https://doi.org/10.5194/egusphere-2024-2896
Preprint. Discussion started: 10 October 2024

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], Zayd Mahmoud Hamdi[1], and Martin Jung[1]

[1]Biogeochemical Integration, Max Planck Institute for Biogeochemistry (MPI BGC), Jena, Germany
[2]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
[3]Chair of Remote Sensing Technology, Technical University of Munich (TUM), Munich, Germany
[4]Environmental Computational Science and Earth Observation Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Sion, Switzerland
[5]Laboratory of Geo-information Science and Remote Sensing, Wageningen University (WU), Wageningen, Netherlands

Correspondence: Basil Kraft (basil.kraft@env.ethz.ch)

Abstract.

Estimating ecosystem-atmosphere fluxes such as evapotranspiration (ET) in a robust manner and at global scale remains a challenge. Machine learning (ML)–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 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 the modeling and upscaling of ET. We compared different types of covariates (meteorological, remote sensing, and plant functional types) and their impact on model performance at the site level in a cross-validation setup. For the upscaling from site to global coverage, we input the best-performing combination of covariates—which was meteorological and remote sensing observations—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 the sequential models diminished with the inclusion of remote-sensing-based predictors. The generated patterns of global ET variability were highly consistent across all 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. In terms of global integrals, the neural network ensembles showed a sizable spread due to training data subsets, which exceeds differences among neural network variants. XGBoost showed smaller ensemble spread compared to neural networks in particular when conditions were poorly represented in the training data.

https://doi.org/10.5194/egusphere-2024-2896
Preprint. Discussion started: 10 October 2024

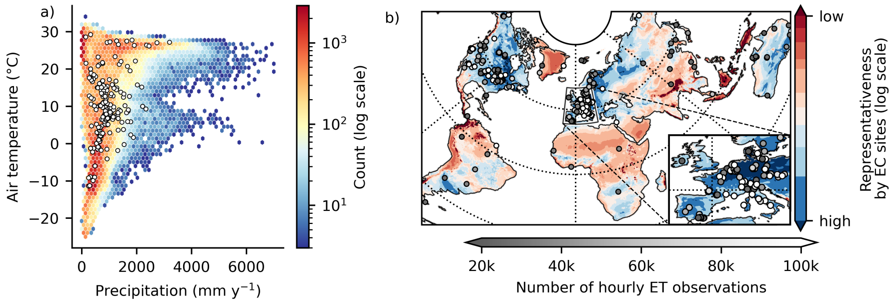

Figure 1. Overview of eddy covariance (EC) sites used in this study: **a)** Distribution of EC sites (white points) and map grid-cells (background color) within the global climate in terms of mean temperature and annual precipitation. **b)** Geographic EC site locations in different gray scales according 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.

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. Approaches such as knowledge-guided ML, which encourage physically consistent results while harnessing the efficiency of ML, or transfer learning, should be 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.

1 Introduction

Measurements of land-atmosphere fluxes of gases, such as water vapor or carbon, are crucial for understanding the interactions between climate and ecosystems. Instruments at eddy covariance (EC) stations measure such fluxes integrated over 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 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. 1).

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 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 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 at all EC station sites 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 using 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 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., instantaneous models that do 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 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 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 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, 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 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. 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 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. For each model, we conduct a feature ablation experiment, where we drop feature groups. The groups considered in addition to meteorology are dynamic state representations, based on remotely sensed observations, and plant functional types (PFTs), which are static descriptors of site vegetation 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).

The key contributions of this study are:

- A systematic comparison of the effectiveness of different ML methods for site-level 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 converted 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 ($R_{\text{in, pot}}$), and time-derivative of potential shortwave irradiation ($\Delta R_{\text{in, pot}}$). In addition, we used remote sensing observations from the moderate imaging spectroradiometer (MODIS) sensor on board both Terra and Aqua satellite platforms, collection v006. 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 retrieved at site level from the MCD43A4 product (Schaaf and Wang, 2015a, spatial resolution of 500 m), and from MCD43C4 for the global data runs (Schaaf and Wang, 2015b, spatial resolution of 0.05°). Additionally, the land surface temperature (LST) was obtained from MOD11A1 at site level (Wan et al., 2015a, spatial resolution of 1 km), and from MOD11C1 globally (Wan et al., 2015b, spatial resolution of 0.05°). Each remote sensing product was interpolated to daily resolution. 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 & Walther et al., 2024; Walther 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.

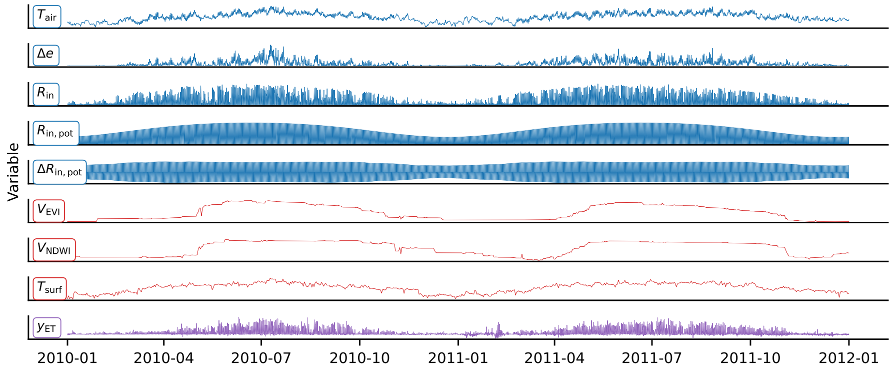

Figure 2. Two-year time series from the Hainich site (DE-Hai) in Germany. Meteorological covariates (hourly): near-surface air temperature (T_{air}), vapor pressure deficit (Δe), shortwave irradiation (R_{in}), potential shortwave irradiation ($R_{\text{in, pot}}$), and time-derivative of potential shortwave irradiation ($\Delta R_{\text{in, pot}}$). 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}).

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 cross-validation setup. The models are trained with different types of covariates: meteorological (`met`), remote sensing (`rs`), and PFTs (`pft`). This experiments with different sets of variables as model inputs, summarized in Tab. 1, gives insights into the relevance of the types of covariates. In total, four covariate setups were tested and combined with five machine-learning models, i.e., twenty models were trained and evaluated at the site level. For the evaluation, we use the Nash-Sutcliffe modeling efficiency (Nash and Sutcliffe, 1970)

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

https://doi.org/10.5194/egusphere-2024-2896
Preprint. Discussion started: 10 October 2024

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 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.

Table 1. The ablation experiment with different covariate groups: Meteorological (`met`, hourly), plant functional type (`pft`, constant), and remote sensing–based (`rs`, daily). Each item corresponds to a unique model setup.

Setup	Covariate groups	Covariates
`met`	Meteorology	T_{air}, Δe, R_{in}, $R_{\text{in, pot}}$, $\Delta R_{\text{in, pot}}$
`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}

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 $\Delta R_{\text{in, pot}}$; enhanced vegetation index V_{EVI}; normalized difference water index V_{NDWI}; land surface temperature T_{surf}; plant functional type S_{PFT}.

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 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.*, the same covariates were used. The remote sensing and PFT covariates were repeated in time to obtain uniform inputs. 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

https://doi.org/10.5194/egusphere-2024-2896
Preprint. Discussion started: 10 October 2024

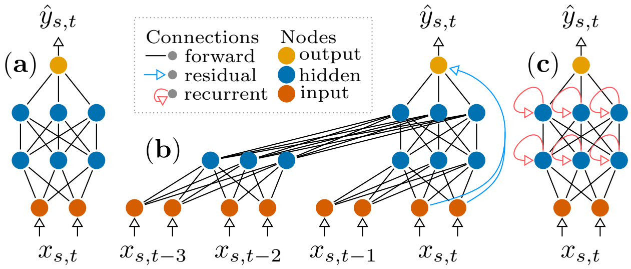

Figure 3. The three neural network layers used in this study: **a)** a feed-forward neural network, **b)** a temporal convolutional network (TCN), which applies causal (i.e., does not consider future time steps) 1D convolutions in the time dimension and **c)** a long short-term memory (LSTM) model, which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

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 was kept identical across different models and architectures. To decrease the dependency between the sets, we ensure that sites in close spatial proximity are part of the same set using clustering of coordinates. 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%), 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.

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.

https://doi.org/10.5194/egusphere-2024-2896
Preprint. Discussion started: 10 October 2024

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, we decided to use only the overall best covariates setup, which was `met+rs`, for all models. We did not use the `linearreg` model for upscaling, as it showed significantly worse performance compared to the non-linear algorithms. For each of the four remaining ML models, we compute an ensemble of eight upscaling products. The members, herein 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 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.

4.1.1 Model performances across scales

The overall site-level performance of hourly ET, shown in Fig. 4, depended more on 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 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 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 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 sequential models. On the seasonal scale and without remote sensing covariates, the sequential models performed best. With remote sensing covariates, 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 interannual variability, all models show very low performance and

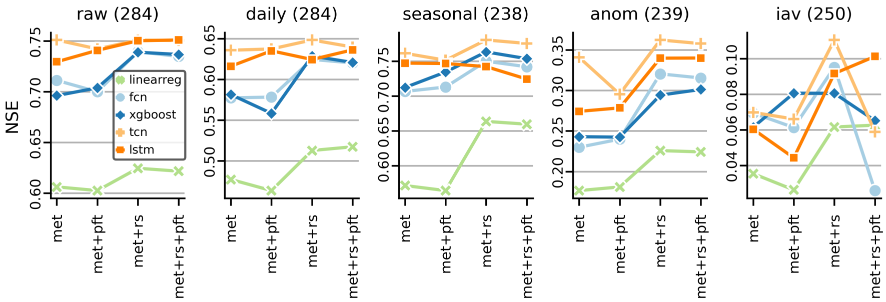

Figure 4. Site-level evaluation for modeled evapotranspiration: Median Nash–Sutcliffe model efficiency (NSE) across sites for different models (lines) and covariates (x-axis) for different temporal scales (panels). The scales shown are `raw` for hourly, `daily` for daily aggregates, `seasonal` for daily 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.

the patterns are less clear; while the `lstm` improved with adding additional covariates, the other models showed a decreased performance with PFT.

The linear models (`linearreg`) fell short consistently because evapotranspiration is characterized by complex interactions and non-linear functions such that the advantages of ML become noteworthy. While the sequential models only marginally improved with adding covariates related to ecosystem state, the non-sequential models improved more prominently. This is, on the one hand, a sanity check for the sequential models: They were able to extract additional information from the temporal meteorological covariates as expected. Still, adding remote sensing covariates improved and stabilized their performance. On the other hand, this shows that remote sensing covariates are useful proxies for ecological memory: The sequential models were able to extract additional information from antecedent covariates, but most of the information seems to be comprised in remote sensing covariates, and thus, the non-sequential models achieved 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 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 to uncommon or extreme conditions. This could be related to processes that are partially observable by remote sensing, but cannot be derived from meteorology, such as forest or crop management and natural disturbances. The low performance on `iav` was also reported by Jung et al. (2019) and Nelson & Walther et al. (2024).

It is notable that adding PFTs as covariates did not improve (and sometimes even harmed) model performance. PFTs have long been criticized for not being representative of the continuous characteristics of ecosystems (Reichstein et al., 2014; Kattge et al., 2011). As our experiments suggest, adding PFTs brings little to no information, while increasing the input features

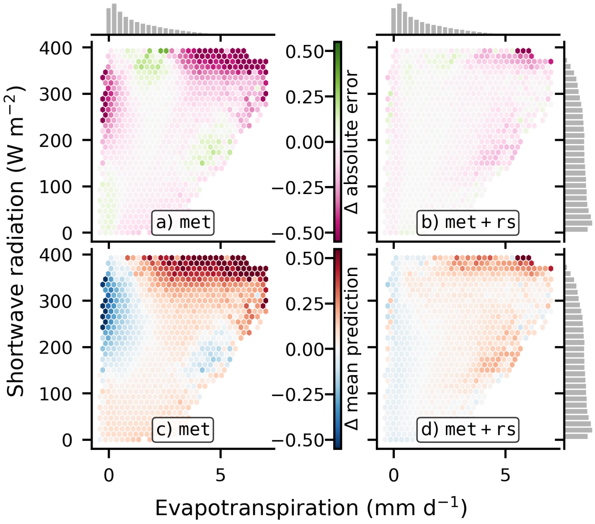

Figure 5. Comparison of a sequential (`lstm`) and a non-sequential (`xgboost`) model in terms of absolute error and mean predicted ET in the space of observed evapotranspiration × shortwave irradiation: Panel **a)** and **b)** show the *difference in absolute error* of the `lstm` minus the `xgboost`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

space; in fact, each of the nine PFTs adds 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, such as soil properties or plant traits. This finding advocates for a comprehensive feature selection to identify more relevant static features and, therefore, to avoid inflating the input dimensionality. Alternatively, or in addition, location embeddings, such as SatCLIP (Klemmer et al., 2024), could help improve model generalizeability by providing a condensed representation of land surface characteristics.

4.1.2 Memory effects matter

As noted before, the difference in model performance between the sequential and non-sequential model shrank when remote sensing observations were added as covariates. We investigate these differences in Fig. 5: As illustrated in the top left panel (Fig. 5a), which shows the absolute error difference between the the `lstm` and `xgboost`, the non-sequential model performed worse (i.e., magenta colors) with high incoming radiation (> 200 Wm^{-2}) paired with either low or high observed ET. To represent these conditions, there must be an implicit knowledge about the water availability learned by the models. It seems that the sequential model was able to learn proxies of wetness from the meteorological time series, but the non-sequential model

https://doi.org/10.5194/egusphere-2024-2896
Preprint. Discussion started: 10 October 2024

was not. Rather, the latter learned an average behavior, which worked well in most instances. As remote sensing covariates were added, the differences were reduced but did not entirely disappear (Fig. 5b). This interpretation is supported by 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 covariates (Fig. 5c), the non-sequential model overestimated ET with high incoming radiation but low observed ET (i.e., blue colors); these are dry conditions that the model failed to identify. On the contrary, large observed ET was underestimated by the non-sequential model (i.e., red colors); these are, supposedly, wet conditions. When adding the remote sensing observations as covariates (Fig. 5d), the differences were reduced significantly. This comparison illustrates why memory effects play a role in modeling ET and how remote sensing covariates are good, but 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 `tcn`, the context length depends on the tuned hyper-parameters: While the `lstm` processes the entire time series sequentially, the `tcn`'s context depends on the number of layers and the kernel size. For `tcn`, 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.

4.2 Scaling evapotranspiration to global coverage

With the models optimized at the site level, we create global ensembles of ET estimates. 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 as covariates. However, when scaling globally, data distribution shifts can (and will) affect different model types in different ways. The shifts evolve 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 while scaling out of the flux station locations.

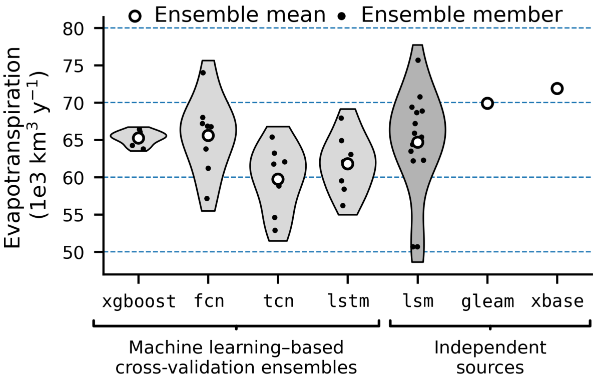

Figure 6. Global annual evapotranspiration (ET) per model. The violin plots represent the density of independent cross-validation runs (black dots), with their mean values across runs displayed as white dots. The data from a number of land surface models (lsm), the GLEAM product (gleam), and FLUXCOM X-Base (xbase) are added as reference.

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 65·10^3 km^3 y^{-1} for the non-sequential models xgboost (65.3±0.9·10^3 km^3 y^{-1}) and fcn (65.6±5.0·10^3 km^3 y^{-1}), in agreement with the land surface models ensemble mean (lsm) of $64.7 \pm 6.9 \cdot 10^3$ km^3 y^{-1}, and to about $60 \cdot 10^3$ km^3 y^{-1} for the sequential models tcn ($59.8 \pm 4.3 \cdot 10^3$ km^3 y^{-1}) and lstm ($61.8 \pm 3.7 \cdot 10^3$ km^3 y^{-1}) on average (Fig. 6). This amounts to a 8% lower annual ET estimate by the sequential models. Both gleam and xbase estimate a larger global ET around 70×10^3 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 xgboost members was considerably smaller.

Overall, the ML models showed consistent patterns of spatial mean, while systematic deviations are evident in mid to low latitudes. This is shown in Fig. 7a, 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 Fig. 7a), 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).

In terms of spatio-temporal patters, all ML models showed a very similar agreement with gleam. This is shown in the diagonal panels in Fig. 8, which represents the monthly ET values of the ML models versus gleam, with pooled temporal

https://doi.org/10.5194/egusphere-2024-2896
Preprint. Discussion started: 10 October 2024

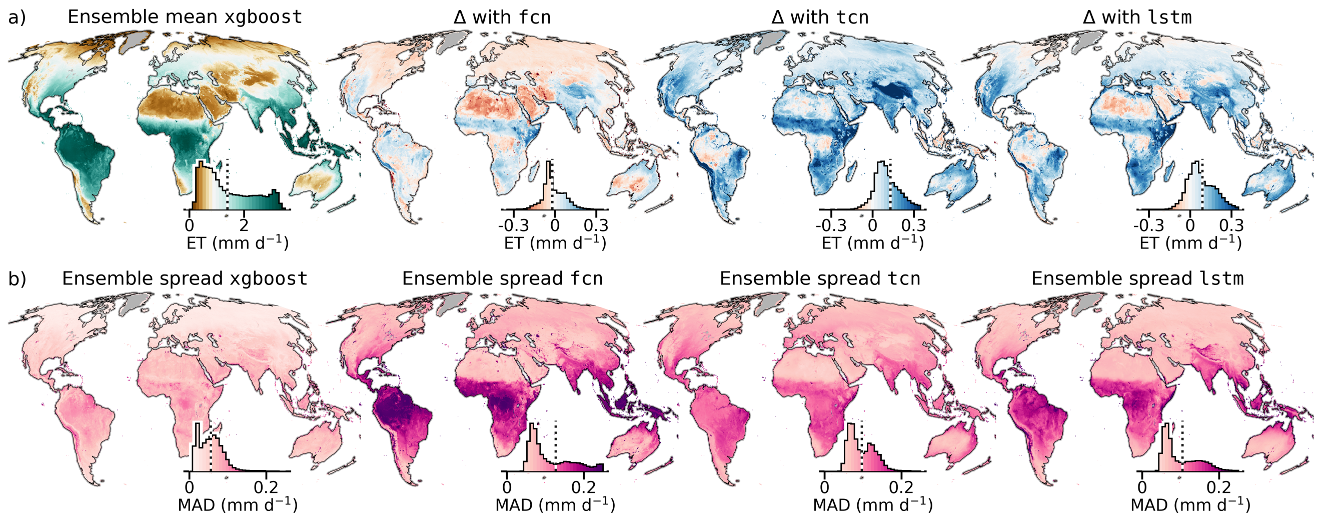

Figure 7. Spatial model evaluation and comparison. **a)** Grid level cross-validation ensemble mean ET for the `xgboost` model is shown in the leftmost map, the difference between `xgboost` and the neural network–based models is shown in the remaining columns. **b)** The grid level median absolute deviation (MAD) per ML model quantifies the cross-validation ensemble uncertainty in $\mathrm{mm\,d^{-1}}$. The map inset histograms represent the distribution of the values weighted by the grid cell area; the median is shown as dashed black line.

and spatial dimension. The linear correlation was between $r = 0.86$ and $r = 0.87$ in all cases. In the upper triangular panels in Fig. 8, 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 found between the sequential models. The weakest relationships were found between `xgboost` and the sequential models ($r = 0.92$ and $r = 0.93$).

The neural network-based models exhibited considerably larger ensemble spread mainly in the tropics, as displayed in Fig. 7b. The figure shows the grid cell median absolute deviation (MAD) per model in $\mathrm{mm\,d^{-1}}$, computed on a monthly scale and averaged across time afterwards. Here, `xgboost` (left-hand side panel) had a low ensemble spread in general, with slightly larger values 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 ensemble spread of the neural networks did not show a strong agreement in terms of spatio-tempoal patterns. The lower triangular panels in Fig. 8 indicate that the association was weak between `tcn` 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 `xgboost` was not much lower with values from $r = 0.59$ to $r = 0.69$.

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 ET predicted by the sequential models compared to the non-sequential models. Third, we have a closer look at the larger ensemble spread of the neural networks when compared to `xgboost`.

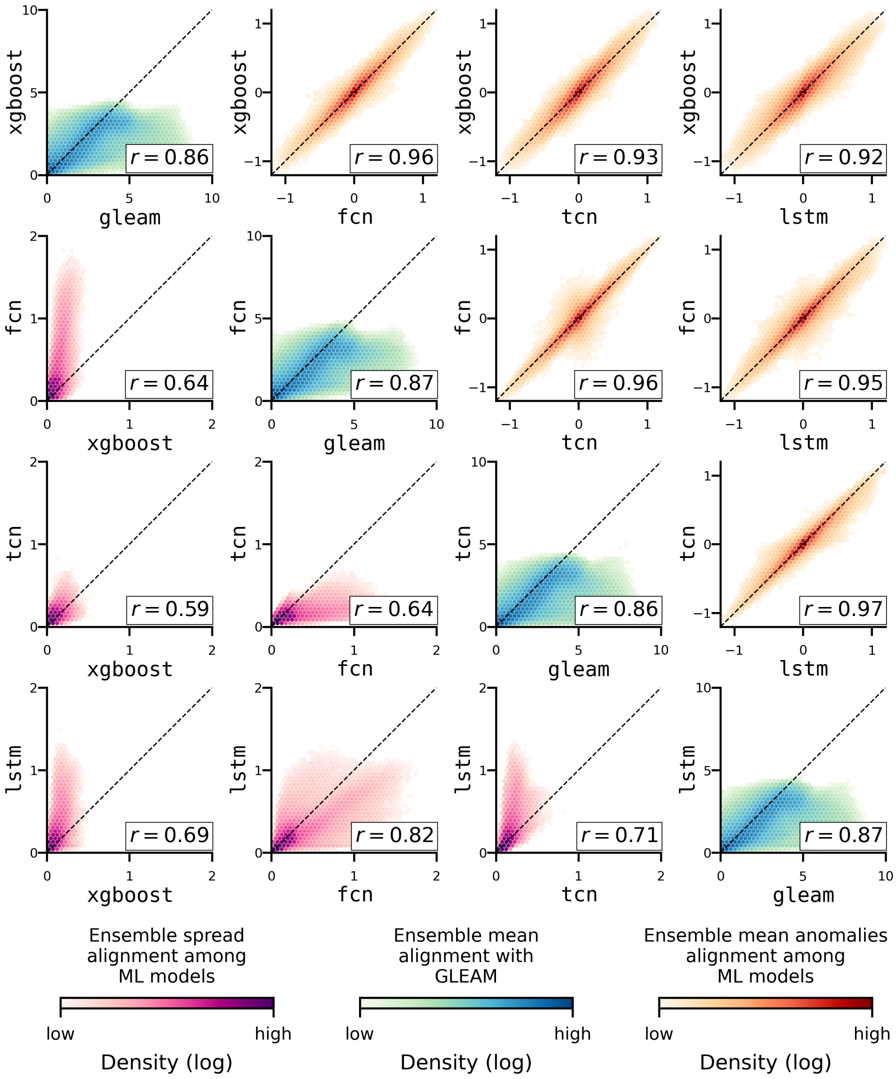

Figure 8. Comparison of monthly ET among machine learning (ML) models and with the GLEAM product (gleam). All panels represent relationships (log density) of pooled spatial and temporal values, axes have units $\mathrm{mm\,d^{-1}}$. The strength of the relationship is quantified with the Pearson correlation (r). The lower triangular (magenta hues) shows 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.

4.2.2 Inconsistencies between annual ET across setups

The global ET estimates, especially by the sequential neural networks, were low compared to our current understanding of global ET magnitude of about $70 \pm 5 \cdot 10^3\ \mathrm{km}^3\ \mathrm{y}^{-1}$ based on a variety of methods (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; Zhang et al., 2023) for which no correction has been applied here. The associated uncertainty for global ET is estimated to be about 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 be added in final version). However, 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 al., 2024). 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.

The inconsistencies between `xbase` and our results were surprising, as the same framework, and in the case of `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 upscaling product 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 influence on the upscaling result, we posit that the PFT had the largest impact. 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 (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.

4.2.3 Lower ET predicted by sequential models

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 (Fig. 4). It is possible that the sequential models learned a better representation of the

https://doi.org/10.5194/egusphere-2024-2896
Preprint. Discussion started: 10 October 2024

Table 2. 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`	`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

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 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 `xgboost` ($r = 0.65$, $\rho = 0.81$), followed by `lstm` and `fcn` ($r = 0.45$, $\rho = 0.31$) and `xgboost` and `fcn` ($r = 0.36$, $\rho = 0.62$). 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) representation of the processes.

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 Barron, 2022; Valmassoi et al., 2023). Additionally, upscaling involves extrapolating into poorly represented ecoclimatic regions (Fig. 1), potentially affecting sequential models more severely due to their reliance on past meteorology. As a reminder, the remote sensing data originates from the same data source for site level and global grids and thus, the distribution shift is less severe compared to the meteorological covariates. Therefore, the models that learn ecosystem state via remote sensing covariates rather than via past meteorology could be more robust. Consequently, we hypothesize that sequential models are more impacted by distribution shifts, which calls for efforts to close this gap.

In contrast to 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. 8), and the small differences in terms of monthly ET anomalies (upper triangular panels in Fig. 8), indicate that the distribution shift from site level to global gridded scale introduces stronger divergence on the mean values than on the relative spatial patterns and temporal dynamics.

4.2.4 Larger ensemble spread by the neural networks

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 `xgboost` model demonstrated considerably lower variability across cross-validation members (Fig. 6). A low ensemble spread is generally deemed advantageous. Nonetheless, it is unclear whether the high robustness of `xgboost` 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 site level (Fig. 4), where the performance of `fcn` and `xgboost` was nearly identical across various scales and data setups. 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 `xgboost`. However, `xgboost` was not an outlier among the models in terms of patterns of spatial ensemble spread (the lower triangular in Fig. 8), alignment with GLEAM (diagonal in Fig. 8), or spatio-temporal patterns of upscaled anomalies (the upper triangular in Fig. 8). 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).

4.3 Lessons learned and outlook

From site-level analyses, it was observed that sequential models generally outperform non-sequential models in ET flux modeling (see 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 state, such as vegetation indices, were incorporated, the performance gap between non-sequential models (e.g., XGBoost and fully-connected feed-forward neural networks) and sequential models narrowed, although latter continued to exhibit superior performance at the anomaly scale. It is conceivable that advanced deep learning 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 only marginal improvements, aligning with our observations that EC flux modeling remains a data-limited challenge with limited benefits from time-domain deep learning techniques to date.

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 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 cross-validation, data handling, and ML configuration in influencing these uncertainties. Notably, similar findings were reported by Zhu et al. (2024).

The XGBoost method resulted in a more robust upscaling ensemble compared to those derived from neural network models (Fig. 7 & 8). 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 EC upscaling. 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).

Nonetheless, we believe that (sequential) deep learning is a promising approach to enhance flux modeling and upscaling, offering 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.

To enhance the performance of deep learning approaches in EC measurement upscaling, leveraging additional data sources could be highly beneficial. Techniques such as transfer learning are particularly effective (Caruana, 1997; Pan and Yang, 2010). By transferring knowledge from one region to another or utilizing richer, related datasets, models can achieve better generalization, especially in data-sparse areas. To address shifts in covariates caused by various factors discussed earlier, domain adaptation methods (He et al., 2023) offer a valuable toolbox for reducing upscaling biases. Additionally, more efficient extrapolation through meta-learning can further improve generalizability in undersampled regions (Nathaniel et al., 2023).

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, which we identified to be currently the main challenge in flux upscaling.

In this study, we focused on the modeling and upscaling of ET. However, we believe these findings are broadly applicable to the upscaling of ecosystem-atmosphere fluxes in general.

5 Conclusions

In this study, we assessed different data setups and ML approaches for modeling ET fluxes at EC sites in a cross-validation setup and assessed the robustness and quality of upscaled ET at a global scale. From our analysis at the site level, we conclude that sequential deep learning approaches can outperform non-sequential models for ET flux modeling.

The sequential models learned memory effects related to water availability, which led to a better representation of arid and wet conditions. 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 feature selection to keep the number of covariates low.

The sequential and non-sequential models yielded 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 models to get a plausible estimate of robustness.

Given the additional complexity of the sequential neural networks and the sequential data handling, the relatively small performance increase at site level, and considering the underestimation of ET globally and the large ensemble spread, we conclude that using such advanced modeling techniques is not a strict requirement for modeling ET globally. Non-sequential machine learning, such as XGBoost, can provide comparably robust predictions across scales when paired with good-quality meteorological and remote sensing covariates.

The potential of upscaling ET to a global scale via modern ML approaches seems to be limited by the information content in the EC site-level data, and hence, 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 covariates or additional EC stations, regional constraints, related data via transfer learning, and the incorporation of prior scientific knowledge could increase both the robustness and physical consistency of the global upscaling products. By embracing these strategies, deep learning has the potential to deliver more precise, robust, and physically grounded predictions across diverse environmental scenarios.

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

https://doi.org/10.5194/egusphere-2024-2896
Preprint. Discussion started: 10 October 2024

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+rs`	`met+pft+rs`
`xgboost`	`max_depth`	$\{6, 8, 10, 12\}$	10	12	8	12
	`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	5	5	5
	`max_delta_step`	$\{1, 5, 10\}$	10	10	5	5
	`# parameters`	derived	208K	640K	132K	1 300K
`fcn`	`num_hidden`	$\{128, 256\}$	128	128	256	256
	`num_layers`	$\{3, 4\}$	3	3	4	3
	`dropout`	$\{0.0, 0.2\}$	0.2	0.2	0.2	0.2
	`learning_rate`	$\{10^{-6}, 10^{-5}, 10^{-4}\}$	10^{-6}	10^{-6}	10^{-5}	10^{-4}
	`weight_decay`	$\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$	10^{-3}	10^{-3}	10^{-3}	10^{-1}
	`# parameters`	derived	17K	19K	134K	71K
`tcn`	`num_hidden`	$\{64, 128, 256\}$	64	256	256	256
	`num_layers`	$\{2, 3, 4\}$	4	3	4	4
	`kernel_size`	$\{4, 8, 16\}$	16	16	4	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^{-6}	10^{-5}	10^{-5}
	`weight_decay`	$\{10^{-3}, 10^{-2}, 10^{-1}\}$	10^{-3}	10^{-1}	10^{-2}	10^{-2}
	`temp. context`	derived	450	210	90	90
	`# parameters`	derived	473K	5 600K	2 000K	2 200K
`lstm`	`num_hidden`	$\{64, 128, 256\}$	128	256	128	64
	`num_layers`	$\{1, 2\}$	2	1	2	1
	`dropout`	$\{0.0, 0.2\}$	0	0.2	0.2	0.0
	`learning_rate`	$\{10^{-6}, 10^{-5}, 10^{-4}\}$	10^{-4}	10^{-6}	10^{-6}	10^{-6}
	`weight_decay`	$\{10^{-3}, 10^{-2}, 10^{-1}\}$	10^{-1}	10^{-3}	10^{-3}	10^{-1}
	`# parameters`	derived	217K	344K	219K	25K

Appendix A

Author contributions. 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, MJ, FG, 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.

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

Acknowledgements. 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

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.

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.

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.

Nathaniel, J., Liu, J., and Gentine, P.: MetaFlux: Meta-learning Global Carbon Fluxes from Sparse Spatiotemporal Observations, Scientific Data, 10, 440, https://doi.org/10.1038/s41597-023-02349-y, 2023.

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 & 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.: 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, 2015a.

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, 2015b.

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, S., Besnard, S., Nelson, J. A., El-Madany, T. S., Migliavacca, M., Weber, U., Carvalhais, N., Ermida, S. L., Brümmer, C., Schrader, F., Prokushkin, A. S., Panov, A. V., and Jung, M.: 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., Jung, M., Migliavacca, M., Poyatos, R., Miralles, D. G., El-Madany, T. S., Galvagno, M., Carrara, A., Arriga, N., Ibrom, A., Mammarella, I., Papale, D., Cleverly, J. R., Liddell, M., Wohlfahrt, G., Markwitz, C., Mauder, M., Paul-Limoges, E., Schmidt, M., Wolf, S., Brümmer, C., Arain, M. A., Fares, S., Kato, T., Ardö, J., Oechel, W., Hanson, C., Korkiakoski, M., Biraud, S., Steinbrecher, R., Billesbach, D., Montagnani, L., Woodgate, W., Shao, C., Carvalhais, N., Reichstein, M., and Nelson, J. A.: 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.

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.