Supporting Information for "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

Contents of this file

1. Text S1 to S4
2. Figures S1 to S18
3. Tables S1

S1. Data pre-processing

The data pre-processing stage was a crucial step in our study, ensuring the reliability and accuracy of our analysis. This process involved several steps:

1.1. Quality Control

Firstly, we applied quality control measures to all datasets. This involved checking for any inconsistencies, errors, or outliers in the data that could potentially skew our results. We used a combination of automated checks and manual review to ensure the integrity

Corresponding author: A. E. Goodwell, Department of Civil Engineering, University of Colorado Denver, USA., (Alison.goodwell@ucdenver.edu)

of our data. Automated checks included algorithms to detect statistical anomalies, while manual review involved visual inspection of the data and cross-checking with source documentation.

1.2. Handling Missing Values

In some datasets, we encountered missing values. To handle these, we used time series imputation methods. The choice of imputation method was dependent on the distribution of the data. For normally distributed data, we used mean imputation. This technique replaces the missing values with the average of the available data for that variable, thus capitalizing on the characteristic symmetric nature of the distribution. Specifically, the variables *Fc*, *SWC*, *Ta*, *TS*, and *Pa* were treated using mean imputation. Conversely, for those variables presenting skewed distributions or characterized by extreme outliers, median imputation was employed. The median, being the middle value of a dataset, is less sensitive to outliers and provides a more robust measure of central tendency for skewed distributions. The variables *WS*, *P*, *NETRAD*, *PPFD* and *RH* were imputed using this method. Through these imputation strategies, we ensured that the integrity of the data distribution was upheld, while concurrently addressing the gaps in our dataset.

Moreover, to address significant missing values in the *PPFD* variable at Br1 and Br3 sites, we employed a linear regression imputation technique using *NETRAD* values as predictors. We first used those part of datasets where *PPFD* and *NETRAD* were concurrently present, using them as training data for individual linear regression models. Once trained, these models were used to predict missing *PPFD* values based on available *NETRAD* values, thus leveraging their linear relationship for accurate imputation.

1.3. Normalization

To ensure efficient learning and to prevent any one variable from dominating others due to scale differences, we normalized all input variables and the output (Fc) data. Specifically, we utilized the "MinMax" scaling technique, where the minimum of feature is made equal to zero and the maximum of feature equal to one. In this method, every feature value is transformed to fall within the range [0,1]. It scales the values to the specific value range without changing the shape of the original distribution. This approach entails subtracting the minimum value of the feature and then dividing by the range of that feature, resulting in a dataset where the minimum and maximum feature values are normalized to lie between 0 and 1. This procedure not only enhances the efficiency of learning algorithms but also aids in preventing potential numerical stability issues.

1.4. Retransformation

The output of all machine learning models was retransformed using the normalization parameters to obtain the final Fc prediction in the original scale. This step is crucial for interpreting the results in the context of the original data.

It's important to note that while these pre-processing steps greatly enhance the quality and usability of the data, they are based on certain assumptions and can introduce some level of uncertainty. However, we applied these methods systematically and transparently to minimize potential biases and ensure the reliability of our results. The full suite of variables used in this study, along with their descriptions and units, is outlined in Table **??** in the main manuscript.

S2. Information Decomposition We use information decomposition to analyze causal interactions in which two sources 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. Any existing IT-based measure can also be defined in terms of combinations of R, U, and S (Figure S1). For example, this partitioning of information implies that the mutual information between the target and each source is the sum of the redundancy and the unique information from the source, i.e. $I(X; Z) = U_{X|Y} + R_{X,Y}$ (Figure S1a). Meanwhile, conditional mutual information, which includes transfer entropy as a special case, is the sum of unique and synergistic components, i.e. $I(X; Z|Y) = U_{X|Y} + S_{X,Y}$ (Figure S1b). Finally, the interaction information, which is symmetric between all three variables, is equivalent to $S_{X,Y} - R_{X,Y}$ (Goodwell & Kumar, 2017, 2015), such that positive or negative interaction information indicates whether synergy or redundancy is dominant (Figure S1c). To simplify notation hereafter, we omit subscripts such that $S_{X,Y} = S$ and $R_{X,Y} = R$ given a particular definition of sources and target. We similarly simplify unique information components to $U_{X|Y} = U_X$ and $U_{Y|X} = U_Y$.

While information decomposition is a useful concept, information theory does not provide formulas to directly determine these quantities. Several studies (Barrett, 2015; Williams & Beer, 2010) defined redundancy measures as the mutual information that the weakest source provides to the target, forcing one unique component to equal zero. Goodwell and Kumar considered that this is actually a maximum bound for redundancy, and applied a "rescaled" redundancy measure in which redundancy is scaled between the minimum and maximum bounds that are defined by information theory. The maximum bound is the minimum mutual information that either source provides to the target,

$R_{max} = min[I(X;Z), I(Y;Z)]$. The minimum bound is zero for cases where the interaction information is positive or $S_R > 0$, i.e. $I(X,Y;Z) > I(X;Z) + I(Y;Z)$. otherwise, if $S - R < 0$, the minimum bound for redundancy is the negative interaction information, in order for synergy to be non-negative. This leads to a definition of the minimum R as $R_{min} = max[0, I(X;Z) + I(Y;Z) - I(X,Y;Z)]$. We then scale redundancy between these bounds based on the normalized information between the source variables:

$$I_s = \frac{I(X;Y)}{\min[H(X),H(Y)]} \tag{1}$$

$$R_s = R_{\min} + I_s(R_{\max} - R_{\min})$$

In general, this definition causes highly correlated sources to be maximally redundant with each other, while independent sources are minimally redundant. A definition for redundancy enables the computation of the other information decomposition components, S, U_X, and U_Y.

S3. Statistical Significance

We compute statistical significance of observed or modeled information theoretic measures using a shuffled surrogates approach. We define a critical value of total information as follows:

$$I_{crit} = I_{\textit{suff, mean}} + 3 \times I_{\textit{suff, stdev}} \tag{2}$$

where $I_{suff,mean}$ and $I_{suff,stdev}$ are the mean and standard deviation of 100 information measures computed with randomly shuffled source data. For example, if the $I(Ta, Ts; Fc) < I_{crit}$, we set all information components to zero and do not do fur-

ther information partitioning. Meanwhile, if $I(Ta, Ts; Fc)$ is statistically significant but $I(Ta; Fc|Ts)$ is not (according to the same shuffled surrogate method), we set the unique component from Ta and the synergistic component to zero, since $I(Ta; Fc|Ts) = U_{Ta} + S$. Then, we define R as $I(Ta; Fc)$, since $I(Ta; Fc) = U_{Ta} + R$, and U_{Ts} is computed as $U_{Ts} = I(Ta; Ts; Fc) - R$. For a case where $I(Ts; Fc|Ta)$ is not statistically significant, we apply a similar process. Finally, if neither conditional term of $I(Ta; Fc|Ts)$ or $I(Ts; Fc|Ta)$ is statistically significant would indicate that the only information component is redundancy. However, we defined that this case never occurs based on our study year period.

S4. Functional Performance

We calculated the individual level (Figures S3, S4, S5, S6, S7) and pairwise level (Figures S8, S9, S10, S11, S12, S14, S15, S16, S17, S18) of functional performance at Ne2, Ne3, Br1, Br3 and GC sites. These sites show similar patterns in mutual information as site Ne1 which presented in the main manuscript. We also 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 (Figures S13-S18) diminish some of the issues observed in the localized models (Figures S8-S12). 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 .

References

Barrett, A. B. (2015). Exploration of synergistic and redundant information sharing in

static and dynamical Gaussian systems. *Physical Review E*, *91*(5). doi: 10.1103/PhysRevE.91.052802

Goodwell, A., & Kumar, P. (2015). Information theoretic measures to infer feedback dynamics in coupled logistic networks. *Entropy*, *17*(11), 7468–7492. doi: 10.3390/e17117468

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

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

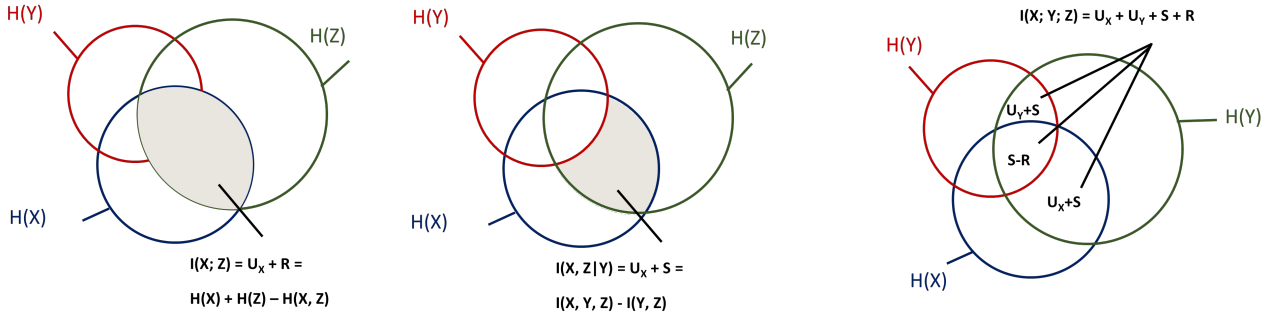

Figure S1. Illustration of information theory metrics. *(a)* Mutual information $I(X;Z)$ is the reduction in uncertainty about Z given knowledge of X. *(b)* Conditional mutual information $I(X;Z|Y)$ is the reduction in uncertainty about Z given knowledge of X, beyond information already provided by Y. *(c)* Multi-variate mutual information $I(X,Y;Z)$ is the total reduction in uncertainty about Z given knowledge of X and Y together, and is composed of four non-negative components of R, U_X, U_Y, and S.

Table S1. Summary of Machine Learning Model Architecture

Attribute	Description/Value
Model Type	Multiple Linear Regression (MLR)
Method	Ordinary Least Squares (OLS)
Implementation	"statsmodels" package in Python
Model Type	Random Forest (RF)
Trees in the Forest	100 (n-estimators)
Max Features	Square root of total features
Structure	Ensemble of Decision Trees
Implementation	"scikit-learn" package in Python
Model Type	Long Short Term Memory Model (LSTM)
Number of LSTM Layers	2
Number of Hidden Units per Layer	9
Dropout Layers	Between LSTM layers
Final Layer Type	Regression (1 unit for Fc)
Sequence Length	12 time steps (half a diurnal cycle)
Batch Size	128
Loss Function	Mean Squared Error (MSE)
Implementation	"torch" package in Python

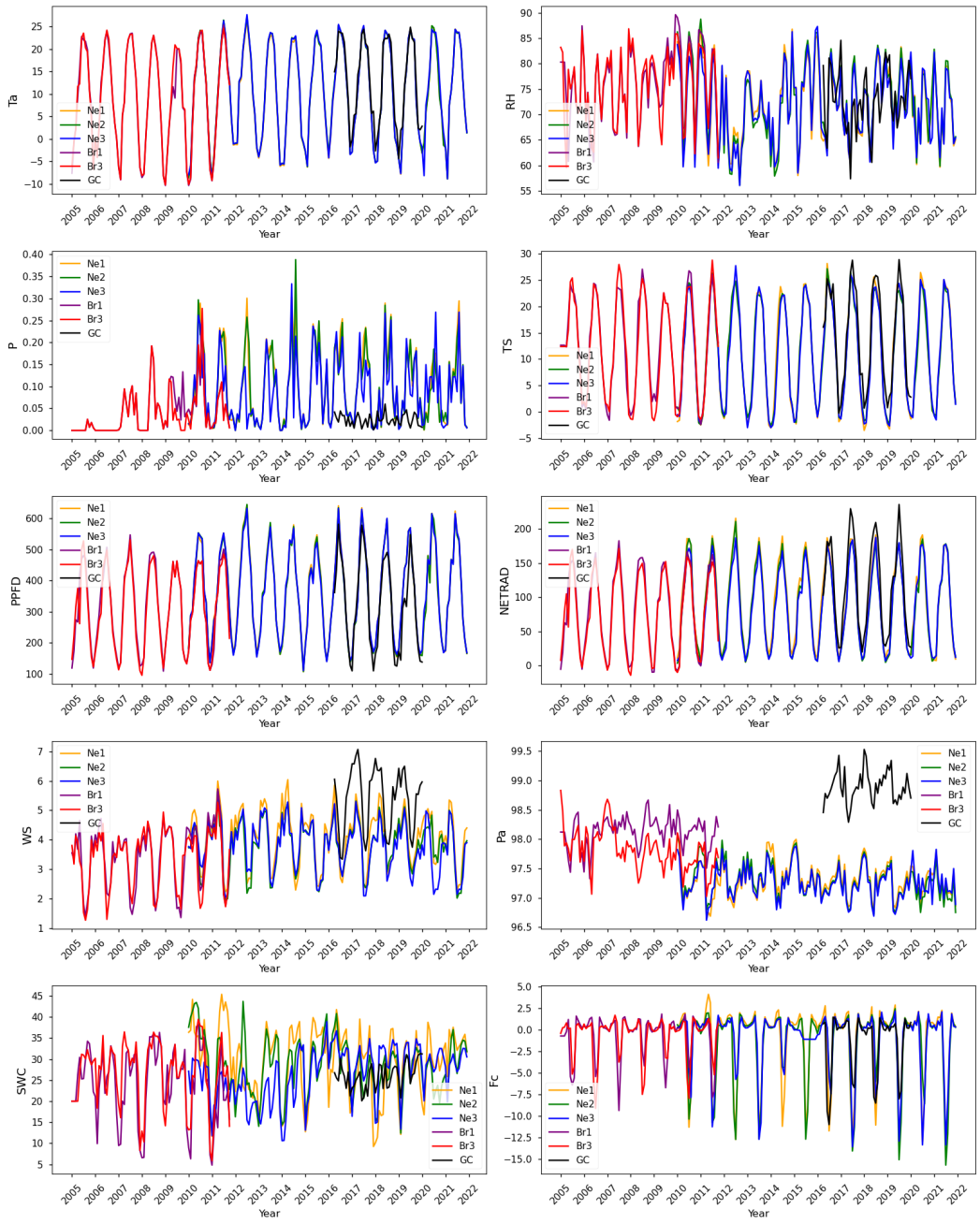

Figure S2. Averaged monthly values of driving variables (air temperature (Ta), relative humidity (RH), precipitation (P), soil temperature (TS), photosynthetic photon flux density ($PPFD$), net radiation ($NETRAD$), wind speed (WS), atmospheric pressure (Pa), soil water content (SWC)) and target variable (Fc) over the study years corresponded to different sites (Ne1, Ne2, Ne3, Br1, Br3, GC). Each site is represented by a unique color.

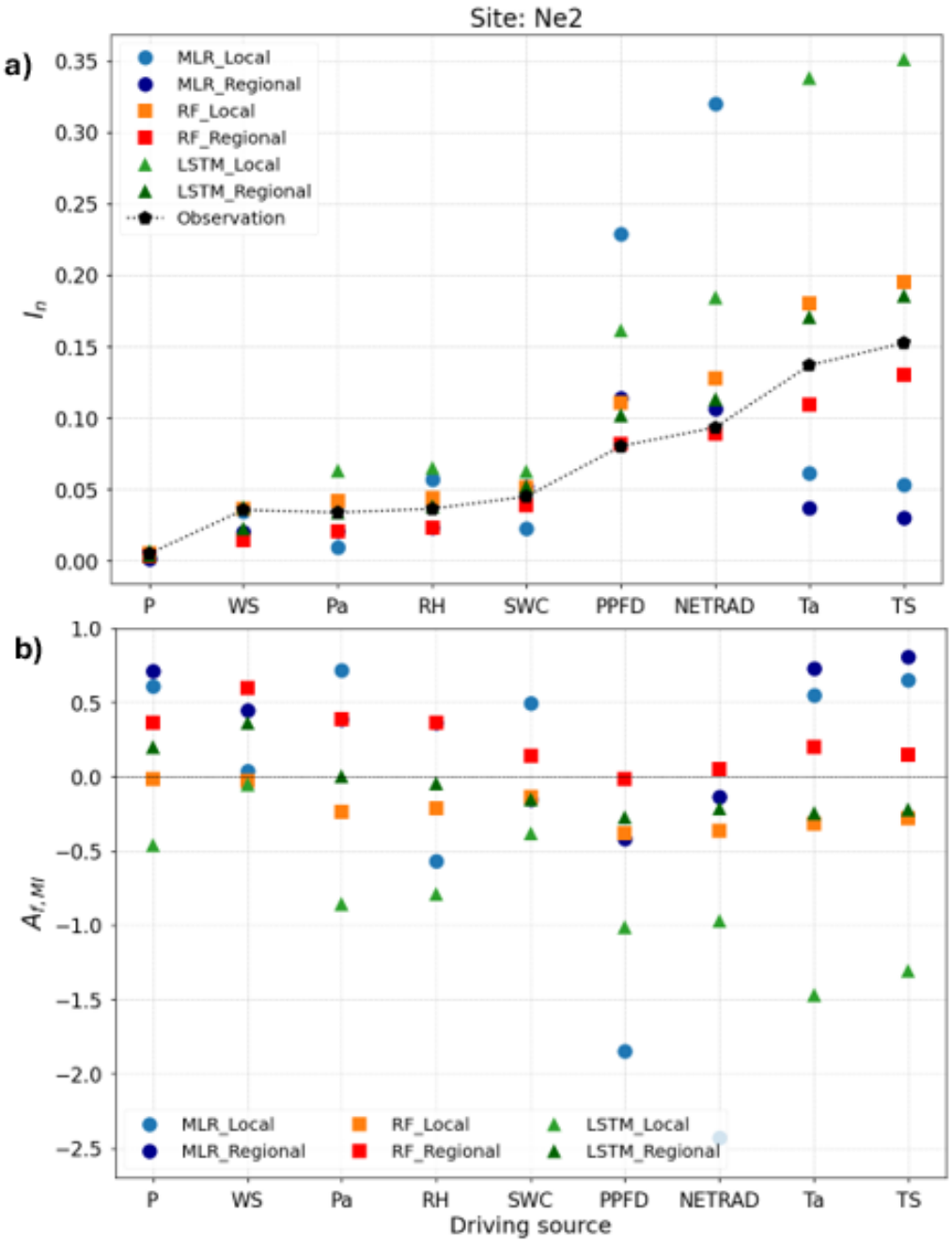

Figure S3. (a) Normalized mutual information (I_n) and (b) Individual source level of functional performance ($A_{f,MI}$) of three different models - Multiple Linear Regression (MLR), Random Forest (RF), and Long Short-Term Memory (LSTM) - under two training experiences, local and regional, at Ne2 site. Each variable is ranked based on the average observed MI across all sites. Observation values are represented with a black dot linked by a dashed line.

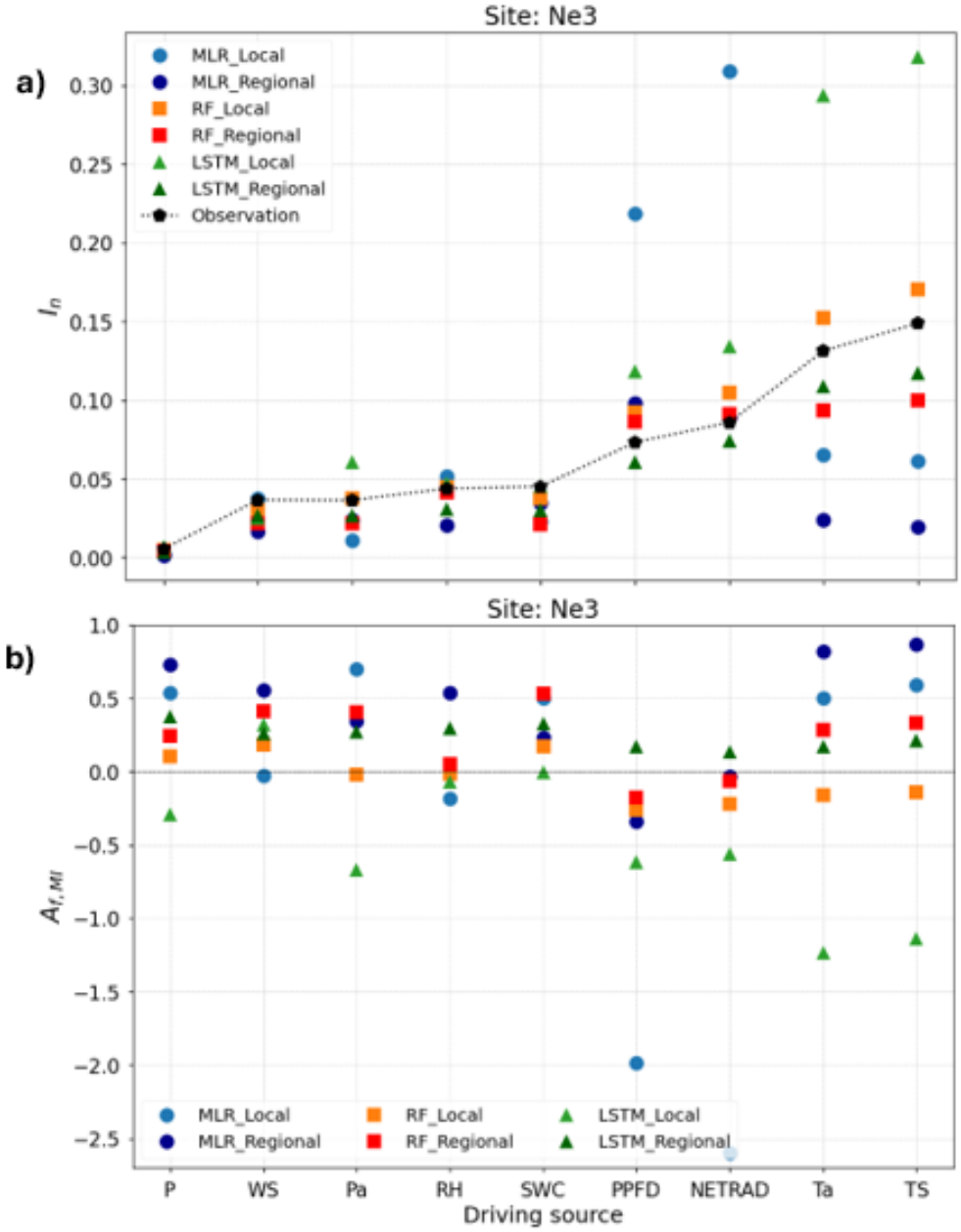

Figure S4. (a) Normalized mutual information (I_n) and (b) Individual source level of functional performance ($A_{f,MI}$) of three different models - Multiple Linear Regression (MLR), Random Forest (RF), and Long Short-Term Memory (LSTM) - under two training experiences, local and regional, at Ne3 site. Each variable is ranked based on the average observed MI across all sites. Observation values are represented with a black dot linked by a dashed line.

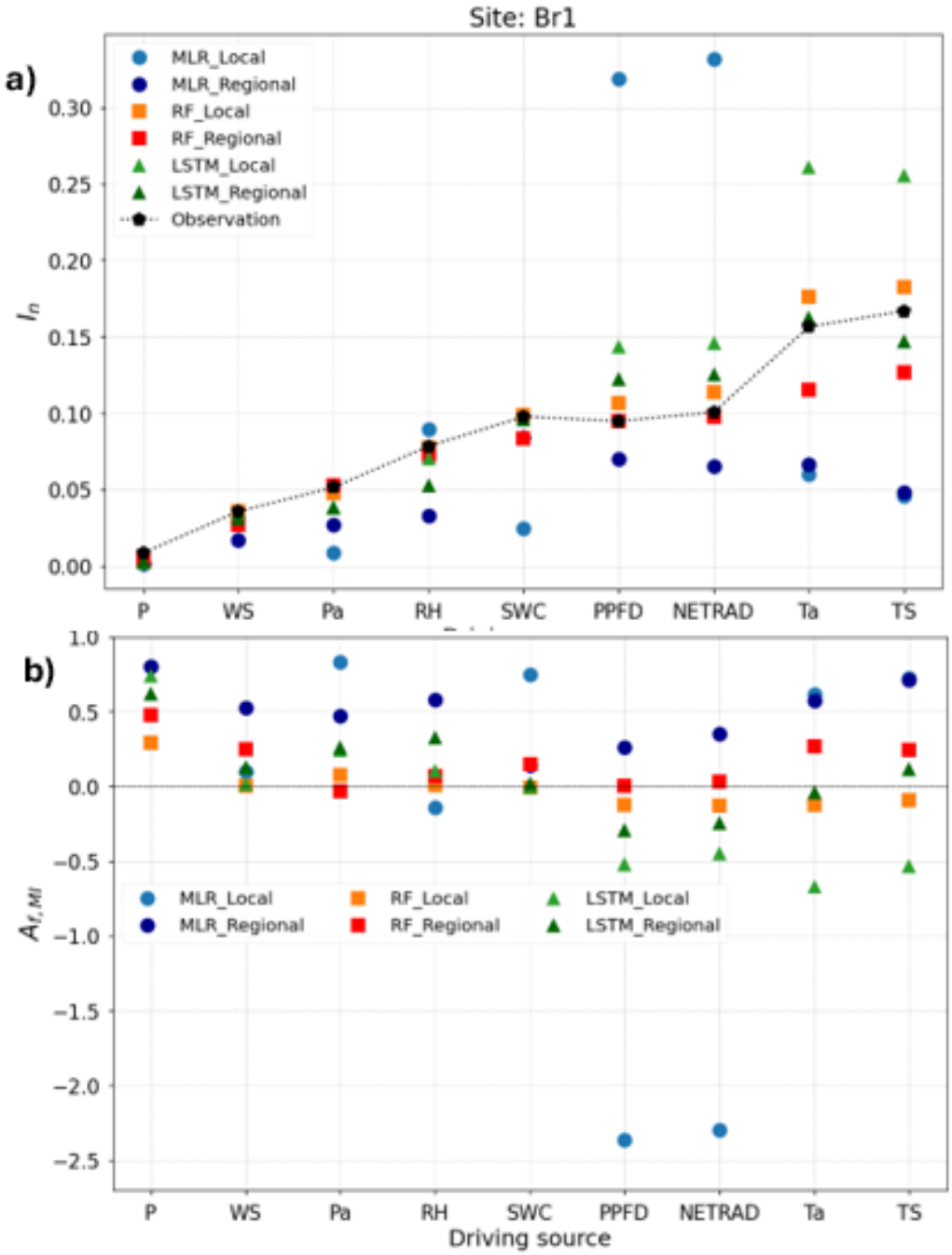

Figure S5. (a) Normalized mutual information (I_n) and (b) Individual source level of functional performance ($A_{f,MI}$) of three different models - Multiple Linear Regression (MLR), Random Forest (RF), and Long Short-Term Memory (LSTM) - under two training experiences, local and regional, at Br1 site. Each variable is ranked based on the average observed MI across all sites. Observation values are represented with a black dot linked by a dashed line.

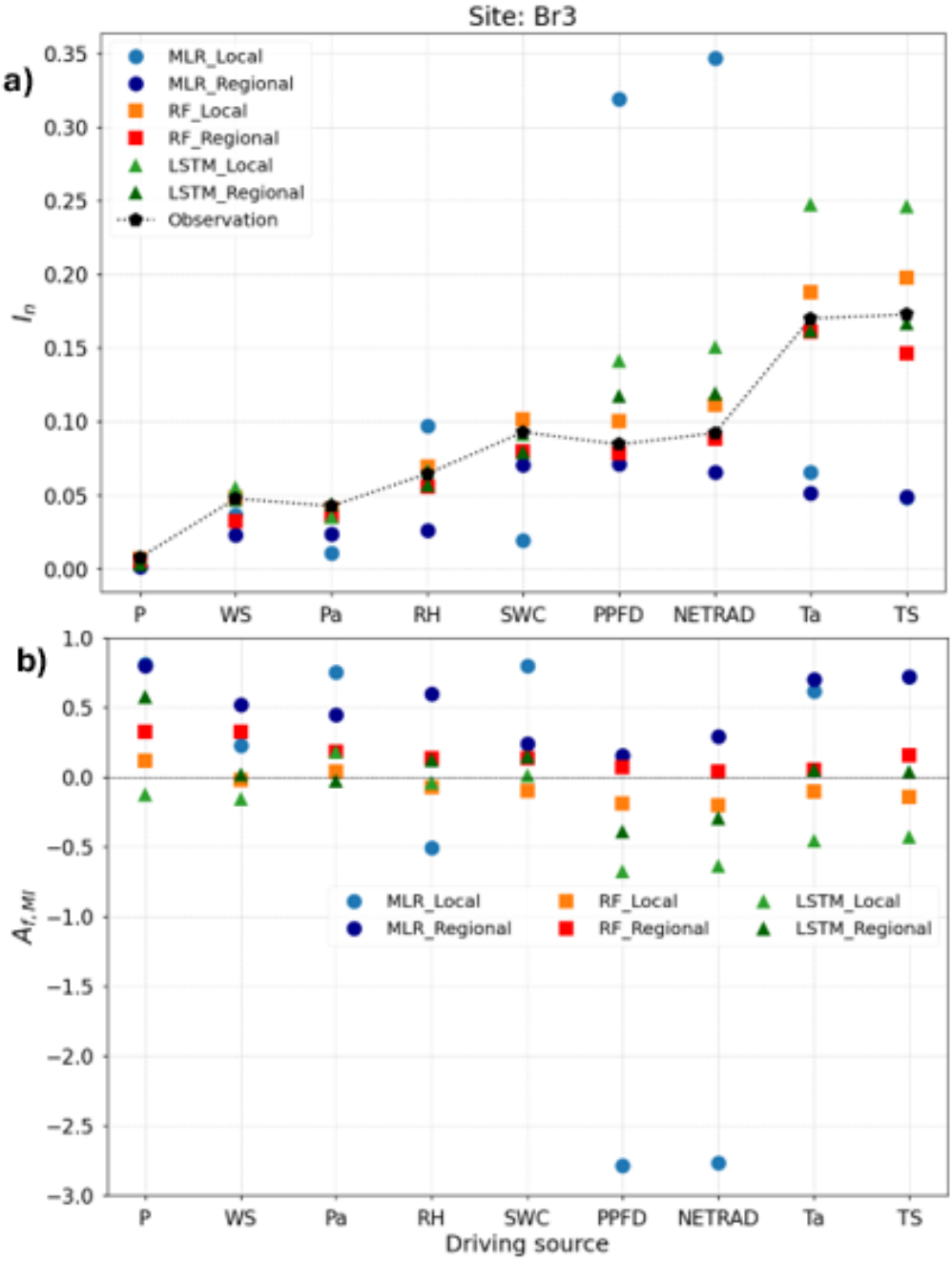

Figure S6. (a) Normalized mutual information (I_n) and (b) Individual source level of functional performance ($A_{f,MI}$) of three different models - Multiple Linear Regression (MLR), Random Forest (RF), and Long Short-Term Memory (LSTM) - under two training experiences, local and regional, at Br3 site. Each variable is ranked based on the average observed MI across all sites. Observation values are represented with a black dot linked by a dashed line.

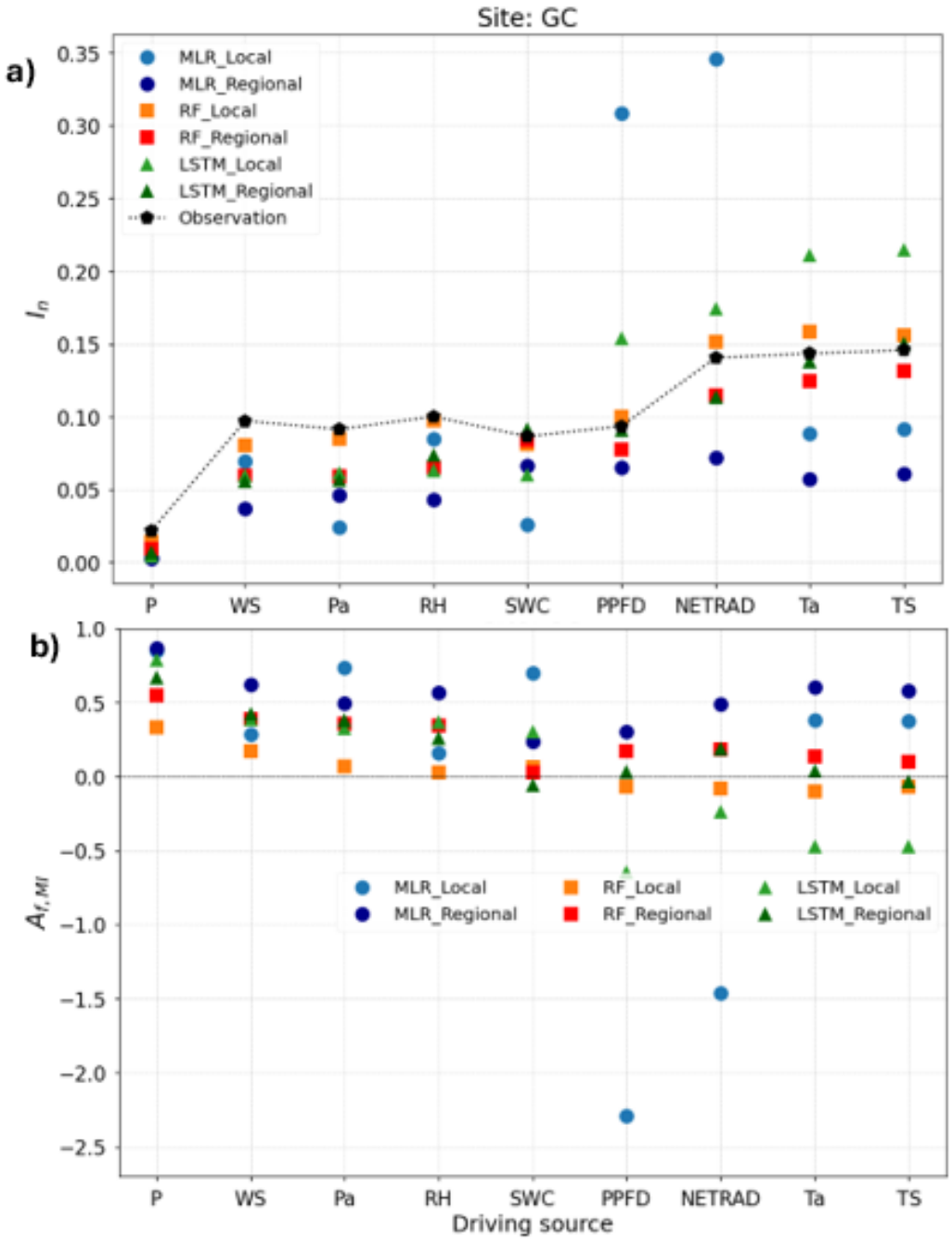

Figure S7. (a) Normalized mutual information (I_n) and (b) Individual source level of functional performance ($A_{f,MI}$) of three different models - Multiple Linear Regression (MLR), Random Forest (RF), and Long Short-Term Memory (LSTM) - under two training experiences, local and regional, at GC site. Each variable is ranked based on the average observed MI across all sites. Observation values are represented with a black dot linked by a dashed line.

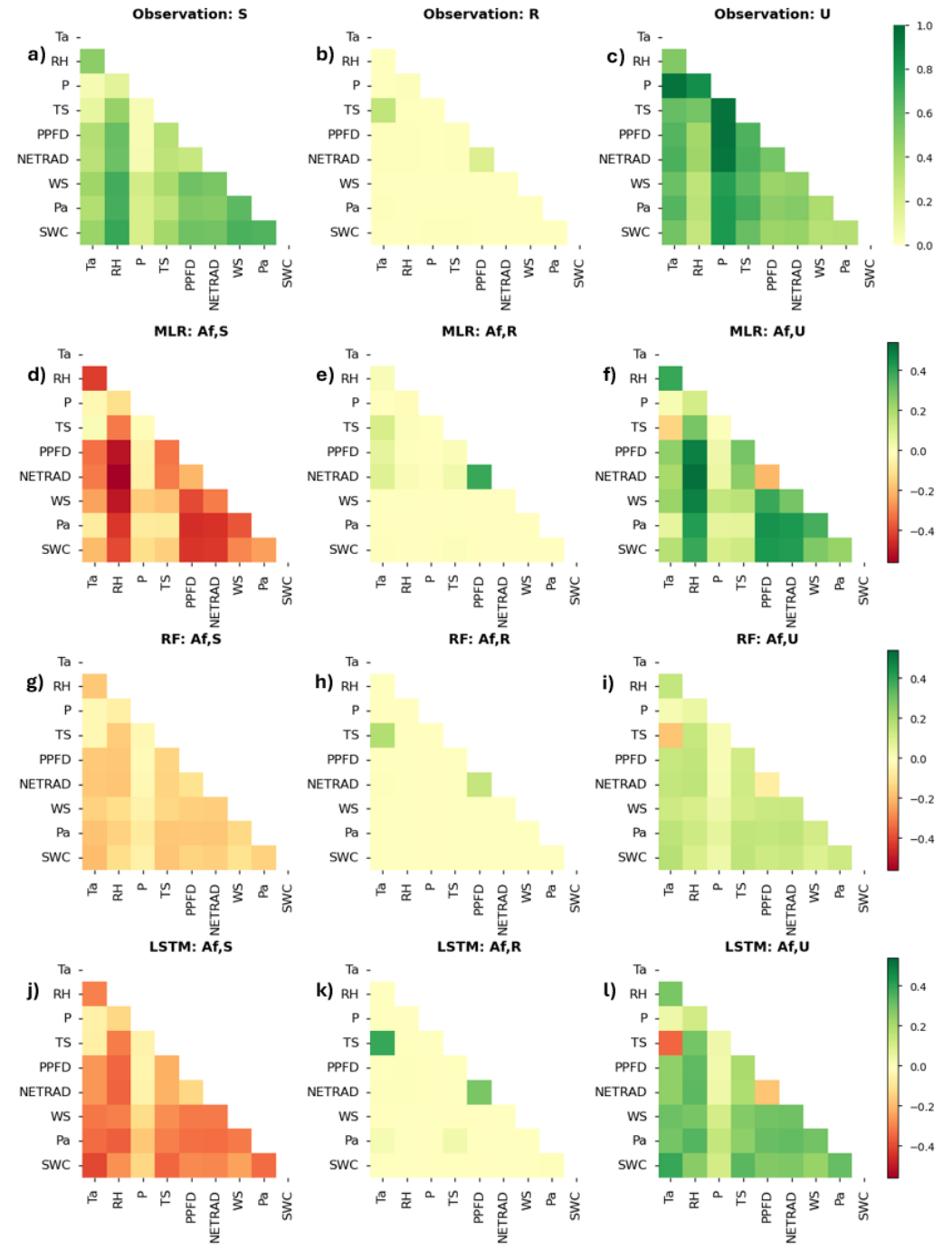

Figure S8. Observed pairwise (a) synergistic ($S_{i,j}$), (b) redundancy ($R_{i,j}$), and (c) uniqueness ($U_{i,j}$) information flow at Ne2 site. Pairwise functional performance of three models under local training experience at Ne2 site. The heat-map represents the relative difference in information 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.

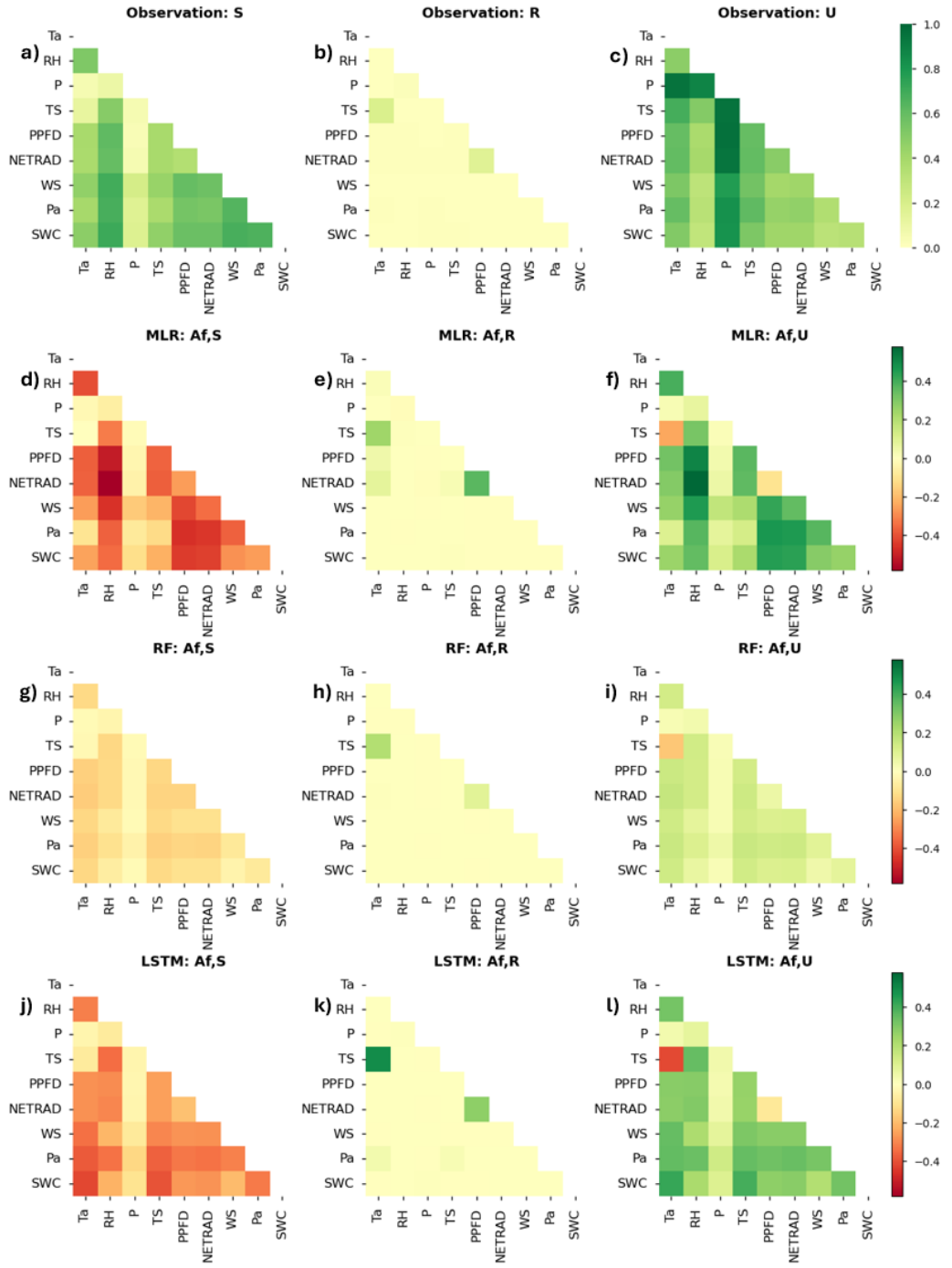

Figure S9. Observed pairwise (a) synergistic ($S_{i,j}$), (b) redundancy ($R_{i,j}$), and (c) uniqueness ($U_{i,j}$) information flow at Ne3 site. Pairwise functional performance of three models under local training experience at Ne3 site. The heat-map represents the relative difference in information 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.

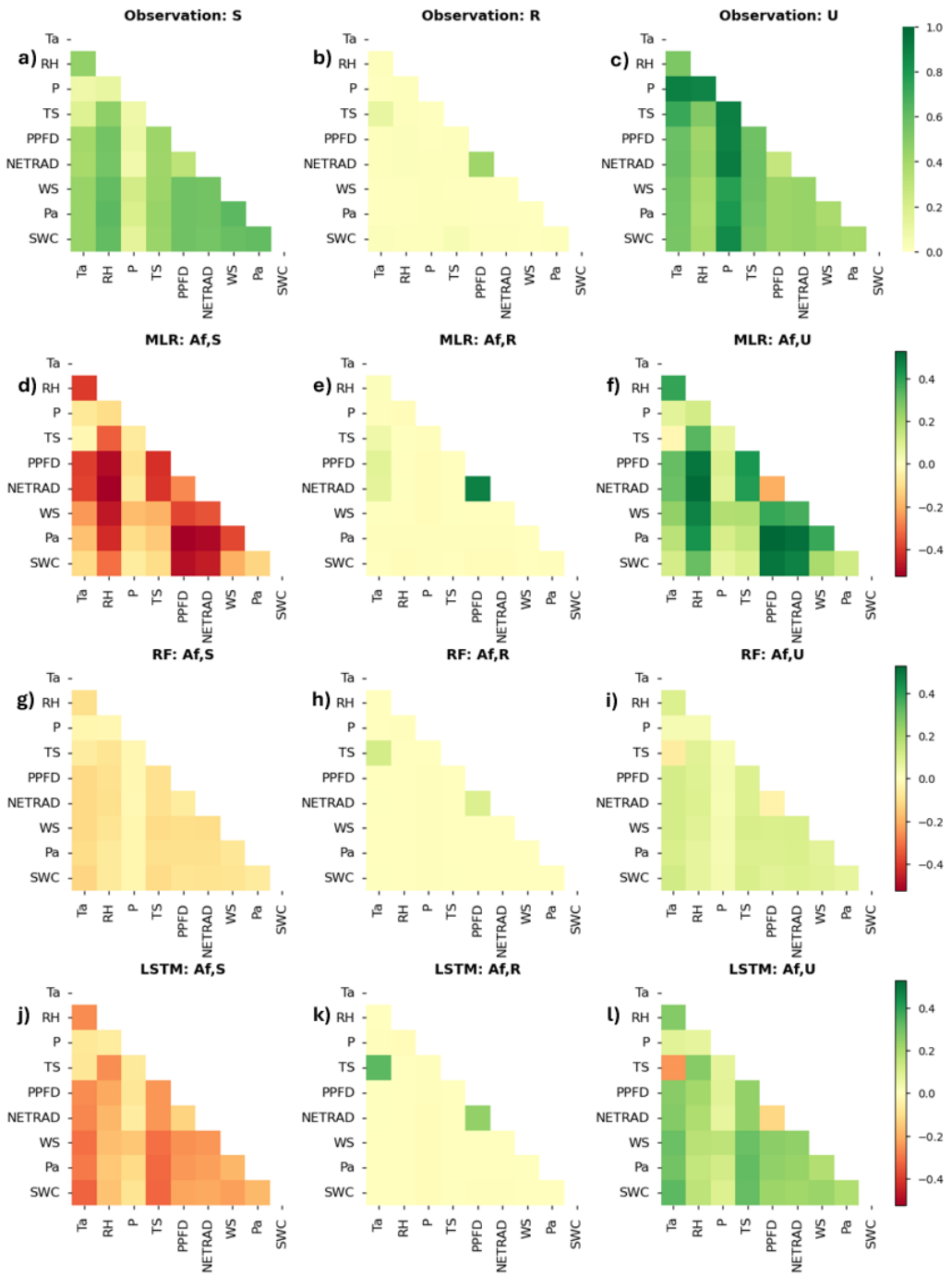

Figure S10. Observed pairwise (a) synergistic ($S_{i,j}$), (b) redundancy ($R_{i,j}$), and (c) uniqueness ($U_{i,j}$) information flow at Br1 site. Pairwise functional performance of three models under local training experience at Br1 site. The heat-map represents the relative difference in information 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.

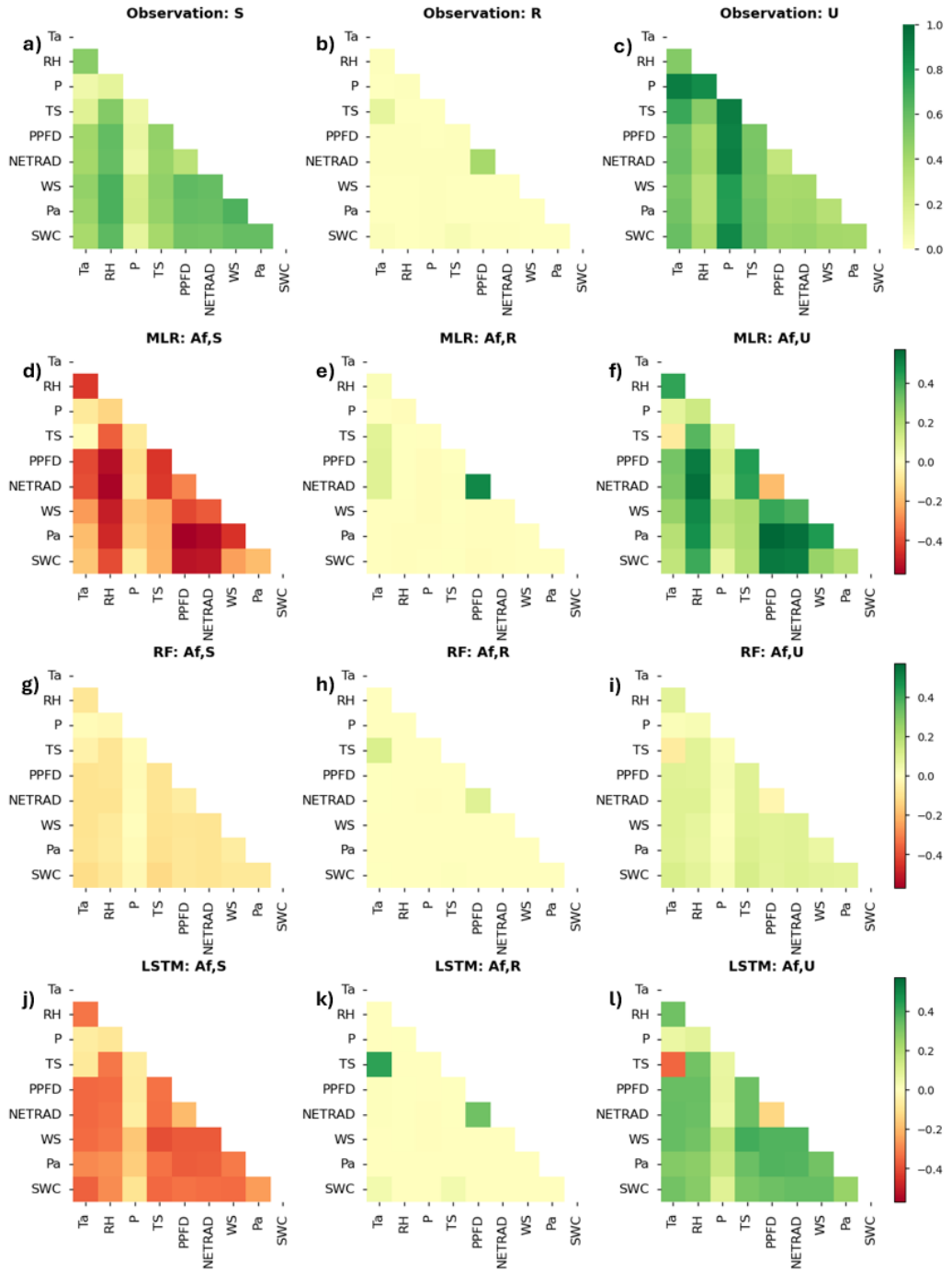

Figure S11. Observed pairwise (a) synergistic ($S_{i,j}$), (b) redundancy ($R_{i,j}$), and (c) uniqueness ($U_{i,j}$) information flow at Br3 site. Pairwise functional performance of three models under local training experience at Br3 site. The heat-map represents the relative difference in information 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.

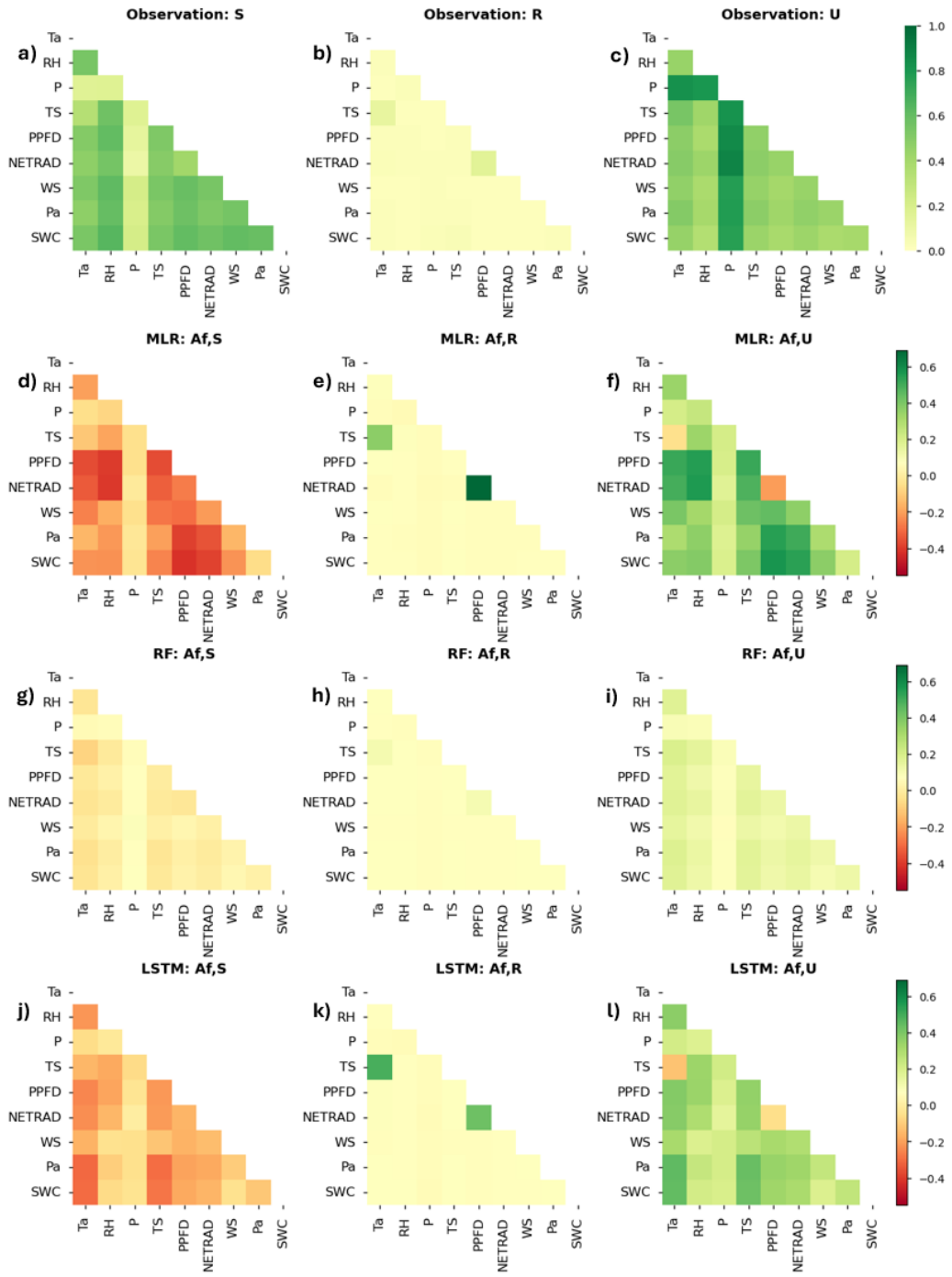

Figure S12. Observed pairwise (a) synergistic ($S_{i,j}$), (b) redundancy ($R_{i,j}$), and (c) uniqueness ($U_{i,j}$) information flow at GC site. Pairwise functional performance of three models under local training experience at GC site. The heat-map represents the relative difference in information 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.

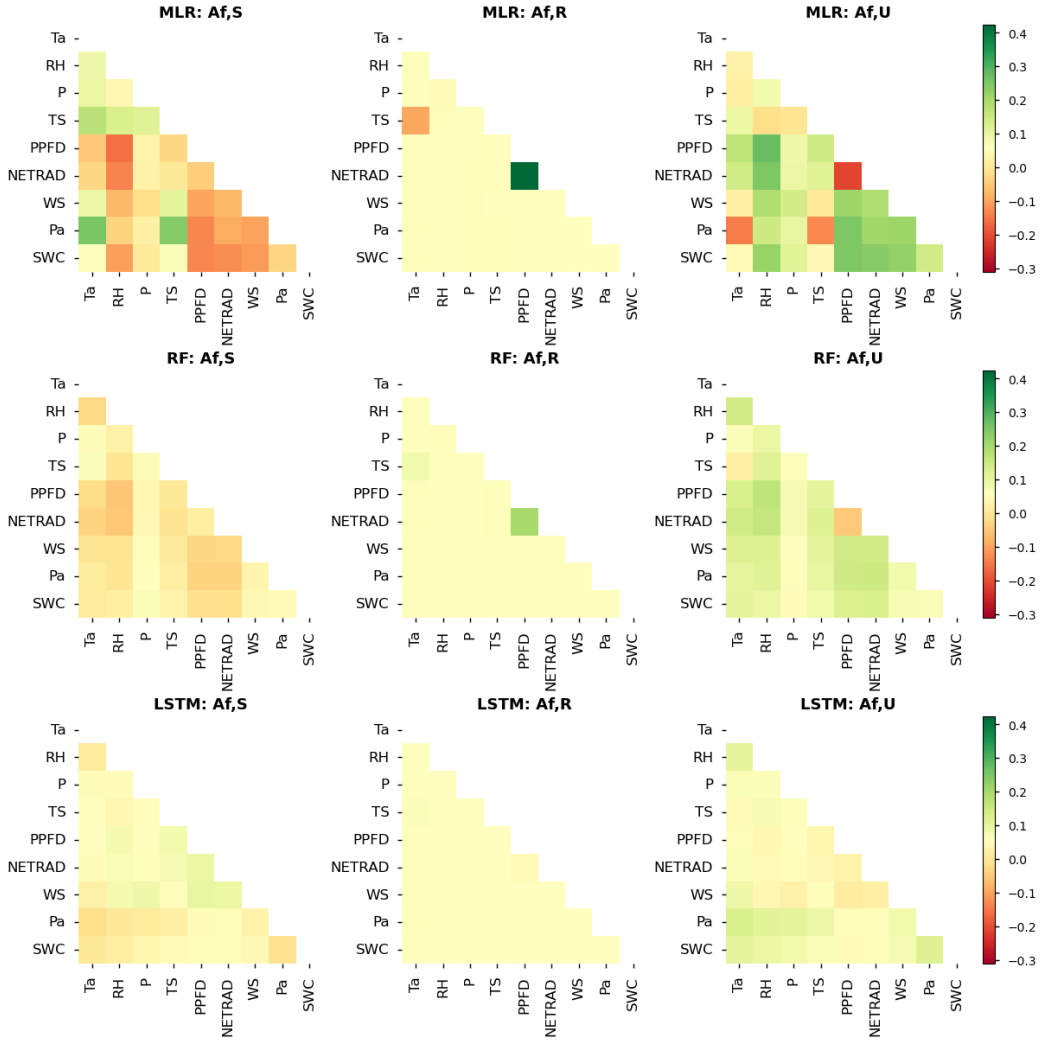

Figure S13. Pairwise functional performance of three models under regional training experience at Ne1 site. The heat-map represents the relative difference in information decomposition partitioning measures ($A_{f,S_{i,j}}$, $A_{f,R_{i,j}}$, and $A_{f,U_{i,j}}$ between modeled and observed data for each pair of forcing variables. Positive values (green) indicate that the model overestimates the information type, while negative values (red) indicate underestimations.

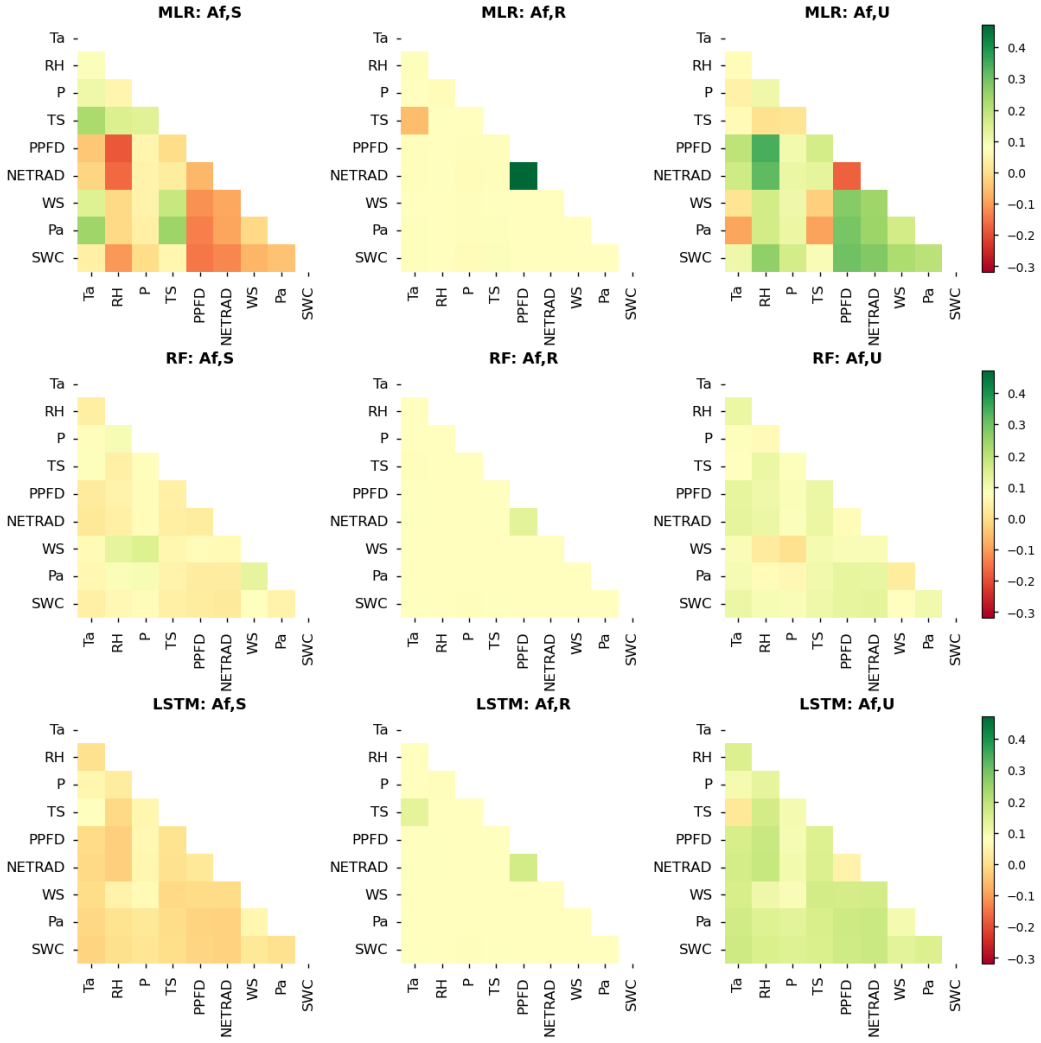

Figure S14. Pairwise functional performance of three models under regional training experience at Ne2 site. The heat-map represents the relative difference in information decomposition partitioning measures ($A_{f,S_{i,j}}$, $A_{f,R_{i,j}}$, and $A_{f,U_{i,j}}$ between modeled and observed data for each pair of forcing variables. Positive values (green) indicate that the model overestimates the information type, while negative values (red) indicate underestimations.

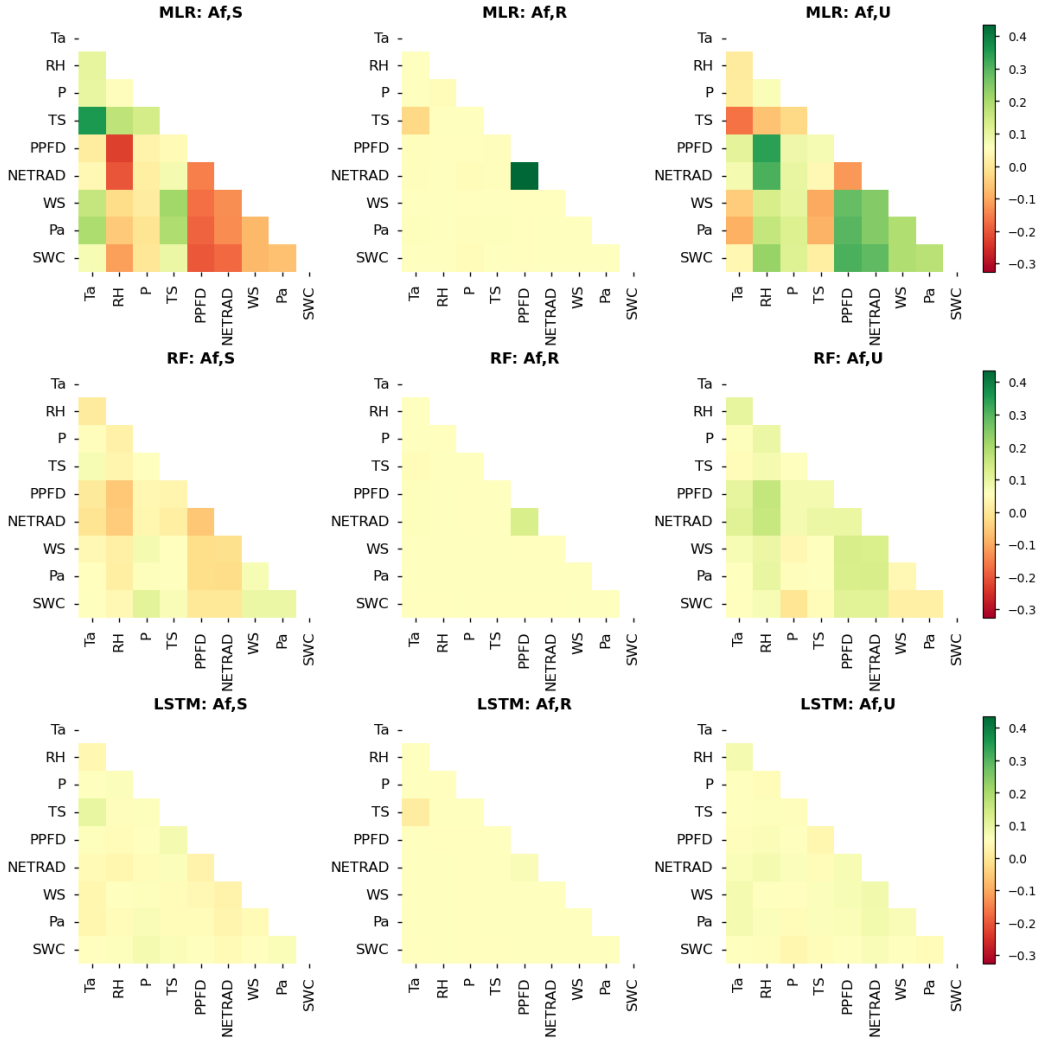

Figure S15. Pairwise functional performance of three models under regional training experience at Ne3 site. The heat-map represents the relative difference in information decomposition partitioning measures ($A_{f,S_{i,j}}$, $A_{f,R_{i,j}}$, and $A_{f,U_{i,j}}$ between modeled and observed data for each pair of forcing variables. Positive values (green) indicate that the model overestimates the information type, while negative values (red) indicate underestimations.

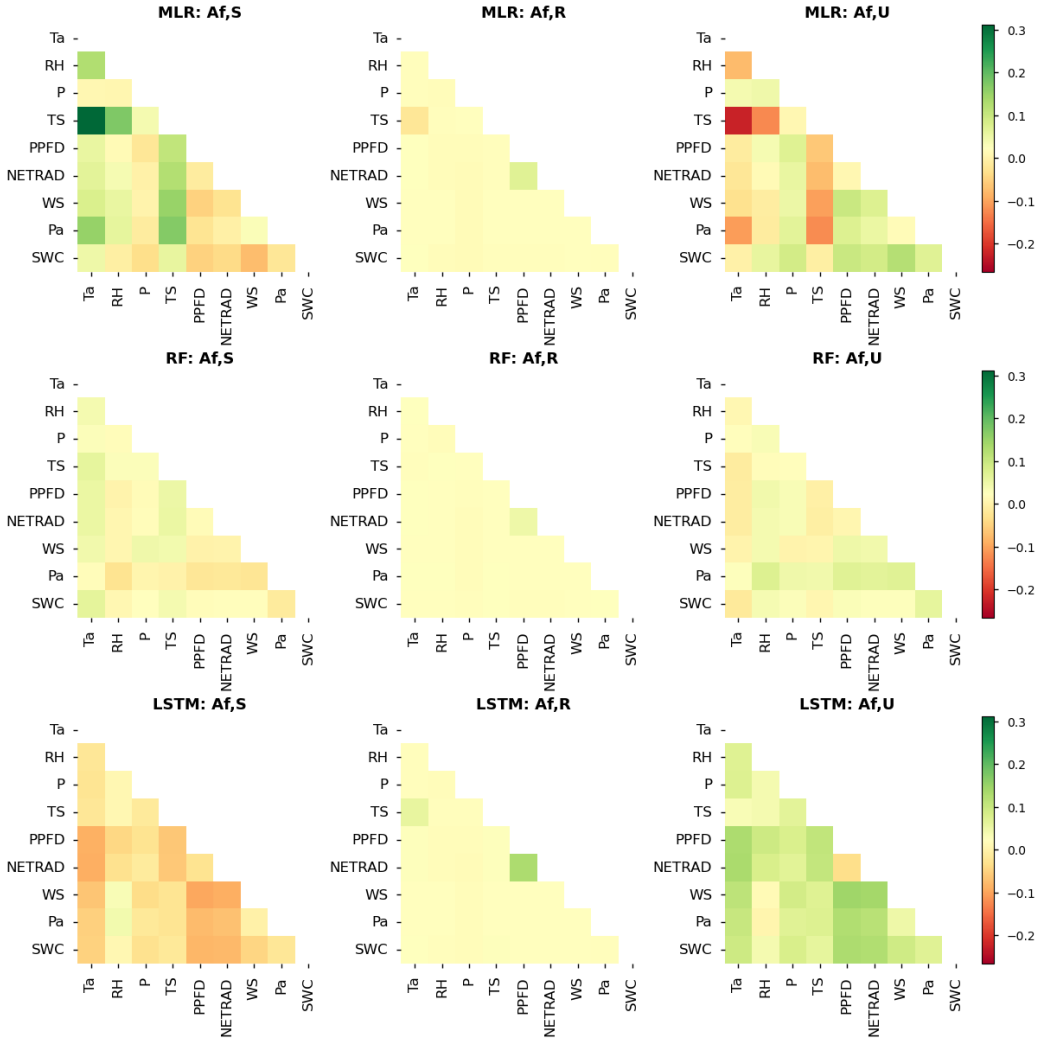

Figure S16. Pairwise functional performance of three models under regional training experience at Br1 site. The heat-map represents the relative difference in information decomposition partitioning measures ($A_{f,S_{i,j}}$, $A_{f,R_{i,j}}$, and $A_{f,U_{i,j}}$ between modeled and observed data for each pair of forcing variables. Positive values (green) indicate that the model overestimates the information type, while negative values (red) indicate underestimations.

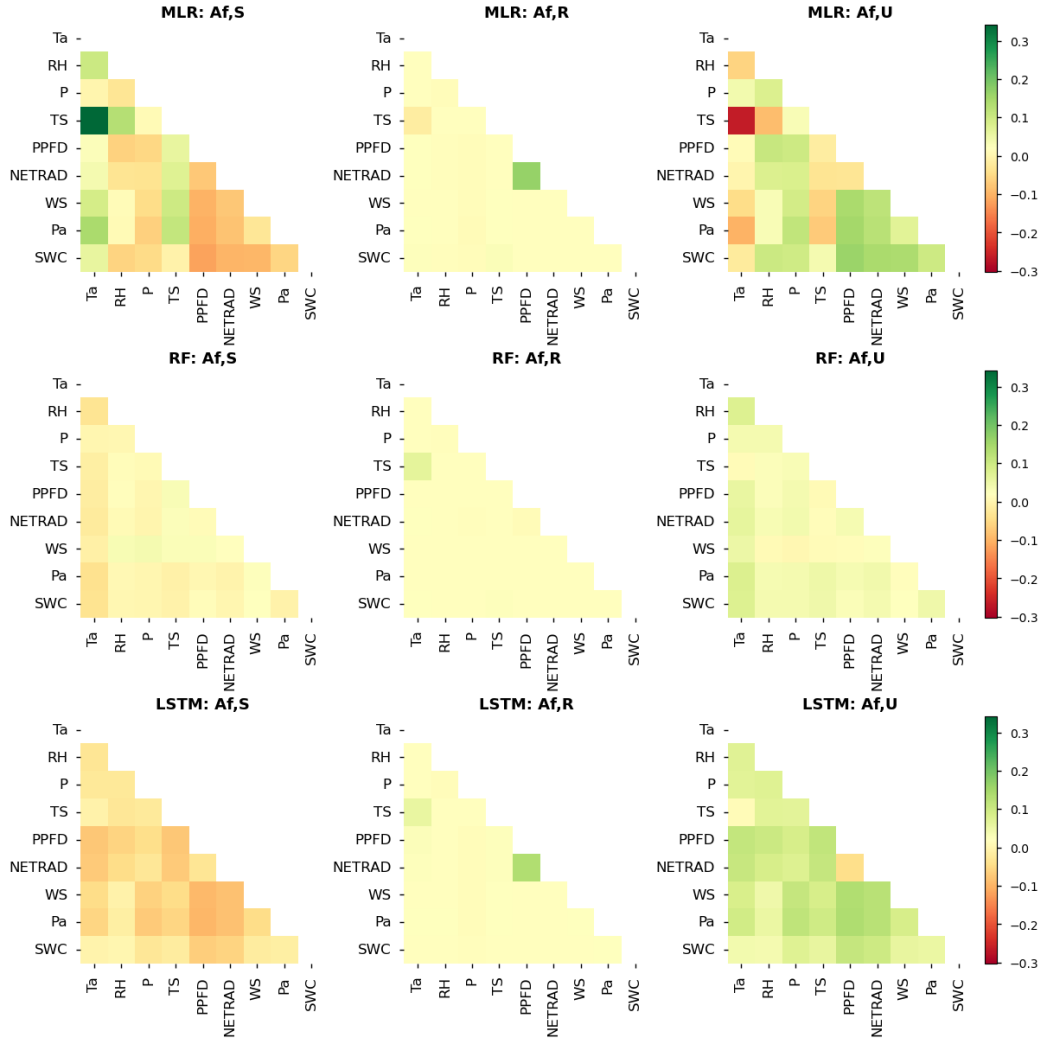

Figure S17. Pairwise functional performance of three models under regional training experience at Br3 site. The heat-map represents the relative difference in information decomposition partitioning measures ($A_{f,S_{i,j}}$, $A_{f,R_{i,j}}$, and $A_{f,U_{i,j}}$ between modeled and observed data for each pair of forcing variables. Positive values (green) indicate that the model overestimates the information type, while negative values (red) indicate underestimations.

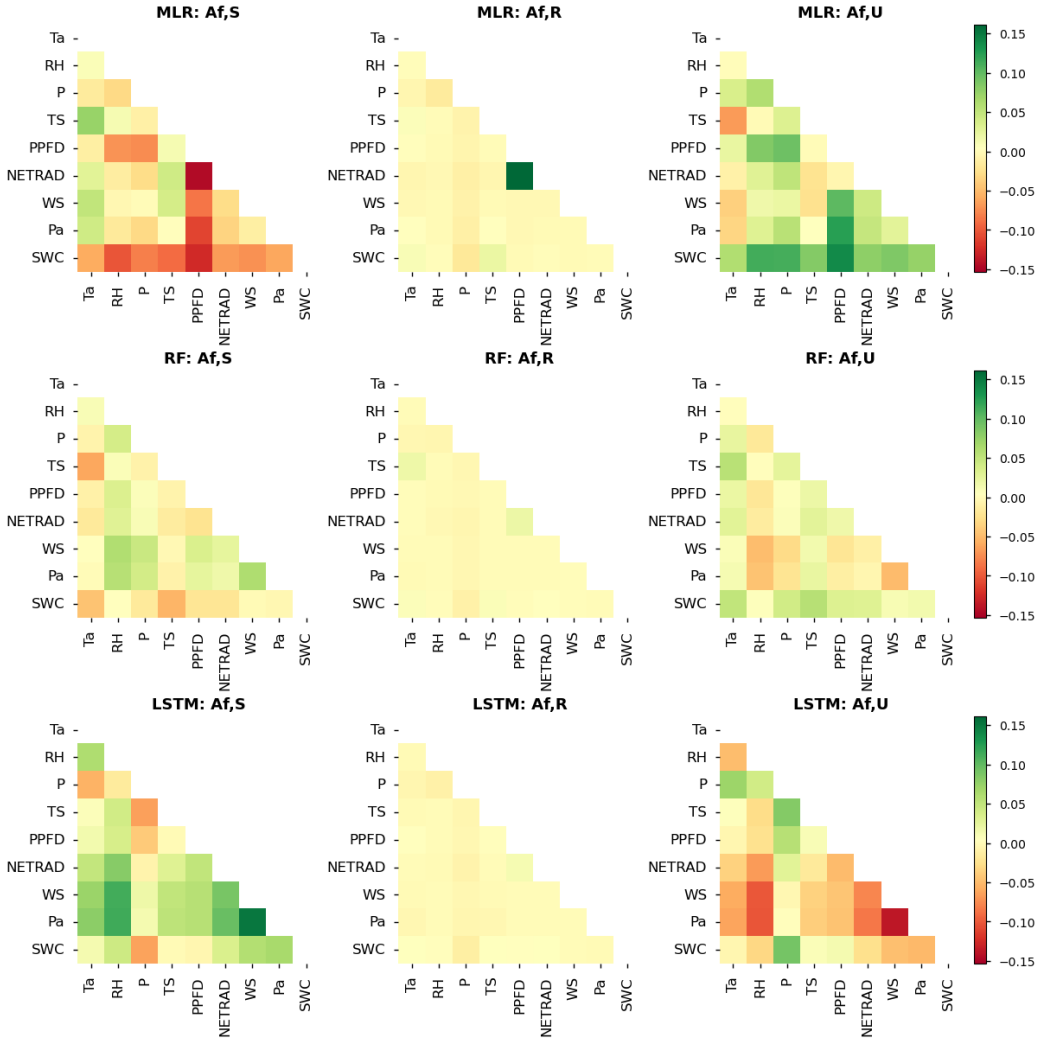

Figure S18. Pairwise functional performance of three models under regional training experience at GC site. The heat-map represents the relative difference in information decomposition partitioning measures ($A_{f,S_{i,j}}$, $A_{f,R_{i,j}}$, and $A_{f,U_{i,j}}$ between modeled and observed data for each pair of forcing variables. Positive values (green) indicate that the model overestimates the information type, while negative values (red) indicate underestimations.