BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Research in Ecology
https://ojs.bilpublishing.com/index.php/re

ARTICLE

Machine Learning and Pattern Analysis Identify Distinctive Influences from Long-term Weekly Net Ecosystem Exchange at Four Deciduous Woodland Locations

David A. Wood*

DWA Energy Limited, Lincoln, LN5 9JP, United Kingdom

ARTICLE INFO

Article history
Received: 25 November 2022
Received in revised form: 18 January 2023
Accepted: 30 January 2023
Published: 8 February 2023

Keywords:
Eddy-covariance
CO_2-flux influences
Multi-fold cross validation
Weekly *NEE* pattern analysis
Site specific *NEE* influences
FLUXNET2015 protocols

ABSTRACT

A methodology integrating correlation, regression (MLR), machine learning (ML), and pattern analysis of long-term weekly net ecosystem exchange (*NEE*) datasets are applied to four deciduous broadleaf forest (DBF) sites forming part of the AmeriFlux (FLUXNET2015) database. Such analysis effectively characterizes and distinguishes those DBF sites for which long-term *NEE* patterns can be accurately predicted using the recorded environmental variables, from those sites cannot be so delineated. Comparisons of twelve *NEE* prediction models (5 MLR; 7 ML), using multi-fold cross-validation analysis, reveal that support vector regression generates the most accurate and reliable predictions for each site considered, based on fits involving between 16 and 24 available environmental variables. SVR can accurately predict *NEE* for datasets for DBF sites US-MMS and US-MOz, but fail to reliably do so for sites CA-Cbo and MX-Tes. For the latter two sites the predicted versus recorded *NEE* weekly data follow a Y ≠ X pattern and are characterized by rapid fluctuations between low and high *NEE* values across leaf-on seasonal periods. Variable influences on *NEE*, determined by their importance to MLR and ML model solutions, identify distinctive sets of the most and least influential variables for each site studied. Such information is valuable for monitoring and modelling the likely impacts of changing climate on the ability of these sites to serve as long-term carbon sinks. The periodically oscillating *NEE* weekly patterns distinguished for sites CA-Cbo and MX-Tes are not readily explained in terms of the currently recorded environmental variables. More detailed analysis of the biological processes at work in the forest understory and soil at these sites are recommended to determine additional suitable variables to measure that might better explain such fluctuations.

*Corresponding Author:
David A. Wood,
DWA Energy Limited, Lincoln, LN5 9JP, United Kingdom;
Email: dw@dwasolutions.com

DOI: https://doi.org/10.30564/re.v4i4.5279

1. Introduction

Carbon dioxide (CO_2) heat and water fluxes from ecosystems need to be accurately measured to establish their potential as long-term carbon sinks [1-3]. The eddy-covariance technique has been refined to provide reliable net ecosystem exchange (*NEE*) CO_2-flux measurements [4-6]. Ecosystems display a wide range of *NEE* seasonal patterns [7] as they are influenced by many environmental, climate, and biological factors. Environmentally, solar radiation, air/soil temperatures, rainfall, soil water contents, and vapor pressure deficit tend to influence *NEE* patterns [8,9]. Periodic disturbances (droughts, wildfires, pests, diseases, industrial activities, etc.) can perturb deciduous broadleaf forest (DBF) *NEE* contributions from the canopy, understory, and soil.

In woodland biomes, canopy height, aerodynamics, mean tree age, tree-stand density, and friction velocity (surface roughness) influence *NEE* patterns [10-13]. Latent and sensible heat fluxes vary diurnally and seasonally due to geographic factors [14,15]. Photosynthetically active radiation (PAR; 400 nm ~ 700 nm wavelengths) and photosynthetic photon flux density (PPFD) are key NEE influencers, for the DBF canopy in particular [16,17]. Satellite-derived information is widely used to extrapolate *NEE* data measured at specific sites across broader geographic areas based on selected environmental variables [18-22], taking into account seasonal, latitudinal and weather-related variations [23-25]. Machine Learning (ML) techniques are often used to fill data gaps in such circumstances [26].

Partitioning *NEE* contributions among the species present in the DBF canopy, understory and soil, taking into account their respective growing seasons (leaf-on versus leaf-off), the impacts of wet and dry seasons, seasonal changes, and extreme events can refine the understanding of *NEE* seasonal patterns at specific sites where such data is recorded [27-33]. Variable correlations and multi-linear regressions (MLR) with *NEE* provide further insight [34,35]. Comparisons of multiple ML algorithms applied to fit *NEE* seasonal patterns [36,37], and data mining techniques to explore the details of *NEE*-environmental-variable relationships typically provide complementary insights to the calculated NEE datasets [38].

NEE determination involves multi-component calculations combining photosynthetic and respiratory ecosystem contributions recorded every thirty minutes, to meet FLUXNET2015 standards, and partitioned between daytime and nighttime recordings [39,40]. The pre-processed and verified *NEE* data are compiled into hourly, daily, weekly, monthly and annual datasets, and then formally released for public-domain analysis [41]. *NEE* calculation and verification are complex and time-consuming. Hence, ML models that accurately predict *NEE* from routinely recorded, site-specific environmental variables are useful for expanding the spatial application of recorded data. However, there is a research gap as not many ML models developed to date accurately replicate multi-year, seasonal *NEE* patterns at specific sites.

The *NEE* computation involves net primary production (*NPP*), taking into account the respiration of autotrophs (Ra) and gross primary production (*GPP*) (Equation (1)) and respiration of forest-litter/soil-based heterotrophs (Rh) (Equation (2)) [42].

$$NPP = GPP - Ra \tag{1}$$

$$NEE = NPP - Rh \tag{2}$$

The daylight hours ($NEE_{daytime}$) contribution is then determined by distinguishing daylight photorespiration (Rp), maintenance respiration (Rm), autotroph growth respiration (Rs), and Rh (Equation (3)).

$$NEE_{daytime} = GPP - Rp - Rm - Rs - Rh \tag{3}$$

The *NEE* components record carbon flux as absorbed (*NEE* +ve; carbon sinks) and/or released (*NEE*-ve; carbon sources) in relation to specific surface areas and periods of time. For daily and weekly time periods, *NEE* is expressed for FLUXNET2015 in $gCm^{-2}d^{-1}$. Weekly information is useful for *NEE*-seasonal-pattern analysis as it provides appropriate granularity for understanding long-term changes impacting specific DBF sites [43].

Four DBF sites with FLUXNET2015-quality, multi-year, eddy covariance datasets from the AmeriFlux database [44] are characterized hourly with MLR and ML models to assess the ability of the recorded environmental variables to reliably predict *NEE* patterns. The analyses specifically:

- Combine correlations, MLR, ML, and pattern analysis to determine whether *NEE* weekly patterns at DBF sites can or cannot be expressed reliably from available environmental variables;
- Establish which available environmental variables have the most influence on multi-year *NEE* pattern fits;
- Identify the specific annual periods at specific DBF sites that the MLR/ML models find difficult to predict.

2. Materials

2.1 FLUXNET Recorded and Processed Data

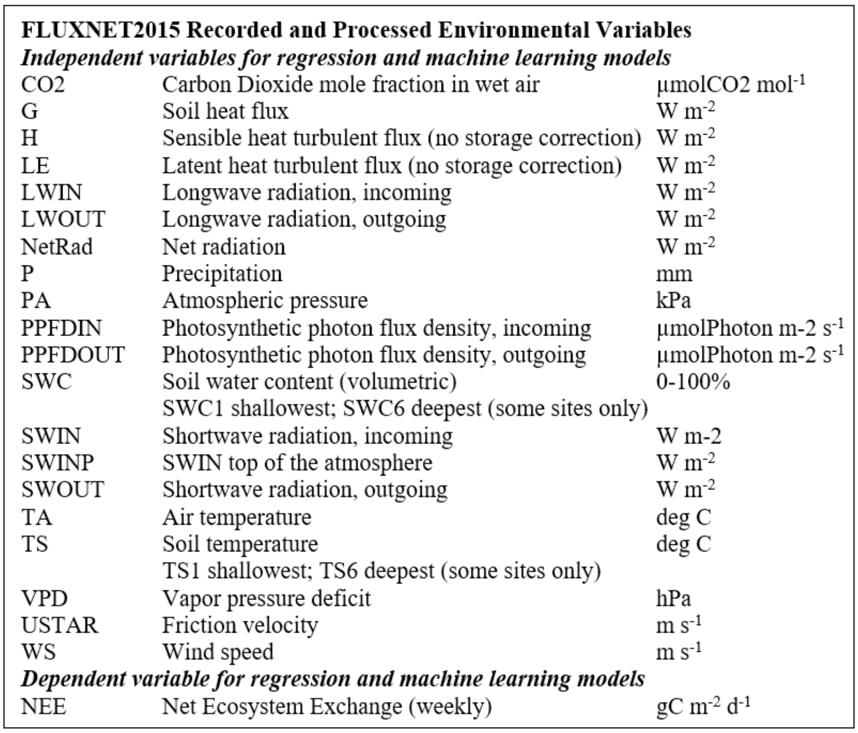
In excess of one thousand eddy-covariance recording sites now constitute the FLUXNET [41] worldwide database [45]. The favored processing pipeline for that data is to the FLUXNET2015 standard [46] rigorously pre-processes, quality controls, and, where possible, gap-fills the recorded data before releasing it for research.

Several hundred eddy-covariance recording sites distributed throughout the Americas belong to the AmeriFlux network [44] supported by the US Department of Energy. The Ameriflux sites make datasets publicly available under license. The data released on specified sites are processed and verified to meet FLUXNET2015 requirements [47]. Weekly datasets from four DBF sites processed to FLUXNET2015 standards were selected for evaluation by this study. These sites are CA-Cbo, MX-Tes, US-MMS, and US-MOz. These sites were selected because continuous data recordings were available at each site for multiple years covering a substantial number of environmental variables without major data gaps. In addition to weekly-averaged *NEE* data, 16 to 24 continuously recorded environmental variables, as part of each site's processed *NEE* datasets, were compiled for use in *NEE* prediction models by this study. 729, 203, 1005, and 739 weekly data records were compiled, respectively for sites CA-Cbo, MX-Tes, US-MMS, and US-MOz for the analysis conducted in this study. Figure 1 defines those variables, the abbreviations applied to them, and their units of measurement.

2.2 Deciduous Woodland AmeriFlux Sites Modelled

Weekly data records of *NEE*-averaged values with associated environmental variables are evaluated for four deciduous broadleaf forest (DBF) sites with seasonal leaf-on and leaf-off periods each year. The sites are CA-Cbo (Canada), MX-Tes (Mexico), US-MMS and US-MOz (U.S.A.) and are described below. These woodland sites are all parts of secondary forests that were at one stage cultivated for agricultural purposes and subsequently allocated for woodland development. The sites form part of the AmeriFlux dataset, the historical data recorded from which now conforms to the FLUXNET2015 protocol.

CA-Cbo (44°19'0"N, 79°50'60"W) is located in the Borden Forest Research Station at 120 m above sea level; an extensive woodland area of Southern Ontario (Canada) close to the ecotone with boreal conifer forests to the north. It constitutes a mixed forest dominated by *Acer rubrum* (red maple; ~50%), *Pinus strobus* (white pine), *Populus grandidentata* (large-tooth aspen), *Fagus grandifolia* (American beech) and *Fraxinus americana* (white ash), evolving as natural regrowth on abandoned farmland since 1916. The forest canopy is currently about 22 m tall and the tree density (living and dead) is greater than



FLUXNET2015 Recorded and Processed Environmental Variables

Independent variables for regression and machine learning models		
CO2	Carbon Dioxide mole fraction in wet air	µmolCO2 mol^{-1}
G	Soil heat flux	W m^{-2}
H	Sensible heat turbulent flux (no storage correction)	W m^{-2}
LE	Latent heat turbulent flux (no storage correction)	W m^{-2}
LWIN	Longwave radiation, incoming	W m^{-2}
LWOUT	Longwave radiation, outgoing	W m^{-2}
NetRad	Net radiation	W m^{-2}
P	Precipitation	mm
PA	Atmospheric pressure	kPa
PPFDIN	Photosynthetic photon flux density, incoming	µmolPhoton m-2 s^{-1}
PPFDOUT	Photosynthetic photon flux density, outgoing	µmolPhoton m-2 s^{-1}
SWC	Soil water content (volumetric) SWC1 shallowest; SWC6 deepest (some sites only)	0-100%
SWIN	Shortwave radiation, incoming	W m-2
SWINP	SWIN top of the atmosphere	W m^{-2}
SWOUT	Shortwave radiation, outgoing	W m^{-2}
TA	Air temperature	deg C
TS	Soil temperature TS1 shallowest; TS6 deepest (some sites only)	deg C
VPD	Vapor pressure deficit	hPa
USTAR	Friction velocity	m s^{-1}
WS	Wind speed	m s^{-1}
Dependent variable for regression and machine learning models		
NEE	Net Ecosystem Exchange (weekly)	gC m^{-2} d^{-1}

Figure 1. Abbreviations and definitions of climatic, environmental, and atmospheric variables were recorded as part of FLUXNET2015 datasets in the AmeriFlux databasee [48]. Note that reliable multi-year recorded data is not available for all these variables at every site. There are data gaps resulting in some variables being omitted at some sites. Data pre-processing involving gap filling and filtering is part of the FLUXNET2015 processing pipeline.

4000 ha^{-1} [49]. The location experiences a warm summer continental climate with a cold winter and substantial precipitation throughout the year; classified as Dfb (Koppen). The mean annual temperature is 6.7 °C and average annual rainfall is 876 mm but that varies substantially from year to year. FLUXNET data has been collected from 1994 to the present day at the Environment Canada facility [50]. More continuous recorded data for twenty potentially influential variables is available for the period from April 2004 to December 2020, and the weekly-averaged data for that period is compiled for *NEE* analysis in this study.

MX-Tes (27°50'41"N, 109°17'52"W) is a dry tropical woodland extending over about 15 ha just east of Rosario de Tesopaco (SE Sonora, Mexico). The recording site is located in almost flat land 460 m above sea level on the western flanks of the Madre Occidental mountains. Up to 80% of annual rainfall tends to occur in the July-September period related to the North American Monsoon but the level of rainfall fluctuates substantially from year to year [30]. It experiences a steppe climate with cold but dry winters and is classified as Bsh (Koppen). The average temperature at the site is 24.3 °C and average annual rainfall is 647 mm. Broadleaf trees reaching > 2 m height cover more than sixty percent of the site with individual trees reaching 10 m. The species present are dominated by *Lysiloma divaricatum* (tepemesquite), *Ipomoea arborescens* (tree morning glory), *Acacia cochliacantha* (boatspine acacia), *Haematoxylum brasiletto* (Mexican logwood), and *Celtis reticulata* (netleaf hackberry). FLUXNET data is only reported from 2004 to 2009 [51], and sixteen potentially influential variables have been compiled from that available data for *NEE* analysis in this study.

US-MMS (39°19'24"N, 86°24'47"W) is a mixed deciduous broadleaf forest located in the Morgan Monroe State Forest (south-central Indiana, U.S.A) at 275 m above sea level displaying ridge and ravine topography. The forest is comprised of about 29 tree species dominated by *Acer saccharum* (sugar maple), *Liriodendron tulipifera* (tulip poplar), *Sassafras albidum* (sassafras), *Quercus alba* (white oak) and *Q. velutina* (black oak), *Lindera benzoin* (spicebush), *Asimina triloba* (pawpaw) and *Cornus florida* (flowering dogwood) with a summer understory of diverse vascular plants contributing substantially to forest litter [52]. The forest has been grown and managed on abandoned farmland since 1929. It experiences a humid, subtropical, mild climate with hot summers and no dry seasons; classified as Cfa (Kloppen). The average mean temperature is 10.9 °C and mean annual rainfall is 1032 mm. FLUXNET data has been collected from this site from 1999 to the present [53] and eighteen potentially influential variables have been compiled from that available data for *NEE* analysis in this study.

US-MOz (38°44'39"N, 92°12'0"W) is a secondary oak-hickory, predominantly broadleaf forest located in the Baskett Wildlife Research and Education Area of the Ozark border region of central Missouri (U.S.A) at 219 m above sea level. Its dominant tree species are *Quercus alba* (white oak), *Q. velutina* (black oak), *Carya ovata* (shagbark), *Acer saccharum* (sugar maple) and *Juniperus virginiana* (eastern red cedar) [54]. The forest has been grown and managed on overgrazed and bankrupt farmland since the 1930s. It experiences a humid, subtropical, mild climate with hot summers and no dry seasons; classified as Cfa (Kloppen). The average mean temperature is 12.1 °C and mean annual rainfall is 986 mm. The forest exists on relatively thin soils which can lead to sever plant-water stress during periods of drought [55]. FLUXNET data has been collected from this site from 2004 to the present [56]. Twenty potentially influential variables have been compiled from that available data (2004 to 2019) for *NEE* analysis in this study.

3. Predicting *NEE* from Recorded Environmental Variables

3.1 Multi-linear Regression (MLR)

MLR techniques expand upon Ordinary Linear Regression (OLR), by applying error minimization algorithms and/or various error penalty functions [57].

MLR models assign values to coefficients C_0 to C_N, (Equation (4)) for each of the *N* independent variables, with C_0, representing a constant term.

$$Y = C_0 + C_1X_1 + C_2X_2 + C_3X_3 \tag{4}$$

where X_1 to X_N are the independent variables and Y is the dependent variable. When applying Equation (4), the models typically assume that no dependencies exist among the influencing variables. In this study five distinct MLR algorithms are applied to predict the *NEE* datasets involving different error functions and in some cases minimizers. These are:

(1) **LR**: applies a basic function for least-squares error determination [58] expressed as Equation (5):

$$J(C_0 + C_1 + C_2 \ldots C_N) = \frac{1}{2m}\sum\left(Y_a^i - Y_p^i\right)^2 \tag{5}$$

where Y_a^i = actual dependent variable value for the i^{th} data record;

Y_p^i = predicted dependent variables value;

m = number of data records;

J = error value to be minimized. A coordinate descent algorithm is applied to minimize J over a large number of iterations.

(2) **Ridge**: adds a penalty term ($\lambda\sum_{j=1}^{N} C_j^2$) to the least-squares function (Equation (6)) in which λ > 0.

$$min_C\ for\ \sum_{i=1}^{m}\left(Y_i - \sum_{j=1}^{N} X_{ij}C_j\right)^2 + \lambda\sum_{j=1}^{N} C_j^2 \qquad (6)$$

This penalty acts to prevent solutions with excessively high coefficient values from being acceptable. With the LR algorithm high coefficient values tend to arise for datasets where some dependencies exist among the independent variables. The Ridge penalty term is an *L2*-regularization function with an *L1* ratio = 0. Ridge can be configured to apply various minimizers.

(3) **LASSO** (least absolute shrinkage and selection operator): applies an absolute penalty function ($\lambda\sum_{j=1}^{N}\left\|C_j\right\|$) (Equation (7)) rather than the squared penalty function used by a ridge.

$$min_C\ for\ \sum_{i=1}^{m}\left(Y_i - \sum_{j=1}^{N} X_{ij}C_j\right)^2 + \lambda\sum_{j=1}^{N}\left\|C_j\right\| \qquad (7)$$

The LASSO penalty term acts to favor solutions with the least number of non-zero C_j values. It does so by eliminating some of the independent variables with the least impact on the dependent-variable predictions. The LASSO penalty term is an *L1*-regularization function with an *L1 ratio* = 1. Note if *λ* = 0 in either Ridge or LASSO the least error term reverts to that of LR. LASSO regression is usually configured to apply a coordinate descent minimizer.

(4) **ElasticNet**: applies a penalty term that involves both *L1*- and *L2*- regularization functions with their relative contributions to the penalty controlled by the α term in Equation (8).

$$min_C\ for\ \frac{\sum_{i=1}^{m}\left(Y_i - \sum_{j=1}^{N} X_{ij}C_j\right)^2}{2N} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{N} C_j^2 + \alpha\sum_{j=1}^{N}\left\|C_j\right\|\right) \quad (8)$$

The α term can be adjusted to provide more flexibility to the penalty term and overcome limitation of Ridge and LASSO. If α = 0 the penalty reverts to an *L2* function (Ridge); whereas if α = 1 it reverts to an *L1* function (LASSO). Therefore, to be effective in a distinctive way from Ridge or LASSO, ElasticNet is usually configured with 0 < α < 1.

(5) **SGDR (Stochastic Gradient Descent Regression):** applies a partial differential of *J* [59] for coefficients C_0 to C_N as expressed in Equation (9).

$$C_0^{k+1} = C_0^k - \alpha\frac{d}{dC_0^k}J\left(C_0^k\right),$$
$$C_1^{k+1} = C_1^k - \alpha\frac{d}{dC_1^k}J\left(C_1^k\right) \ldots\ C_N^{k+1} = C_N^k - \alpha\frac{d}{dC_N^k}J\left(C_N^k\right) \qquad (9)$$

where *k* = iteration number of the gradient descent optimizer and *α* = learning rate.

Differentiation of Equation (9) derives the C_0 (Equation (10)) and C_1 to C_N (Equation 11)) values for the next iteration.

$$C_0^{k+1} = C_0^k - \alpha\frac{1}{m}\sum\left(Y_a^i - Y_p^i\right) \qquad (10)$$

$$C_1^{k+1} = C_1^k - \alpha\frac{1}{m}\sum\left(Y_a^i - Y_p^i\right)X_1^i\ \ldots.$$
$$C_N^{k+1} = C_N^k - \alpha\frac{1}{m}\sum\left(Y_a^i - Y_p^i\right)X_N^i \qquad (11)$$

α is tuned to suit the dataset, with higher α values generating larger changes in the values of C_0 to C_N from one iteration to the next. Various penalty functions can be applied with SGDR (i.e., *L1*, *L2* regularization terms).

3.2 Machine Learning Models Evaluated

A diverse suite of widely used ML models is applied in attempts to determine accurate *NEE* predictions from the influencing variables available for the four DBF sites considered (Table 1). The models are run in customized Python code applying publicly available SciKit Learn functions [60]. These mathematical formulations on which the selected ML models are based are well established and are available in the citations provided in Table 1, together with examples of their more recent applications in ecological research.

The ML (and most of the MLR) models require tuning their hyperparameters to suit the data variable distributions and relationships of each DBF dataset. The model structures and optimized tunable control values applied to

Table 1. ML models adopted for NEE prediction of four DBF site weekly-averaged datasets.

Machine Learning Models Applied to Predict NEE at DBF Sites With Influencing Variables				
Model	Code	Type	Originator(s)	Examples of Use in Ecology Studies
Adaptive Boosting	ADA	Boosted Tree ensemble	[61]	[62]
Decision Tree	DT	Single tree	[63]	[64]
K-Nearest Neighbor	KNN	Data Matching	[65]	[66]
Multi-Layer Perceptron	MLP	Artificial Neural Network	[67]	[68,69]
Random Forest	RF	Tree ensemble	[70]	[37, 71-72]
Support vector Regressor	SVR	Hyperplane Fit	[73]	[74-75]
Extreme Gradient Boosting	XGB	Boosted Tree ensemble	[76]	[77-78]

each model are described in Table 2. The optimized values were established using a combination of trial and error, grid search [79] and Bayesian optimizers [80].

Determining the appropriate number of data records from the datasets to assign to training and validation subsets is an important requirement for MLR and ML model evaluation. This is typically defined as the split percentages (A% training: B% validation; where A+B=100%) or "splits". The splits applied need to allocate sufficient data records to training such that the models adequately fit the full range of data present, and, at the same time, allocate sufficient data records to validation to verify that the trained models can generate reproducible and reliable predictions based on random allocations.

3.3 Multi-fold Cross Validation Analysis

Evaluating data subset splits for MLR/ML models can be conducted by trial and error but it is time consuming and tends to lack statistical rigor. The K-fold cross-validation technique offers a powerful and statistically robust tool for evaluating various percentage splits between training and validation subsets [81], particularly when applied with multiple values of K. The technique divides, at random, a dataset into K subsets of equivalent size. The value of K can vary, but for most datasets it is worthwhile applying the techniques with 4 <= K <= 15. For small datasets (less than about 200 data records) the larger K values in that range tend to allocate too few data records (less than about 20) to each validation subset, introducing more statistical dispersion into the results. Each K-fold analysis sequentially assigns one of the K subsets for validation and the remaining K-1 subsets for training. The sequence is continued until each of the K subsets is evaluated once as the validation subset. For a 4-fold evaluation that involves four separate executions, for a 15-fold evaluation

Table 2. MLR and ML Model architectures and control parameters applied.

NEE Prediction Models Applied	**Hyperparameter Values Applied**
Regression Models	
Ordinary Least Squares Regression (LR)	Fit Intercept = true
LASSO	Alpha= 0.0001 (0.1 for CA-Cbo); coordinate descent optimization; tolerance =0.0001; L1 regularization; L1 ratio =1.0; Fit intercept = true
ElasticNet	Alpha= 0.0001; optimization solver = auto; tolerance =0.0001; L1 ratio =0.01; Fit intercept = true
Ridge	Alpha= 0.1 (1.0 for CA-Cbo); optimization solver = auto; tolerance =0.001; L2 regularization (L1 ratio =0.0); Fit intercept = true
Stochastic Gradient Descent (SGDR)	Alpha= 0.0001; Loss function = episilon insensitive; learning rate = invscaling; L2 regularization; L1 ratio = 0.15; Fit intercept = true
Machine Learning Models	
Adaptive Boosting (ADA)	Number of estimators=1000; learning rate =0.01; loss function = linear base estimator is DT with depth =15; splitter =best
Decision Tree (DT)	Maximum depth = None; splitter =best; splitting criterion = mse
K Nearest Neighbour (KNN)	Number of nearest neighbours assessed K = 8 to 11 (depending on the site); distance metric = Minkowski with p = 2 (Euclidian) for US-Moz and MX-Tes) and p=1(Manhattan) for US-MMS and CA-Cbo; neighbour selection algorithm = auto

Multi-layer Perceptron (MLP)	3 hidden layers with 100, 50 and 25 neurons; Learning rate = adaptive
	with initial learning rate = 0.001; Solver = adam; alpha=0.01
	(=1.0 MX-Tes); activation fn. = tanh for US-MMS and MX-Tes;
	activation fn. = relu for US-MOz and CA-Cbo;

Random Forest (RF)	Number of estimators = 1000; maximum depth = 100; Splitting criterion = mse

Support Vector Regressor (SVR)	Kernel = rbf;
	CA-Cbo: C = 7; gamma = 0.22; MX-Tes: C = 0.3; gamma = 0.1;
	US-MZo: C = 21; gamma = 0.22; US-MMS: C = 12; gamma = 0.23;

Extreme Gradient Boosting (XGB)	Number of estimators=1500; Maximum depth = 5 (=4 MX-Tes); eta = 0.01; Subsample = 0.6 (=0.7 MX-Tes); Columns sampled per tree = 0.8

that involves fifteen separate executions. The error analysis is then compiled to establish the mean and standard deviation of the error metric used (in this study MAE). As the initial k-fold division into subsets is random, it makes sense to repeat the evaluation for each K-fold in several separate runs to improve statistical confidence in the results. For instance, repeating a 10-fold cross validation analysis three times, with three distinct initial random splits of the datasets into ten subsets, result in thirty distinct cases, leading to a more robust mean and standard deviation.

Multi-K-fold cross validation analysis is a powerful tool. By comparing the means and standard deviations of the multi-K-fold results the appropriate splits of data records between ML training and validation subsets can be more precisely determined. When this approach is used to select the optimum data splits between training and validation subsets, it tends to substantially improve the prediction accuracy of the MLR/ML models applied and minimizes the impacts of data over-fitting. The multi-K-fold cross validation analysis applied to the four DBF datasets is presented and interpreted in Section 4.2.

3.4 Filtering FLUXNET2015 Data Records

FLUXNET data records over time typically include some data gaps, due either to equipment/recording failures or unreliable data being recorded. The FLUXNET2015 processing pipeline [46] is designed to identify these and gap fill the data, if possible. In this study, for the four weekly-averaged datasets compiled (CA-Cbo, MX-Tes, US-MMS and US-MOz) any data record with missing values for the influencing data variables selected has been omitted.

Prior to MLR and ML analysis, the values of each data variable are normalized (Equation (12)) within each dataset to a range of –1 to +1. Normalization avoids the impacts of variable scaling biases.

$$X_i{*}= 2{*}[(X_i\text{- } Xmin)/(Xmax\text{-}Xmin)]\text{-}1 \qquad (12)$$

where X_i is the i^{th} data point within the distribution of variable *X* values and *Xmin*, *Xmax* and

X_i* are the minimum, maximum and normalized values relating to that variable distribution.

The error performance measures employed to assess the performance of MLR and ML models in this study are mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), correlation coefficient (R) and correlation coefficient squared (R^2). These terms are defined in Appendix A. The work-flow sequence recommended and adopted for high-level characterization and analysis of multi-year, weekly-averaged *NEE* FLUXNET2015 datasets, recorded in conjunction with potentially influential environmental variables is summarized in Figure 2. It combines correlation, MLR, ML and pattern analysis, and applied systematically, it provides a useful approach for benchmarking and comparing sites at different locations involving related ecosystems.

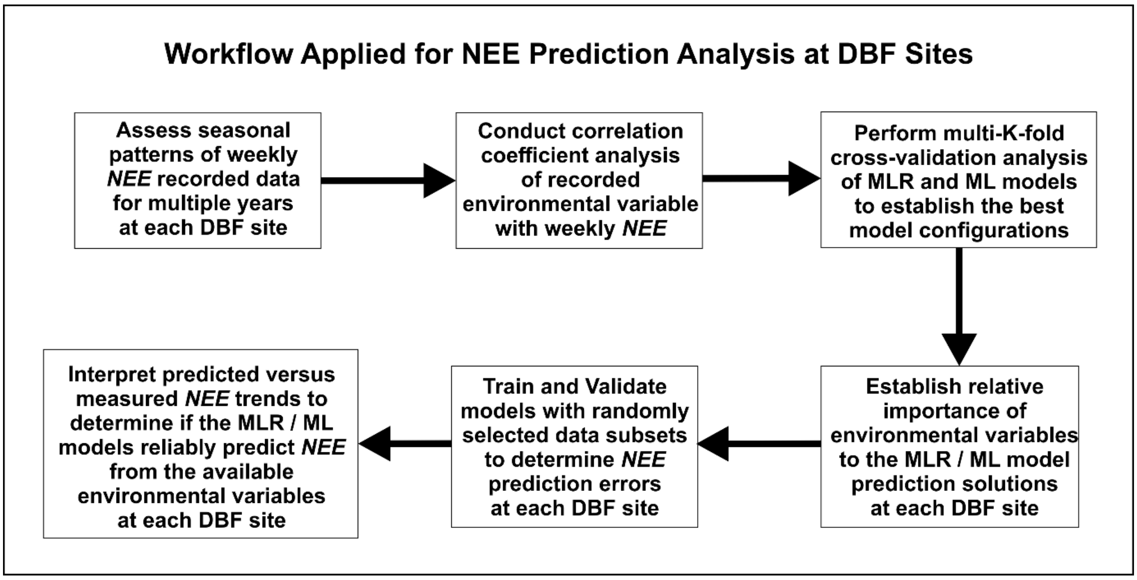

Workflow Applied for NEE Prediction Analysis at DBF Sites
Assess seasonal patterns of weekly NEE recorded data for multiple years at each DBF site
Conduct correlation coefficient analysis of recorded environmental variable with weekly NEE
Perform multi-K-fold cross-validation analysis of MLR and ML models to establish the best model configurations
Establish relative importance of environmental variables to the MLR / ML model prediction solutions at each DBF site
Train and Validate models with randomly selected data subsets to determine NEE prediction errors at each DBF site
Interpret predicted versus measured NEE trends to determine if the MLR / ML models reliably predict NEE from the available environmental variables at each DBF site

Figure 2. Work flow description of multi-year, weekly *NEE* pattern analysis proposed and executed in this study of four DBF sites.

4. Results

4.1 *NEE* Weekly Patterns and Variable Correlations

Repetitive seasonal variations are apparent at each of the four sites when weekly-averaged *NEE* data from each site are considered over multiple years. The *NEE* annual ranges are more extreme but the seasonal patterns are more regular for sites US-MMS and US-MOz (Figure 3A) located within the same climatic zone. On the other hand, the magnitudes of the *NEE* weekly-averaged peaks and troughs vary substantially from year to year at the CA-Cbo site (Figure 3B) with some abrupt alternations between peaks and troughs within the summer season. In contrast, the *NEE* range displayed at the MX-Tes site is quite narrow with few weekly data points falling outside the –0.5 to +0.5 $gCm^{-2}d^{-1}$ range, however, its seasonal *NEE* pattern is similar to that shown in Figure 3B for the CA-Cbo site.

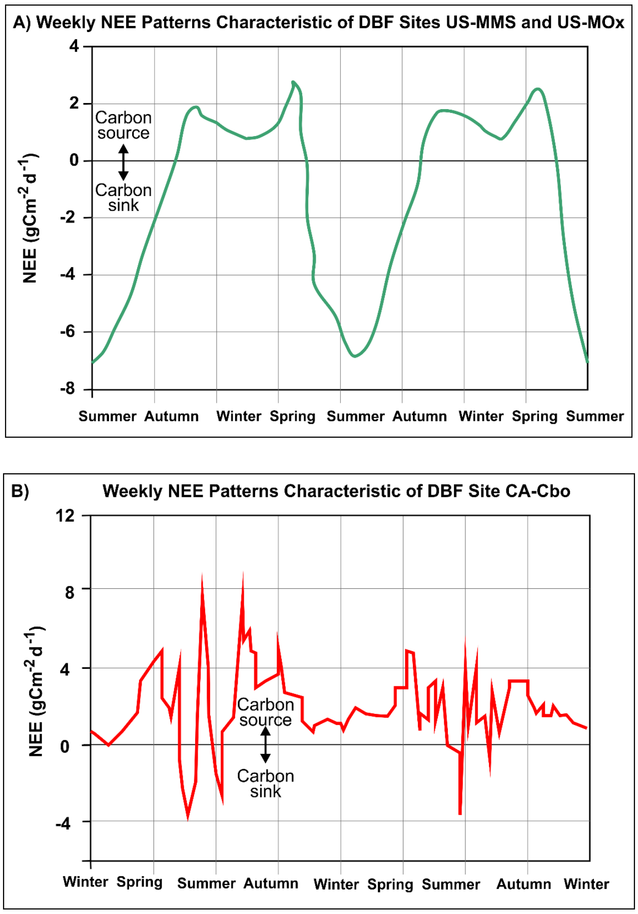


Figure 3. Characteristic seasonal patterns of net ecosystem exchange (*NEE*) weekly data sequences for: a) AmeriFlux sites and US-MMS; and (B) CA-Cbo. Note that the pattern for site MX-Tes is similar to that of CACbo but on a more compressed *NEE* scale of =1 to +0.25 $gCm^{-2}d^{-1}$ and with the majority of the weeks recording values below 0.

Pearson (R) [82] and Spearman (p) [83] correlation coefficients between influencing variables and multi-year *NEE* weekly patterns can be usefully compared to benchmark variable relationships in different ecosystems [34,35]. Both correlation coefficients are calculated using the same formula (Appendix A): R using actual distribution values; p using their rank positions. Figure 4 displays such a comparison for the four sites of interest revealing distinctive *NEE*-influencing variable relationships. For sites US-MMS and US-MOz (Figures 4A and 4B) many of the recorded variables show high negative correlation coefficients (< –0.5) with *NEE*, with variables WS, USTAR, CO_2 and SWC showing moderate positive correlations (between +0.2 and +0.5). Moreover, there is generally good agreement between R and p values at both sites, implying that the linear parametric variable distribution relationships assumed in the R calculations [84] are a reasonable approximation for the distributions recorded at those sites [85].

Because the p calculation involves the rank positions of the data variable values in its distribution range, making it more suitable for assessing both parametric and non-parametric relationships [86]. The p values are more suitable than the R values in quantifying relationships between variable distributions that involve a degree of non-linearity [87,88]. Most variables recorded at sites US-MMS and US-MOz display slightly lower p than R values, although for CO_2 the reverse is the case (Figures 4A and 4B). Such minor differences imply that a minor degree of non-linearity participates in the relationships between most variable distributions and *NEE.*

In contrast, the R and p values for site CA-Cbo (Figure 4C) are both low with only SWC just reaching values of –0.2. Moreover, for several variables (VPD, PPFDout, SWC, LE and H) there are substantial differences between the calculated p and R values for this site. These relationships imply the *NEE* distribution cannot be easily or well described in terms of the potentially "influential" variables measured at that site, and that several of the poor relationships that do exist are substantially non-linear.

The R and p values for site MX-Tes (Figure 4D) are between –0.5 and +0.4, thereby falling about midway between sites US-MMS and US-MOz on the one hand, and site CA-Cbo on the other. However, the p and R values for most of the variables at site MX-Tes are quite different, highlighting that their relationships with *NEE* are non-parametric with a strong degree of non-linearity. Taken together, the plots displayed in Figures 3 and 4 generally provide high-level characterization of *NEE* recording sites considering multi-year, weekly-averaged data.

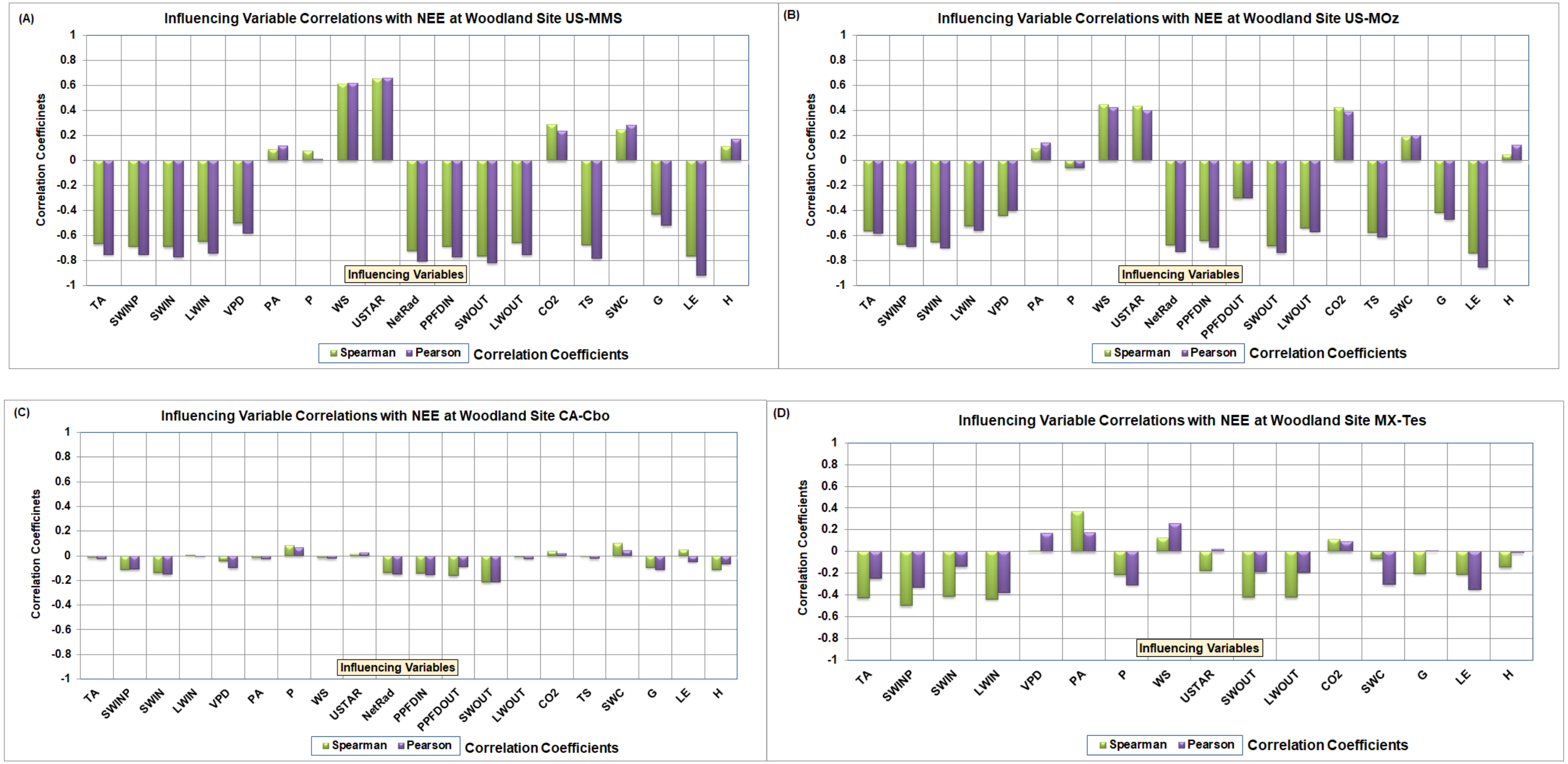


Figure 4. Correlations of influential variables with *NEE* based on weekly datasets for woodland sites: (A) US-MMS; (B) US-MOz; (C) CA-Cbo; and (D) MX-Tes. Variable abbreviations are defined in Figure 1.

4.2 *NEE* Prediction Models

Five MLR models and seven ML models were carefully configured and applied to the compiled datasets for each of the four DBL woodland sites considered. Prior to describing and comparing those model results in detail, graphics are displayed for predicted *NEE* versus recorded *NEE,* including each weekly-averaged data record in the multi-year sequence, for the best performing MLR and ML models applied to each site. Figure 5 compares those results for sites US-MMS and US-MOz, whereas Figure 6 compares those results for sites CA-Cbo and MX-Tes. It is apparent from Figure 5 that for sites US-MMS and US-MOz the models generate similar results, with site US-MOz achieving slightly low prediction errors. The Ridge MLR models (Figures 5A and 5C) provide credible predictions with predicted versus recorded *NEE* values approximating Y=X patterns, albeit with a degree of scatter. However, the SVR models (Figures 5B and 5D) noticeably improve upon the Ridge model results, following Y=X patterns (R^2 > 0.95)with less dispersion (MAE = 0.3463 $gCm^{-2}d^{-1}$ for US-MMS; MAE = 0.2343 $gCm^{-2}d^{-1}$ for US-MMS).

On the other hand, the *NEE*-prediction model performances are less impressive for sites CA-Cbo and MX-Tes (Figure 6). What particularly stands out, for both sites, is that the predicted versus recorded *NEE* patterns for both MLR and ML models deviate substantially from Y = X patterns. Moreover, they do so in a significantly systematic way, with *NEE* predictions overestimating the lowest (most negative) values and underestimating the highest (most positive) values. This leads to relatively high MAE and RMSE values in relation to the *NEE* range recorded and low R^2 values, despite clear linear patterns between predicted and recorded *NEE* values being established. In fact, the majority of the data points recorded at both sites are predicted with reasonable accuracy (i.e., those situated between recorded *NEE* of –1 and +3 $gCm^{-2}d^{-1}$ for site CA-Cbo, and those situated between recorded *NEE* of –0.6 and 0 $gCm^{-2}d^{-1}$ for site MX-Tes). The prediction problem, using the influential variables recorded is associated with the *NEE* peaks and troughs for these two sites. Although the ML models improve prediction performance slightly compared to the MLR models they do not resolve the Y ≠ X pattern issue.

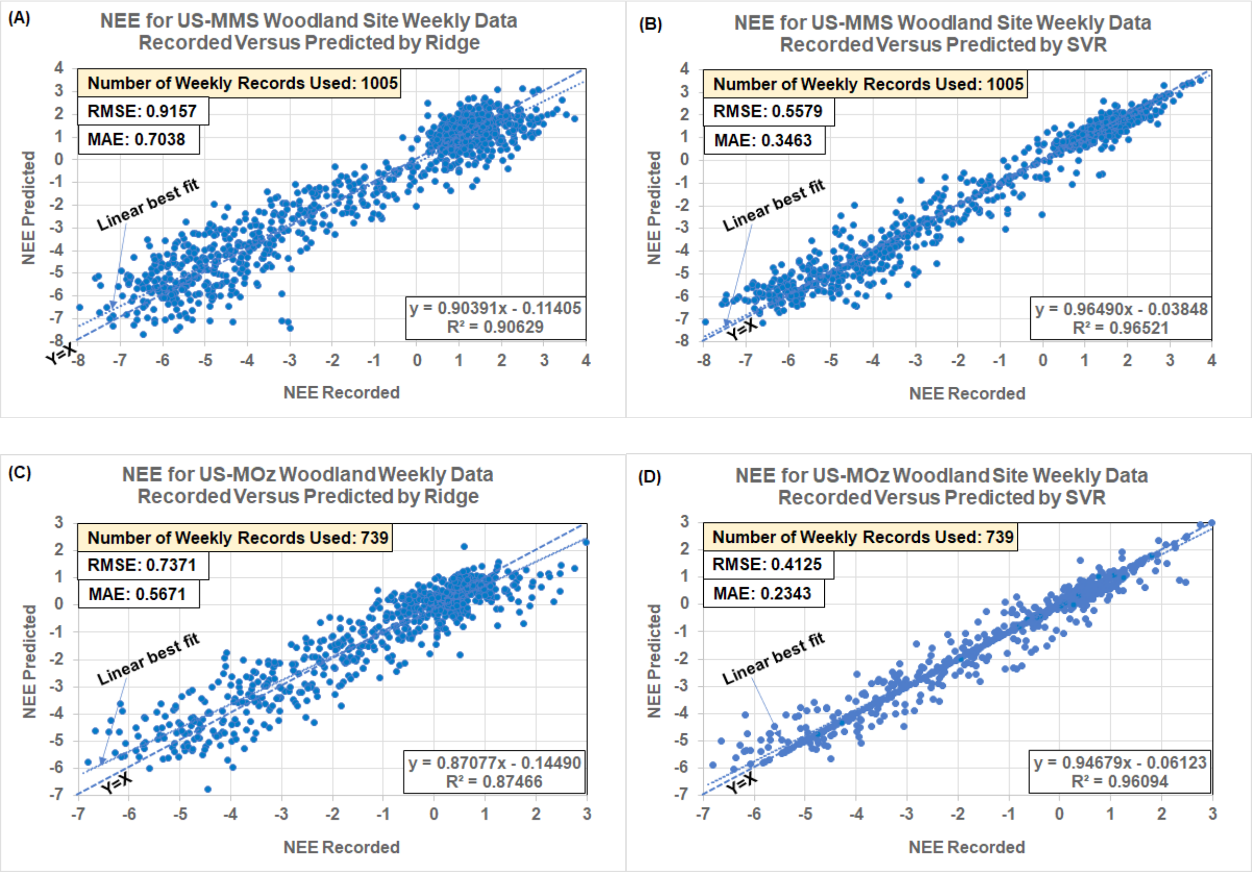


Figure 5. Predicted versus recorded *NEE* for: (A) US-MMS Ridge Solution; (B) US-MMS SVR Solution; (C) US-MOz Ridge Solution; and (D) US-MOz SVR Solution. *NEE* values displayed in $gCm^{-2}d^{-1}$.

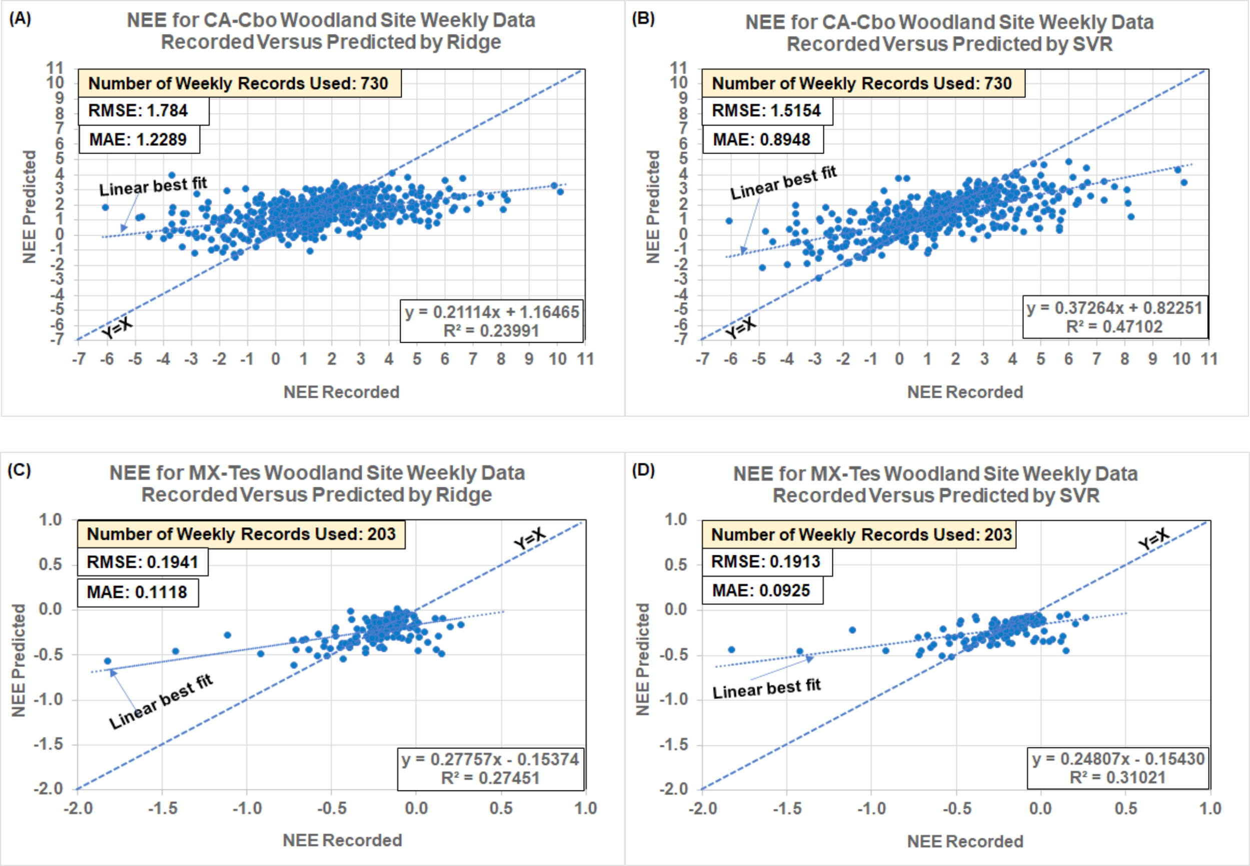


Figure 6. Predicted versus recorded NEE for: (A) CA-Cbo Ridge Solution; (B) CA-Cbo SVR Solution; (C) MX-Tes Ridge Solution; and (D) MX-Tes SVR Solution. NEE values displayed in gCm^{-2}d^{-1}.

4.3 Multiple-K-fold Cross Validation of MLR and ML Models

K-fold cross-validation analysis (Section 3.3) offers a highly effective technique for establishing the reliability and reproducibility of the *NEE* prediction models (MLR and ML) by systematically and randomly running multiple cases using different training/validation splits of each dataset. Additionally, it helps to guard against being misled by the tendency of some models to overfit to varying degrees the training subsets used to fit the models. The K-fold analysis results relating to each of the sites considered, including 4-, 5-, 10- and 15-fold configurations, are listed in Tables 3 to 6. The results of the 10-fold cross-validation configuration are also displayed graphically in Figure 7. This technique is a reliable way of determining which models generate the most reliable *NEE* predictions at each DBF site with minimum prediction errors.

It is apparent from Tables 3 to 6 that the SVR model generates the best *NEE* prediction accuracy for all sites, although its performance is matched by the MLP model for site US-MOz. Taking into account the mean MAE values and their standard deviations the 10-fold and 15-fold configurations provide the most accurate results with the least dispersion for the models applied to each site. This indicates that the models work best with at least 90% of the data records allocated to the training subset.

With respect to the MLR models applied to datasets from sites US-MMS and US-MOz (Tables 3 and 4), four models generate almost identical prediction models for each K-fold configuration. The SGDR model generates inferior predictions for those sites. For site CA-Cbo (Table 5), the LASSO model slightly outperforms the Ridge model providing the best MLR *NEE* predictions, whereas the SGDR model provides the poorest *NEE* predictions. However, for the MX-Tes site the SGDR model provides the best MLR *NEE* predictions (Table 6).

For the ML models, the SVR, XGB, MLP and RF models generate better *NEE* prediction accuracy than the ADA, KNN and DT models for sites US-MMS, US-MOz and CA-Cbo (Tables 3 to 5). However, for site MX-Tes the ADA and KNN models provide comparable *NEE* predictions to the SVR, XGB, MLP and RF models (Table 6). The DT model provides the worst *NEE* prediction results and is outperformed by all MLR models as well as the other ML models.

Table 3. Cross-validation (4-, 5-, 10-, and 15-fold) of *NEE* prediction-error analysis for woodland site US-MMS in terms of mean absolute error (MAE) displaying results for 5 MLR models and 7 ML models. Best-performing SVR solution is shown in bold.

K-Fold Cross Validation NEE Prediction Errors (MAE) for Woodland Site US-MMS								
Mean Absolute Error (MAE)	4-Fold		5-Fold		10-Fold		15-Fold	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Regression								
LR	0.7186	0.0304	0.7196	0.0355	0.7180	0.0547	0.7185	0.0629
LASSO	0.7184	0.0305	0.7194	0.0356	0.7179	0.0548	0.7183	0.0630
RIDGE	0.7181	0.0305	0.7190	0.0357	0.7177	0.0548	0.7181	0.0630
ELASTICNET	0.7182	0.0305	0.7191	0.0357	0.7177	0.0548	0.7181	0.0630
SGDR	0.7291	0.0341	0.7307	0.0385	0.7274	0.0546	0.7266	0.0645
Machine Learning								
ADA	0.5457	0.0279	0.5443	0.0316	0.5374	0.0439	0.5354	0.0568
DT	0.7757	0.0503	0.7891	0.0413	0.8017	0.0888	0.7933	0.0913
KNN	0.5476	0.0184	0.5428	0.0229	0.5401	0.0365	0.5367	0.0514
MLP	0.5127	0.0264	0.5414	0.0391	0.5008	0.0483	0.4972	0.0557
RF	0.5451	0.0276	0.5476	0.0328	0.5403	0.0485	0.5380	0.0615
SVR	0.4982	0.0161	0.4962	0.0205	0.4894	0.0400	0.4856	0.0504
XGB	0.5119	0.0187	0.5092	0.0233	0.5020	0.0405	0.5013	0.0560

Table 4. Cross-validation (4-, 5-, 10-, and 15-fold) of *NEE* prediction-error analysis for woodland site US-MOz in terms of mean absolute error (MAE), displaying results for 5 MLR models and 7 ML models. Best-performing MLP and SVR solutions are shown in bold.

K-Fold Cross Validation NEE Prediction Errors (MAE) for Woodland Site US-MOz								
Mean Absolute Error (MAE)	4-Fold		5-Fold		10-Fold		15-Fold	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Regression								
LR	0.5858	0.0256	0.5861	0.0421	0.5839	0.0545	0.5842	0.0692
LASSO	0.5854	0.0255	0.5859	0.0422	0.5836	0.0544	0.5839	0.0688
ElasticNet	0.5847	0.0253	0.5852	0.0423	0.5831	0.0541	0.5833	0.0684
RIDGE	0.5841	0.0252	0.5848	0.0424	0.5829	0.0539	0.5831	0.0682
SGDR	0.6280	0.0287	0.6286	0.0424	0.6260	0.0535	0.6252	0.0704
Machine Learning								
ADA	0.5368	0.0312	0.5374	0.0304	0.5207	0.0470	0.5169	0.0561
DT	0.7310	0.0542	0.7621	0.0554	0.7061	0.0725	0.7214	0.0863
KNN	0.5662	0.0260	0.5630	0.0356	0.5547	0.0614	0.5510	0.0714
MLP	0.4955	0.0307	0.4931	0.0328	0.4744	0.0469	0.4698	0.0491
RF	0.5313	0.0292	0.5342	0.0357	0.5182	0.0499	0.5146	0.0576
SVR	0.4879	0.0246	0.4883	0.0310	0.4745	0.0441	0.4692	0.0520
XGB	0.4907	0.0199	0.4899	0.0287	0.4825	0.0490	0.4756	0.0514

Table 5. Cross-validation (4-, 5-, 10-, and 15-fold) of *NEE* prediction-error analysis for woodland site CA-Cbo in terms of mean absolute error (MAE), displaying results for 5 MLR models and 7 ML models. Best-performing SVR solution is shown in bold.

K-Fold Cross Validation NEE Prediction Errors (MAE) for Woodland Site CA-Cbo								
Mean Absolute Error (MAE)	4-Fold		5-Fold		10-Fold		15-Fold	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Regression								
LR	1.2723	0.0864	1.2715	0.1120	1.2674	0.1532	1.2673	0.2174
LASSO	1.2524	0.0993	1.2539	0.1165	1.2512	0.1570	1.2522	0.2177
ElasticNet	1.2706	0.0873	1.2698	0.1124	1.2659	0.1533	1.2658	0.2172
RIDGE	1.2597	0.0931	1.2605	0.1139	1.2576	0.1561	1.2585	0.2166
SGDR	1.3070	0.1095	1.3103	0.1196	1.3011	0.1573	1.3024	0.2197
Machine Learning								
ADA	1.1465	0.1033	1.1508	0.1284	1.1394	0.1695	1.1375	0.2231
DT	1.6126	0.1495	1.6411	0.1893	1.5927	0.2457	1.6179	0.3002
KNN	1.2509	0.0870	1.2538	0.1218	1.2373	0.1661	1.2388	0.2281
MLP	1.2525	0.1003	1.2567	0.1384	1.1994	0.1623	1.2252	0.2291
RF	1.1477	0.0979	1.1488	0.1281	1.1460	0.1617	1.1429	0.2127
SVR	1.1105	0.0758	1.1159	0.1127	1.1029	0.1496	1.1005	0.2061
XGB	1.1240	0.0928	1.1152	0.1148	1.1090	0.1599	1.1085	0.2094

Table 6. Cross-validation (4-, 5-, 10-, and 15-fold) *NEE* prediction-error analysis for woodland site MX-Tes in terms of mean absolute error (MAE), displaying results for 5 MLR models and 7 ML models. Best-performing SVR solution is shown in bold.

K-Fold Cross Validation NEE Prediction Errors (MAE) for Woodland Site MX-Tes								
Mean Absolute Error (MAE)	4-Fold		5-Fold		10-Fold		15-Fold	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Regression								
LR	0.1342	0.0215	0.1335	0.0194	0.1277	0.0323	0.1257	0.0436
LASSO	0.1229	0.0208	0.1236	0.0192	0.1192	0.0324	0.1184	0.0452
ElasticNet	0.1337	0.0214	0.1330	0.0193	0.1273	0.0323	0.1254	0.0436
RIDGE	0.1316	0.0210	0.1312	0.0192	0.1259	0.0324	0.1243	0.0437
SGDR	0.1146	0.0214	0.1165	0.0259	0.1159	0.0376	0.1124	0.0483
Machine Learning								
ADA	0.1176	0.0229	0.1189	0.0222	0.1170	0.0341	0.1158	0.0479
DT	0.1623	0.0213	0.1743	0.0309	0.1587	0.0506	0.1646	0.0641
KNN	0.1104	0.0234	0.1113	0.0241	0.1086	0.0337	0.1082	0.0464
MLP	0.1135	0.0212	0.1153	0.0219	0.1127	0.0340	0.1126	0.0468
RF	0.1161	0.0230	0.1169	0.0250	0.1148	0.0328	0.1142	0.0478
SVR	0.1036	0.0240	0.1044	0.0262	0.1038	0.0348	0.1038	0.0482
XGB	0.1233	0.0216	0.1228	0.0232	0.1214	0.0338	0.1213	0.0464

Figure 7 highlights the superiority of the SVR prediction model taking into account the datasets from all four sites considered. Note also that the hyperparameters applied to the SVR model need to be optimized for each of those datasets.

4.4 MLR/ML Model Training/Validation

Additional insight into the *NEE*-prediction model performances at each DBF site is provided by assessing training, validation and full dataset results for an individual, ***randomly*** selected case contributing to the multi-K-fold cross-validation analysis. Tables 7 to 10 display the results for Case X, one of the thirty cases contributing to the 10-fold cross validation analysis, applied to each of the four sites considered. Case X results therefore involve a split of data records comprising 90 percent assigned to the training subset and ten percent assigned to the validation subset. Each site generates distinct prediction results and prediction errors for Case X. These results are plotted separately for each site studied in Tables 7 to 10. Those tables also present the computational times taken to execute each MLR and ML model at each site.

The Case X results of the regression models for the four sites are consistent, with the *NEE* prediction performance for its validation subset being slightly better in terms of MAE, RMSE and R^2 values. Moreover, the *NEE* prediction performances for the full datasets are close to those of the training subset for the MLR models indicating that no overfitting has occurred. The MLR models all take between 4 and 6 seconds to execute for the four datasets assessed. The MLR models can therefore be executed rapidly and be relied upon to generate consistent results for each data subset assessed.

On the other hand, several of the ML models show clear evidence of overfitting with respect to Case X applied to each site (Tables 7 to 10). The ADA, DT, KNN, RF and XGB models fit the training subset with very low error values (MAE and RMSE close to or at zero, and R^2 values close to or at 1.0), whereas the error values associated with the validation subsets are much higher, and the errors achieved for the full datasets are substantially higher than for the training subsets. Such outcomes are indicative of overfitting. On the other hand, for the SVR models applied to Case X for each site, errors achieved for the training subset, the validation subset and the full dataset are in closer agreement. This is consistent with the SVR model not overfitting the dataset, and thereby generating consistently high *NEE* prediction accuracy for the validation subsets compared to the other ML models.

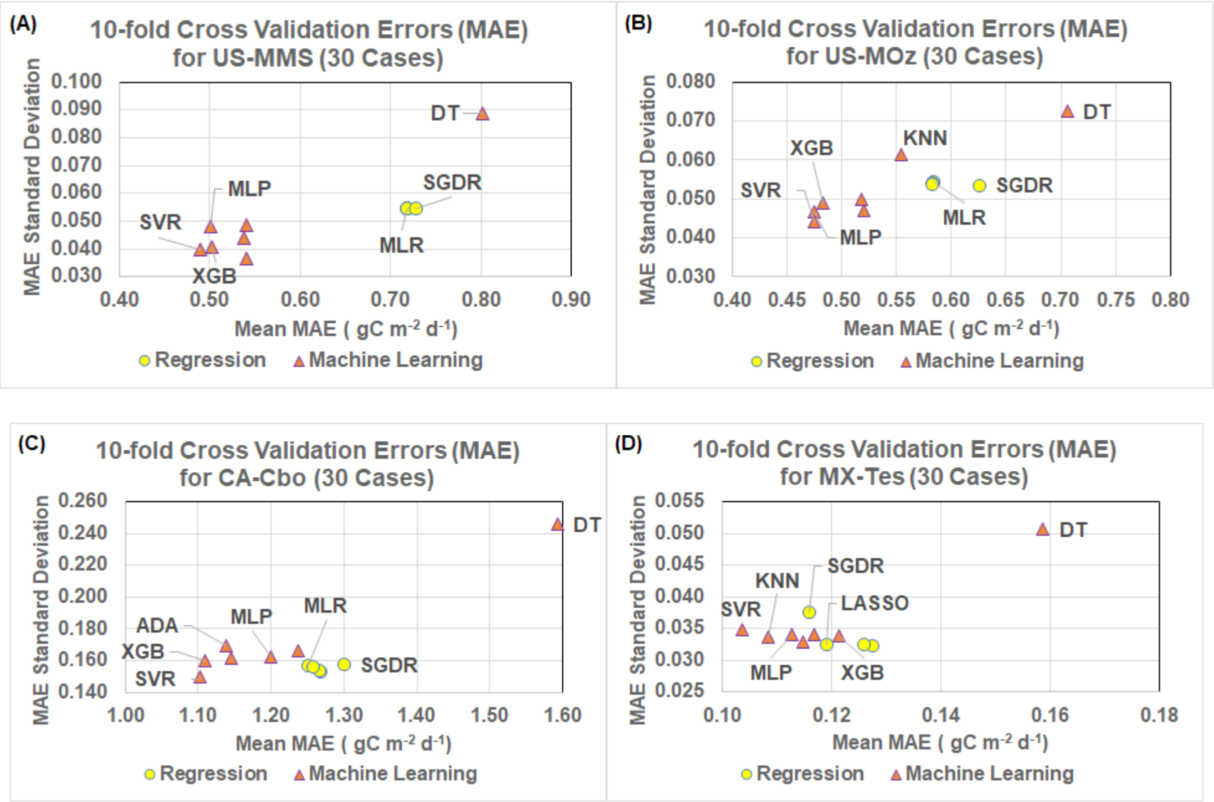


Figure 7. *NEE* weekly prediction results from cross-validation analysis (10-fold only displayed) applied to each MLR and ML model evaluated for sites (A) US-MMS; (B) US-MOz; (C) CA-Cbo: and (D) MX-Tes.

Table 7. Case X (one example of the thirty 10-fold cross-validation cases evaluated) results for training /validation/ full dataset *NEE* prediction performances for MLR/ML models applied to woodland site US-MMS. The best performing SVR solution is shown in bold.

NEE Forecasting Performance for Training and Validation Analysis Applied to the Full Dataset for Site US-MMS										
	Case X Training Subset (90% of Data Records)			Case X Validation Subset (10% of Data Records)			Case X Full Dataset (100% of Data Records)			
Model	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	Ex Time
Regression										
LR	0.9028	0.9272	0.7116	0.9323	0.8027	0.6336	0.9063	0.9155	0.7038	4.8
LASSO	0.9028	0.9272	0.7115	0.9322	0.8032	0.6339	0.9063	0.9155	0.7037	5.2
ElasticNet	0.9028	0.9272	0.7116	0.9320	0.8043	0.6347	0.9063	0.9156	0.7038	6.2
Ridge	0.9028	0.9273	0.7115	0.9320	0.8045	0.6348	0.9063	0.9157	0.7038	4.3
SGDR	0.8996	0.9425	0.7236	0.9256	0.8411	0.6586	0.9027	0.9328	0.7171	4.8
Machine Learning										
ADA	1.0000	0.0129	0.0031	0.9424	0.7403	0.5424	0.9938	0.2350	0.0573	221.0
DT	1.0000	0.0000	0.0000	0.8738	1.0957	0.8092	0.9865	0.3474	0.0813	4.8
KNN	1.0000	0.0000	0.0000	0.9431	0.7356	0.5561	0.9939	0.2332	0.0559	5.4
MLP	0.9806	0.4140	0.2989	0.9548	0.6557	0.5194	0.9779	0.4443	0.3211	250.0
RF	0.9907	0.2876	0.1975	0.9408	0.7502	0.5640	0.9854	0.3619	0.2344	174.6
SVR	0.9658	0.5501	0.3302	0.9591	0.6235	0.4911	0.9652	0.5579	0.3463	7.6
XGB	0.9976	0.1467	0.1111	0.9528	0.6700	0.5003	0.9928	0.2539	0.1502	74.9

Notes: (1) RMSE and MAE are expressed in terms of the measured weekly NEE range for the site: -8.37437 to 3.73218 gC m^{-2} d^{-1}. (2) Execution times (Ex Time) are expressed in seconds and include full 10-fold cross validation analysis.

Table 8. Case X (one example of the thirty 10-fold cross-validation cases evaluated) results for training/validation/full dataset *NEE* prediction performances for MLR/ML models applied to woodland site US-MOz. The best performing MLP solution is shown in bold.

NEE Forecasting Performance for Training and Validation Analysis Applied to the Full Dataset for Site US-MOz										
	Case X Training Subset (90% of Data Records)			Case X Validation Subset (10% of Data Records)			Case X Full Dataset (100% of Data Records)			
Model	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	Ex Time
Regression										
LR	0.8731	0.7437	0.5725	0.8885	0.6647	0.5194	0.8749	0.7362	0.5672	4.5
LASSO	0.8731	0.7438	0.5725	0.8883	0.6650	0.5189	0.8749	0.7362	0.5671	5.8
ElasticNet	0.8730	0.7440	0.5723	0.8876	0.6672	0.5189	0.8748	0.7367	0.5670	5.7
Ridge	0.8729	0.7443	0.5724	0.8871	0.6686	0.5194	0.8746	0.7371	0.5671	4.7
SGDR	0.8464	0.8182	0.6223	0.8373	0.8028	0.6152	0.8461	0.8167	0.6216	4.6
Machine Learning										
ADA	1.0000	0.0129	0.0031	0.8891	0.6627	0.4150	0.9898	0.2101	0.0443	267.4
DT	1.0000	0.0000	0.0000	0.8060	0.8766	0.6111	0.9822	0.2774	0.0612	4.7
KNN	1.0000	0.0000	0.0000	0.8711	0.7144	0.4527	0.9882	0.2261	0.0453	4.6
MLP	0.9844	0.2606	0.1868	0.9329	0.5153	0.4002	0.9798	0.2962	0.2082	65.1
RF	0.9833	0.2694	0.1955	0.8984	0.6343	0.3890	0.9756	0.3250	0.2149	207.3
SVR	0.9636	0.3984	0.2153	0.9310	0.5226	0.4057	0.9607	0.4125	0.2343	6.5
XGB	0.9979	0.0953	0.0739	0.9114	0.5924	0.3710	0.9900	0.2081	0.1037	65.7

Notes: (1) RMSE and MAE are expressed in terms of the measured weekly NEE range for the site: -7.21799 to 2.99522 gC m^{-2} d^{-1}. (2) Execution times (Ex Time) are expressed in seconds and include full 10-fold cross validation analysis.

The MLP model results for Case X show little evidence of overfitting for sites US-MMS (Table 7), US-MOz (Table 8) and MX-Tes (Table 10) for which the MLP model rivals the SVR model in terms of the *NEE* prediction accuracy it achieves. On the other hand, for site CA-Cbo (Table 9) the MLP model results show signs of overfitting, and it generates the poorest *NEE* prediction accuracy for the validation subset of Case X.

The Case X results (Tables 7 to 10) confirm the findings of the K-fold cross-validation analysis that the SVR and MLP models (except for site CA-Cbo in the case of the MLP model) generate better *NEE* prediction accuracy than the other MLR and ML models assessed for the four DBF woodland site datasets. The SVR models also rival the regression models in terms of their fast execution times, making it an accurate, fast and dependable model to apply to each of the weekly datasets considered.

4.5 Relative Influences of Recorded Variables on *NEE* Predictions

Insight into the relative influence of the measured environmental variables available for each woodland site can be obtained by comparing the regression coefficients of the MLR model solutions [34,35], the support vector coefficients of the SVR model solutions and the Gini coefficients of the DT and ensemble model (ADA, RF, XGB) solutions. Bar-chart analysis is presented in Figure 8 to compare variable influences on the 10-fold cross-validation solutions for sites US-MMS and US-MOz. Figures 8A and 8C display the variable influence comparisons based on regression coefficients for those sites. It is apparent that the poorer performing SGDR model assigns distinctive significance to the influencing variables compared to the other four regression models. Considering the LR, LASSO, ElasticNet and Ridge model results, in descending order of significance, those model solutions indicate that the most important influences for those sites are:

- **US-MMS**: H, SWINF, LE, LWINF, TA, VPDF, GF, PF, WSF, SWOUT and LWOUT
- **US-MOz**: SWOUT, PPFDIN, SWINF, TA, H, LWINF, VPDF, LWOUT, TS

On the other hand, the regression models make the least use of the following variables for the solutions they generate for those sites:

- **US-MMS**: PAF, USTAR and PPF
- **US-MOz**: PAF, WSF, PF, USTAR, NETRAD and CO_2

Figures 8B and 8D show quite distinctive variable importance assigned to these two sites by the ML models. The tree and ensemble models all assign overriding significance to LE (> 50% weight) and low significance to the other variables. In the case of the XGB model, it also assigns minor importance to TS for site US-MMS and SWINP for site US-MOz which helps it to provide more accurate *NEE* predictions for those sites than the other tree and ensemble models.

Table 9. Case X (one example of the thirty 10-fold cross-validation cases evaluated) results for training/validation/full dataset *NEE* prediction performances for MLR/ML models applied to woodland site CA-Cbo. The best performing SVR solution is shown in bold.

NEE Forecasting Performance for Training and Validation Analysis Applied to the Full Dataset for Site CA-Cbo										
	Case X Training Subset			Case X Validation Subset			Case X Full Dataset			
Model	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	Ex Time
Regression										
LR	0.2411	1.7762	1.2446	0.2516	1.7760	1.1212	0.2429	1.7762	1.2322	5.0
LASSO	0.2086	1.8139	1.2375	0.2103	1.8244	1.1173	0.2096	1.8149	1.2255	4.5
ElasticNet	0.2410	1.7763	1.2436	0.2490	1.7791	1.1222	0.2426	1.7766	1.2315	5.3
Ridge	0.2360	1.7822	1.2403	0.2307	1.8006	1.1261	0.2362	1.7840	1.2289	4.4
SGDR	0.1936	1.8309	1.2827	0.1898	1.8479	1.1761	0.1941	1.8326	1.2721	6.0
Machine Learning										
ADA	0.9993	0.0541	0.0150	0.3857	1.6090	0.9565	0.9372	0.5117	0.1093	267.0
DT	1.0000	0.0000	0.0000	0.1977	1.8389	1.1375	0.9187	0.5819	0.1139	5.1
KNN	1.0000	0.0000	0.0000	0.2721	1.7515	1.0513	0.9263	0.5543	0.1053	4.5
MLP	0.9383	0.5065	0.3577	0.1554	1.8867	1.1520	0.8591	0.7663	0.4372	29.7
RF	0.8916	0.6713	0.4342	0.3407	1.6669	0.9855	0.8359	0.8269	0.4894	56.4
SVR	0.4497	1.5126	0.8911	0.4370	1.5405	0.9280	0.4489	1.5154	0.8948	5.4
XGB	0.9880	0.2231	0.1693	0.3517	1.6530	1.0192	0.9236	0.5643	0.2545	43.6

Notes: (1) RMSE and MAE are expressed in terms of the measured weekly NEE range for the site: -6.05399 to 10.1207 gC m^{-2} d^{-1}. (2) Execution times (Ex Time) are expressed in seconds and include full 10-fold cross validation analysis.

Table 10. Case X (one example of the thirty 10-fold cross-validation cases evaluated) results for training/validation/full dataset *NEE* prediction performances for MLR/ML models applied to woodland site MX-Tes. The best performing MLP solution is shown in bold.

NEE Forecasting Performance for Training and Validation Analysis Applied to the Full Dataset for Site MX-Tes										
	Case X Training Subset			Case X Validation Subset			Case X Full Dataset			
Model	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	Ex Time
Regression										
LR	0.2796	0.1990	0.1147	0.0713	0.1457	0.0928	0.2732	0.1942	0.1124	4.4
LASSO	0.2662	0.2009	0.1101	0.1613	0.1384	0.0824	0.2662	0.1953	0.1072	4.5
ElasticNet	0.2795	0.1990	0.1146	0.0739	0.1455	0.0925	0.2733	0.1942	0.1123	4.4
Ridge	0.2793	0.1991	0.1141	0.0859	0.1445	0.0915	0.2737	0.1941	0.1118	4.4
SGDR	0.2008	0.2096	0.1121	0.0189	0.1497	0.0965	0.1961	0.2042	0.1105	4.3
Machine Learning										
ADA	1.0000	0.0008	0.0001	0.2181	0.1337	0.0788	0.9655	0.0430	0.0082	25.4
DT	1.0000	0.0001	0.0000	0.0538	0.2166	0.1392	0.9064	0.0697	0.0144	4.5
KNN	1.0000	0.0000	0.0000	0.1193	0.1419	0.0769	0.9599	0.0456	0.0080	4.5
MLP	0.2377	0.2047	0.1090	0.2570	0.1303	0.0782	0.2477	0.1983	0.1058	40.9
RF	0.8796	0.0814	0.0450	0.1567	0.1388	0.0805	0.8472	0.0890	0.0487	46.7
SVR	0.2980	0.1965	0.0940	0.1601	0.1385	0.0792	0.2948	0.1913	0.0925	4.4
XGB	0.9979	0.0108	0.0083	0.0858	0.1445	0.0807	0.9563	0.0476	0.0158	12.8

Notes: (1) RMSE and MAE are expressed in terms of the measured weekly NEE range for the site: -1.82451 to 0.422752 gC m^{-2} d^{-1}. (2) Execution times (Ex Time) are expressed in seconds and include full 10-fold cross validation.

The best performing SVR model assigns relative importance to the influencing variables that bear closer relationships to those of the MLR models than the tree/ensemble models. In descending order of significance, the SVR solutions indicate that the most important influences for those sites are:

- **US-MMS**: LE, H, SWINF, SWOUT, LWINF, TA, VPDF and PF
- **US-MOz**: SWOUT, TA, SWINF, LWINF, VPDF, LWOUT, PPFDIN, PPFDOUT, H and LWOUT

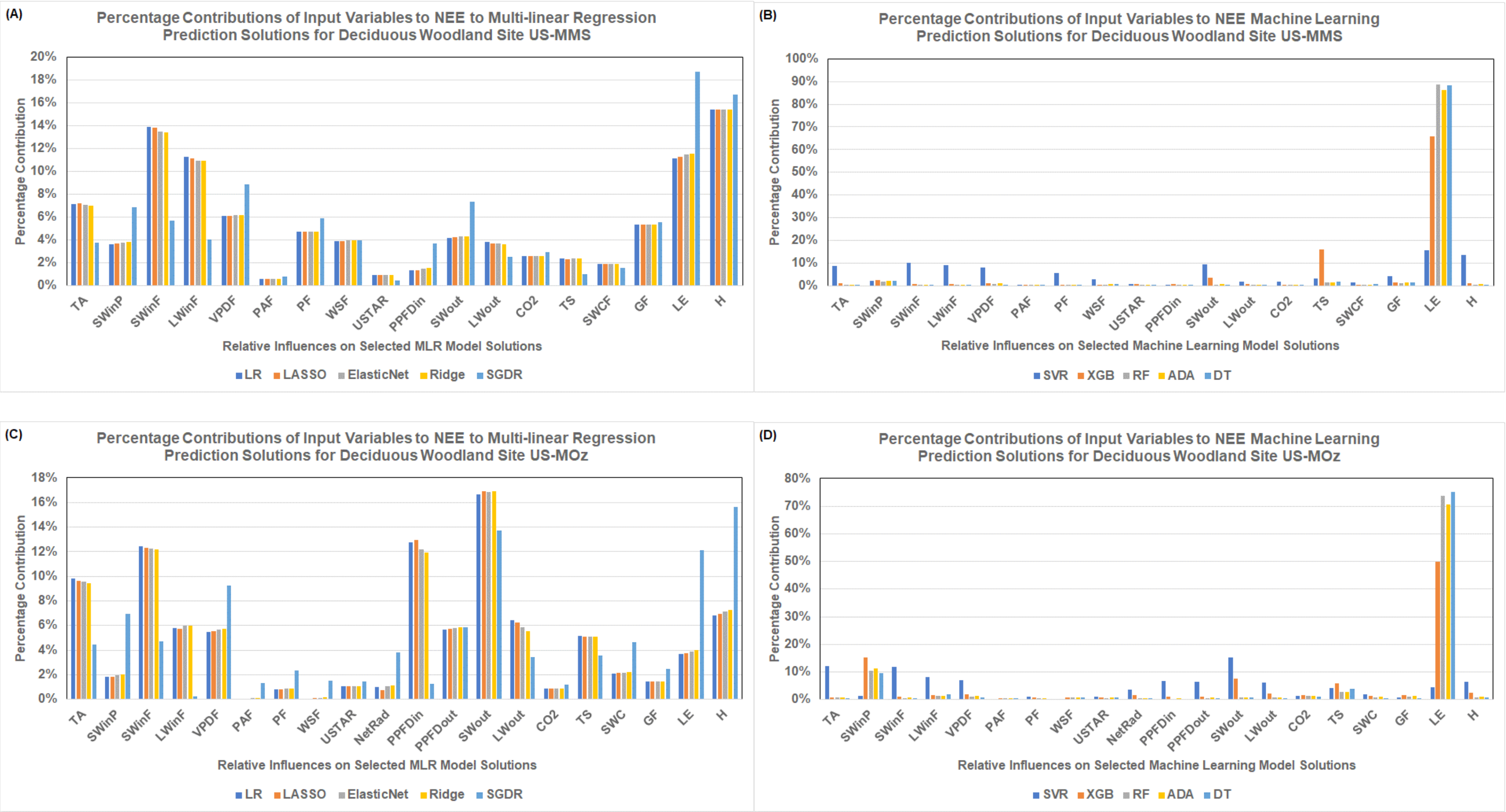
For the CA-Cbo dataset, the MLR models (Figure 9A) assign the most importance (in descending order) to variables SWOUT, PPFDOUT, WSF, LE, H and TA, and the least significance (in ascending order) to PAF, PF, SWC and SWINP for the 10-fold cross-validation solutions. The LASSO model assigns zero weights to variables SWINF, LWINF, PDF, PF, NetRad and PPFDIN, allowing it to assign more weight to preferred variables SWOUT, PPFDOUT, LE and TS than the other regression variables. That distinction enables the LASSO model to provide slightly better *NEE* predictions for CA-Cbo than the other MLR models.

The tree and ensemble ML models assign the most weight to variables LE, SWOUT, TS, WSF and SWC for CA-Cbo 10-fold cross-validation solutions (Figure 9B). Those models assign low but relatively similar levels of importance to the other variables for those solutions. In contrast, the better performing SVR model assigns substantially higher weights to variables SWOUT, PPFDOUT, TA and LWINF than the other ML models for its CA-Cbo solution (Figure 9B).

For the MX-Tes dataset, the MLR models (Figure 9C) assign the most importance (in descending order) to variables WSF, SWINP, SWINF, H, VPDF and SWOUT, and the least significance (in ascending order) to USTAR, LWOUT and CO_2 (except for the SGDR model) for the 10-fold cross-validation solutions. The SGDR model for this site achieves more accurate regression solutions than the other MLR models by assigning more weight to WSF, SWINP and CO_2 and less weight to H and LE.

The tree-ensemble 10-fold model solutions applied to the MX-Tes dataset assign higher weights to variables SWCF, USTAR and WSF and relatively even weights to the remaining variables (Figure 9D). In contrast, the better performing SVR model assigns more weight (in descending order) to PF, VPDF, TA, H, SWINP, SWOUT and PAF and less weight (in ascending order) to USTAR, SWINF and GF than the other ML models.

This analysis indicates, as should be expected due to their distinct locations and ecosystems, that different sets of the recorded variables are exerting the most influence at each site. Such information provides useful insight to assist in better understanding the ecosystem dynamics of specific sites.


(A) Percentage Contributions of Input Variables to NEE to Multi-linear Regression Prediction Solutions for Deciduous Woodland Site US-MMS

(B) Percentage Contributions of Input Variables to NEE Machine Learning Prediction Solutions for Deciduous Woodland Site US-MMS

(C) Percentage Contributions of Input Variables to NEE to Multi-linear Regression Prediction Solutions for Deciduous Woodland Site US-MOz

(D) Percentage Contributions of Input Variables to NEE Machine Learning Prediction Solutions for Deciduous Woodland Site US-MOz

Figure 8. The influences of measured environmental variables on the 10-Fold cross-validation *NEE*-prediction solutions of MLR and ML models for woodland sites: (A) US-MMS MLR models; (B) US-MMS ML models; (C) US-MOz MLR models; and, (D) US-MOz ML models.

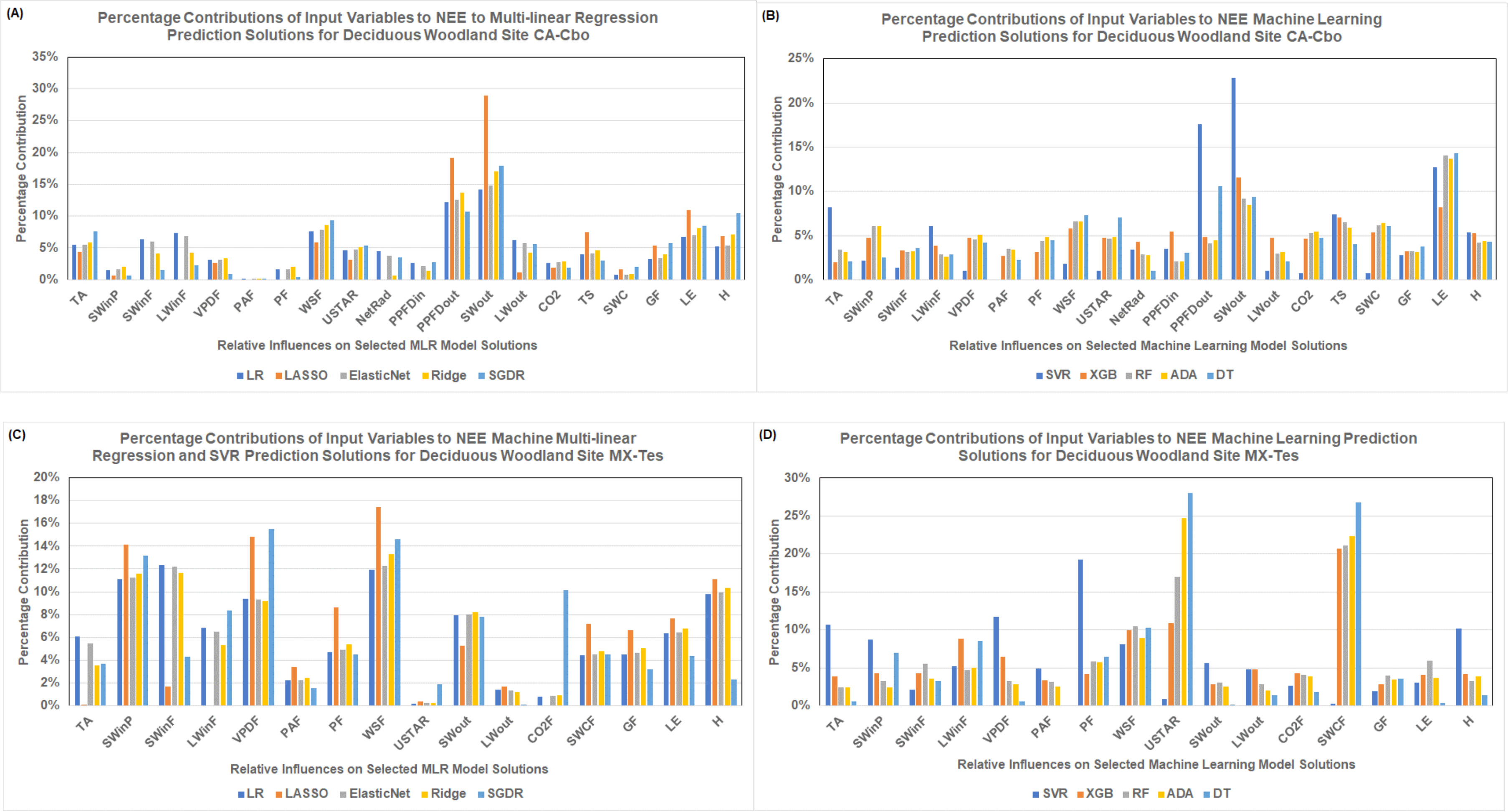


Figure 9. The influences of measured environmental variables on the 10-Fold cross-validation *NEE*-prediction solutions of MLR and ML models for woodland sites: (A) CA-Cbo MLR models; (B) CA-Cbo ML models; (C) MX-Tes MLR models; and, (D) MX-Tes ML models.

5. Discussion

5.1 Implications of Modelled *NEE* Patterns for Deciduous Woodland Ecosystems

The results presented identify distinctive *NEE* patterns and influences at the four DBF sites evaluated in terms of their weekly-averaged FLUXNET2015 datasets. Sites US-MMS and US-MOz display more systematic variations (Figure 3A) in their seasonal patterns that are easier to fit and explain in terms of the recorded environmental variables than sites CA-Cbo (Figure 3B) and MX-Tes.

Sites US-MMS and US-MOz (Figure 3A) display distinctive spring *NEE* peaks in April of each year, descending rapidly into a summer *NEE* trough (in July at US-MMS and May/June at US-MOz). The *NEE* pattern then rises to an October/November peak followed by a shallow winter trough at those two sites. These *NEE* patterns are consistent with leaf-on and leaf-off periods and seasonal weather fluctuations reflected in the environmental-variable recordings at those sites. MLR and ML models have little difficulty in fitting the *NEE* weekly patterns at those two sites with the environmental variables available.

At sites CA-Cbo (Figure 3B) and MX-Tes the weekly *NEE* patterns are seasonally more complex and variables but within narrower *NEE* annual ranges. A spring peak occurs in May each year at CA-Cbo, developing into a summer *NEE* trough (June/July), followed by an autumn peak (September/October) and winter trough (January/February). During that leaf-on summer season, several short-lived alternations between *NEE* peaks (+4 to +10) and troughs (–4 to –5) occur. The MLR and ML models struggle to accurately fit those summer peaks and troughs (Figures 6A and 6B). The short-term, summer *NEE* fluctuations cannot be explained by the environmental variables considered for this site, and are likely to be a consequence of biological processes.

At the MX-Tes site, *NEE* declines slowly across the winter reaching a low in April each year. During the leaf-on season, *NEE* oscillates displaying peaks (+0.25) and troughs (–1) during the wet season (late June to early September), and subsequently rises to an autumn high, which may occur at any time between late September and early December. From that high across the early winter the *NEE* pattern falls erratically following a more regular and gentle downward pattern from January/February to April each year. The peak-trough oscillations at MX-Tes do not correspond to changes in the environmental variables considered, making the fits of the MLR and ML model inaccurate (Figures 6C and 6D). At this site, data analysis indicates that the summer and autumn *NEE* variations are not responses to short-term fluctuations in the recorded environmental variables.

Further studies of the biological processes at work in the canopy, understory and topsoil during the leaf-on season are required at sites CA-Cbo and MX-Tes in attempts to explain the *NEE* oscillations at those sites. Possible processes to consider include microbial/fungal/insect cycles, and potential lag times in those cycles related to prior season environmental conditions or fluctuations in ground-water levels and/or soil temperatures. Additional soil temperature (TS) and soil water content data have been recorded at CA-Cbo. The variables TS and SWC, used by the *NEE* prediction models are for the shallowest soil depths. However, six TS (TS1 to TS6 deepest) and SWC (SWC1 to SWC6 deepest) values have been recorded at different soil depths. MLR and ML analysis was repeated to include four ratios of the TS and SWC recordings at different depths (TS1/TS6, TS1/TS2, SWC1/SWC6, and SWC1/SWC2).

The MLR and ML models for CA-Cbo were re-run with a 24-variable dataset including the four additional TS and SWC ratios. The results revealed very slight improvements in the *NEE* predictions: the Ridge 10-fold cross-validation model MAE decreased from 1.2576 $gCm^{-2}d^{-1}$ to 1.2530 $gCm^{-2}d^{-1}$ for the 20-variable and 24-variable models, respectively; the SVR 10-fold cross-validation model MAE decreased from 1.1029 $gCm^{-2}d^{-1}$ to 1.0843 $gCm^{-2}d^{-1}$ for the 20-variable and 24-variable models, respectively. However, for the SVR and Ridge 24-variable models the predicted versus recorded *NEE* relationship remains skewed from the Y = X pattern to similar degrees to that of the 20-variable models. These results imply that short-term fluctuations in SWC and TS cannot explain the *NEE* summer season peak-trough oscillations observed at the site.

SWC is a highly influential variable at the MX-Tes site (Figure 9D), however, SWC and TS measurements were not taken at multiple depths, so it is not possible to determine the role of groundwater level fluctuations in the *NEE* peak-trough oscillations at that site. Verduzco et al. (2015) identified a threshold level of about 350-400 mm of wet season precipitation that caused this site to switch from a net carbon source to a net carbon sink. That threshold was only exceeded in about half of the years assessed. Measuring SWC and TS at different soil depths would probably help to establish prior-season influences at the MX-Tes site.

5.2 Economic and Climate Change Implications of *NEE* Models

Reliable *NEE* prediction models based on a broad set of continuously recorded environmental variables for specific ecological sites are necessary to provide an understanding of the carbon flux driving mechanisms and how they vary seasonally at that site. Seasonal patterns in the weekly recorded *NEE* data understood in terms of fluctuating environmental and biological conditions through ML prediction models, will help to establish the capabilities of specific sites to act as reliable carbon sinks over time. Such information also helps in the understanding of how climate change, unusual natural events, and other anthropogenic activities will likely influence a site's *NEE* weekly pattern. Reliable and credible multi-year *NEE* prediction models are also necessary, and require regular auditing, to justify whether a particular site is entitled to receive carbon credit payments (i.e., it is a verifiable carbon sink with cumulative *NEE* values across the seasons < 0. There are inaccuracies in net-carbon account balances reported for some forest sites, which can result in substantial over-payments under existing commercial forest carbon protocol (CFCP) offsets [89]. One issue is that some sites fail to accurately record soil efflux of carbon and ecosystem respiration. Transparent ML models that use a set of recorded variables that can accurately predict a site's weekly *NEE* seasonal patterns would increase confidence in the carbon fluxes reported by the site.

The methodology proposed and applied in this study using weekly *NEE* seasonal data patterns can be used to rapidly establish whether a site is able or not to explain its seasonal *NEE* fluctuations in terms of the environmental variables it records. The methodology can help to identify sites requiring recordings of additional variables in order to better predict their long-term *NEE* patterns, and how those patterns might change as the sites are subjected to specific changes in environmental conditions.

6. Conclusions

Weekly net ecosystem exchange (*NEE*) patterns, combined with large suites of environmental variables, recorded as FLUXNET2015 (AmeriFlux) datasets over multiple years for four deciduous broadleaf forest (DBF) sites located in Canada, Mexico, and the United States can be assessed to characterize the varied influences on DBF sites that impact their potential as long-term carbon sinks. A proposed methodology that integrates considerations, correlations, multi-linear regression (MLR), machine learning (ML), and pattern analysis provides valuable insight into long-term variable relationships with *NEE* at DBF sites. Specifically, that methodology can distinguish sites for which long-term *NEE* patterns can be readily explained and predicted in terms of the onsite environmental variables currently recorded, from those that cannot.

Correlation-coefficients analysis distinguishes predominantly parametric from non-parametric relationships between environmental variables and *NEE* distributions. MLR models exploit the linear-variable relationships in the datasets, whereas the ML models exploit more complex non-linear-variable relationships to provide better predictions of the long-term, weekly, *NEE* patterns. Multi-fold cross-validation analysis, with repeated runs, determines the reproducibility of MLR and ML models, making it easier to avoid or minimize the overfitting tendencies of some ML models. 10-fold and 15-fold analyses with multiple runs provide the most dependable *NEE* prediction models for the site datasets assessed. From twelve MLR+ML models evaluated the support vector regression (SVR) model consistently generates the lowest prediction errors for the weekly data for each of the four DBF sites considered.

Models for two DBF sites (US-MMS and US-MOz) accurately predict their long-term, weekly *NEE* patterns by exploiting 18 and 20 of the recorded environmental variables available, respectively. Importantly, the SVR models for those sites deliver predicted (Y) versus recorded (X) *NEE* patterns that closely follow Y=X relationships. In contrast, the available datasets from two other DBF sites (CA-Cbo and MX-Tes) fail to adequately predict their long-term, weekly *NEE* patterns by exploiting up to 24 and 16 of the recorded environmental variables available, respectively. Significantly, the SVR models for those two sites deliver predicted versus recorded *NEE* patterns that deviate substantially from Y=X relationships.

Analysis of the relative influence of each recorded environmental variable on the prediction solutions of certain MLR and ML models to reproduce the measured multi-year *NEE* patterns, reveals that different sets of variables exert the most influence at each of the sites studied. Such information is useful in focusing attention on the likely impacts of climate change on each site's potential as a carbon sink. Detailed *NEE* pattern analysis for sites CA-Cbo and MX-Tes reveals rapid oscillations between high and low *NEE* values across specific seasonal periods including the leaf-on seasons at both sites. Large week-to-week swings in *NEE* cannot be explained or accurately modelled in terms of the recorded environmental variable variations. Such swings suggest that other biological cycles are at play at those sites, possibly in the soil and forest understory that require further investigation to identify their drivers. Such work should establish additional variables to record at those sites that would make it possible to establish more accurate and dependable *NEE*-prediction models for those sites. Such models are essential to quantify and justify the future potential of such sites as reliable carbon sinks as environmental conditions change over time.

Conflict of Interest

The author has no conflicts of interest associated with this study.

Funding

No funding was received for this study.

Acknowledgment

Thank you to AmeriFlux and its operating teams for their efforts in providing detailed eddy covariance measurement the sites studied and making them publicly available for research under AmeriFlux CC-BY-4.0 License.

References

[1] Baldocchi, D.D., Hicks, B.B., Meyers, T.P., 1988. Measuring biosphere-atmosphere exchanges of biologically related gases with micro meteorological methods. Ecology. 69, 1331-1340.

[2] Swinbank, W.C., 1951. The measurement of vertical transfer of heat and water vapor by eddies in the lower atmosphere. Journal of Meteorology. 8(3), 135-145.
DOI: https://doi.org/10.1175/1520-0469(1951)008<0135:TMOVTO>2.0.CO;2

[3] Valentini, R. (editor), 2003. Fluxes of carbon, water and energy of European forests. Ecological studies. Springer : Berlin, Heidelberg.
DOI: https://doi.org/10.1007/978-3-662-05171-9

[4] Goulden, M.L., Munger, W., Fan, S.M., et al., 1996. Measurements of carbon sequestration by long-term eddy covariance: methods and a critical evaluation of accuracy. Global Change Biology. 2(3), 169-182.
DOI: https://doi.org/10.1111/j.1365-2486.1996.tb00070.x

[5] Barnhart, B.L., Eichinger, W.E., Prueger, J.H., 2012. A new eddy-covariance method using empirical mode decomposition. Boundary Layer Meteorology. 145(2), 369-382.
DOI: https://doi.org/10.1007/s10546-012-9741-6

[6] Baldocchi, D.D., 2020. How eddy covariance flux measurements have contributed to our understanding of global change biology. Global Change Biology. 26, 242-260.

[7] Baldocchi, D., Chu, H., Reichstein, M., 2018. Inter-annual variability of net and gross ecosystem carbon fluxes: A review. Agriculture and Forest Meteorology. 249, 520-533.
DOI: https://doi.org/10.1016/j.agrformet.2017.05.015

[8] Monteith, J.L., 1972. Solar radiation and productivity in tropical ecosystems. Journal of Applied Ecology. 9(3), 747.
DOI: https://doi.org/10.2307/2401901

[9] Saigusa, N., Yamamoto, S., Murayama, S., et al., 2002. Gross primary production and net ecosystem exchange of a cool-temperate deciduous forest estimated by the eddy covariance method. Agricultural and Forest Meteorology. 112(3-4), 203-215.
DOI: https://doi.org/10.1016/S0168-1923(02)00082-5

[10] Sellers, P.J., Berry, J.A., Collatz, G.J., et al., 1992. Canopy reflectance, photosynthesis, and transpiration. III. A reanalysis using improved leaf models and a new canopy integration scheme. Remote Sensing of Environment. 42(3), 187-216.
DOI: https://doi.org/10.1016/0034-4257(92)90102-P

[11] Chu, H., Baldocchi, D.D., Poindexter, C., et al., 2018. Temporal dynamics of aerodynamic canopy height derived from eddy covariance momentum flux data across North American flux networks. Geophysical Research Letters. 45, 9275-9287.
DOI: https://doi.org/10.1029/2018GL079306

[12] Gough, C.M., Curtis, P.S., Hardiman, B.S., et al., 2016. Disturbance, complexity, and succession of net ecosystem production in North America's temperate deciduous forests. Ecosphere. 7(6), e01375.
DOI: https://doi.org/10.1002/ecs2.1375

[13] Holtmann, A., Huth, A., Pohl, F., et al., 2021. Carbon sequestration in mixed deciduous forests: The influence of tree size and species composition derived from model experiments. Forests. 12, 726.
DOI: https://doi.org/10.3390/f12060726

[14] Falge, E., Aubinet, M., Bakwin, P., et al., 2005. FLUXNET Marconi conference gap-filled flux and meteorology data, 1992–2000 [Internet] [cited 2023 Jan 15]. Available from: https://catalog.data.gov/dataset/fluxnet-marconi-conference-gap-filled-flux-and-meteorology-data-1992-2000.

[15] Neog, P., Kumar, A., Srivastava, A.K., et al., 2005. Estimation and application of Bowen ratio fluxes over crop surfaces—An overview. Journal of Agricultural Physics. 5(1), 36-45.

[16] Yuan, W., Liu, S., Zhou, G., et al., 2007. Deriving a light use efficiency model from eddy covariance flux data for predicting daily gross primary production across biomes. Agricultural and Forest Meteorology. 143(3-4), 189-207.
DOI: https://doi.org/10.1016/J.AGRFORMET.2006.12.001

[17] Ge, S., Smith, R.G., Jacovides, C.P., et al., 2011. Dynamics of photosynthetic photon flux density (PPFD) and estimates in coastal northern California. Theoretical and Applied Climatology. 105, 107-118.

DOI: https://doi.org/10.1007/s00704-010-0368-6

[18] Kia, S.H., Milton, E.J., 2015. Hyper-temporal remote sensing for scaling between spectral indices and flux tower measurements. Applied Ecology and Environmental Research. 13(2), 465-487.
DOI: https://doi.org/10.15666/aeer/1302_465487

[19] Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment. 8(2), 127-150.

[20] Tang, X., Wang, Z., Liu, D., et al., 2012. Estimating the net ecosystem exchange for the major forests in the northern United States by integrating MODIS and AmeriFlux data. Agricultural and Forest Meteorology. 156, 75-84.
DOI: https://doi.org/10.1016/j.agrformet.2012.01.003

[21] Niu, B., He, Y., Zhang, X., et al., 2016. Tower-based validation and improvement of MODIS gross primary production in an alpine swamp meadow on the Tibetan Plateau. Remote Sensing. 8(7), 592.
DOI: https://doi.org/10.3390/rs8070592

[22] Xu, C., Qu, J.J., Hao, X., et al., 2020. Monitoring soil carbon flux with in-situ measurements and satellite observations in a forested region. Geoderma. 378,114617.
DOI: https://doi.org/10.1016/j.geoderma.2020.114617

[23] Zhou, X., Wang, X., Tong, L., et al., 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 (China). 24(12), 2104-2112. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23534206

[24] Valentini, R., Matteucci, A.J., Dolman, E.D., et al., 2000. Respiration as the main determinant of carbon balance in European forests. Nature. 404(6780), 861-865.
DOI: https://doi.org/10.1038/35009084

[25] Gudasz, C., Karlsson, J., Bastviken, D., 2021. When does temperature matter for ecosystem respiration? Environmental Research Communications. 3, 121001.
DOI: https://doi.org/10.1088/2515-7620/ac3b9f

[26] Zhu, S., Clement, R., McCalmont, J., et al., 2022. Stable gap-filling for longer eddy covariance data gaps: A globally validated machine-learning approach for carbon dioxide, water, and energy fluxes. Agricultural and Forest Meteorology. 314(1), 108777.
DOI: https://doi.org/10.1016/j.agrformet.2021.108777

[27] Duman, T., Schäfer, K.V.R., 2018. Partitioning net ecosystem carbon exchange of native and invasive plant communities by vegetation cover in an urban tidal wetland in the New Jersey Meadowlands (USA). Ecological Engineering. 114, 16-24.
DOI: https://doi.org/10.1016/J.ECOLENG.2017.08.031

[28] Rödig, E., Huth, A., Bohn, F., et al., 2017. Estimating the carbon fluxes of forests with an individual-based forest model. Forest Ecosystems. 4, 4.
DOI: https://doi.org/10.1186/s40663-017-0091-1

[29] Churkina, G., Schimel, D., Braswell, B.H., et al., 2005. Spatial analysis of growing season length control over net ecosystem exchange. Global Change Biology. 11(10), 1777-1787.
DOI: https://doi.org/10.1111/j.1365-2486.2005.001012.x

[30] Verduzco V.S., Garatuza-Payán, J., Yépez, E.A., et al., 2015. Variations of net ecosystem production due to seasonal precipitation differences in a tropical dry forest of northwest Mexico. Journal of Geophysical Research: Biogeosciences. 120(10), 2081-2094.
DOI: https://doi.org/10.1002/2015JG003119

[31] Griffis, T., Roman, D., Wood, J., et al., 2020. Hydrometeorological sensitivities of net ecosystem carbon dioxide and methane exchange of an Amazonian palm swamp peatland. Agricultural and Forest Meteorology. 295, 108167.
DOI: https://doi.org/10.1016/j.agrformet.2020.108167

[32] Besnard, S., Carvalhais, N., Arain, M.A., et al., 2019. Memory effects of climate and vegetation affecting net ecosystem CO_2 fluxes in global forests. PLoS ONE. 14(2), e0211510.
DOI: https://doi.org/10.1371/journal.pone.0211510

[33] Mendes, K.R., Campos, S., da Silva, L.L., et al., 2020. Seasonal variation in net ecosystem CO_2 exchange of a Brazilian seasonally dry tropical forest. Scientific Reports. 10, 9454.
DOI: https://doi.org/10.1038/s41598-020-66415-w

[34] Wood, D.A., 2022. Machine learning and regression analysis reveal different patterns of influence on net ecosystem exchange at two conifer woodland sites. Research in Ecology. 4(2), 24-50.
DOI: https://doi.org/10.30564/re.v4i2.4552

[35] Wood, D.A., 2022. Net ecosystem exchange comparative analysis of the relative influence of recorded variables in well monitored ecosystems. Ecological Complexity. 50, 100998.
DOI: https://doi.org/10.1016/j.ecocom.2022.100998

[36] Cai, J., Xu, K., Zhu, Y., et al., 2020. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. Applied Energy. 262, 114566.
https://doi.org/10.1016/j.apenergy.2020.114566

[37] Abbasian, H., Solgia, E., Hosseini, S.M., et al., 2022.

Modeling terrestrial net ecosystem exchange using machine learning techniques based on flux tower measurements. Ecological Modelling. 446, 109901.
DOI: https://doi.org/10.1016/j.ecolmodel.2022.109901

[38] Wood, D.A., 2021, Net ecosystem carbon exchange prediction and data mining with an optimized data-matching algorithm achieves useful knowledge-based learning relevant to environmental carbon storage. Ecological Indicators. 124, 107426.
DOI: https://doi.org/10.1016/j.ecolind.2021.107426

[39] Kirschbaum, M.U., Mueller, R., 2001. Net Ecosystem Exchange: Workshop Proceedings, Cooperative Research Centre for Greenhouse Accounting [Internet] [cited 2001 April 18-20]. Available from: https://www.kirschbaum.id.au/NEE_Workshop_Proceedings.pdf.

[40] Reichstein, M., Falge, E.M., Baldocchi, D.D., et al., 2005. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm. Global Change Biology. 11, 1424-1139.
DOI: https://doi.org/10.1111/j.1365-2486.2005.001002.x

[41] FLUXNET, 2023. International Network of Eddy Covariance Measurement Sites [Internet] [cited 2023 Jan 15]. Available from: https://fluxnet.org/.

[42] Luyssaert, S., Reichstein, M., Schulze, E.D., et al., 2009. Toward a consistency cross-check of eddy covariance flux–Based and biometric estimates of ecosystem carbon balance. Global Biogeochemical Cycles. 23, 13.
DOI: https://doi.org/10.1029/2008GB003377

[43] Fei, X., Jin, Y., Zhang, Y., et al., 2017. Eddy covariance and biometric measurements show that a savanna ecosystem in Southwest China is a carbon sink. Scientific Reports. 7, 41025.
DOI: https://doi.org/10.1038/srep41025

[44] AmeriFlux, 2022. AmeriFlux Management Project [Internet] [cited 2013 Jan 15]. Available from: https://ameriflux.lbl.gov/about/ameriflux-management-project/.

[45] Baldocchi, D., Falge, E., Gu, L., et al., 2001. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem–Scale carbon dioxide, water vapor, and energy flux densities. Bulletiin of the American Meterorological Society. 82, 2415-2434.

[46] Pastorello, G., Trotta, C., Canfora, E., et al., 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Scientific Data. 7, 225.
DOI: https://doi.org/10.1038/s41597-020-0534-3

[47] Ameriflux, 2022. Flux/met data processing pipeline overview [Internet] [cited 2023 Jan 15]. Available from: https://ameriflux.lbl.gov/data/data-processing-pipelines/.

[48] Ameriflux, 2022. Data Variable Descriptions for the FLUXNET Product [Internet] [cited 2023 Jan 15]. Available from: https://ameriflux.lbl.gov/data/about-data/data-variables/.

[49] Teklemariam, T., Staebler, R., Barr, A.G., 2009. Eight years of carbon dioxide exchange above a mixed forest at Borden, Ontario. Agricultural and Forest Meteorology. 149, 2040-2053.
DOI: https://doi.org/10.1016/j.agrformet.2009.07.011

[50] Staebler, R., 2022. AmeriFlux FLUXNET-1F CA-Cbo Ontario-mixed deciduous, borden forest site, Ver. 3-5. AmeriFlux AMP, (Dataset).
DOI: https://doi.org/10.17190/AMF/1854365

[51] Yepez, E.A., Garatuza, J., 2021. AmeriFlux FLUXNET-1F MX-Tes Tesopaco, secondary tropical dry forest, Ver. 3-5. AmeriFlux AMP, (Dataset).
DOI: https://doi.org/10.17190/AMF/1832156

[52] Welch, N.T., Belmont, J.M., Randolph, J.C., 2007. Summer ground layer biomass and nutrient contribution to above-ground litter in an Indiana temperate deciduous forest. The American Midland Naturalist. 157(1), 11-26.

[53] Novick, K., Phillips, R., 2022. AmeriFlux FLUXNET-1F US-MMS Morgan Monroe State Forest, Ver. 3-5. AmeriFlux AMP (Dataset).
DOI: https://doi.org/10.17190/AMF/1854369

[54] Gu, L., Pallardy, S., Hosman, K.P., et al., 2016. Impacts of precipitation variability on plant species and community water stress in a temperate deciduous forest in the central US. Agricultural and Forest Meteorology. 217, 120-136.
DOI: https://doi.org/10.1016/J.AGRFORMET.2015.11.014

[55] Gu, L., Pallardy, S.G., Hosman, K.P., et al., 2015. Drought-influenced mortality of tree species with different predawn leaf water dynamics in a decade-long study of a central US Forest. Biogeosciences. 12(10), 2831-2845.
DOI: https://doi.org/10.5194/bg-12-2831-2015

[56] Wood, J., Gu, L., 2021. AmeriFlux FLUXNET-1F US-MOz Missouri Ozark Site, Ver. 3-5. AmeriFlux AMP, (Dataset).
DOI: https://doi.org/10.17190/AMF/1854370

[57] Harrell, F.E., 2015. Regression Modeling Strategies. Second Edition. Springer: Switzerland.
DOI: https://doi.org/10.1007/978-3-319-19425-7

[58] Stigler, S.M., 1981. Gauss and the invention of least squares. The Annals of Statistics. 9(3), 465-474.
DOI: https://doi.org/10.1214/aos/1176345451

[59] Bottou, L., 1998. Online algorithms and stochastic approximations. Online Learning and Neural Networks. Cambridge University Press: UK.

[60] SciKit Learn, 2023. Supervised and Unsupervised Machine Learning Models in Python [Internet] [cited 2023 Jan 15]. Available from: https://scikit-learn.org/stable/.

[61] Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 55, 119-139.
DOI: https://doi.org/10.1006/jcss.1997.1504

[62] Chan, J.C.W., Paelinckx, D., 2008. Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sensing of Environment. 112(6), 2999-3011.
DOI: https://doi.org/10.1016/j.rse.2008.02.011

[63] Quinlan, J.R., 1986. Induction of decision trees. Machine Learning. 1, 81-106.
DOI: https://doi.org/10.1007/BF00116251

[64] Debeljak, M., Džeroski, S., 2011. Decision trees in ecological modelling. Modelling Complex Ecological Dynamics. Springer: Berlin, Heidelberg. pp. 197-209.

[65] Fix, E., Hodges Jr., J.L., 1951. Discriminatory analysis, nonparametric discrimination: consistency properties. International Statistical Review. 57(3), 238-240.

[66] Fu, Y., He, H.S., Hawbaker, T.J., et al., 2019. Evaluating k-Nearest Neighbor (kNN) imputation models for species-level aboveground forest biomass mapping in northeast China. Remote Sensing. 11, 2005.
DOI: https://doi.org/10.3390/rs11172005

[67] Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review. 65(6), 386-408.
DOI: https://doi.org/10.1037/h0042519

[68] Eshel, G., Dayalu, A., Wofsy, S.C.C., et al., 2019. Listening to the forest: An artificial neural network-based model of carbon uptake at Harvard Forest. Journal of Geophysical Research: Biogeosciences. 124, 461-478.
DOI: https://doi.org/10.1029/2018JG004791

[69] Safa, B., Arkebauer, T.J., Zhu, Q., et al., 2019. Net Ecosystem Exchange (NEE) simulation in maize using artificial neural networks. IFAC Journal of Systems and Control. 7, 100036.
DOI: https://doi.org/10.1016/j.ifacsc.2019.100036

[70] Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 20(8), 832-844.
DOI: https://doi.org/10.1109/34.709601

[71] Zhou, Q., Fellows, A., Flerchinger, G.N., et al., 2019. Examining interactions between and among predictors of net ecosystem exchange: A machine learning approach in a semi-arid landscape. Scientific Reports. 9, 2222.
DOI: https://doi.org/10.1038/s41598-019-38639-y

[72] Huang, N., Wang, L., Zhang, Y., et al., 2021. Estimating the net ecosystem exchange at global FLUXNET sites using a random forest model. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 14, 9826-9836.
DOI: https://doi.org/10.1109/JSTARS.2021.3114190

[73] Cortes, C., Vapnik, V., 1995. Support-Vector networks. Machine Learning. 20(3), 273-297.
DOI: https://doi.org/10.1007/BF00994018

[74] IIlie, I., Dittrich, P., Carvalhais, N., et al., 2017. Reverse engineering model structures for soil and ecosystem respiration: The potential of gene expression programming. Geoscientific Model Development. 10(9), 3519-3545.
DOI: https://doi.org/10.5194/gmd-10-3519-2017

[75] Li, Z., Chen, C., Nevins, A., et al., 2021. Assessing and modeling ecosystem carbon exchange and water vapor flux of a pasture ecosystem in the temperate climate-transition zone. Agronomy. 11, 2071.
DOI: https://doi.org/10.3390/agronomy11102071

[76] Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Krishnapuram, Balaji; Shah, et al. (editors). Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 August 13-17; San Francisco, CA, USA. New York: Association for Computing Machinery. p. 785-794.
DOI: https://doi.org/10.1145/2939672.2939785

[77] Yan, S., Wu, L., Zhang, F., et al., 2021. A novel hybrid WOA-XGB model for estimating daily reference evapotranspiration using local and external meteorological data: Applications in arid and humid regions of China. Agricultural Water Management. 244, 106594.
DOI: https://doi.org/10.1016/j.agwat.2020.106594

[78] Liu, J., Zuo, Y., Wang, N., et al., 2021. Comparative analysis of two machine learning algorithms in predicting site-level net ecosystem exchange in major biomes. Remote Sensing. 13, 2242.
DOI: https://doi.org/10.3390/rs13122242

[79] SciKit Learn, 2023. GridSearchCV: Exhaustive Search Over Specified Parameter Values for an

Estimator in Python [Internet] [cited 2023 Jan 15]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[80] SciKit Learn, 2023. Bayesian Optimization of Hyperparameters in Python [Internet] [cited 2023 Jan 15]. Available from: https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html.

[81] SciKit Learn, 2023. Cross-validation: Evaluating Estimator Performance [Internet] [cited 2023 Jan 15]. Available from: https://scikit-learn.org/stable/modules/cross_validation.html.

[82] Pearson, K., 1894. On the dissection of asymmetrical frequency curves. Philosophical Transactions of the Royal Society of London. 185, 71-110.

[83] Spearman, C., 1904. The proof and measurement of association between two things. American Journal of Psychology. 15(1), 72-101.
DOI: https://doi.org/10.2307/1412159

[84] Lawrence, I., Lin, K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 255-268.
DOI: https://doi.org/10.2307/2532051

[85] Boddy, R., Smith, G., 2009. Statistical Methods in Practice: For scientists and technologists. John Wiley & Sons Ltd: Chichester, UK. pp. 95-96.

[86] Wayne, D.W., 1990. Spearman rank correlation coefficient. Applied Nonparametric Statistics (2nd ed). PWS-Kent: Boston. pp. 58-365.

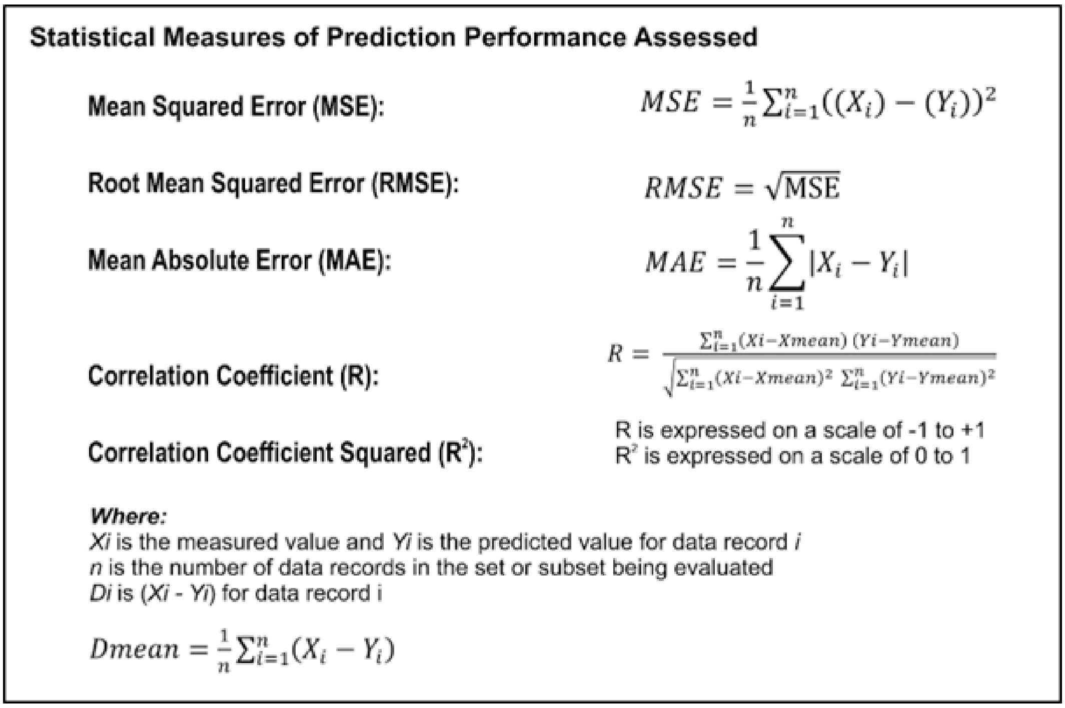
[87] Artusi, R., Verderio, P., Marubini, E., 2002. Bravais-Pearson and Spearman correlation coefficients: meaning, test of hypothesis and confidence interval. The International Journal of Biological Markers. 17(2),148-151.
DOI: https://journals.sagepub.com/doi/pdf/10.1177/172460080201700213

[88] Myers, L., Sirois, M.J., 2004. Spearman correlation coefficients, differences between. Encyclopedia of Statistical Sciences. John Wiley & Sons: UK.
DOI: https://doi.org/10.1002/0471667196.ess5050

[89] Marino, B.D.V., Bautista, N., 2022. Commercial forest carbon protocol over-credit bias delimited by zero-threshold carbon accounting. Trees, Forests and People. 7, 100171.
DOI: https://doi.org/10.1016/j.tfp.2021.100171

Appendix A. Error Metrics Used to Assess MLR and ML models

The statistical metrics used in this study to determine correlation coefficients and quantify prediction errors are defined in Figure A.



Statistical Measures of Prediction Performance Assessed

Mean Squared Error (MSE): $MSE = \frac{1}{n}\sum_{i=1}^{n}((X_i) - (Y_i))^2$

Root Mean Squared Error (RMSE): $RMSE = \sqrt{\text{MSE}}$

Mean Absolute Error (MAE): $MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i|$

Correlation Coefficient (R): $R = \frac{\sum_{i=1}^{n}(Xi - Xmean)\,(Yi - Ymean)}{\sqrt{\sum_{i=1}^{n}(Xi - Xmean)^2\ \sum_{i=1}^{n}(Yi - Ymean)^2}}$

Correlation Coefficient Squared (R^2):

R is expressed on a scale of -1 to +1
R^2 is expressed on a scale of 0 to 1

Where:
Xi is the measured value and *Yi* is the predicted value for data record *i*
n is the number of data records in the set or subset being evaluated
Di is (*Xi* - *Yi*) for data record i

$Dmean = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)$

Figure A. Definitions of prediction error measures applied.