.: FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities, Bulletin of the American Meteorological Society, 82, 2415–2434, https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2, 2001.

Bambach, N., Kustas, W., Alfieri, J., Prueger, J., Hipps, L., McKee, L., Castro, S., Volk, J., Alsina, M., and McElrone, A.: Evapotranspiration uncertainty at micrometeorological scales: the impact of the eddy covariance energy imbalance and correction methods, Irrigation Science, 40, 445–461, https://doi.org/10.1007/s00271-022-00783-1, 2022.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M. A., Baldocchi, D., Bonan, G. B., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K. W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, F. I., and Papale, D.: Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate, Science, 329, 834–838, https://doi.org/10.1126/science.1184984, 2010.

Besnard, S., Carvalhais, N., Arain, M. A., Black, A., Brede, B., Buchmann, N., Chen, J., Clevers, J. G. P. W., Dutrieux, L. P., Gans, F., Herold, M., Jung, M., Kosugi, Y., Knohl, A., Law, B. E., Paul-Limoges, E., Lohila, A., Merbold, L., Roupsard, O., Valentini, R., Wolf, S., Zhang, X., and Reichstein, M.: Memory Effects of Climate and Vegetation Affecting Net Ecosystem CO2 Fluxes in Global Forests, PLOS ONE, 14, e0211 510, https://doi.org/10.1371/journal.pone.0211510, 2019.

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, https://doi.org/10.5194/essd-10-1327-2018, 2018.

Breiman, L.: Random forests, Machine learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Camps-Valls, G., Tuia, D., Zhu, X. X., and Reichstein, M.: Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences, Wiley, Hoboken, NJ, 1st edition edn., 2021.

Caruana, R.: Multitask Learning, Machine Learning, 28, 41–75, https://doi.org/10.1023/A:1007379606734, 1997.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pp. 785–794, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/2939672.2939785, 2016.

Drought 2018 Team and ICOS Ecosystem Thematic Centre: Drought-2018 Ecosystem Eddy Covariance Flux Product for 52 Stations in FLUXNET-Archive Format, https://www.icos-cp.eu/data-products/YVR0-4898, 2020.

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, https://doi.org/10.1088/1748-9326/acbbe0, 2023.

FLUXCOM: FLUXCOM Global Energy and Carbon Fluxes, https://fluxcom.org/, 2017.

Gao, B.-c.: NDWI—A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space, Remote Sensing of Environment, 58, 257–266, https://doi.org/10.1016/S0034-4257(96)00067-3, 1996.

Grusson, Y. and Barron, J.: Challenges in reanalysis products to assess extreme weather impacts on agriculture: Study case in southern Sweden, PLOS climate, 1, e0000 063, https://doi.org/0.1371/journal.pclm.000006, 2022.

He, H., Queen, O., Koker, T., Cuevas, C., Tsiligkaridis, T., and Zitnik, M.: Domain Adaptation for Time Series Under Feature and Label Shifts, in: Proceedings of the 40th International Conference on Machine Learning, edited by Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., vol. 202 of *Proceedings of Machine Learning Research*, pp. 12 746–12 774, PMLR, https://proceedings.mlr.press/v202/he23b.html, 2023.

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global Gridded Soil Information Based on Machine Learning, PLOS ONE, 12, e0169 748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 Global Reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

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

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., and Ferreira, L. G.: Overview of the Radiometric and Biophysical Performance of the MODIS Vegetation Indices, Remote Sensing of Environment, 83, 195–213, https://doi.org/10.1016/S0034-4257(02)00096-2, 2002.

Jung, M., Reichstein, M., and Bondeau, A.: Towards Global Empirical Upscaling of FLUXNET Eddy Covariance Observations: Validation of a Model Tree Ensemble Approach Using a Biosphere Model, Biogeosciences, 6, 2001–2013, https://doi.org/10.5194/bg-6-2001-2009, 2009.

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., Bonan, G., Cescatti, A., Chen, J., de Jeu, R., Dolman, A. J., Eugster, W., Gerten, D., Gianelle, D., Gobron, N., Heinke, J., Kimball, J., Law, B. E., Montagnani, L., Mu, Q., Mueller, B., Oleson, K., Papale, D., Richardson, A. D., Roupsard, O., Running, S., Tomelleri, E., Viovy, N., Weber, U., Williams, C., Wood, E., Zaehle, S., and Zhang, K.: Recent Decline in the Global Land Evapotranspiration Trend Due to Limited Moisture Supply, Nature, 467, 951–954, https://doi.org/10.1038/nature09396, 2010.

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, https://doi.org/10.1029/2010JG001566, 2011.

Jung, M., Reichstein, M., Schwalm, C. R., Huntingford, C., Sitch, S., Ahlström, A., Arneth, A., Camps-Valls, G., Ciais, P., Friedlingstein, P., Gans, F., Ichii, K., Jain, A. K., Kato, E., Papale, D., Poulter, B., Raduly, B., Rödenbeck, C., Tramontana, G., Viovy, N., Wang, Y.-P., Weber, U., Zaehle, S., and Zeng, N.: Compensatory Water Effects Link Yearly Global Land CO2 Sink Changes to Temperature, Nature, 541, 516–520, https://doi.org/10.1038/nature20780, 2017.

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., and Reichstein, M.: The FLUXCOM Ensemble of Global Land-Atmosphere Energy Fluxes, Scientific data, 6, 1–14, https://doi.org/10.1038/s41597-019-0076-8, 2019.

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S., Bodesheim, P., Carvalhais, N., Chevallier, F., Gans, F., Goll, D. S., Haverd, V., Köhler, P., Ichii, K., Jain, A. K., Liu, J., Lombardozzi, D., Nabel, J. E. M. S., Nelson, J. A., O'Sullivan, M., Pallandt, M., Papale, D., Peters, W., Pongratz, J., Rödenbeck, C., Sitch, S., Tramontana, G., Walker, A., Weber, U., and Reichstein, M.: Scaling Carbon Fluxes from Eddy Covariance Sites to Globe: Synthesis and Evaluation of the FLUXCOM Approach, Biogeosciences, 17, 1343–1365, https://doi.org/10.5194/bg-17-1343-2020, 2020.

Jung, M., Nelson, J., Migliavacca, M., El-Madany, T., Papale, D., Reichstein, M., Walther, S., and Wutzler, T.: Technical Note: Flagging Inconsistencies in Flux Tower Data, Biogeosciences Discussions, pp. 1–45, https://doi.org/10.5194/bg-2023-110, 2023.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V.: Machine Learning for the Geosciences: Challenges and Opportunities, IEEE Transactions on Knowledge and Data Engineering, 31, 1544–1554, https://doi.org/10.1109/TKDE.2018.2861006, 2019.

Kattge, J., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P. B., Wright, I. J., Cornelissen, J. H. C., Violle, C., Harrison, S. P., Van BODEGOM, P. M., Reichstein, M., Enquist, B. J., Soudzilovskaia, N. A., Ackerly, D. D., Anand, M., Atkin, O., Bahn, M., Baker, T. R., Baldocchi, D., Bekker, R., Blanco, C. C., Blonder, B., Bond, W. J., Bradstock, R., Bunker, D. E., Casanoves, F., Cavender-Bares, J., Chambers, J. Q., Chapin Iii, F. S., Chave, J., Coomes, D., Cornwell, W. K., Craine, J. M., Dobrin, B. H., Duarte, L., Durka, W., Elser, J., Esser, G., Estiarte, M., Fagan, W. F., Fang, J., Fernández-Méndez, F., Fidelis, A., Finegan, B., Flores, O., Ford, H., Frank, D., Freschet, G. T., Fyllas, N. M., Gallagher, R. V., Green, W. A., Gutierrez, A. G., Hickler, T., Higgins, S. I., Hodgson, J. G., Jalili, A., Jansen, S., Joly, C. A., Kerkhoff, A. J., Kirkup, D., Kitajima, K., Kleyer, M., Klotz, S., Knops, J. M. H., Kramer, K., Kühn, I., Kurokawa, H., Laughlin, D., Lee, T. D., Leishman, M., Lens, F., Lenz, T., Lewis, S. L., Lloyd, J., Llusià, J., Louault, F., Ma, S., Mahecha, M. D., Manning, P., Massad, T., Medlyn, B. E., Messier, J., Moles, A. T., Müller, S. C., Nadrowski, K., Naeem, S., Niinemets, Ü., Nöllert, S., Nüske, A., Ogaya, R., Oleksyn, J., Onipchenko, V. G., Onoda, Y., Ordoñez, J., Overbeck, G., Ozinga, W. A., Patiño, S., Paula, S., Pausas, J. G., Peñuelas, J., Phillips, O. L., Pillar, V., Poorter, H., Poorter, L., Poschlod, P., Prinzing, A., Proulx, R., Rammig, A., Reinsch, S., Reu, B., Sack, L., Salgado-Negret, B., Sardans, J., Shiodera, S., Shipley, B., Siefert, A., Sosinski, E., Soussana, J.-F., Swaine, E., Swenson, N., Thompson, K., Thornton, P., Waldram, M., Weiher, E., White, M., White, S., Wright, S. J., Yguel, B., Zaehle, S., Zanne, A. E., and Wirth, C.: TRY – a Global Database of Plant Traits, Global Change Biology, 17, 2905–2935, https://doi.org/10.1111/j.1365-2486.2011.02451.x, 2011.

Klemmer, K., Rolf, E., Robinson, C., Mackey, L., and Rußwurm, M.: SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery, 2024.

Kraft, B., Jung, M., Körner, M., Requena Mesa, C., Cortés, J., and Reichstein, M.: Identifying Dynamic Memory Effects on Vegetation State Using Recurrent Neural Networks, Frontiers in Big Data, 2, 2019.

Kraft, B., Besnard, S., and Koirala, S.: Emulating Ecological Memory with Recurrent Neural Networks, in: Deep Learning for the Earth Sciences, chap. 18, pp. 269–281, John Wiley & Sons, Ltd, https://doi.org/10.1002/9781119646181.ch18, 2021.

Kraft, B., Jung, M., Körner, M., Koirala, S., and Reichstein, M.: Towards Hybrid Modeling of the Global Hydrological Cycle, Hydrology and Earth System Sciences, 26, 1579–1614, https://doi.org/10.5194/hess-26-1579-2022, 2022.

Kraft, B., Schirmer, M., Aeberhard, W. H., Zappa, M., Seneviratne, S. I., and Gudmundsson, L.: CH-RUN: a deep-learning-based spatially contiguous runoff reconstruction for Switzerland, Hydrology and Earth System Sciences, 29, 1061–1082, https://doi.org/10.5194/hess-29-1061-2025, 2025.

Lipton, Z. C., Berkowitz, J., and Elkan, C.: A Critical Review of Recurrent Neural Networks for Sequence Learning, arXiv:1506.00019 [cs], 2015.

Malistov, A. and Trushin, A.: Gradient Boosted Trees with Extrapolation, in: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), pp. 783–789, https://doi.org/10.1109/ICMLA.2019.00138, 2019.

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C.: GLEAM v3: Satellite-Based Land Evaporation and Root-Zone Soil Moisture, Geoscientific Model Development, 10, 1903–1925, https://doi.org/10.5194/gmd-10-1903-2017, 2017.

Mauder, M., Jung, M., Stoy, P., Nelson, J., and Wanner, L.: Energy balance closure at FLUXNET sites revisited, Agricultural and Forest Meteorology, 358, 110 235, https://doi.org/https://doi.org/10.1016/j.agrformet.2024.110235, 2024.

Migliavacca, M., Sonnentag, O., Keenan, T. F., Cescatti, A., O'Keefe, J., and Richardson, A. D.: On the Uncertainty of Phenological Responses to Climate Change, and Implications for a Terrestrial Biosphere Model, Biogeosciences, 9, 2063–2083, https://doi.org/10.5194/bg-9-2063-2012, 2012.

Nakagawa, R., Chau, M., Calzaretta, J., Keenan, T., Vahabi, P., Todeschini, A., Bassiouni, M., and Kang, Y.: Upscaling Global Hourly GPP with Temporal Fusion Transformer (TFT), arXiv preprint arXiv:2306.13815, 2023.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Nelson, J. A., Carvalhais, N., Cuntz, M., Delpierre, N., Knauer, J., Ogée, J., Migliavacca, M., Reichstein, M., and Jung, M.: Coupling Water and Carbon Fluxes to Constrain Estimates of Transpiration: The TEA Algorithm, Journal of Geophysical Research: Biogeosciences, 123, 3617–3632, https://doi.org/10.1029/2018JG004727, 2018.

Nelson, J. A., Pérez-Priego, O., Zhou, S., Poyatos, R., Zhang, Y., Blanken, P. D., Gimeno, T. E., Wohlfahrt, G., Desai, A. R., Gioli, B., Limousin, J.-M., Bonal, D., Paul-Limoges, E., Scott, R. L., Varlagin, A., Fuchs, K., Montagnani, L., Wolf, S., Delpierre, N., Berveiller, D., Gharun, M., Marchesini, L. B., Gianelle, D., Šigut, L., Mammarella, I., Siebicke, L., Black, T. A., Knohl, A., Hörtnagl, L., Magliulo, V., Besnard, S., Weber, U., Carvalhais, N., Migliavacca, M., Reichstein, M., and Jung, M.: Ecosystem Transpiration and Evaporation: Insights from Three Water Flux Partitioning Methods across FLUXNET Sites, Global Change Biology, 26, 6916–6930, https://doi.org/10.1111/gcb.15314, 2020.

Nelson & Walther, Gans, F., Kraft, B., Weber, U., Novick, K., Buchmann, N., Migliavacca, M., Wohlfahrt, G., Šigut, L., Ibrom, A., Papale, D., Göckede, M., Duveiller, G., Knohl, A., Hörtnagl, L., Scott, R. L., Zhang, W., Hamdi, Z. M., Reichstein, M., Aranda-Barranco, S., Ardö, J., Op de Beeck, M., Billdesbach, D., Bowling, D., Bracho, R., Brümmer, C., Camps-Valls, G., Chen, S., Cleverly, J. R., Desai, A., Dong, G., El-Madany, T. S., Euskirchen, E. S., Feigenwinter, I., Galvagno, M., Gerosa, G., Gielen, B., Goded, I., Goslee, S., Gough, C. M., Heinesch, B., Ichii, K., Jackowicz-Korczynski, M. A., Klosterhalfen, A., Knox, S., Kobayashi, H., Kohonen, K.-M., Korkiakoski, M., Mammarella, I., Mana, G., Marzuoli, R., Matamala, R., Metzger, S., Montagnani, L., Nicolini, G., O'Halloran, T., Ourcival, J.-M., Peichl, M., Pendall, E., Ruiz Reverter, B., Roland, M., Sabbatini, S., Sachs, T., Schmidt, M., Schwalm, C. R., Shekhar, A., Silberstein, R., Silveira, M. L., Spano, D., Tagesson, T., Tramontana, G., Trotta, C., Turco, F., Vesala, T., Vincke, C., Vitale, D., Vivoni, E. R., Wang,

Y., Woodgate, W., Yepez, E. A., Zhang, J., Zona, D., and Jung, M.: X-BASE: the first terrestrial carbon and water flux products from an extended data-driven scaling framework, FLUXCOM-X, EGUsphere, 2024, 1–51, https://doi.org/10.5194/egusphere-2024-165, 2024.

Ogle, K., Barber, J. J., Barron-Gafford, G. A., Bentley, L. P., Young, J. M., Huxman, T. E., Loik, M. E., and Tissue, D. T.: Quantifying Ecological Memory in Plant and Ecosystem Processes, Ecology Letters, 18, 221–235, https://doi.org/10.1111/ele.12399, 2015.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K.: Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499, 2016.

Pan, S., Pan, N., Tian, H., Friedlingstein, P., Sitch, S., Shi, H., Arora, V. K., Haverd, V., Jain, A. K., Kato, E., et al.: Evaluation of global terrestrial evapotranspiration using state-of-the-art approaches in remote sensing, machine learning and land surface modeling, Hydrology and Earth System Sciences, 24, 1485–1509, https://doi.org/10.5194/hess-24-1485-2020, 2020.

Pan, S. J. and Yang, Q.: A Survey on Transfer Learning, IEEE Transactions on Knowledge and Data Engineering, 22, 1345–1359, https://doi.org/10.1109/TKDE.2009.191, 2010.

Parker, W. S.: Reanalyses and Observations: What's the Difference?, Bulletin of the American Meteorological Society, 97, 1565 – 1572, https://doi.org/10.1175/BAMS-D-14-00226.1, 2016.

Pastore, A. and Carnini, M.: Extrapolating from neural network models: a cautionary tale, Journal of Physics G: Nuclear and Particle Physics, 48, 084 001, https://doi.org/10.1088/1361-6471/abf08a, 2021.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K. J., Cinti, B. D., de Grandcourt, A., Ligne, A. D., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., di Tommasi, P., Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., Gharun, M., Gianelle, D., Gielen, B., Gioli, B., Gitelson, A., Goded, I., Goeckede, M., Goldstein, A. H., Gough, C. M., Goulden, M. L., Graf, A., Griebel, A., Gruening, C., Grünwald, T., Hammerle, A., Han, S., Han, X., Hansen, B. U., Hanson, C., Hatakka, J., He, Y., Hehn, M., Heinesch, B., Hinko-Najera, N., Hörtnagl, L., Hutley, L., Ibrom, A., Ikawa, H., Jackowicz-Korczynski, M., Janouš, D., Jans, W., Jassal, R., Jiang, S., Kato, T., Khomik, M., Klatt, J., Knohl, A., Knox, S., Kobayashi, H., Koerber, G., Kolle, O., Kosugi, Y., Kotani, A., Kowalski, A., Kruijt, B., Kurbatova, J., Kutsch, W. L., Kwon, H., Launiainen, S., Laurila, T., Law, B., Leuning, R., Li, Y., Liddell, M., Limousin, J.-M., Lion, M., Liska, A. J., Lohila, A., López-Ballesteros, A., López-Blanco, E., Loubet, B., Loustau, D., Lucas-Moffat, A., Lüers, J., Ma, S., Macfarlane, C., Magliulo, V., Maier, R., Mammarella, I., Manca, G., Marcolla, B., Margolis, H. A., Marras, S., Massman, W., Mastepanov, M., Matamala, R., Matthes, J. H., Mazzenga, F., McCaughey, H., McHugh, I., McMillan, A. M. S., Merbold, L., Meyer, W., Meyers, T., Miller, S. D., Minerbi, S., Moderow, U., Monson, R. K., Montagnani, L., Moore, C. E., Moors, E., Moreaux, V., Moureaux, C., Munger, J. W., Nakai, T., Neirynck, J., Nesic, Z., Nicolini, G., Noormets, A., Northwood, M., Nosetto, M., Nouvellon, Y., Novick, K., Oechel, W., Olesen, J. E., Ourcival, J.-M., Papuga, S. A., Parmentier, F.-J., Paul-Limoges, E., Pavelka, M., Peichl, M., Pendall, E., Phillips, R. P., Pilegaard, K., Pirk, N., Posse, G., Powell, T., Prasse, H., Prober, S. M., Rambal, S., Rannik, Ü., Raz-Yaseef, N., Rebmann, C., Reed, D., de Dios, V. R., Restrepo-Coupe, N., Reverter, B. R., Roland, M., Sabbatini, S., Sachs, T., Saleska, S. R., Sánchez-Cañete, E. P., Sanchez-Mejia, Z. M., Schmid, H. P., Schmidt,

M., Schneider, K., Schrader, F., Schroder, I., Scott, R. L., Sedlák, P., Serrano-Ortíz, P., Shao, C., Shi, P., Shironya, I., Siebicke, L., Šigut, L., Silberstein, R., Sirca, C., Spano, D., Steinbrecher, R., Stevens, R. M., Sturtevant, C., Suyker, A., Tagesson, T., Takanashi, S., Tang, Y., Tapper, N., Thom, J., Tomassucci, M., Tuovinen, J.-P., Urbanski, S., Valentini, R., van der Molen, M., van Gorsel, E., van Huissteden, K., Varlagin, A., Verfaillie, J., Vesala, T., Vincke, C., Vitale, D., Vygodskaya, N., Walker, J. P., Walter-Shea, E., Wang, H., Weber, R., Westermann, S., Wille, C., Wofsy, S., Wohlfahrt, G., Wolf, S., Woodgate, W., Li, Y., Zampedri, R., Zhang, J., Zhou, G., Zona, D., Agarwal, D., Biraud, S., Torn, M., and Papale, D.: The FLUXNET2015 Dataset and the ONEFlux Processing Pipeline for Eddy Covariance Data, Scientific Data, 7, 225, https://doi.org/10.1038/s41597-020-0534-3, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019.

Reichstein, M., Bahn, M., Mahecha, M. D., Kattge, J., and Baldocchi, D. D.: Linking Plant and Ecosystem Functional Biogeography, Proceedings of the National Academy of Sciences, 111, 13 697–13 702, https://doi.org/10.1073/pnas.1216065111, 2014.

Reichstein, M., Besnard, S., Carvalhais, N., Gans, F., Jung, M., Kraft, B., and Mahecha, M.: Modelling Landsurface Time-Series with Recurrent Neural Nets, in: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 7640–7643, https://doi.org/10.1109/IGARSS.2018.8518007, 2018.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep Learning and Process Understanding for Data-Driven Earth System Science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Reichstein, M., Ahrens, B., Kraft, B., Camps-Valls, G., Carvalhais, N., Gans, F., Gentine, P., and Winkler, A. J.: Combining system modeling and machine learning into hybrid ecosystem modeling, in: Knowledge Guided Machine Learning, pp. 327–352, Chapman and Hall/CRC, 2022.

Rußwurm, M. and Körner, M.: Temporal Vegetation Modelling Using Long Short-Term Memory Networks for Crop Identification from Medium-Resolution Multi-spectral Satellite Images, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1496–1504, IEEE Computer Society, https://doi.org/10.1109/CVPRW.2017.193, 2017.

Schaaf, C. and Wang, Z.: MCD43A2 MODIS/Terra+Aqua BRDF/Albedo Quality Daily L3 Global - 500m V006. NASA EOSDIS Land Processes Distributed Active Archive Center, https://doi.org/10.5067/modis/mcd43a2.006, 2015a.

Schaaf, C. and Wang, Z.: MCD43A4 MODIS/Terra+Aqua BRDF/Albedo Nadir BRDF Adjusted Ref Daily L3 Global - 500m V006. NASA EOSDIS Land Processes Distributed Active Archive Center, https://doi.org/10.5067/modis/mcd43a4.006, 2015b.

Schaaf, C. and Wang, Z.: MCD43C4 MODIS/Terra+Aqua BRDF/Albedo Nadir BRDF-Adjusted Ref Daily L3 Global 0.05Deg CMG V006. NASA EOSDIS Land Processes Distributed Active Archive Center, https://doi.org/10.5067/modis/mcd43c4.006, 2015c.

Stoy, P. C., Mauder, M., Foken, T., Marcolla, B., Boegh, E., Ibrom, A., Arain, M. A., Arneth, A., Aurela, M., Bernhofer, C., Cescatti, A., Dellwik, E., Duce, P., Gianelle, D., van Gorsel, E., Kiely, G., Knohl, A., Margolis, H., McCaughey, H., Merbold, L., Montagnani, L., Papale, D., Reichstein, M., Saunders, M., Serrano-Ortiz, P., Sottocornola, M., Spano, D., Vaccari, F., and Varlagin, A.: A data-driven analysis of energy balance closure across FLUXNET research sites: The role of landscape scale heterogeneity, Agricultural and Forest Meteorology, 171-172, 137–152, https://doi.org/https://doi.org/10.1016/j.agrformet.2012.11.004, 2013.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., et al.: Predicting Carbon Dioxide and Energy Fluxes across Global FLUXNET Sites with Regression Algorithms, Biogeosciences (Online), 13, 4291–4313, https://doi.org/10.5194/bg-13-4291-2016, 2016.

Upton, S., Reichstein, M., Gans, F., Peters, W., Kraft, B., and Bastos, A.: Constraining biospheric carbon dioxide fluxes by combined top-down and bottom-up approaches, Atmospheric Chemistry and Physics, 24, 2555–2582, https://doi.org/10.5194/acp-24-2555-2024, 2024.

Valmassoi, A., Keller, J. D., Kleist, D. T., English, S., Ahrens, B., Ďurán, I. B., Bauernschubert, E., Bosilovich, M. G., Fujiwara, M., Hersbach, H., Lei, L., Löhnert, U., Mamnun, N., Martin, C. R., Moore, A., Niermann, D., Ruiz, J. J., and Scheck, L.: Current Challenges and Future Directions in Data Assimilation and Reanalysis, Bulletin of the American Meteorological Society, 104, E756 – E767, https://doi.org/10.1175/BAMS-D-21-0331.1, 2023.

Van Houdt, G., Mosquera, C., and Nápoles, G.: A Review on the Long Short-Term Memory Model, Artificial Intelligence Review, 53, 5929–5955, https://doi.org/10.1007/s10462-020-09838-1, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention Is All You Need, in: Advances in Neural Information Processing Systems, pp. 5998–6008, 2017.

Walther & Besnard, Nelson, El-Madany, Migliavacca, Carvalhais, W., Ermida, Brümmer, Schrader, Prokushkin, Panov, and Jung: Technical Note: A View from Space on Global Flux Towers by MODIS and Landsat: The FluxnetEO Data Set, Biogeosciences, 19, 2805–2840, https://doi.org/10.5194/bg-19-2805-2022, 2022.

Wan, Z., Hook, S., and Hulley, G.: MOD11A1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006 [Data Set]. NASA EOSDIS Land Processes Distributed Active Archive Center, https://doi.org/10.5067/modis/mod11a1.006, 2015a.

Wan, Z., Hook, S., and Hulley, G.: MOD11C1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 0.05Deg CMG V006. NASA EOSDIS Land Processes Distributed Active Archive Center, https://doi.org/10.5067/modis/mod11c1.006, 2015b.

Warm Winter 2020 Team and ICOS Ecosystem Thematic Centre: Warm Winter 2020 Ecosystem Eddy Covariance Flux Product for 73 Stations in FLUXNET-Archive Format—Release 2022-1, https://www.icos-cp.eu/data-products/2G60-ZHAK, https://doi.org/10.18160/2G60-ZHAK, 2022.

Xiao, J., Ollinger, S. V., Frolking, S., Hurtt, G. C., Hollinger, D. Y., Davis, K. J., Pan, Y., Zhang, X., Deng, F., Chen, J., et al.: Data-Driven Diagnostics of Terrestrial Carbon Dynamics over North America, Agricultural and Forest Meteorology, 197, 142–157, https://doi.org/10.1016/j.agrformet.2014.06.013, 2014.

Xu, T., Guo, Z., Liu, S., He, X., Meng, Y., Xu, Z., Xia, Y., Xiao, J., Zhang, Y., Ma, Y., et al.: Evaluating Different Machine Learning Methods for Upscaling Evapotranspiration from Flux Towers to the Regional Scale, Journal of Geophysical Research: Atmospheres, 123, 8674–8690, https://doi.org/10.1029/2018JD028447, 2018.

Zhang, W., Nelson, J. A., Miralles, D. G., Mauder, M., Migliavacca, M., Poyatos, R., Reichstein, M., and Jung, M.: A New Post-Hoc Method to Reduce the Energy Imbalance in Eddy Covariance Measurements, Geophysical Research Letters, 51, e2023GL107 084, https://doi.org/10.1029/2023GL107084, 2024.

Zhao, W. L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., Lin, C., Li, X., and Qiu, G. Y.: Physics-Constrained Machine Learning of Evapotranspiration, Geophysical Research Letters, 46, 14 496–14 507, https://doi.org/10.1029/2019GL085291, 2019.

Zhu, S., Quaife, T., and Hill, T.: Uniform upscaling techniques for eddy covariance FLUXes (UFLUX), International Journal of Remote Sensing, 45, 1450–1476, https://doi.org/10.1080/01431161.2024.2312266, 2024.