Causal Drivers of Land-Atmosphere Carbon Fluxes from Machine Learning Models and Data

Mozhgan A. Farahani[1], **Allison E. Goodwell**[1,2]

[1]University of Colorado Denver, Department of Civil Engineering
[2]Prairie Research Institute, University of Illinois at Urbana-Champaign

[2]Prairie Research Institute, 615 E. Peabody Dr. MC-650, Champaign, IL 61820

Key Points:

- Information theory measures describe individual and joint causal relationships in observed versus modeled vertical carbon dioxide fluxes.
- Three machine learning models overestimate unique information from sources at the expense of synergistic, or pairwise information.
- Regionally trained models have improved functional performance that is not always captured by traditional predictive performance metrics.

Corresponding author: Allison Goodwell, `goodwel2@illinois.edu`

Abstract

Interactions among atmospheric, root-soil, and vegetation processes drive carbon dioxide fluxes (Fc) from land to atmosphere. Eddy covariance measurements are commonly used to measure Fc at sub-daily timescales and validate process-based and data-driven models. However, these validations do not reveal process interactions, thresholds, and key differences in how models replicate them. We use information theory-based measures to explore multivariate information flow pathways from forcing data to observed and modeled hourly Fc, using flux tower datasets in the Midwestern U.S. in intensively managed corn-soybean landscapes. We compare Multiple Linear Regressions (MLR), Long-Short Term Memory (LSTM), and Random Forests (RF) to evaluate how different model structures use information from combinations of sources to predict Fc. We extend a framework for model predictive performance and functional performance, which examines the full suite of dependencies from all forcing variables to the observed or modeled target. Of the three model types, RF exhibited the highest functional and predictive performance. Regionally trained models demonstrate lower predictive but higher functional performance compared to site-specific models, suggesting superior reproduction of observed relationships. This study shows that some metrics of predictive performance encapsulate functional behaviors better than others, highlighting the need for multiple metrics of both types. This study improves our understanding of carbon fluxes in an intensively managed landscape, and more generally provides insight into how model structures and forcing variables translate to interactions that are well versus poorly captured in models.

Plain Language Summary

In an agricultural landscape, exchanges of carbon dioxide between the land and atmosphere occur due to photosynthesis and respiration, and depend on weather, soil, and vegetation conditions. In modeling, predictive performance focuses on the relationship between observed and modeled outputs, while functional performance considers the relationships between interacting inputs and outputs. We compare several performance measures for three different machine learning models that simulate sub-daily carbon fluxes. We look at how drivers such as solar radiation, soil moisture, temperature, humidity, and rainfall provide information to carbon fluxes, and whether different machine learning models also capture these interactions. In other words:

Air, soil, and plants drive carbon's upward path,
Models are detectives, interpreting their math.
With information theory, we map data's travel courses,
To see how models find or miss carbon's causal sources.

1 Introduction

The ecohydrologic system constitutes a complex web of interactions between water, soil, and vegetation. The exchange of carbon dioxide (CO_2) between the land and atmosphere plays a significant role in the Earth's surface temperature balance, and is one of these key process affected by hydrological and ecological feedback (Liang et al., 2020). In terrestrial ecosystems, the carbon exchange rate is mainly controlled by the photosynthesis - respiration process. Complex and nonlinear drivers such as meteorology, soils, vegetation, and available energy cause vertical carbon fluxes to be highly variable in space and time and challenging to measure and model (Huang et al., 2017; He et al., 2018; Chen et al., 2020; Dou & Yang, 2018). Several approaches have been developed to understand current and future terrestrial carbon flux over the past several decades involving field observations (Falge et al., 2002; Xiao et al., 2011), large-scale remote sensing (Xiao et al., 2019), process-based modeling (D. Wang et al., 2011; Dunkl et al., 2021), or a combination of these methods (Vetter et al., 2008; Jung et al., 2011). We take a data-driven approach to explore the predictability of the net CO_2 exchange rate, also known as Net Ecosystem CO_2 exchange (NEE), in agricultural landscapes in the Midwest U.S. NEE is the net carbon balance between photosynthetic CO_2 gain and respiratory CO_2 losses from plants and animals, and we use Fc as the nomenclature for NEE measured at an eddy covariance flux tower.

In this system, causal interactions need to be detected to understand interrelated processes at multiple spatial and temporal scales (Runge et al., 2019; Bollt et al., 2018). From a modeling perspective, this involves "intervening" in the system and manipulating model structures, parameters, or inputs, and observing the resulting model behavior relative to observations (Goodwell et al., 2020). Specifically, a causal model evaluation framework should consider dependencies between inputs or source variables and the target, or the "functional performance" relative to observed interactions (Goodwell & Bassiouni, 2022; Bassiouni & Vico, 2021; Ruddell et al., 2019). This is particularly crucial for machine learning and deep learning models, where relationships between inputs and outputs are not transparent. Understanding how these models learn, or fail to learn, the dependencies we observe in nature to predict an output is vital (Goodfellow et al., 2016). Meanwhile, predictive performance measures capture features of the relationship between the observed and modeled target output variable. In this study, we focus on the functional and predictive performance of data-driven models of hourly Fc.

Information theory (IT) measures, which characterize uncertainty and reductions in uncertainty based on probability distributions (Cover & Thomas, 2012; Shannon, 1948), have been employed in various geoscience contexts to measure complexity, dependencies, and driving or causal mechanisms (Balasis et al., 2013). Previous applications characterized ecohydrological process networks that reveal ecosystem behaviors (Ruddell & Kumar, 2009a; Franzen et al., 2020; Goodwell & Kumar, 2017; Ruddell et al., 2019; Sendrowski & Passalacqua, 2017). Recent applications of IT-based measures in hypothesis testing frameworks (Nearing et al., 2016, 2018) and to evaluate the functional performance of models based on a selection of sources (Sendrowski et al., 2018; Ruddell et al., 2019; Tennant et al., 2020; Moges et al., 2022; Bassiouni & Vico, 2021; Goodwell & Bassiouni, 2022) have shown great potential to better understand how models capture causal interactions in various Earth systems. However, these studies tend to consider a small subset of sources or a single modeled process. In this study, we take a more comprehensive view of complex ecohydrologic models and analyze information flow through the entire model. This allows for identification of potential sources of model error and insights into the relationships between different components of the model. This can lead to a better understanding of the model's behavior and performance, and ultimately, more accurate predictions of ecological and hydrological processes.

ML techniques have shown to be more effective and adaptable relative to mechanistic or semi-empirical model approaches, providing a complementary strategy to predict carbon fluxes at local to global scales (Dou & Yang, 2018; Dou et al., 2018). Machine learning (ML) algorithms construct empirical models based on the patterns contained in data and are very data adaptive because no assumption and functional forms need to be prescribed (Jung et al., 2011). ML has been used for interpolation for gap-filling carbon flux data and climatic driving factors based on flux tower measurements (Moffat et al., 2007; Ooba et al., 2006), decreasing the predictive errors of carbon fluxes from the land surface models (T. Wang et al., 2012), and upscaling carbon fluxes of terrestrial ecosystems from site to regional and global scales (Papale et al., 2015). Several studies similarly indicate the ability of ML to reproduce complex ecohydrological patterns, particularly in relation to flux tower measurements (Q. Zhou et al., 2019; Tramontana et al., 2020; Reichstein et al., 2019). Specifically, Q. Zhou et al. applied a ML approach to estimate NEE using variables such as the fraction of photosynthetically active radiation (PAR), leaf area index (LAI), soil moisture, downward solar radiation, precipitation, and mean air temperature. Tramontana et al. developed an ANN model to estimate NEE based on the light-use efficiency concept and used a comprehensive dataset of soil and micrometeorological variables as flux drivers.

While machine learning models tend to make better predictions than traditional models, they are often not trusted by the hydrologic community due to their black-box nature (Welchowski et al., 2022). By characterizing information flow pathways and comparing models beyond predictive performance, we can gain insights into their process representations (Goodwell & Bassiouni, 2022). This is particularly important when using a certain model to extrapolate in an unknown future climate, where a model with better process representations may be more trustworthy to apply to an unseen scenario. In this paper, we apply our IT-based model evaluation framework to three ML models, Long Short Term Memory (LSTM), Random Forest (RF), and multiple linear regression (MLR) to characterize how these models reproduce observed dependencies in terms of individual, pairwise and more multivariate interactions to predict sub-daily Fc. Recurrent Neural Networks (RNN) with LSTM are deep learning models that can successfully learn long-range temporal dependencies between time steps of sequence data (Hochreiter & Schmidhuber, 1997a; Sutskever et al., 2014; Kratzert et al., 2018, 2019). Meanwhile, the RF is a classical ML method that is known for its capacity to handle large datasets, resist the negative impacts of noise and overfitting (Breiman, 2001), and rank the significance of input variables (Leroux et al., 2017; Meng et al., 2021). RFs have been extensively applied in ecological classification and regression tasks (Meyer et al., 2019; Reitz et al., 2021; Q. Zhou et al., 2019). We use MLR as a simple model with which to compare the more complex ML models. We develop both locally and regionally trained models to compare model responses to larger training datasets that span multiple sites.

This paper is organized as follows. Section 2 describes the study site, datasets used, machine learning model development, and model evaluation. Section 3 presents the results of MLR, RF, and LSTM models. Section 4 provides a discussion, and Section 5 is a conclusion.

2 Materials and Methods

2.1 Site Description and Data

The data for this study was collected from multiple flux tower sites in maize/soybean landscapes in the Upper Midwest Corn Belt. The Goose Creek flux tower in central Illinois (Figure 1a) is part of the NSF-funded Critical Interface Network (CINet) project (`https://cinet.ncsa.illinois.edu/`), and collects 15-minute fluxes and meteorological variables at a 25m height, along with vegetation and soil properties. The Goose Creek site has been extensively studied using Lidar topography and high-resolution modeling of nutrient and carbon fluxes (Yan et al., 2019; Dutta et al., 2017; Woo & Kumar, 2017), and footprint modeling has been applied to study how landscape heterogeneity influences evapotranspiration fluxes (Hernandez Rodriguez et al., 2023). For this study, the 15-minute data was resampled to hourly resolution to match with other sites.

We also use data from 5 maize-soybean rotation sites in the FLUXNET2015 (Pastorello et al., 2020) dataset (Table 1), which provides over 1500 site-years of quality-controlled datasets for various landscapes. We used the AmeriFlux version of the hourly carbon flux data and meteorological variables for sites US-Ne1 (Mead - irrigated continuous maize site), US-Ne2 (Mead - irrigated maize-soybean rotation site), and US-Ne3 (Mead - rainfed maize-soybean rotation site). These sites are located within 1.6 km of each other at the University of Nebraska Agricultural Research and Development Center near Mead, Nebraska. Additionally, we used the hourly measurements of sites US-Br1 and US-Br3, located in adjacent maize and soybean fields in central Iowa. The farming systems, associated tillage, and nutrient management practices for maize/soybean production at these sites are typical of those throughout the Upper Midwest Corn Belt.

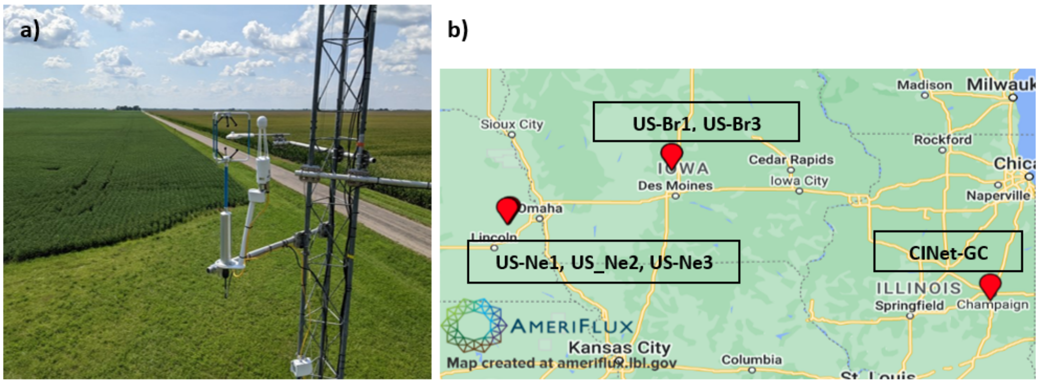

Figure 1: *(a)* At a 25m height eddy covariance flux tower in Central Illinois, observed fluxes originate from up to a 10km surrounding region, dominated by a patchwork of maize and soybean fields. *(b)* Three flux tower sites are located in maize/soybean systems.

Table 1: Characteristics of flux tower sites. MAT, ($°C$) is Mean Annual Temperature. MAP (mm) is Mean Annual Precipitation.

Site ID	Name	MAT	MAP	Year	Reference
US-Ne1	Mead-irrigated continuous maize	10.07	790.37	2010-2021	(Suyker, 2022a)
US-Ne2	Mead-irrigated maize-soybean rotation	10.08	788.89	2010-2021	(Suyker, 2022b)
US-Ne3	Mead-rainfed maize-soybean rotation	10.11	783.68	2010-2021	(Suyker, 2022c)
US-Br1	Brooks Field Site 10-Ames	8.95	842.33	2005-2011	(Prueger & Parkin, 2016a)
US-Br3	Brooks Field Site 11-Ames	8.9	846.6	2005-2011	(Prueger & Parkin, 2016b)
CINet-GC	Goose Creek flux tower	10	900	2016-2020	(Hernandez Rodriguez et al., 2023)

The forcing variables selected for this study (Table 2) are expected to influence the dynamics of Fc between the land and atmosphere, through direct or indirect influence on photosynthesis, respiration, and other biogeochemical processes. Specifically:

- Ta and TS: Soil and air temperatures influence both photosynthetic rates and microbial respiration. For example, it has been found that plant respiration increases more than photosynthesis as temperature rises, which indicates that a substantial temperature increase could turn an ecosystem from a carbon source to a sink (X. Zhou et al., 2012). Meanwhile, other studies have determined that this relationship is more complex when aspects such as changing rainfall and atmospheric CO_2 concentrations are considered (Drewry et al., 2010a, 2010b; Le et al., 2011).

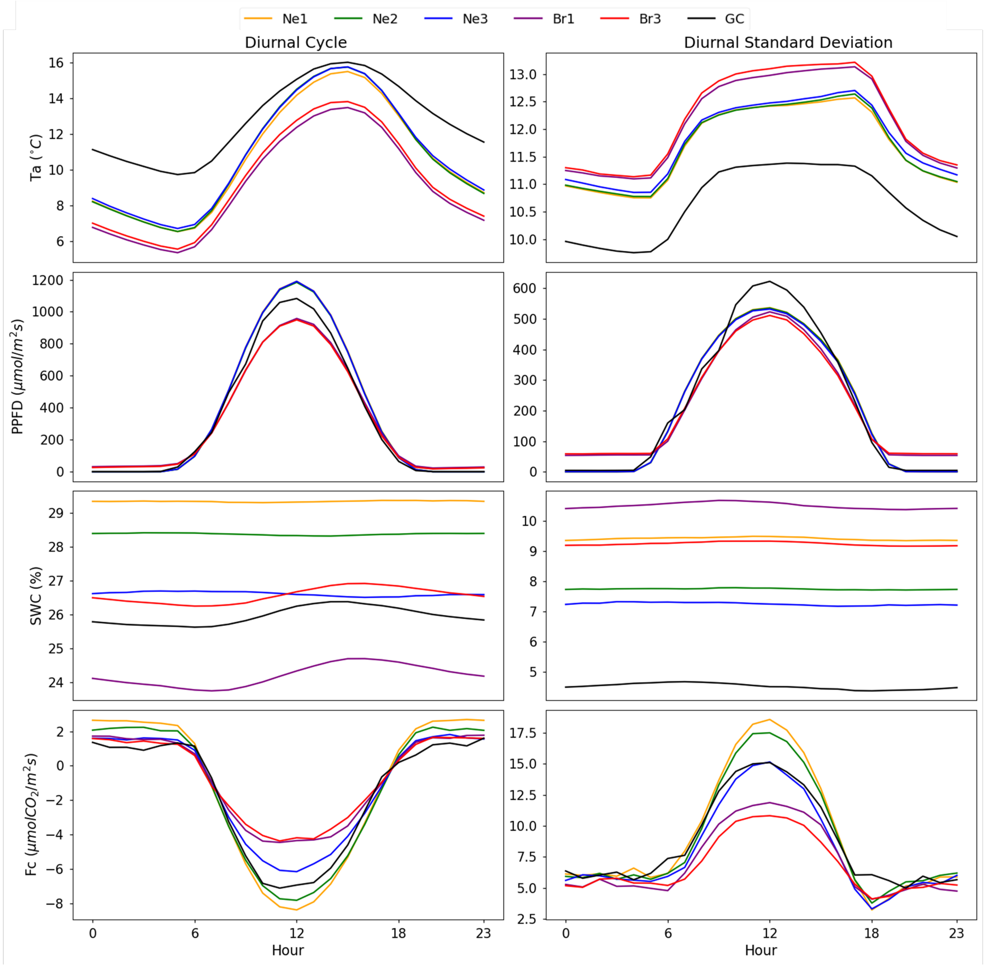

Figure 2: Diurnal cycle (left panel) and diurnal standard deviation cycle (right panel) of air temperature (Ta), photosynthetic photon flux density ($PPFD$), soil water content (SWC)) and carbon flux (Fc) over the study years corresponded to different sites (Ne1, Ne2, Ne3, Br1, Br3, GC). Each site is represented by a unique color.

Table 2: The full suite of variables used in this study.

Variable Description	Symbol	Unit
Carbon dioxide (CO_2) flux	Fc	$\mu mol CO_2/m^2 s$
Relative humidity	RH	%
Air temperature	Ta	$^\circ C$
Wind speed	WS	m/s
Atmospheric pressure	Pa	kPa
Precipitation	P	mm
Net radiation	$NETRAD$	W/m^2
Incoming photosynthetic photon flux density	$PPFD$ *	$\mu mol Photons/m^2 s$
Soil water content (volumetric)	SWC	%
Soil temperature	TS	$^\circ C$

* PAR: Photosynthetically Active Radiation ($\mu mol/m^2 s$) in the CINet-GC site

- RH: Humidity levels can impact plant transpiration and stomatal conductance, thereby influencing carbon uptake during photosynthesis.
- P and SWC: Water availability affects photosynthesis, and scarcity can lead to stress conditions, slowing down carbon sequestration.
- $PPFD$ and $NETRAD$: These radiation variables influence the energy balance and are related to the amount of light available for photosynthesis, which is a primary driver for carbon uptake in plants.
- WS: While not a direct factor, wind speed can affect plant transpiration rates, humidity levels, and even the mixing of carbon dioxide in the atmospheric layer.
- Pa: Changes in atmospheric pressure can impact gas exchange rates, indirectly affecting Fc.

We undertook rigorous data pre-processing (SI section S1) to ensure the reliability of our analysis. This involved applying quality control measures to all datasets, and identifying and removing any outliers or erroneous patterns. We encountered missing values in some datasets, which we imputed using time series imputation methods. We note that imputation is based on certain assumptions and can introduce uncertainty, which is discussed along with the results.

2.2 Model Development and Experimental Design

In this study, we develop three ML models to predict Fc: Multiple Linear Regression (MLR), Long Short Term Memory (LSTM), and Random Forest (RF). Each of these models offers unique advantages and capabilities. To ensure efficient learning, all input driving variables and the output (Fc) data were normalized by subtracting the mean and dividing by the standard deviation (Minns & Hall, 1996). The output of all ML models was retransformed using the normalization parameters to obtain the final Fc prediction.

The setup of ML models necessitates the optimization of hyperparameters, a task we performed via a combination of grid search and cross-validation techniques. Grid search encompasses defining a range of possible parameter values and evaluating the model's performance for each combination. Cross-validation helps to evaluate the model's generalization ability by partitioning the data into training and validation sets. We used a 5-fold cross-validation approach to search over the hyperparameter grid, where the data were split into 5 subsets of equal size, and each subset was used once for validation while the remaining 4 subsets were used for training. This process was repeated multiple times with different partitions to ensure a robust estimate of the model's performance.

The ML architectures (refer to SI, Table S1) used in this study worked well for all sites in comparison to observation and were therefore chosen to be applied here without further tuning. However, a systematic sensitivity analysis of the effects of different hyperparameters was not performed in our study and could be explored in more detail in terms of their effect on predictive and functional performance.

2.2.1 Multiple Linear Regression Model

MLR assumes a linear function of the independent variables to predict the dependent variable. The simplicity, interpretability, and ease of use of MLR make it a popular choice for many applications. However, it assumes a linear relationship between the dependent and independent variables and is sensitive to outliers and multicollinearity. In our study, MLR provides a baseline for comparison with the more complex RF and LSTM models. We adopted the Ordinary Least Squares (OLS) method for model fitting, which optimizes the model by minimizing the sum of the squared residuals.

2.2.2 Random Forest Model

The Random Forest (RF) model is a powerful ensemble learning algorithm that generates predictions by combining the outputs of multiple decision trees. Each of these trees is constructed using a randomly selected subset of the features and data samples, which helps to prevent overfitting. The final prediction is then derived by averaging the outputs from all the trees. In a decision tree, each node represents a feature in our data, each branch represents a decision rule, and each leaf represents an outcome. The root node, the topmost node in a tree, corresponds to the best predictor. Decisions are made by walking down the tree from the root to a leaf node.

The RF model is highly regarded for its accuracy, resilience to noise and outliers, and its ability to handle high-dimensional data with nonlinear relationships and missing values (Breiman, 2001), making it a suitable choice for our study to predict Fc. However, due to its complexity, interpreting the model can be challenging, and the computational cost can increase significantly with the number of trees in the forest. The performance of the RF model is significantly influenced by the fine-tuning of hyperparameters. The n-estimators (set to 100 in this study) parameter represents the number of trees in the forest and a trade-off between computation time and model performance. The max-depth parameter (set to 9, total number of features) controls the complexity of the model, playing a crucial role in preventing overfitting. The max-features parameter (set to 3), denoting the number of features to consider at each split (the maximum depth of each tree), can significantly impact the model's performance and is typically set to the square root of the total number of features. It is also worth noting that the random-state (set to 42) parameter ensures the consistency and reproducibility of our results.

2.2.3 Long Short Term Memory Model

LSTM is a specialized form of the Artificial Recurrent Neural Network (RNN) architecture, which is designed to remember long-term dependencies in sequential data. This capability is achieved through a unique arrangement of memory cells and three types of gates: the input gate, output gate, and forget gate. These components work together to selectively retain or discard information over time, making LSTM particularly adept at time-series prediction tasks (Hochreiter & Schmidhuber, 1997b). We choose LSTM for its capacity to model temporal dependencies in time series data, a vital characteristic for accurate carbon flux prediction. We operate the LSTM in sequence-to-sequence mode, in which any length of input sequence generates an equally long output sequence. We chose a constant sequence length of 12 hourly time steps. This is based on the diurnal cycle of environmental patterns, including temperature and light, that significantly affect Fc (Figure 2).

The design and training of LSTM models necessitate careful selection of various parameters. These include the number of layers in the network, the number of hidden units per layer, the learning rate, and the sensitivity of back-propagation to residuals between predicted and observed outputs. Additionally, the presence or absence of dropout layers, which help prevent overfitting, must be considered. To find an optimal model architecture, we conducted a series of experiments at different sites, manually adjusting different architectures (e.g., one or two LSTM layers or 5, 10, 15, or 20 cell/hidden units). The chosen architecture consists of a two-layer LSTM network, with each layer having a cell/hidden state length of 9, as number of driving source variables (Table 2). Dropout layers are added between the LSTM layers to prevent overfitting (Srivastava et al., 2014), and a regression layer with a single unit is added for the target variable (Fc).

During the training of LSTMs, each iteration step typically works with a subset (called a batch or mini-batch) of the available training data. In our case, the batch size is defined to be 128, and each sample in the batch consists of the Fc value and the driving variables of the 12 preceding time steps. The loss function, calculated as the average of the Mean Squared Error (MSE) of simulated and observed Fc of these 128 samples, is computed in every iteration step. For faster convergence, it is advantageous to have random samples in one batch. In traditional ecohydrological model calibration, the number of iteration steps defines the total number of model runs performed during calibration. The corresponding term for neural networks is called an “epoch”, which is defined as the period in which each training sample is used once for updating the model parameters. For instance, if the dataset consists of 1000 training samples and the batch size is 10, one epoch would consist of 100 iteration steps.

2.2.4 Experimental Setup

Our experimental design involves two main experiments aimed at evaluating the performance of our ML models in predicting Fc.

Local models for each site: This experiment tests the general ability of our MLMs to predict Fc at individual sites. We trained separate models for each site (Table 1) using the first 80% of the studied years as training data and the last 20% of studied years as the testing period. This resulted in six separately trained networks, one for each site.

Regional model: We train a regional model on a large dataset with data from all sites, to learn general patterns and relationships between input and output data. In this, we grouped all sites for the definition of the study region and used the combined data of 80% randomly selected for the entire period of all sites. We then test the model on each of the sites separately. The regional experiment is motivated by the idea that deep learning models perform better when trained with large amounts of data (Hestness et al., 2017; Schmidhuber, 2015) and regional models could be a potential solution for prediction in sites without flux tower measurements (Hrachowitz et al., 2013; Sivapalan, 2003). Having a large training dataset allows the model to learn more generalized and abstract patterns and relationships between input and output data. For instance, if two sites behave similarly, but one lacks high precipitation events or extended drought periods in the calibration period, while having these events in the validation period, the ML model can learn the response behavior to those extremes and use this knowledge in the first site.

2.3 Model Evaluation Framework

We gauge model performance both in terms of predictive accuracy and ability to encapsulate functional relationships. In this context, we consider two types of performance measures: predictive performance, which assesses the model's ability to accurately predict outcomes, and functional performance, which evaluates the model's ability to capture the underlying functional relationships between variables (Nearing et al., 2020; Goodwell & Bassiouni, 2022; Bassiouni & Vico, 2021). Predictive performance metrics include quantitative measures of the discrepancy between the model's predictions and the actual values, while functional performance can be assessed using various methods, including sensitivity analysis, partial dependence plots, and information-theoretic measures. We use a combination of several predictive and functional performance measures to evaluate the performance of ML models at different granularities.

2.3.1 Predictive Performance

We use Nash–Sutcliffe Efficiency (*NSE*) (Nash & Sutcliffe, 1970), Kling-Gupta Efficiency (*KGE*) (Gupta et al., 2009), and Shannon Entropy (H) (Shannon, 1948), an information theory (IT)-based measure to evaluate model predictive performance. Both *NSE* and *KGE* are widely recognized in hydrology for their effectiveness in assessing the quality of modeled predictions in relation to observed data. On the other hand, the entropy metric quantifies the uncertainty inherent in the model's predictions relative to observations. These metrics provide different perspectives on prediction errors.

The NSE is a normalized statistic that quantifies the relative magnitude of the residual variance, often referred to as "noise", in comparison to the variance of the measured data, or "information" (Nash & Sutcliffe, 1970). It is computed as follows:

$$\mathrm{NSE}(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \tag{1}$$

where n is the number of observations, $\overline{\hat{y}}$ is the mean of modeled values and y_i and $\hat{y}_i$ are the observed and modeled values, respectively. The NSE ranges from $-\infty$ to 1. An *NSE* of 1 signifies a perfect match between modeled and observed data. An *NSE* of 0 indicates that the model's predictions are as accurate as the mean of the observed data. A negative *NSE* occurs when the observed mean is a better predictor than the model.

The *KGE* is defined by the following equation:

$$\mathrm{KGE}(y,\hat{y}) = 1 - \sqrt{(r(y,\hat{y})-1)^2 + (\alpha(y,\hat{y})-1)^2 + (\beta(y,\hat{y})-1)^2}, \tag{2}$$

where r is the Pearson correlation coefficient between the observed (y_i) and modeled values ($\hat{y}_i$), defined as:

$$r(y,\hat{y}) = \frac{\sum_{i=1}^{n}(y_i-\overline{y})(\hat{y}_i-\overline{\hat{y}})}{\sqrt{\sum i = 1^{n}(y_i-\overline{y})^2(\hat{y}_i-\overline{\hat{y}})^2}} \tag{3}$$

Here, n is the number of observations, and $\overline{y}$ and $\overline{\hat{y}}$ are the mean of observed and modeled values, respectively. The variability ratio, α, is the ratio of the standard deviation of modeled values ($\sigma_{\hat{y}}$) to observed values (σ_y). β, the bias ratio, is the ratio of the mean of modeled values ($\overline{\hat{y}}$) to observed values ($\overline{y}$). Similar to NSE, *KGE* values range between $-\infty$ and 1, where 1 represents a perfect fit.

The *NSE* and *KGE* can be more or less suitable depending on the characteristics of the data and the objectives of the model (Knoben et al., 2019). *NSE* is based on the mean squared error and is particularly sensitive to the ability of the model to reproduce the variance of the data around its mean. Consequently, a model's consistent over- or underestimation can influence the *NSE* value. If the model consistently over- or underestimates the data, this will strongly affect the *NSE*. On the other hand, *KGE* also includes the correlation between observed and simulated data in addition to bias and variability. This enables *KGE* to adeptly identify patterns of over- or underestimation. Moreover, the breakdown of the *KGE* into its components can provide valuable insights into the model's strengths and weaknesses. A model might have a high *KGE*, but a low *NSE* if it reproduces the overall dynamics of the data (which *KGE* assesses) well but fails to capture the variance around the mean (which *NSE* emphasizes) accurately. Conversely, a model might have a high *NSE*, indicating a good reproduction of the observed data's variance, but a low *KGE* if there are biases or variability issues.

IT is based on Shannon Entropy (Shannon, 1948), $H(X) = -\sum p(x)\log_2 p(x)$, where $p(x)$ is a probability distribution function (pdf). $H(X)$ is a measure of uncertainty of the random variable X, or the missing information that would lead to its full predictability. Here we consider the normalized difference in entropy between observed and modeled Fc as another predictive performance measure:

$$A_H = 1 - \frac{H(Fc_{mod})}{H(Fc_{obs})} \tag{4}$$

A_H indicates how well the model captures the uncertainty that exists in the observed Fc and it ranges from $-\infty$ to 1. The values of $A_H = -\infty$ never occurs in this case as $H(Fc_{obs}) \neq 0$. $A_H = 0$ represents the "best" performance where the model exactly replicates the observed uncertainty. Positive values of A_H indicate that the modeled entropy ($H(Fc_{mod})$) is lower than the observed entropy ($H(Fc_{obs})$). In other words, the model output is less uncertain, or more predictable, than the observed data. Conversely, negative values of A_H indicate that the model's outputs are more uncertain than the observed data. To compute pdfs, we discretize observed and modeled variables in $N = 100$ equally sized bins spanning the minimum and maximum values of observed output data.

2.3.2 Functional Performance

We also use IT to quantify the information shared between forcing variables, model outputs, and observations, which can be interpreted as a measure of the model's functional performance (Nearing et al., 2020). This perspective shifts the focus from uncertainty quantification to information quantification. We explore how various model types use information from driving variables (Table 2) to predict an output, or "target" variable, which here is *Fc*. The functional performance of a model indicates the extent to which this information use is similar to or different from observed dependencies. We take a multi-level IT-based approach to evaluate the functional performance of our models. We will characterize complex process linkages between forcing variables or other available information sources and *Fc* to assess the model's ability in capturing the relationships between the driving variables and the target variable. We consider functional performance at several different levels, specifically for individual source-target relationships, pairs of sources, and all combinations of sources, or the whole model level.

For an individual source (X, here a forcing variable), and target (Y, here Fc), we consider reductions in uncertainty, or gains in information, in the form of mutual information as follows:

$$I(X;Y) = \sum p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) = H(X) - H(X|Y) \tag{5}$$

where $I(X;Y)$ measures the reduction in uncertainty Y given the knowledge of X with units of bits. $I(X;Y)$ is symmetric with respect to X and Y, and for independent variables, $I(X;Y) = 0$, while for fully dependent variables, $I(X;Y) = min[H(X), H(Y)]$. In other words, mutual information is upper bounded by the minimum uncertainty of variables involved. We calculate functional performance for individual sources based on mutual information as follows:

$$\begin{aligned} I_n(X;Z) &= \frac{I(X;Z)}{H(Z)} \\ A_{f,MI} &= 1 - \frac{I_n(X;Fc_{mod})}{I_n(X;Fc_{obs})} \end{aligned} \tag{6}$$

where $I_n(X;Y)$ is the normalized MI, $H(Z)$ is the the entropy of the target variable (Fc), $I_n(X;Fc_{obs})$ and $I_n(X;Fc_{mod})$ are normalized MI of observed and modeled target variable (Fc) respectively. This captures the extent to which modeled mutual information matches that of the observed target variable. $A_{f,MI}$ value close to zero represents the "best" performance where the model most closely replicates the observed mutual information. This can be used to assess how a model may be overestimating (negative $A_{f,MI}$ value) or underestimating (positive $A_{f,MI}$ value) the influence of certain drivers, and identify the most important drivers to include in a model.

In a more multivariate context, transfer entropy (TE) and partial information decomposition (PID) have been used to characterize interactions at different scales (Goodwell et al., 2020). TE (Schreiber, 2000) is a specific instance of conditional mutual information, which quantifies the information transferred to a target, Y_t, from a sequence of historical states of another variable, given the knowledge of its own past states. In hydrologic modeling research, TE has been used to validate and diagnose missing process connections in a delta model (Sendrowski et al., 2018), evaluate a multi-hypothesis ecohydrological modeling framework (Bennett et al., 2019), select time aggregations and lags toward ML applications (Tennant et al., 2020), and characterize the functional performance of a multi-layer canopy model (Ruddell et al., 2019). However, a TE-based analysis only highlights pairwise causal connections and does not address the feature of joint or simultaneous forcing from multiple drivers. Instead, we use PID to to characterize joint influences from multiple source variables to a target (Williams & Beer, 2010; Goodwell et al., 2020). For example, previous studies have compared how stomatal optimization models respond to soil water supply and atmospheric demand (Bassiouni & Vico, 2021), how simple to complex models behave under different source dependencies (Goodwell & Bassiouni, 2022), and stomatal model representations of physiological limits on transpiration (Hawkins et al., 2022). We consider two sources, or model forcing variables, that provide information to a target variable, which could be an observation or a model output. In a system where two sources share information from X and Y with a target Z, the total information quantity, $I(X,Y;Z)$, can be partitioned into synergistic (S), unique (U), and redundant (R) components. This partitioning is as follows:

$$I(X,Y;Z) = S_{X,Y} + R_{X,Y} + U_{X|Y} + U_{Y|X} \tag{7}$$

Here, $S_{X,Y}$ is synergistic information or joint information that is provided only when both sources are known together. $R_{X,Y}$ is redundant information or overlapping information that both sources provide individually. $U_{X|Y}$ and $U_{Y|X}$ terms indicate unique information that individuals influence when one source provides information that is not provided by the other. We use a partitioning method described in Goodwell and Kumar to obtain these components of the total information (refer to SI section S2 for more details). We normalize components by dividing each by the total mutual information $I(X,Y;Z)$, such that all information components add up to 1, and a given component indicates the fraction of reduced uncertainty in Z that can be attributed to that information type. These IT-based measures R, U, and S characterize different types of causal relationships between variables. They are particularly useful to interpret multivariate interactions, such as the Fc-related processes of interest here.

For computing mutual information and information partitioning components, we used different number of bins, based on the range of observed and modeled data (i.e., the difference between the maximum and minimum values). We calculated the number of bins for the model by taking the ratio of the range of the model to the range of the observation, multiplied by the number of bins in the observations ($N = 100$). This method effectively scales the number of bins based on the relative range of the model and observed data, with the assumption that a wider range would need more bins to capture the data distribution effectively. We compute statistical significance of observed or modeled IT measures using a shuffled surrogates approach (Ruddell & Kumar, 2009b). Details on these methods are provided in SI, Section S3.

We use PID to calculate the pairwise functional performance in terms of redundancy, synergy, and unique information and "overall" information partitioning for a given pair of sources. We consider the pairwise functional performance as the relative difference in an information flow measure for modeled versus observed data, separated into different components related to information partitioning measures S, R, and U, (Equation 7), respectively as $A_{f,S}$, $A_{f,R}$, and $A_{f,U}$ (Goodwell & Bassiouni, 2022). For example:

$$A_{f,S_{i,j}} = S(X_i, X_j; Z_{mod}) - S(X_i, X_j; Z_{obs}); \quad \text{for } i \neq j \tag{8}$$

where X_i and X_j indicate two source variables. The same concept applies for R. For unique information, we consider the sum of the two unique components ($U_X + U_Y$). A positive value indicates that the model overestimates a particular component at the expense of a different information type. The partitioning functional performance for a pair of sources is defined as the sum of the absolute values of the three pairwise measures as follows:

$$A_{f,Ipart_{i,j}} = |A_{f,S_{i,j}}| + |A_{f,R_{i,j}}| + |A_{f,U_{i,j}}| \tag{9}$$

This measure ranges from 0, for a model that exactly reproduces the observed information components, to 2, for a model that entirely substitutes one type of information for another or a combination of other information types. For instance, if the observed system shows that $U = 1$ (all information is unique), but a model system estimates $S = 1$ (that all information is synergistic), this leads to $A_{f,S} = 1$, $A_{f,U} = -1$ and $A_{f,Ipart} = 2$. While the individual source level identifies how the ranking of modeled variable importance differs from observations, this pairwise level identifies how the model is interpreting information provided by combinations of sources.

At the highest "whole model" level of analysis, we calculate average overall functional performance across all individual ($A_{f,MI}$) and pairs of sources ($A_{f,Ipart}$) as follows:

$$A_{f,MI,tot} = \frac{\sum_{i=1}^{n}(1 - |A_{f,MI_i}|)}{n}, \tag{10}$$

and

$$A_{f,Ipart,tot} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(2 - A_{f,Ipart_{i,j}})}{(n^2 - n)}, \tag{11}$$

where n is the number of source variables. $A_{f,MI,tot}$ ranges from $-\infty$ to 1 and $A_{f,Ipart,tot}$ ranges from 0 to 2. We note that these measures are the originally defined individual and pairwise performance measures subtracted from 1 or 2, in order to align higher values with "best" model performance. In other words, a value of 1 (or 2 for $A_{f,Ipart,tot}$) now corresponds to a perfect match of modeled values to the observed data (Table 3). This level of functional performance metrics gauges the model's overall ability to replicate the observed interactions. Figure 3 and Table 3 indicate the different levels of functional and predictive performance analysis.

Table 3: Summary of predictive and functional performance metrics.

Metric	**Range**	**Best Performance**	**Eq. No.**	**Description**
NSE	-∞ to 1	1	1	Nash-Sutcliffe Efficiency (predictive)
KGE	-∞ to 1	1	2	Kling-Gupta Efficiency (predictive)
A_H	-∞ to 1	0	4	Normalized difference in entropy between observed and modeled (predictive)
$A_{f,MI}$	-∞ to 1	0	6	MI difference for individual source (functional)
$A_{f,S_{i,j}}$, $A_{f,R_{i,j}}$, $A_{f,U_{i,j}}$	-1 to 1	0	8	Information partitioning components difference for a pair of sources (functional)
$A_{f,Ipart_{i,j}}$	0 to 2	0	9	Overall information component difference for a pair of sources (functional)
$A_{f,MI,tot}$	-∞ to 1	1	10	Average functional performance of individual source level across all driving sources
$A_{f,Ipart,tot}$	0 to 2	2	11	Average functional performance across all pairs of sources for overall information partitioning

3 Results

3.1 Predictive Performance

NSE and *KGE* values are higher for local relative to regional training across all ML models and sites (Figure 4a). This implies that local training allows the models to

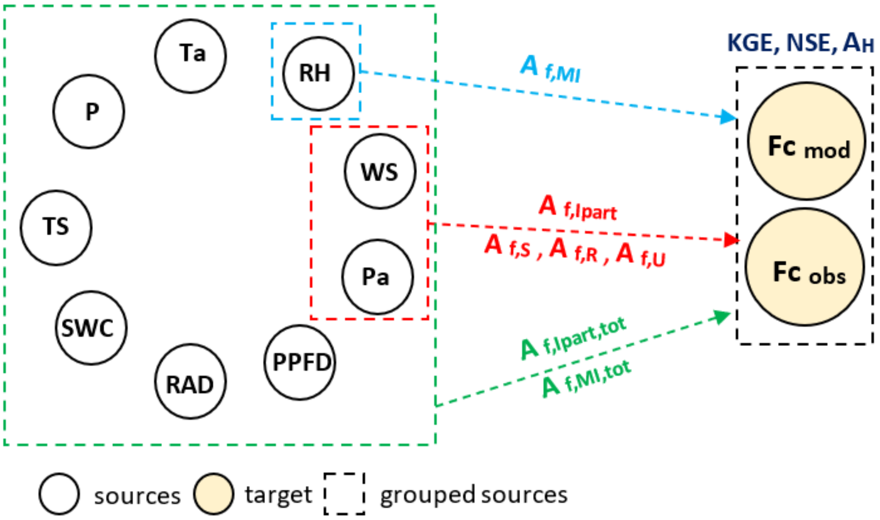

Figure 3: Illustration of functional and predictive performance. Nodes represent driving sources and target variables, and arrows represent different levels of functional performance. Predictive performance (NSE and KG and A_H) measure agreement between observed and modeled values (Equations 1, 2, and 4). Blue, red, and green links show relationships that can be captured by functional performance metrics at different levels (Table 3).

better capture certain characteristics of each site. The regional model performance may stem from the limitations of this study, mainly a relatively small number of sites and site-years. A more extensive dataset encompassing multiple sites over varied temporal spans may provide the model with a broader range of conditions and variability, enabling it to generalize more effectively.

Meanwhile, we find that the A_H of local models is higher than that of regional models (Figure 4b). A negative A_H occurs when $H_{mod} > H_{obs}$. This means that regional models actually introduce greater variability or uncertainty in Fc relative to observations. It is important to note that a negative A_H does not indicate "inferior" performance, since values close to zero represent "best" performance where the models reproduce the observed $H(Fc)$. While regional models over-estimate uncertainty in Fc, locally trained models underestimate uncertainty to a similar degree (Figure 4b).

When comparing performances of the three different models, RF (square markers in Figure 4a) consistently exhibits higher NSE and KGE values across all sites and both training experiences. This indicates the robustness of the RF model irrespective of the scale of the training data. Moreover, RF generally performs well in capturing the uncertainty in the observed Fc in both local and regional scales (square markers, Figure 4b). RF models have the best A_H performance for both regional and local models, indicating their ability to replicate the observed entropy of Fc.

MLR (circle markers in Figure 4) performance varies highly between sites. For some sites, the NSE values are very low, especially for regional training, suggesting MLR does not capture the specific behaviors of those sites effectively. The negative NSE values indicate that a mean predictor would have been better for most sites. Meanwhile, KGE values fall closer to the 1:1 line of Figure 4a, indicating that the KGE metric does not distinguish as many differences between regional and local training. Similarly, A_H for

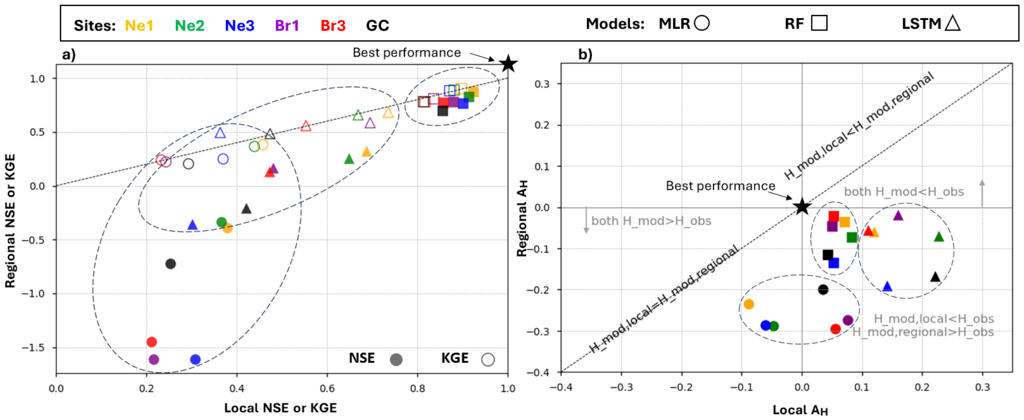

Figure 4: Predictive performance, (a) *NSE* (filled markers) and *KGE* (empty markers), and (b) the normalized difference in entropy between observed and modeled values (A_H) of three different models (MLR, RF, and LSTM, marker shapes) trained on local and regional data for six different sites (Table 1). Colors denote sites. The 1:1 line indicates equal performance for local and regional models.

the MLR model has the most spread between the study sites. For Nebraska sites (Ne1, Ne2, and Ne3), MLR has negative A_H values, which suggests that MLR model's outputs for these sites are more uncertain compared to the observed data. On the other hand, MLR for the other sites show positive A_H values.

LSTM (triangle markers in Figure 4a) results in *NSE* and *KGE* values between those of RF and MLR. For some sites, performance is close to that of the RF. This suggests that LSTMs can model temporal patterns at individual sites to some extent, and is always better than a mean predictor, but it never outperforms the RF model given the same training data. Given that LSTMs can model temporal sequences, the varied performance suggests that while some regional patterns are temporal, others might be non-sequential. We also find similar behaviour for LSTM as RF in capturing the entropy of observed *Fc*, except for more variability between sites. When models are trained locally, LSTM models tend to produce outputs that are less uncertain, or more predictable, than the observed data ($A_H > 0$). When models are trained regionally, LSTM outputs are more uncertain than observations. This difference between local and regional training for both LSTM and RF indicates that the regional training enables the model to produce more variable outputs, while local training leads to a more restricted range of *Fc*.

3.2 Functional Performance

At the individual and pairwise level, we focus on a single site, Ne1, as the site with the highest predictive performance and few gaps in forcing variables (*WS* and *NETRAD*). Other sites show similar patterns in mutual information and information decomposition measures, and we present full results for these in the Supplementary Information (SI Figures S3-S18).

3.2.1 Individual Source Level

Each variable is ranked based on the average observed *MI* across all sites (Figure 5a, black line). *TS* and *Ta* share the most information with *Fc*, indicating a strong dependence on fluctuations in both air and soil temperatures. The next variables that share information with *Fc* are radiation variables, *NETRAD* and *PPFD*. Meanwhile, precipitation (P) is a very weak predictor of Fc, which is expected since sub-daily precipitation contains many zero-values, leading to low entropy. Instead, we see that SWC shares more information with Fc, indicating that moisture available to roots and soil is important. Meteorological variables Pa, RH, and WS are relatively weak individual predictors. Models either overestimate or underestimate these mutual information values, resulting in a different ranking of variables for each model type (Figure 5a).

We use $A_{f,MI}$ to assess the extent to which mutual information matches with the observed target variable at Ne1 site (Figure 5b) and at other sites (SI Figure S3-S6). Higher absolute $A_{f,MI}$ values suggest that the modeled value is far from the observed value. If $A_{f,MI}$ is negative, the model overestimates the mutual information of observed Fc (an overly deterministic model), and if $A_{f,MI}$ is positive, the model underestimates observed mutual information (an overly random model).

The MLR model tends to underestimate mutual information (positive $A_{f,MI}$) for TS, Ta, SWC, Pa, WS, and P while overestimating for $NETRAD$, and $PPFD$, particularly for local training (Figure 5b, blue circles). MLR also shows the largest spread in over and underestimates of mutual information. The LSTM model for local training has a negative $A_{f,MI}$ for the most relevant drivers, but this is improved under regional training (Figure 5b, green triangles). The RF models closely replicate observed mutual information for both regional and local training (Figure 5b, red and orange squares). This highlights the power of RF in capturing the intricacies and dependencies within Fc regardless of the scale of the training data. Here we discuss the model representation of individual forcing variables.

- TS, Ta: While local and regional MLR model greatly underestimates the influence of temperature variables, the locally trained LSTM model overestimates it to a similar degree. In other words, the local LSTM model correctly identifies these as top sources of information to Fc, but to a more extreme degree, while the MLR models do not consider temperature as a top source.
- $NETRAD$ and $PPFD$: For local MLR, $A_{f,MI}$ is large and negative, indicating that the model overestimates the influence of radiation variables, and interprets them as the most important forcing variables instead of temperatures. However for the regional MLR, $A_{f,MI}$ is close to zero, indicating that the regional model mitigates this over-estimation. The only model that slightly underestimates mutual information from these variables is the regionally trained LSTM.
- SWC and P: Precipitation is a very weak driver according to both observations and models (Figure 5a), but models nearly always underestimate mutual information. They also underestimate information from SWC, except for the regionally trained LSTM. This indicates that models may lack sensitivity to moisture variability.
- WS: Across all models, the $A_{f,MI}$ values are fairly consistent, small, and positive, indicating all models slightly underestimate the influence of wind speed.
- Pa and RH: The locally trained MLR model shows the worst performance in terms of both over and under-estimating information from these variables.

These patterns are similar for other sites and under regional training (SI Figures S3-S6). This consistency suggests that the observed MI behaviors are not merely site-specific but possibly representative of broader environmental interactions. The key takeaway is that all models overestimate the influence of certain drivers at the expense of others, but to different degrees. This understanding can be useful to refine models or test the sensitivity of certain drivers. However, this level of analysis may omit drivers that provide information jointly rather than individually.

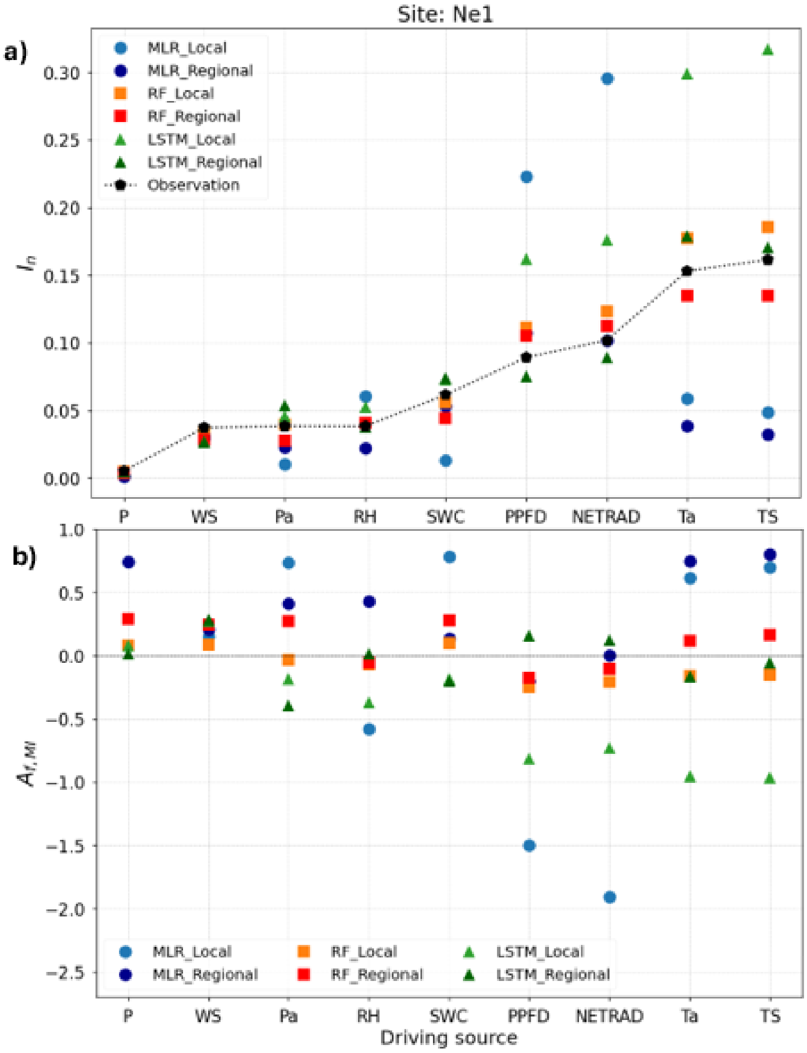

Figure 5: (a) Normalized mutual information (I_n) and (b) functional performance for individual variables ($A_{f,MI}$), Equation 6, for Multiple Linear Regression (MLR), Random Forest (RF), and Long Short-Term Memory (LSTM) models, under local and regional training at Ne1 site. Each variable is ranked (order on x-axis) based on the average observed MI across all sites (black line).

3.2.2 Pairwise and Model Level

In the observed data, most variable pairs provide synergistic (S) or unique information (U) to *Fc* (Figure 6a-c). The only pairs that provide a large fraction of redundant information (R) are closely related pairs (*Ta*, *TS*) and (*PPFD*, *NETRAD*). However, we note that their redundancy is still less than 0.5 as a fraction of total information, and the other half of the information they provide is U. Precipitation (P) provides the most U when paired with other variables (Figure 6c), but as found in the previous analysis of individual sources, the actual amount of information it provides is very small due to its low entropy. Meanwhile, *Ta* tends to provide the next highest fraction of U when paired with other sources, while *RH* and *WS* to provide S along with other sources. In general, regardless of the amount of information that sources provide, here we find that they mainly provide unique and synergistic information types.

All models tend to underestimate S (negative $A_{f,S}$, Figure 6d,g,j) for most variable pairs, at the expense of overestimating U (positive $A_{f,U}$, Figure 6f,i,l). For example, in the MLR model, *RH* greatly underestimates S and overestimates U when paired with other variables (Figure 6d,f). While the underestimation of synergistic relationships is widespread, the overestimation of redundancy only tends to occur for the most correlated variable pairs, specifically (*Ta*, *TS*) and (*PPFD*, *NETRAD*). This indicates that models rely excessively on these correlations, which results in an overemphasis in R. In other words, the observed relationship between these variables is not as redundantly informative for *Fc* as the model predicts, but they are instead more unique predictors.

Essentially, depending on the variable pair, the model either uses information uniquely where observations show a synergistic type of relationship, or uses information redundantly where observations show both unique and redundant contributions. The MLR model shows the largest trade-off between S and U partitioning performances (Figure 6d,f), followed by LSTM. Meanwhile, MLR is the only model that does not overestimate R provided by (*Ta*, *TS*), and in fact captures all information types accurately for this pair. However, we note that this MLR model also greatly underestimates the individual information components shared by each of these variables to the target (Figure 5). In other words, the MLR model greatly underestimates the importance of these temperature variables as predictors of *Fc*, but does reflect the mechanism by which they jointly provide information.

While broad patterns in information decomposition components are similar between models, there are several differences. For example, consider the (*SWC*, *Ta*) pair (bottom corner in all Figure 6 panels). For MLR, the information components are reproduced fairly accurately. For RF, U is overestimated at the expense of S to a minor degree. For LSTM, this occurs to a higher degree and R is also slightly overestimated. Meanwhile the MLR model greatly overestimates U from the pair (*RH*, *NETRAD*) at the expense of *S*, while the other two models have a similar but less extreme pattern.

When we consider the combined partitioning performance, $A_{f,Ipart}$ for each variable pair, the RF model has the best model performance, as it shows more $A_{f,Ipart}$ values close to zero (Figure 7). The MLR shows the most variability between pairs of sources, such that some pairs have very good functional partitioning performance and others have values of $A_{f,Ipart}$ greater than 1, indicating that over half of the information decomposition is misrepresented by the model. *RH*, *NETRAD*, and *PPFD* have particularly poor functional performance when combined with other sources for the MLR model. The LSTM model also has lower functional partitioning performance relative to RF, but behavior is more even between pairs of variables. Precipitation (P) always has the best functional performance when paired with other variables, but it is the weakest source and provides very little information regarding *Fc* for either models or observations.

When we consider other sites (SI Figures S7-S12), we find similar patterns in pairwise functional performance, specifically the overestimation of U at the expense of S and overestimation of R for correlated source pairs. However, we find that regionally trained models diminish some of the issues observed in the localized models. The broader dataset that regional training offers seems to provide a more balanced representation, allowing models to discern patterns beyond local-specific interactions. The regional model also corrects the balance between synergy and unique contributions, leading to a more accurate representation of how these variables interact. This trend is especially evident in the LSTM model, which demonstrates enhanced functional performance under regional training (SI Figures S13-S18). In terms of site differences, we find that regional LSTM model has the best model performance at Ne1 and Ne3 sites and RF model has the best performance among other sites.

When we calculate average overall functional performance at individual level ($A_{f,MI,tot}$), we find patterns that are similar to the average pairwise functional performance ($A_{f,Ipart,tot}$) (Figure 8). Specifically, local RF models perform slightly better than regional RF models on the individual level, while regional MLR and LSTM models generally perform better than the local models (Figure 8a). However, at the pairwise level, regional models consistently outperform their local equivalents (Figure 8b). This contrasts with trends observed in the predictive performance metrics (Figure 4), where local training led to higher NSE values relative to regional training.

Among all models, the RF model demonstrates the best performance, both at individual and pairwise levels (square markers in Figure 8). For individual sources, local RF models have better performance than the regional models. But when considering pairwise relationships, the regional RF model shows superior performance. On the other hand, the MLR model exhibits the lowest performance values at the individual level but performs more similarly to LSTM when considering pairwise relationships. The regional LSTM model also shows good performance at both the individual and pairwise levels. However, the performance of the local LSTM model varies more across different sites at the individual level, while the pairwise performance is more consistent for the regional model. This analysis highlights that changes in one aspect of functional performance do not necessarily translate to similar changes in other aspects.

3.3 Relationship between Predictive and Functional Performance

The relationship between predictive performance and functional performance provides insights into how a model balances replicating the observed data and its ability to capture observed relationships. As an illustration, we first focus on two key metrics: the KGE, representing predictive performance, and the $A_{f,Ipart,tot}$, indicating functional performance (Figure 9). For the 6 sites, two training types, and 3 model types, we have 36 total model runs for this comparison. All models show higher functional performance under regional training, but differences in KGE are on a site-by-site basis. The Ne1 site tends to be the highest performing site for all models in terms of KGE, but varies between models for $A_{f,Ipart,tot}$.

The functional and predictive performances for RF are both high relative to other models, and there is little variability between sites. However, there is an apparent trade-off between functional and predictive performance, in that sites with the highest KGE tend to have lower $A_{f,Ipart,tot}$. Meanwhile, there is a slight positive trend for locally trained LSTM and MLR models, where higher functional and predictive performances go together (Figure 9).

A correlation analysis shows that while functional and predictive performance measures tend to be correlated to each other (Figure 10a,c), there are fewer statistically significant ($p < 0.05$) correlations between the two types (Figure 10b). This correlation analysis is based on all 36 model cases (3 ML models, regional and local, and 6 sites) so

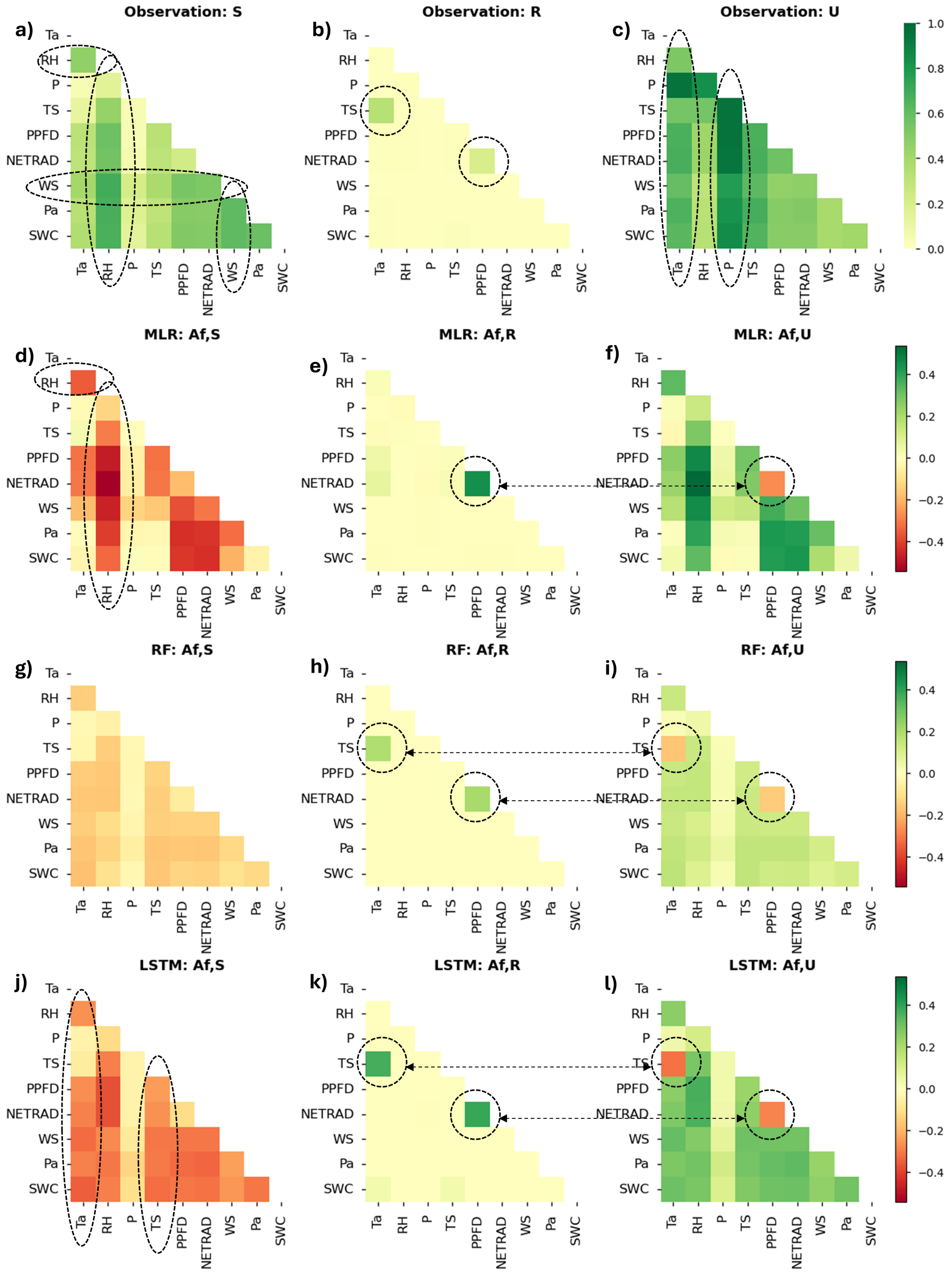

Figure 6: Observed pairwise (a) synergistic ($S_{i,j}$), (b) redundancy ($R_{i,j}$), and (c) uniqueness ($U_{i,j}$) information flow at Ne1 site. Pairwise functional performance of three models under local training experience at Ne1 site. The heat-map represents the relative difference in information decomposition partitioning measures ($A_{f,S_{i,j}}$, $A_{f,R_{i,j}}$, and $A_{f,U_{i,j}}$ between modeled and observed data for each pair of forcing variables. Positive values (green) in (d)-(l) indicate that the model overestimates the information type, while negative values (red) indicate underestimations.

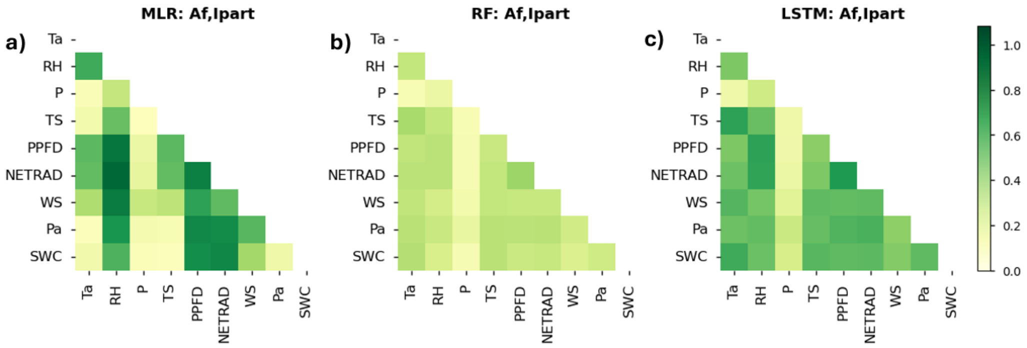

Figure 7: Pairwise functional partitioning performance $A_{f,Ipart_{i,j}}$ for (a) MLR, (b) RF, and (c) LSTM models under local training experience at Ne1 site. Values close to zero indicate optimal partitioning performance for a given pair.

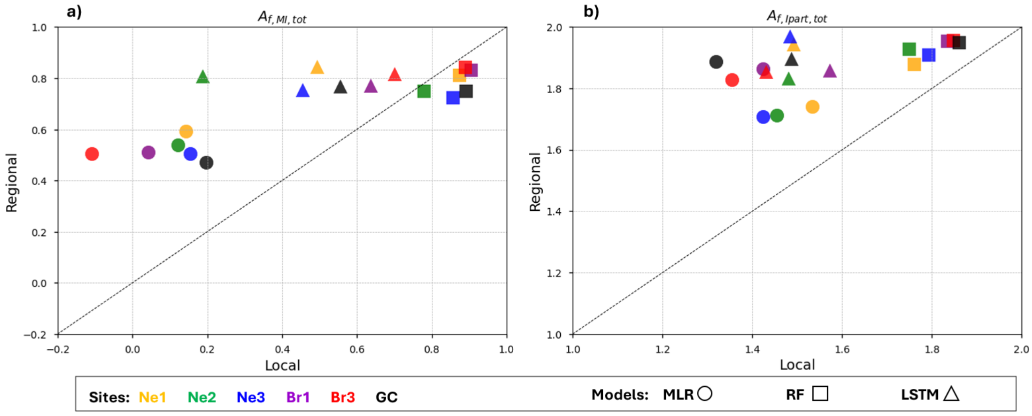

Figure 8: The whole model functional performance for (a) across all individual sources, $A_{f,MI,tot}$ and (b) across all pairs of sources, $A_{f,Ipart,tot}$), of three model types under two training experiences, local and regional, for six sites.

does not distinguish trends for a single model type or training experience. As illustrated in Figure 9, there may be a stronger correlation within a given model type and training. We split the KGE into its three constituent components, where high values of each term indicate "best" model performance. Similarly, the A_H measure of entropy and functional performance metrics are scaled so that high values indicate best performance, and positive correlations are easy to interpret.

Predictive performance metrics are positively correlated, except for the α, or variability, term of KGE with NSE and A_H. We find that the correlation component (r) is most correlated to the total KGE. Meanwhile, β and α terms are less correlated to KGE, and individual KGE components are less correlated to each other. This indicates that the correlation between observed and modeled Fc is the most predictive of KGE for these models. Meanwhile, both β and r terms are highly correlated with NSE. This highlights that the NSE is sensitive to the bias between model and observations and their correlation. The two full model functional performance metrics are also positively correlated (Figure 10c), indicating that models with high performance in terms of individual sources also reproduce pairwise relationships well.

In terms of correlations between functional and predictive measures (Figure 10b), 5 of the 12 possible correlations are positive and the other 7 are non-statistically significant, indicating that higher predictive performance is generally but not always associated with higher functional performance. The KGE α, or variability, component shows the highest correlation with functional measures, followed by the total KGE. This leads us to interpret that α is the most indicative of functional performance, and is the basis for the correlation between KGE and the functional measures. This indicates that models that reproduce the standard deviation of observed Fc, upon which α is based, also tend to reproduce observed forcing-Fc relationships at both a pairwise and individual level. Meanwhile, A_H, which is based on the difference in entropies of observed and modeled Fc, does not have a statistically significant correlation with functional performance. This illustrates that a model can reproduce the entropy of the observation, but not reproduce the distribution or functional relationships. In other words, the entropy is a summary statistic that does not necessarily indicate whether the model correctly replicates other features of the distribution of the data. No functional performance measures are correlated to the NSE, the β, or bias component of KGE, or A_H. This could be related to the linearity of these predictive performance measures that may not reflect nonlinear and joint interactions detected with mutual information. Additionally, we note that IT-based measures consider the distribution of the data but not the actual values, such that an IT measure would not capture a constant bias between two variables.

4 Discussion

Many machine learning approaches have been applied across major sub-domains of Earth system science and are increasingly being integrated into operational schemes and used to discover patterns, improve our understanding, and benchmark physically-based models. Ideally, ML models generate predictive models devoid of any presumptions on the underlying ecological structure or the mathematical representation of processes and interactions in an ecosystem. However, this lack of presumptions is correlated to a lack of understanding of whether and how these models are capturing functional relationships that exist in nature. The results of this study emphasize that functional performance—how accurately models capture the underlying relationships between variables—can be paired with more traditional metrics of model performance. By evaluating both functional and predictive aspects and their interrelationship, we can obtain a wider perspective on the strengths and limitations of different machine learning models. This multi-tiered approach not only can be used to explore the behavioral ranges for both machine learning and process-based models but also guides model development by highlighting model deficiencies based on information flow pathways that would not be apparent based on existing measures.

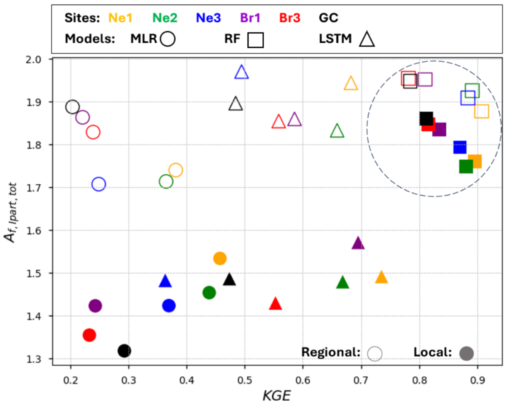

Figure 9: Predictive performance (KGE) and the overall model level of functional performance ($A_{f,Ipart,tot}$) of three model types under two training experiences, local (filled markers) and regional (empty markers).

Since ML-predicted fluxes can be used as benchmarks for physical land-surface and climate model evaluation (Q. Zhou et al., 2019; Anav et al., 2015; Best et al., 2015), it is valuable to understand nuances in their behavior.

While earlier studies on the CO_2 balance of vegetated surfaces applied linear regression for estimating the carbon fluxes (Jensen et al., 1996; Xu & Qi, 2001; Burrows et al., 2005), artificial neural network (ANN) and the support vector machine (SVM) methods have also been used to estimate terrestrial carbon fluxes and interpret the nonlinear relationship between ecosystem-based carbon fluxes and environment variables based on eddy covariance measurements (Papale & Valentini, 2003; Dou & Yang, 2018). For example, an ANN was able to filter out noise, predict the seasonal and diurnal variation of carbon fluxes, and extract patterns such as increased respiration in spring during root growth, which was formerly not well represented in carbon cycle models (Papale & Valentini, 2003). In this study, the Random Forest model showed both the highest functional and predictive performances, confirming that its better predictions really are associated with better process representations. The RF's non-parametric nature means it makes fewer assumptions about the underlying relationships between variables, thus enabling it to proficiently model intricate, non-linear interactions. Meanwhile, linear regression had a wide spread in performance levels between individual sites, and greatly overestimated the influence of radiation drivers that are highly linearly correlated to carbon flux. The LSTM model performance varied greatly between local and regional training, indicating that its functional performance benefited from training data from multiple sites.

Complex and nonlinear drivers such as meteorology, soils, vegetation, and available energy cause Fc to be highly variable in space and time and challenging to measure and model (Huang et al., 2017; He et al., 2018; Chen et al., 2020; Dou & Yang, 2018). Several approaches have been developed to understand current and future terrestrial carbon flux over the past several decades involving field observations (Falge et al., 2002; Xiao et al., 2011), large-scale remote sensing (Xiao et al., 2019), process-based modeling (D. Wang et al., 2011; Dunkl et al., 2021), or a combination of these methods (Vetter et al., 2008; Jung et al., 2011). Our study sheds further light on how forcing variables provide information to observed carbon fluxes. We found that temperature and radiation variables are most highly informative of Fc, followed by moisture-related variables such as RH and SWC. While many variables have a diurnal pattern, including Fc, we find that forcing variables tend to provide synergistic or unique information, rather than redundant information,

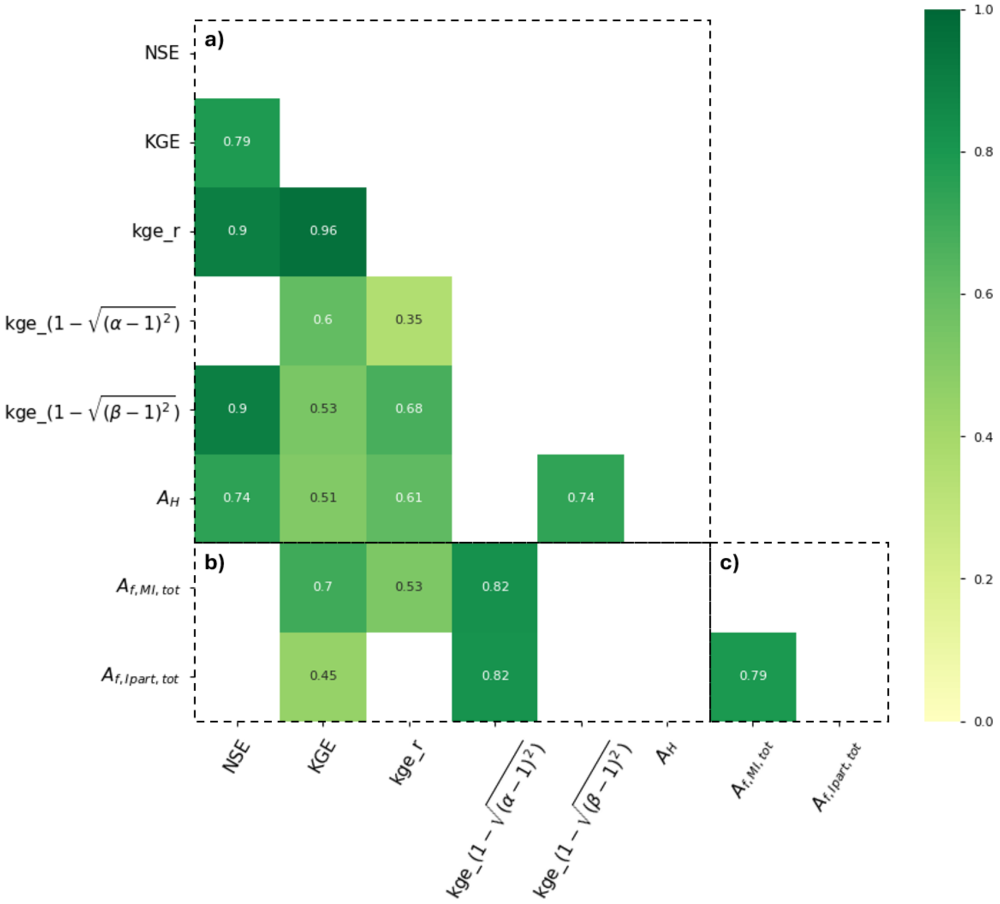

Figure 10: Correlation (p-$value < 0.05$) between performance metrics listed in Table 3 (scaled so that larger values always correspond to best performance), for the 36 model runs performed in this study. (a) and (c) separate correlations within predictive and functional categories, respectively, while (b) shows correlations between functional and predictive metrics.

indicating that the overlap in information content is relatively low. Meanwhile, RH is relatively weak as an individual source, but we found that it provides synergistic information when paired with many other sources. This indicates that the relevance of a variable like RH could be underestimated in an analysis that did not consider multivariate interactions, since it is a weak individual source but enhances the information content of other sources. In terms of modeling, we found that MLR, the simplest model, overestimates information from radiation variables and underestimates information from temperatures. This suggests that MLR captures the strongly linear diurnal pattern between energy availability and carbon flux, but misses a stronger but more nonlinear relationship with temperature due to the limitations in its parameterization. Finally, the tendency of all models to underestimate information from SWC indicates that water availability to plants is a complex driver of Fc that is difficult to capture in a functional form.

We note several limitations and assumptions that could be improved in future work. Future research could delve deeper into variations between sites, exploring what site-specific features influence model performance. One of the uncertainties of using flux tower measurements to estimate Fc is the impact of shifting land cover on the accuracy of the observations. The land-atmosphere exchange fluxes that generate carbon flux are influenced by the dynamic upwind surface area, called the flux footprint, which can exhibit spatial heterogeneities (Hernandez Rodriguez et al., 2023; Leclerc & Foken, 2014). As a result, fluxes from different sources can mix at the observation point, introducing uncertainty into the measurements. Meanwhile, this study assumes that the mix of crop types between sites and between observation time points leads to similar causal interactions between forcing variables and carbon flux. We also did not specifically focus on the optimization of hyperparameters within each ML model, which could have an effect on functional and predictive performance. Moreover, the precision and general quality of the forcing variables and Fc are important as they have underlying uncertainties and have been gap-filled, and our interpolation methods may have more effect on some model structures than others and future research could explore how models use information encoded in forcing data (Farahani et al., 2022). We also note that the MLR performance can be significantly influenced by multicollinearity among the forcing variables, and we did not test for this aspect. In terms of data size, we only considered six locations and approximately 50-site years, so further studies could more specifically consider the effect of increasingly large and diverse training datasets on model functional behaviors. Finally, the models evaluated represent just a fraction of the available algorithms, and we do not consider a wider range of ML and process-based models.

While predictive and functional metrics tend to be positively correlated, there are cases where a model change could be made that appears to improve predictions, but sacrifices a functional relationship. For example, the finding that regionally trained models tend to have improved functional performance indicates that these models can discern patterns beyond local-specific interactions. However, in this study the predictive performance of regional models was somewhat lower relative to single-site models, potentially marking a trade-off between functional and predictive performance. A "perfect" model should replicate all functional relationships as they are observed, but it still may not have perfect predictive performance due to missing information. In other words, the forcing variables simply do not contain all the information necessary to make a perfect accurate prediction. In this way, information-based metrics of functional performance provide a type of upper bound for predictive performance. This underscores the need for a nuanced approach to model selection. For an ungauged site with no validation data, a regionally trained model is likely the most applicable since it has a stronger functional performance and can reproduce processes as they are observed. The LSTM model was the most responsive to changes in training data size, which could relate to its complexity and need for many datasets to learn time-dependent interactions.

5 Conclusion

Predictive accuracy is just one facet of modeling complex ecohydrologic systems. Meanwhile, functional performance metrics capture how a model grasps the intricate relationships among variables. In order to use models for prediction in unseen conditions, and compare between machine learning and physically based model structures, we need to ensure that models don't just predict well, but also understand and represent the underlying processes effectively. In other words, understanding the why and how behind predictions can be as vital as the predictions themselves. In this study, the Random Forest model emerged as a consistently reliable model in terms of both predicting carbon fluxes and reproducing observed functional relationships at multiple levels. Meanwhile, a simple linear regression will overestimate the influence of variables with the most linear relationships to the target outcome. All models in this study had the common feature of underestimating synergistic interactions and overestimating unique ones. This indicates that all models are not quite capturing information flows at higher levels, where multiple sources provide information to the target jointly, and indicates that even the models with the highest predictive performance could be improved. Similarly, while performance measures tend to be correlated, no single performance measure captures the effect of all the others. This study advocates for a combined approach to model evaluation and validation, which considers both predictive performance and how the model captures interactions in the ecohydrologic system.

Acknowledgments

M.A. Farahani and A.E. Goodwell acknowledge funding from NSF Grant EAR #2012850 for the Critical Interface Network for Intensively Managed Landscapes *(CINet)* and the NASA New Investigator Grant #80$NSSC$21K0934. Python codes for analyses presented here are available on at `https://github.com/allisongoodwell/Farahani_CarbonML2023`.

References

Anav, A., Friedlingstein, P., Beer, C., Ciais, P., Harper, A., Jones, C., … Zhao, M. (2015). Spatiotemporal patterns of terrestrial gross primary production: A review. *Reviews of Geophysics*, *53*(3), 785-818. doi: https://doi.org/10.1002/2015RG000483

Balasis, G., Donner, R. V., Potirakis, S. M., Runge, J., Papadimitriou, C., Daglis, I. A., … Kurths, J. (2013). *Statistical mechanics and information-theoretic perspectives on complexity in the Earth system* (Vol. 15) (No. 11). doi: 10.3390/e15114844

Bassiouni, M., & Vico, G. (2021). Parsimony versus predictive and functional performance of three stomatal optimization principles in a big-leaf framework. *New Phytologist*, 0–2. doi: 10.1111/nph.17392

Bennett, A., Nijssen, B., Ou, G., Clark, M., & Nearing, G. (2019). Quantifying process connectivity with transfer entropy in hydrologic models. *Water Resources Research*, *55*(6), 4613-4629. doi: 10.1029/2018WR024555

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., … Vuichard, N. (2015). The plumbing of land surface models: Benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), 1425 - 1442. doi: https://doi.org/10.1175/JHM-D-14-0158.1

Bollt, E. M., Sun, J., & Runge, J. (2018). Introduction to focus issue: Causation inference and information flow in dynamical systems: Theory and applications. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *28*(7), 075201. doi: 10.1063/1.5046848

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. doi: 10.1023/A:1010933404324

Burrows, E. H., Bubier, J. L., Mosedale, A., Cobb, G. W., & Crill, P. M. (2005). Net ecosystem exchange of carbon dioxide in a temperate poor fen: a comparison of automated and manual chamber techniques. *Biogeochemistry*, *76*(1), 21–45.

Chen, N., Wang, A., An, J., Zhang, Y., Ji, R., Jia, Q., … Guan, D. (2020). Modeling canopy carbon and water fluxes using a multilayered model over a temperate meadow in inner mongolia. *International Journal of Plant Production*, *14*(1), 141-154. doi: 10.1007/s42106-019-00074-4

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Dou, X., & Yang, Y. (2018). Comprehensive evaluation of machine learning techniques for estimating the responses of carbon fluxes to climatic forces in different terrestrial ecosystems. *Atmosphere*, *9*(3). doi: 10.3390/atmos9030083

Dou, X., Yang, Y., & Luo, J. (2018). Estimating forest carbon fluxes using machine learning techniques based on eddy covariance measurements. *Sustainability*, *10*(1). doi: 10.3390/su10010203

Drewry, D. T., Kumar, P., Long, S., Bernacchi, C., Liang, X. Z., & Sivapalan, M. (2010a). Ecohydrological responses of dense canopies to environmental variability: 1. Interplay between vertical structure and photosynthetic pathway. *Journal of Geophysical Research: Biogeosciences*, *115*(4). doi: 10.1029/2010JG001340

Drewry, D. T., Kumar, P., Long, S., Bernacchi, C., Liang, X. Z., & Sivapalan, M. (2010b). Ecohydrological responses of dense canopies to environmental variability: 2. Role of acclimation under elevated CO_2. *Journal of Geophysical Research: Biogeosciences*, *115*(4), 1–22. doi: 10.1029/2010JG001341

Dunkl, I., Spring, A., Friedlingstein, P., & Brovkin, V. (2021). Process-based analysis of terrestrial carbon flux predictability. *Earth System Dynamics*, *12*(4), 1413–1426. doi: 10.5194/esd-12-1413-2021

Dutta, D., Wang, K., Lee, E., Goodwell, A., Woo, D., Wagner, D., & Kumar, P. (2017). Characterizing vegetation canopy structure using airborne remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(2), 1160–1178. doi: 10.1109/TGRS.2016.2620478

Falge, E., Baldocchi, D., Tenhunen, J., Aubinet, M., Bakwin, P., Berbigier, P., … Wofsy, S. (2002). Seasonality of ecosystem respiration and gross primary production as derived from fluxnet measurements. *Agricultural and Forest Meteorology*, *113*(1), 53-74. (FLUXNET 2000 Synthesis) doi: https://doi.org/10.1016/S0168-1923(02)00102-8

Farahani, M. A., Vahid, A., & Goodwell, A. (2022). Evaluating ecohydrological model sensitivity to input variability with an information-theory-based approach. *Entropy*, *24*(7). doi: 10.3390/e24070994

Franzen, S. E., Farahani, M. A., & Goodwell, A. (2020). Information flows: Characterizing precipitation-streamflow dependencies in the Colorado headwaters with an information theory approach. *Water Resources Research*, *56*(10), e2019WR026133. doi: https://doi.org/10.1029/2019WR026133

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. (`http://www.deeplearningbook.org`)

Goodwell, A., & Bassiouni, M. (2022). Source relationships and model structures determine information flow paths in ecohydrologic models. *Water Resources Research*, *58*(9). doi: https://doi.org/10.1029/2021WR031164

Goodwell, A., Jiang, P., Ruddell, B. L., & Kumar, P. (2020). Debates—does information theory provide a new paradigm for Earth science? Causality, interaction, and feedback. *Water Resources Research*, *56*(2). doi: https://doi.org/10.1029/2019WR024940

Goodwell, A., & Kumar, P. (2017). Temporal information partitioning: Characterizing synergy, uniqueness, and redundancy in interacting environmental variables. *Water Resources Research*, 5920–5942. doi: 10.1002/2016WR020218

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1), 80-91. doi: https://doi.org/10.1016/j.jhydrol.2009.08.003

Hawkins, L. R., Bassouni, M., Anderegg, W. R. L., Venturas, M. D., Good, S. P., Kwon, H. J., … Still, C. J. (2022). Comparing model representations of physiological limits on transpiration at a semi-arid ponderosa pine site. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2021MS002927. doi: https://doi.org/10.1029/2021MS002927

He, L., Li, J., Harahap, M., & Yu, Q. (2018). Scale-specific controller of carbon and water exchanges over wheat field identified by ensemble empirical mode decomposition. *International Journal of Plant Production*, *12*(1), 43-52. doi: 10.1007/s42106-017-0005-8

Hernandez Rodriguez, L., Goodwell, A., & Kumar, P. (2023). Inside the flux footprint: The role of organized land cover heterogeneity on the dynamics of observed land-atmosphere exchange fluxes. *Agricultural and Forest Meteorology*. doi: http://dx.doi.org/10.2139/ssrn.4034618

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., …

Zhou, Y. (2017). *Deep learning scaling is predictable, empirically.*

Hochreiter, S., & Schmidhuber, J. (1997a). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735-1780. doi: 10.1162/neco.1997.9.8.1735

Hochreiter, S., & Schmidhuber, J. (1997b). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735-1780. doi: 10.1162/neco.1997.9.8.1735

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., . . . Cudennec, C. (2013). A decade of predictions in ungauged basins (pub)—a review. *Hydrological Sciences Journal*, *58*(6), 1198-1255. doi: 10.1080/02626667.2013.803183

Huang, C.-W., Domec, J.-C., Ward, E. J., Duman, T., Manoli, G., Parolari, A. J., & Katul, G. G. (2017). The effect of plant water storage on water fluxes within the coupled soil–plant system. *New Phytologist*, *213*(3), 1093-1106. doi: 10.1111/nph.14273

Jensen, L., Mueller, T., Tate, K., Ross, D., Magid, J., & Nielsen, N. (1996). Soil surface co2 flux as an index of soil respiration in situ: A comparison of two chamber methods. *Soil Biology and Biochemistry*, *28*(10), 1297-1306. doi: https://doi.org/10.1016/S0038-0717(96)00136-8

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., . . . Williams, C. (2011). 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*(G3). doi: https://doi.org/10.1029/2010JG001566

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? comparing nash–sutcliffe and kling–gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. doi: 10.5194/hess-23-4323-2019

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. doi: 10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. doi: 10.5194/hess-23-5089-2019

Le, P. V. V., Kumar, P., & Drewry, D. T. (2011). Implications for the hydrologic cycle under climate change due to the expansion of bioenergy crops in the midwestern united states. *Proceedings of the National Academy of Sciences*, *108*(37), 15085-15090. doi: 10.1073/pnas.1107177108

Leclerc, M. Y., & Foken, T. (2014). *Footprints in micrometeorology and ecology* (Vol. 239). Springer.

Leroux, L., Bégué, A., Seen, D. L., Jolivot, A., & Kayitakire, F. (2017). Driving forces of recent vegetation changes in the Sahel: Lessons learned from regional and local level analyses. *Remote Sensing of Environment*, *191*, 38–54.

Liang, J., Guo, Q., Zhang, Z., Zhang, M., Tian, P., & Zhang, L. (2020). *Influence of complex terrain on near-surface turbulence structures over loess plateau* (Vol. 11) (No. 9). doi: 10.3390/atmos11090930

Meng, Y., Yang, M., Liu, S., Mou, Y., Peng, C., & Zhou, X. (2021). Quantitative assessment of the importance of bio-physical drivers of land cover change based on a random forest method. *Ecological Informatics*, *61*, 101204.

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications–moving from data reproduction to spatial prediction. *Ecological Modelling*, *411*, 108815.

Minns, A. W., & Hall, M. J. (1996). Artificial neural networks as rainfall-runoff models. *Hydrological Sciences Journal*, *41*(3), 399-417. doi: 10.1080/02626669609491511

Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr,

A. G., . . . Stauch, V. J. (2007). Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agricultural and Forest Meteorology*, *147*(3), 209-232.

Moges, E., Ruddell, B. L., Zhang, L., Driscoll, J. M., Norton, P., Perez, F., & Larsen, L. G. (2022). Hydrobench: Jupyter supported reproducible hydrological model benchmarking and diagnostic tool. *Frontiers in Earth Science*, *10*. doi: 10.3389/feart.2022.884766

Nash, J., & Sutcliffe, J. (1970). River flow forecasting through conceptual models part i — a discussion of principles. *Journal of Hydrology*, *10*(3), 282-290. doi: https://doi.org/10.1016/0022-1694(70)90255-6

Nearing, G. S., Mocko, D. M., Peters-Lidard, C. D., Kumar, S. V., & Xia, Y. (2016). Benchmarking NLDAS-2 soil moisture and evapotranspiration to separate uncertainty contributions. *Journal of Hydrometeorology*(2013), 160113112628008,. doi: 10.1175/JHM-D-15-0063.1

Nearing, G. S., Ruddell, B. L., Bennett, A. R., Prieto, C., & Gupta, H. V. (2020). Does Information Theory Provide a New Paradigm for Earth Science? Hypothesis Testing. *Water Resources Research*, *56*(2), 1–8. doi: https://doi.org/10.1029/2019WR024918

Nearing, G. S., Ruddell, B. L., Clark, M. P., Nijssen, B., & Peters-Lidard, C. (2018). Benchmarking and process diagnostics of land models. *Journal of Hydrometeorology*, *19*(11), 1835 - 1852. doi: 10.1175/JHM-D-17-0209.1

Ooba, M., Hirano, T., Mogami, J.-I., Hirata, R., & Fujinuma, Y. (2006). Comparisons of gap-filling methods for carbon flux dataset: A combination of a genetic algorithm and an artificial neural network. *Ecological Modelling*, *198*(3), 473-486.

Papale, D., Black, T. A., Carvalhais, N., Cescatti, A., Chen, J., Jung, M., . . . Ráduly, B. (2015). Effect of spatial sampling from european flux towers for estimating carbon and water fluxes with artificial neural networks. *Journal of Geophysical Research: Biogeosciences*, *120*(10), 1941-1957. doi: 10.1002/2015JG002997

Papale, D., & Valentini, R. (2003). A new assessment of european forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Global Change Biology*, *9*(4), 525-535. doi: 10.1046/j.1365-2486.2003.00609.x

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., . . . Papale, D. (2020). The FLUXNET2015 dataset and the oneflux processing pipeline for eddy covariance data. *Scientific Data*, *7*(1), 225. doi: 10.1038/s41597-020-0534-3

Prueger, J., & Parkin, T. (2016a). Ameriflux base us-br1 brooks field site 10- ames. *AmeriFlux AMP, (Dataset)*. doi: https://doi.org/10.17190/AMF/1246038

Prueger, J., & Parkin, T. (2016b). Ameriflux base us-br3 brooks field site 11- ames. *AmeriFlux AMP, (Dataset)*. doi: https://doi.org/10.17190/AMF/1246039

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743). doi: 10.1038/s41586-019-0912-1

Reitz, O., Graf, A., Schmidt, M., Ketzler, G., & Leuchner, M. (2021). Upscaling net ecosystem exchange over heterogeneous landscapes with machine learning. *Journal of Geophysical Research: Biogeosciences*, *126*(2), e2020JG005814.

Ruddell, B. L., Drewry, D. T., & Nearing, G. S. (2019). Information Theory for Model Diagnostics: Structural Error is Indicated by Trade-Off Between Functional and Predictive Performance. *Water Resources Research*, *55*(8), 6534–6554. doi: 10.1029/2018WR023692

Ruddell, B. L., & Kumar, P. (2009a). Ecohydrologic process networks: 1. Identification. *Water Resources Research*, *45*(3), 1–23. doi: 10.1029/2008WR007279

Ruddell, B. L., & Kumar, P. (2009b). Ecohydrologic process networks: 1. Identification. *Water Resources Research*, *45*(3), 1–22. doi: 10.1029/2008WR007279

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., . . . Zscheischler, J. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, *10*(1), 2553. doi: 10.1038/s41467-019-10105-3

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. doi: 10.1016/j.neunet.2014.09.003

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, *85*(2), 461. doi: 10.1103/PhysRevLett.85.461

Sendrowski, A., & Passalacqua, P. (2017). Process connectivity in a naturally prograding river delta. *Water Resources Research*, *53*(3), 1841–1863. doi: 10.1002/2016WR019768

Sendrowski, A., Sadid, K., Meselhe, E., Wagner, W., Mohrig, D., & Passalacqua, P. (2018). Transfer entropy as a tool for hydrodynamic model validation. *Entropy*, *20*(1). doi: 10.3390/e20010058

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *196*(4), 519–520. doi: 10.1016/S0016-0032(23)90506-5

Sivapalan, M. (2003). Prediction in ungauged basins: A grand challenge for theoretical hydrology. *Hydrological Processes*, *17*, 3163 - 3170. doi: 10.1002/hyp.5155

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, *15*(1), 1929–1958.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to sequence learning with neural networks* (Vol. 27; Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger, Eds.). Curran Associates, Inc. doi: https://doi.org/10.48550/arXiv.1409.3215

Suyker, A. (2022a). Ameriflux base us-ne1 mead - irrigated continuous maize site. *AmeriFlux AMP, (Dataset)*. doi: https://doi.org/10.17190/AMF/1246084

Suyker, A. (2022b). Ameriflux base us-ne2 mead - irrigated maize-soybean rotation site. *AmeriFlux AMP, (Dataset)*. doi: https://doi.org/10.17190/AMF/1246085

Suyker, A. (2022c). Ameriflux base us-ne3 mead - rainfed maize-soybean rotation site. *AmeriFlux AMP, (Dataset)*. doi: https://doi.org/10.17190/AMF/1246086

Tennant, C., Larsen, L., Bellugi, D., Moges, E., Zhang, L., & Ma, H. (2020). The utility of information flow in formulating discharge forecast models: A case study from an arid snow-dominated catchment. *Water Resources Research*, *56*(8), e2019WR024908. doi: 10.1029/2019WR024908

Tramontana, G., Migliavacca, M., Jung, M., Reichstein, M., Keenan, T. F., Camps-Valls, G., . . . Papale, D. (2020). Partitioning net carbon dioxide fluxes into photosynthesis and respiration using neural networks. *Global Change Biology*, *26*(9), 5235-5253. doi: https://doi.org/10.1111/gcb.15203

Vetter, M., Churkina, G., Jung, M., Reichstein, M., Zaehle, S., Bondeau, A., . . . Heimann, M. (2008). Analyzing the causes and spatial pattern of the European 2003 carbon flux anomaly using seven models. *Biogeosciences*, *5*(2), 561–583. doi: 10.5194/bg-5-561-2008

Wang, D., Ricciuto, D., Post, W., & Berry, M. W. (2011). Terrestrial ecosystem carbon modeling. In D. Padua (Ed.), *Encyclopedia of parallel computing* (p. 2034-2039). Boston, MA: Springer US. doi: 10.1007/978-0-387-09766-4_395

Wang, T., Brender, P., Ciais, P., Piao, S., Mahecha, M. D., Chevallier, F., . . . Vaccari, F. P. (2012). State-dependent errors in a land surface model across biomes inferred from eddy covariance observations on multiple timescales. *Ecological Modelling*, *246*, 11-25.

Welchowski, T., Maloney, K. O., Mitchell, R., & Schmid, M. (2022). Techniques to improve ecological interpretability of black-box machine learning models. *Journal of Agricultural, Biological and Environmental Statistics*, *27*(1), 175-197. doi: 10.1007/s13253-021-00479-7

Williams, P. L., & Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Woo, D. K., & Kumar, P. (2017). Role of micro-topographic variability on the distribution of inorganic soil-nitrogen age in intensively managed landscape. *Water Resources Research*, *53*(10), 8404-8422. doi: 10.1002/2017WR021053

Xiao, J., Chevallier, F., Gomez, C., Guanter, L., Hicke, J. A., Huete, A. R., ... Zhang, X. (2019). Remote sensing of the terrestrial carbon cycle: A review of advances over 50 years. *Remote Sensing of Environment*, *233*, 111383. doi: https://doi.org/10.1016/j.rse.2019.111383

Xiao, J., Davis, K. J., Urban, N. M., Keller, K., & Saliendra, N. Z. (2011). Upscaling carbon fluxes from towers to the regional scale: Influence of parameter variability and land cover representation on regional flux estimates. *Journal of Geophysical Research: Biogeosciences*, *116*(G3).

Xu, M., & Qi, Y. (2001). Soil-surface CO_2 efflux and its spatial and temporal variations in a young ponderosa pine plantation in Northern California. *Global Change Biology*, *7*(6), 667-677. doi: https://doi.org/10.1046/j.1354-1013.2001.00435.x

Yan, Q., Le, P. V. V., Woo, D. K., Hou, T., Filley, T., & Kumar, P. (2019). Three-dimensional modeling of the coevolution of landscape and soil organic carbon. *Water Resources Research*, *55*(2), 1218-1241. doi: 10.1029/2018WR023634

Zhou, Q., Fellows, A., Flerchinger, G. N., & Flores, A. N. (2019). Examining interactions between and among predictors of net ecosystem exchange: A machine learning approach in a semi-arid landscape. *Scientific Reports*, *9*(1), 2222. doi: 10.1038/s41598-019-38639-y

Zhou, X., Wang, X., Tong, L., Zhang, H., Lu, F., Zheng, F., ... Ouyang, Z. (2012). Soil warming effect on net ecosystem exchange of carbon dioxide during the transition from winter carbon source to spring carbon sink in a temperate urban lawn. *Journal of Environmental Sciences*, *24*(12), 2104-2112. doi: https://doi.org/10.1016/S1001-0742(11)61057-7