BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

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

ARTICLE

Machine Learning and Regression Analysis Reveal Different Patterns of Influence on Net Ecosystem Exchange at Two Conifer Woodland Sites

David A. Wood*

DWA Energy Limited, Lincoln, United Kingdom

ARTICLE INFO

Article history
Received: 22 March 2022
Received in revised form: 19 April 2022
Accepted: 26 April 2022
Published: 18 May 2022

Keywords:
Eddy covariance
FLUXNET2015
Weekly *NEE* trends
Variable importance
Correlation comparisons
NEE prediction

ABSTRACT

Variations in net ecosystem exchange (*NEE*) of carbon dioxide, and the variables influencing it, at woodland sites over multiple years determine the long term performance of those sites as carbon sinks. In this study, weekly-averaged data from two AmeriFlux sites in North America of evergreen woodland, in different climatic zones and with distinct tree and understory species, are evaluated using four multi-linear regression (MLR) and seven machine learning (ML) models. The site data extend over multiple years and conform to the FLUXNET2015 pre-processing pipeline. Twenty influencing variables are considered for site CA-LP1 and sixteen for site US-Mpj. Rigorous k-fold cross validation analysis verifies that all eleven models assessed generate reproducible *NEE* predictions to varying degrees of accuracy. At both sites, the best performing ML models (support vector regression (SVR), extreme gradient boosting (XGB) and multi-layer perceptron (MLP)) substantially outperform the MLR models in terms of their *NEE* prediction performance. The ML models also generate predicted versus measured *NEE* distributions that approximate cross-plot trends passing through the origin, confirming that they more realistically capture the actual *NEE* trend. MLR and ML models assign some level of importance to all influential variables measured but their degree of influence varies between the two sites. For the best performing SVR models, at site CA-LP1, variables air temperature, shortwave radiation outgoing, net radiation, longwave radiation outgoing, shortwave radiation incoming and vapor pressure deficit have the most influence on *NEE* predictions. At site US-Mpj, variables vapor pressure deficit, shortwave radiation incoming, longwave radiation incoming, air temperature, photosynthetic photon flux density incoming, shortwave radiation outgoing and precipitation exert the most influence on the model solutions. Sensible heat exerts very low influence at both sites. The methodology applied successfully determines the relative importance of influential variables in determining weekly *NEE* trends at both conifer woodland sites studied.

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

DOI: https://doi.org/10.30564/re.v4i2.4552

1. Introduction

The balance of photosynthesis and respiration in the life existing above- and below-ground in biomes around the world determine the rate of exchange of carbon dioxide (CO_2) between biosphere and atmosphere. The measurement of CO_2 fluxes is essential for determining the locations of naturally occurring carbon sinks and sources [1]. Indeed, in addition to carbon, the fluxes of water and energy [2] are required to understand the dynamics of ecosystem -atmosphere exchanges [3]. Such requirements inspired the development of eddy-covariance measurements [4], and the progressive improvement of the technique and a refinement of measurements involved [5,6]. That technique has facilitated regular half-hourly measurements of Net Ecosystem Exchange (*NEE*), also referred to as CO_2 flux measurements, being recorded at many hundreds of sites around the world.

The *NEE* recordings have made it possible to identify distinctive and somewhat variable trends in carbon exchange occurring in different biomes [7]. This is not surprising as *NEE* is influenced by multiple factors. Influences include the intensity of solar radiation at specific sites [8], variations in weather and climate, geographic position, species mixture and degree and frequency of disturbance (wildfires, pests, diseases and anthropogenic activities involving soil exposure). This requires a substantial number of influential variables to be monitored as part of flux-tower recording projects.

The rates of biosphere photosynthetic and respiratory processes are, to an extent determined by certain specific environmental variables, including near-ground level atmospheric temperatures [9], various solar-radiation attributes, soil-water concentrations, soil-temperature trends and woodland-canopy conditions [10], specifically canopy height and aerodynamic conditions [11] and tree stand density [12]. Heat energy being released to or absorbed from the atmosphere varies substantially in the biosphere on a diurnal basis [13], both in terms of latent heat (related to phase changes) and sensible heat (related to changes in temperature). Such heat changes influence CO_2 fluxes between atmosphere and biomes. The Bowen Ratio (sensible heat/latent heat) is about 0.1 at the ocean surface in the tropics at one extreme rising to about 10.0 in desert environments and can vary substantially over the seasonal and crop cycles in some ecosystems [14]. Photosynthetically-active radiation (PAR) is incident light (wavelength from 400 nm to 700 nm) representing the spectral fraction stimulating photosynthesis and thereby influencing *NEE* trends [15]. Photosynthetic photon-flux density (PPFD) comprises PAR's photon-flux density [16]. PPFD is often measured as an *NEE*-influencing variable as it represents the component of PAR-zone light that reaches the biosphere canopy.

In order to extend the spatial and temporal understanding of *NEE* variations recorded at individual flux-tower sites they are frequently correlated with satellite-recorded spectral measurements [17]. Normalized difference vegetation index (NDVI) provides a satellite-derived remote-sensing measurement, that assesses whether a region includes live green vegetation based on infrared linear combinations [18]. Some of the mentioned variables, including NDVI and PPFD, can be monitored spatially using the moderate-resolution-imaging-spectroradiometer (MODIS) and land-remote-sensing-satellite (Landsat) datasets [19], particularly across woodland terrains [20].

Respiratory processes are known to fluctuate more substantially in response to ecosystem temperature changes than photosynthetic processes. This can result in significant seasonal *NEE* fluctuations in certain biomes [21], and, to an extent, explains why some ecosystems become more effective carbon sinks at lower latitudes [22]. Eddy-covariance recordings can become unreliable in rapidly changing weather and abrupt fluctuations in atmospheric conditions. Such conditions negatively impact the ability to record reliable and continuous datasets at some sites. Eddy-correlation techniques and machine-learning (ML) algorithms can, in such circumstances, assist in providing realistic data-gap replacement values [23].

Certain species communities are able to deliver unique seasonal components to *NEE*, and these can sometimes be usefully distinguished by partitioning species-related CO_2 fluxes [24,25]. Lengths of active growth seasons of specific species and whole biomes influence *NEE* and can fluctuate with local changes in climate. For instance, evergreen and deciduous forests often exhibit distinctive *NEE* trends both seasonally and longer term, partly related to fluctuating respiratory contributions [26]. The multiple influencing factors identified and the wide spectrum of climates and latitudes in which ecosystems exist explains why multi-year *NEE* trends tend to be complex, and in some cases difficult to understand. Attempts to distinguish the key influential factor determining fluctuations in seasonal [27] and annual *NEE* trends are important but fraught with uncertainties. Correlation and regression analysis [28], application of various ML models [29,30], and data mining techniques [31] can provide useful insight to these relationships.

In this study, two evergreen woodland eddy-covariance recording sites in North America, forming part of the AmeriFlux dataset [32] processed to FLUXNET2015 [33] requirements over multiple years, were modelled to predict *NEE* weekly-averaged trends in terms of multiple influencing variables measured at those sites. The key ob-

jectives of the study were to: (1) compare the abilities of four multi-linear regression and seven ML models to accurately model the multi-year *NEE* trends at the two sites; (2) identify the relative contributions of the influential variables to selected model solutions; and, (3) interpret the significance of the model results for *NEE* prediction approaches for woodland sites more generally.

2. Materials and Methods

2.1 *NEE* Determination

The *NEE* involves a defined calculation [34] designed to distinguish carbon becoming fixed in land-based biomes by organic processes, including above-ground photosynthetic activity and below-ground microbial activity, from carbon being released autotrophically and/or heterotrophically into the atmosphere, especially by respiration. In order to record that information accurately requires the separate measurement of daytime and nighttime fluxes relating to a biome's carbon uptake and respiration [35]. At most monitoring sites such measurements are recorded every half-hour. Those recordings are then preprocessed to verify data quality and infill, when possible, data gaps and errors. The pre-processed data are then compiled as hourly-, daily- and weekly- averaged values, enabling diurnal, monthly and seasonal trends to be routinely monitored.

NEE values need to be computed from the key recorded components mentioned in a two-step sequence as described by Equations (1) and (2) [36].

$$NPP = GPP - Ra \quad (1)$$

$$NEE = NPP - Rh \quad (2)$$

where, NPP = Net Primary Production accounting for carbon from photosynthesis less autotrophic respiration; GPP = Gross Primary Production; a measure of carbon generated and retained (i.e., at least temporarily fixed within the biome) as a result of photosynthesis;

Ra = carbon released to the atmosphere as a result of plant (autotrophic) respiration: and,

Rh = carbon released to the atmosphere as a result of microbial (heterotrophic) respiration.

Daytime respiration is typically subdivided into four components to provide more accurate measurements, so that it is more usefully described by Equation (3).

$$NEE_{daytime} = GPP - Rp - Rm - Rs - Rh \quad (3)$$

where, *Rp*= photorespiration;

Rm = maintenance respiration;

Rs = autotrophic synthesis, also referred to as growth respiration; and,

Rh = faunal/microbial heterotrophic respiration.

For daily and weekly analysis these variables are all measured in units of $gC\ m^{-2}\ d^{-1}$.

These variables are measured as magnitudes of absorbed/released carbon associated with a designated surface area over specified unit time periods. *NEE* data calculated and averaged on a weekly basis, as used in this study, are typically compiled and reported in $gC\ m^{-2}\ d^{-1}$ units. On the other hand, hourly *NEE* data are typically compiled and reported in $\mu molCO_2\ m^{-2}\ s^{-1}$ units. An *NEE* value below zero (negative *NEE*) signifies time periods during which carbon is absorbed, and at least temporarily, retained by the combined above-ground and below-ground components of a biome on a net basis. However, an *NEE* value above zero (positive *NEE*) signifies time periods during which carbon is released from a biome into the atmosphere on a net basis. *NEE* trends, accurately recorded over multiple years make it possible to elucidate whether a specific biome is acting as a long-term carbon source or carbon sink [37]. Due to variable weather conditions and climatic changes year-on-year some biomes can act as carbon sinks in some years and carbon sources in others. Hence, to reliably quantify the average magnitude of carbon stored on an annual basis in biomes considered to be long-term carbon sinks, meticulous recording of *NEE* data over multiple years is essential. The same is true for monitoring carbon flux responses to local changes in climate.

2.2 FLUXNET Variable Recording

FLUXNET is an organization that operates a worldwide network of micrometeorological tower sites [33]. It includes more than one thousand eddy covariance measurement sites distributed globally covering most climatic zones and a wide range of biomes. Eddy covariance was adopted as the favored method for measuring trace-gas fluxes (ecosystems to/from atmosphere) more than two decades ago [38]. FLUXNET2015 [39] is the most up-to-date, publicly available, FLUXNET dataset. It is hosted by the Lawrence Berkeley National Laboratory (U.S.A.) and delivers improved data-quality-control protocols combined with the more rigorous data-processing pipeline that deals more effectively with data gaps.

AmeriFlux [32] comprises a network of about 560 ecosystem-monitoring sites, built up since the mid-1990s and distributed throughout the Americas (Central, North and South), recording CO_2, water, and energy fluxes by applying the eddy covariance techniques. The United States Department of Energy, through the auspices of the AmeriFlux Management Project, supports the AmeriFlux network. In early 2022, 14 of the Ameriflux sites offered public data processed to the FLUXNET2015 standard [40], including the two evergreen woodland sites evaluated in this study, one in British Columbia, Canada (CA-LP1) and

another in New Mexico, United States of America (US-Mpj). These sites are described in detail in Section 2.3.

FLUXNET-designated sites are obliged to record certain variables using approved techniques. They are also encouraged, where possible, to report on other relevant variables [41]. Twenty such variables are available for site CA-LP1 and sixteen for site US-Mpj. Figure 1 lists those variables together with their measurement units and the abbreviations applied to them in this study.

2.3 Woodland Sites Evaluated

Weekly-averaged *NEE* and influencing variable data are compiled in this study for two evergreen-conifer woodland sites, CA-LP1 and US-Mpj that form part of the AmeriFlux datasets conforming to FLUXNET2015 protocols. These were the only sites included in that dataset located in conifer woodlands in early 2022.

The needle-leaf pine forest site in western Canada (**CA-LP1**) has recorded *NEE* data from 2007 to present [42]. It is located in British Columbia at latitude 55.1119 ^{o}N and longitude 122.8414 ^{o}W at +751 m elevation relative to sea level (Csa Koppen climate zone experiencing dry hot summers which are tending to become hotter with climate change [42]). Greater than sixty percent of the trees are lodge-pole pine up to 15 m high. The understory consists of *Vaccinium spp*. (mosses) and *Cladonia spp*. (moss-like lichens). Since before FLUXNET monitoring began, the trees have been under attack from the mountain-pine beetle [43]. This has resulted in progressive damage to many trees disturbing site attributes such as tree density overtime, as under normal conditions lodge-pole pines grow in dense stands.

The pinyon-juniper woodland is a biome characteristic of upland desert sites in the Western United States. Site **US-Mpj** close to Mountainair in New Mexico has recorded *NEE* data from 2007 to present [44] with published data available from 2008 to 2020. The woodland is situated on a mesa at +2196 m relative to sea level at latitude 34.4385 ^{o}N and longitude 106.2377 ^{o}W (Bsk Koppen climate zone experiencing a steppe climate with a relatively warm winter and dry hot summers which are tending to become hotter with climate change [44]). It is owned and managed by Heritage Land Conservancy. The tree species *Pinus edulis* (Engelm.) (an erect pine growing to about 21 m height) and *Juniperus monosperma* (Engelm.) Sarg. (a shrubby conifer growing to about 7 m height) make up about 95%

FLUXNET 2015 Processed Variables Considered

Dependent variable for regression and machine learning models		
NEE	Net Ecosystem Exchange (weekly)	gC m^{-2} d^{-1}
Independent variables for multi-linear 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)	0-100%
SWIN	Shortwave radiation, incoming	W m^{-2}
SWINP	Shortwave radiation at top of the atmosphere	W m^{-2}
SWOUT	Shortwave radiation, outgoing	W m^{-2}
TA	Air temperature	deg C
TS	Soil temperature	deg C
VPD	Vapor pressure deficit	hPa
USTAR	Friction velocity	m s^{-1}
WS	Wind speed	m s^{-1}

Figure 1. Variables evaluated in this study including their measurement units.

of the percent of the trees present on the site, with overall tree cover varying from about 30% to >60% forming a woodland/savanna. The herbaceous/ shrubby understory, including sages, artemisias and various grass species exist typically more than 2 m below the tree canopy. The site does experience periodic droughts and low grade wild fires [45].

Weekly recorded *NEE* data for the two sites are displayed in Figure 2, revealing clear seasonal variations with a certain amount of scatter in both cases. The data assessed in this study for site CA-LP1 are curtailed in 2015 because there are substantial gaps in recorded data of some influencing variables in several of the subsequent years.

Table 1 provides a statistical summary of the variables assessed for the 323 pre-processed data records compiled for site CA-LP1. This includes 20 of the independent variables listed and defined in Figure 1.

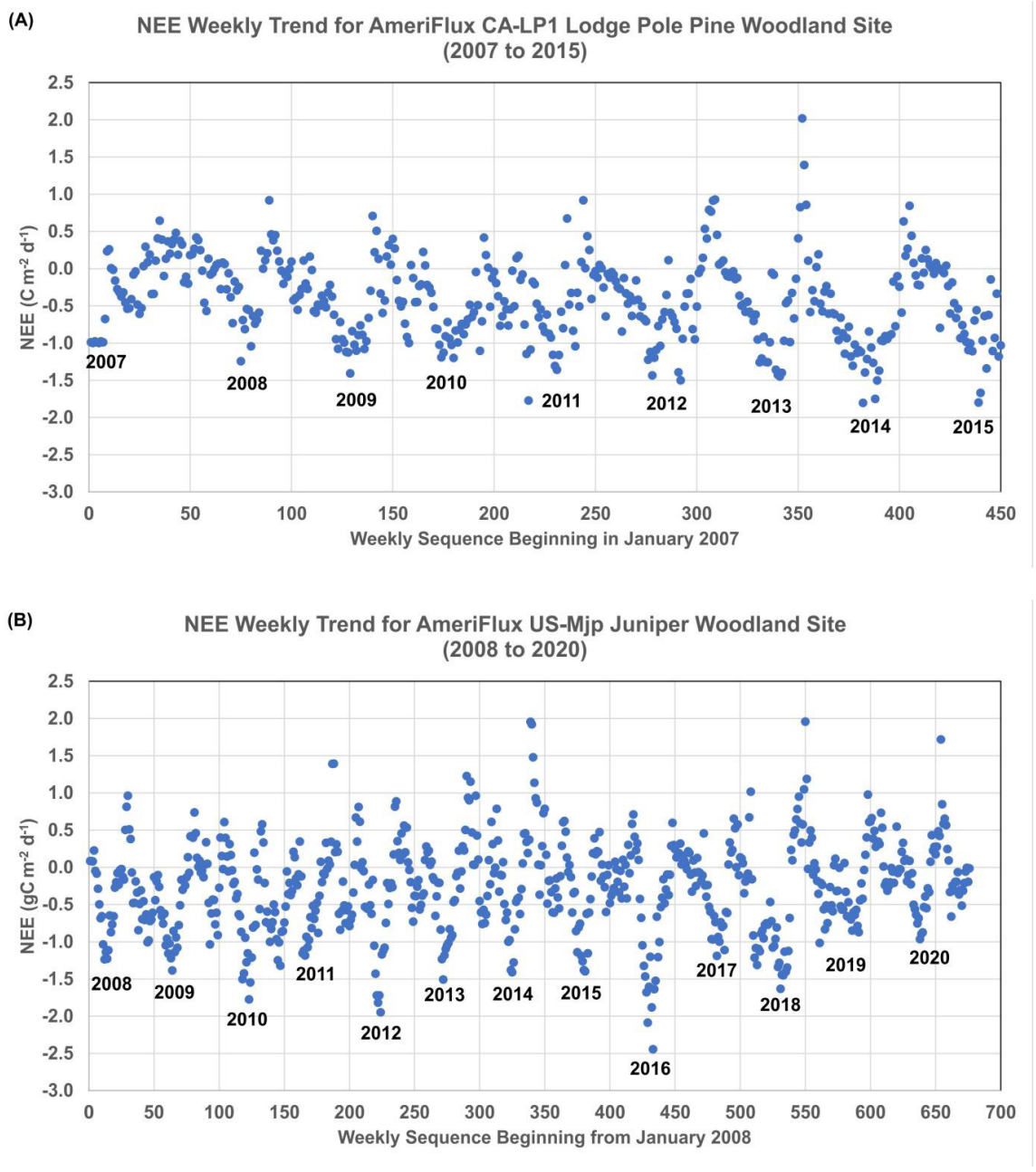

Figure 2. Calculated *NEE* trends from AmeriFlux recorded data for sites: (A) CA-LP1; and (B) US-Mpj. The data periods sampled represent time intervals over which the most continuous recordings of the independent variables exist at each site.

Table 1. Statistical description of data variables for the 323 data records compiled for woodland site CA-LP1.

Statistical Summary of Variables compiled for the CA-LP1 Pre-Processed Dataset					
Variable	**Units**	**Min**	**Mean**	**Max**	**Standard Deviation**
323 weekly data records					
Dependent Variable					
NEE	gC m^{-2} d^{-1}	–1.80	–0.44	2.02	0.56
20 Independent Variables					
TA	deg C	–25.37	6.14	21.77	8.93
SWINP	W m^{-2}	53.50	303.43	484.10	142.08
SWIN	W m^{-2}	11.26	143.80	317.97	81.27
LWIN	W m^{-2}	155.85	287.00	354.15	35.99
VPD	hPa	0.20	4.48	17.66	3.40
PA	kPA	90.46	92.32	93.67	0.47
P	mm	0.00	1.56	9.03	1.70
WS	m s^{-1}	1.13	2.26	4.09	0.43
USTAR	m s^{-1}	0.12	0.37	0.87	0.09
NetRad	W m^{-2}	–36.21	75.94	190.22	60.36
PPFDIN	µmolPhoton m^{-2} s^{-1}	21.08	341.44	814.21	201.73
PPFDOUT	µmolPhoton m^{-2} s^{-1}	2.47	19.54	70.14	13.25
SWOUT	W m^{-2}	1.58	15.77	37.02	7.55
LWOUT	W m^{-2}	199.47	341.31	423.86	44.43
CO_2	µmolCO_2 mol^{-1}	372.51	393.03	412.65	6.75
TS	deg C	–5.54	6.54	19.52	6.37
SWC	0-100%	2.30	9.21	15.48	2.51
G	W m^{-2}	–33.74	1.48	41.26	9.31
LE	W m^{-2}	0.06	21.28	59.17	14.31
H	W m^{-2}	–27.68	40.63	117.49	35.65
Note: See Figure 1 for definitions of the variables					

Tables 2 provides a statistical summary of the variables assessed for the 624 pre-processed data records compiled for site US-Mpj. This includes 16 of the independent variables listed and defined in Figure 1. Independent variables PPFDOUT, TS, SWC and G included for site CA-LP1 were not available for site US-Mpj.

Comparisons of the Pearson correlation coefficients (R) [46] and the Spearman correlation coefficients (p) [47] between influencing variables and calculated *NEE* are displayed for CA-LP1 and US-Mpj in Figure 3. R assumes linear/parametric distribution relationships between the variables it assesses [48,49], whereas p makes no such assumptions as it uses the rank positions of the data points in the variable distributions it assesses [50]. By avoiding linear/parametric assumptions [51] p is of more general relevance in determining whether two variable distributions can be meaningfully expressed as functions of each other, in parametric or non-parametric, and linear or non-linear terms [52].

Table 2. Statistical description of data variables for the 624 data records compiled for woodland site US-Mpj.

Statistical Summary of US-Mpj Pre-Processed Dataset					
Variable	**Units**	**Min**	**Mean**	**Max**	**Standard Deviation**
624 weekly data records					
Dependent Variable					
NEE	gC m^{-2} d^{-1}	–2.09	–0.26	1.96	0.61
20 Independent Variables					
TA	deg C	–5.14	11.23	26.58	7.76
SWINP	W m^{-2}	198.29	352.51	484.64	101.16
SWIN	W m^{-2}	0.26	236.82	390.56	74.95
LWIN	W m^{-2}	199.26	274.02	365.44	41.89
VPD	hPa	1.02	9.54	29.12	5.53
PA	kPA	76.95	78.10	78.85	0.36
P	mm	0.00	0.97	16.80	1.66
WS	m s^{-1}	2.00	3.56	5.77	0.67
USTAR	m s^{-1}	0.26	0.52	0.88	0.10
NetRad	W m^{-2}	–98.41	110.93	215.95	57.42
PPFDIN	µmolPhoton m^{-2} s^{-1}	1.77	443.39	723.34	142.41
SWOUT	W m^{-2}	2.59	30.21	60.43	8.57
LWOUT	W m^{-2}	278.95	375.03	471.12	47.47
CO_2	µmolCO_2 mol^{-1}	367.45	394.92	423.32	10.43
LE	W m^{-2}	3.40	29.40	107.69	18.17
H	W m^{-2}	–2.48	72.05	158.36	39.59
Note: See Figure 1 for definitions of the variables					

Figure 3. Pearson and Spearman correlation coefficients compared between influencing variables and NEE (weekly data) for sites: (A) CA-LP1 with all 20 variables listed in Figure 1 recorded; and (B) US-Mpj with 16 of the Figure 1 variables recorded (PPFDout, Ts, SWC and G are not available for this site).

The R and p values for sites CA-LP1 and US-Mpj are in relatively close agreement between all the influencing and *NEE* (Figure 3). This implies that parametric relationships predominate for these datasets and the degree of non-linearity involved in these variable relationships is relatively low. MLR models should be expected to provide reasonable *NEE* predictions if the measured variables capture the key influences impacting *NEE* at these sites.

3. *NEE* Prediction Models

3.1 Regression: Alternative Multi-linear Methods

Regression models considering multiple influential or input variables in attempts to predict the values of a dependent variable are termed multi-linear regression (MLR). Prior to applying machine learning (ML) methods it is typically worthwhile applying MLR methods to establish whether regression can generate accurate predictions for the dependent variable of interest. Although MLR methods simplistically assume linear relationships between the independent and dependent variables [53], that assumption can often provide quite accurate predictions for complex systems. MLR methods build on classical linear regression [54] by employing an optimizer to minimize the regression errors. Standard MLR methods apply coordinate descent as their optimization algorithm. Such a model used in this study is referred to as the LR model.

LR determines the coefficient values C_1 to C_N, together with a constant value C_0, that when applied to the set of influencing variables to provide the most accurate predic-

tions of dependent variable *Y*, by minimizing error term *J*, according to the linear Equation (4).

$$Y = C_0 + C_1X_1 + C_2X_2 + C_3X_3 + \dots C_NX_N + J \quad (4)$$

By finding the coefficient values that minimize an error function (*J*) leads to the determination of the most accurate dependent variable values that can be determined with Equation (4). A least-squares error function [55] is used in LR, as defined by Equation (5).

$$J(C_0 + C_1 + C_2 \dots C_N) = \frac{1}{2m}\sum \left(Y_i^a - Y_i^p\right)^2 \quad (5)$$

in which, and Y_i^p represent the actual and predicted values, respectively, for the dependent variable relating to the i^{th} data record, and *m* represents the number of data records in the dataset. In order to determine the optimum minimum value of *J,* multiple iterations of Equation (5) are required.

The LR method assumes independence among the influencing variables, and if dependences do occur they can magnify errors by resulting in multiple high coefficient values. The Ridge regression method addresses this by imposing an additional penalty or regularization term to the least-squares minimization function. This acts to limit the magnitude of the regression coefficients by progressively penalizing the residual sum of squares as the coefficients considered become larger. The least-squares minimization function for Ridge regression is expressed as Equation (6).

$$min_w\, for\, \sum_{i=1}^{m}\left(Y_i - \sum_{j=1}^{p} X_{ij}w_j\right)^2 + \lambda\sum_{j=1}^{p} w_j^2 \quad (6)$$

in which, $\lambda\sum_{j=1}^{p} w_j^2$ is the regularization or penalty function. In Ridge regression that penalty function includes each tested coefficient squared thereby inhibiting higher coefficient values in the optimum solution. This makes it an L2 regularization term with an L1 ratio of 0. When $\lambda = 0$ zero the penalty component of Equation (3) is removed and the regression reverts to LR. So, $\lambda > 0$ is required for a Ridge regression but if $\lambda >> 0$ the penalty will become too large and the data will be underfitted leading to reduced accuracy. With an appropriate λ value Ridge regression limits the effects of collinearity/dependency among the influencing variables and also reduces the risks of overfitting the data compared to LR. A range of optimizers are available to minimize the Ridge regression error function and in this study it is configured to automatically choose the best solver available.

Least absolute shrinkage and selection operator (LASSO) regression is an alternative method with a different regularization term added to the least-squares minimization function. Its configuration acts to preferentially select solutions with the least number of non-zero variable coefficients (i.e., it tends to disregard those variables with the least influence on the dependent variable). The least-squares minimization function for LASSO regression is expressed as Equation (7).

$$min_w\, for\, \sum_{i=1}^{m}\left(Y_i - \sum_{j=1}^{p} X_{ij}w_j\right)^2 + \lambda\sum_{j=1}^{p}\|w_j\| \quad (7)$$

in which, $\lambda\sum_{j=1}^{p}\|w_j\|$ is the regularization or penalty function. In contrast to Ridge regression, the penalty function for LASSO regression includes the absolute value of each tested coefficient, thereby acting to reduce the least important variable coefficients to 0 and introducing a degree of feature selection. This makes LASSO regularization an L1 term with an L1 ratio of 1. LASSO regression typically applies coordinate descent optimization.

An alternative approach is to adopt gradient-descent optimizers (GD) [56]. These employ a partial differential of *J* for coefficients C_0 to C_N as defined in Equation (8).

$$C_0^{k+1} = C_0^k - \alpha\frac{d}{dC_0^k}J\left(C_0^k\right) \quad (8)$$

in which, *k* indicates a particular epoch in the optimizer's execution, and α specifies the learning rate. By differentiating each term in Equation (8), C_0 to C_N coefficient values for the next GD epoch are derived according to Equation (9) for C_0, and Equation (10) for C_1 to C_N.

$$C_0^{k+1} = C_0^k - \alpha\frac{1}{m}\sum\left(Y_i^a - Y_i^p\right) \quad (9)$$

$$C_1^{k+1} = C_1^k - \alpha\frac{1}{m}\sum\left(Y_i^a - Y_i^p\right)X_1^i \dots. \quad (10)$$

If a small value of α is applied, the adjustments made in each epoch to coefficient C_0 to C_N values are small, so it is appropriate to optimize the α value applied to suit a specific dataset. Various penalty functions can be applied to the error functions of linear regression models utilizing GD optimizers. The stochastic gradient descent algorithm is applied as an alternative regression method (SGDR) in this study with an L2 penalty involved in its error function.

The four regression methods described (LR, LASSO, Ridge and SGDR) [57] are applied to the two AmeriFlux woodland datasets considered in this study (CA-LP1 with 20 variables and 323 data records (Table 1); US-Mpj with 16 variables and 624 data records (Table 2)). This assesses the relative benefits of applying the different regression error / penalty functions and optimization methods to these specific datasets.

3.2 Machine Learning Methods Applied

In addition to the four MLR algorithms, seven ML algorithms are applied to the woodland datasets evaluated in this study. These ML algorithms were executed in Python and customize publicly available codes [58]. The algorithms are selected specifically because they have proven capabilities of being able to successfully process and evaluate complex independent/dependent variable relationships.

The models are listed alphabetically. The first citation associated with each ML method refers to the original developers of the technique. The subsequent citations for each ML method refer to ecological studies that have applied the specific ML methods.

- Adaptive boosting-ADA [59,60]
- Decision tree- DT [61,62]
- K-nearest neighbor-KNN [63,64]
- Multi-layer perceptron-MLP (artificial neural network) [65-67]
- Random forest-RF [30,68-70]
- Support-vector regressor-SVR [71-73]
- Extreme gradient boosting-XGB [74-76]

These ML algorithms apply distinctive methodologies, making it useful to compare their results when applied to complex datasets. They can be categorized as regression-based (SVR), single-tree (DT), ensemble-tree (ADA, RF, XGB), data-matching (KNN) and neural-network (MLP) algorithms. These ML algorithms are widely used and their applications extensively published (see citations provided), so their mathematical methodologies are not repeated here. Nevertheless these models need to be appropriately tuned and configured to suit specific datasets. This requires establishing values for their hyperparameters that optimize their performance for the woodland datasets evaluated (Table 3). That optimization was achieved for this study using GridSearchCV [77] and Bayesian optimization [78] techniques.

Another requirement to determine when executing regression and ML algorithms is the optimum percentage of data records ("splits") in the datasets to allocate to the training and validation subsets to achieve reliable and reproducible prediction results. If the splits are too much in favor of the training subsets then the small randomly selected validation subsets generate a range of prediction results with too much variation (high standard deviations). On the other hand, if the splits are too much in favor of the validation subsets then model training tends to be in-

Table 3. Hyperparameters optimized for MLR and ML models applied to predict *NEE* from measured input variables at two evergreen conifer AmeriFlux sites.

Learning Algorithms	**Control Parameter Values Applied**
Regression	
Ordinary Least Squares Regression (LR)	Fit Intercept = true
LASSO	Alpha= 0.001; coordinate descent optimization; tolerance =0.0001; L1 regularization; L1 ratio =1.0; Fit intercept = true
Ridge	Alpha= 0.1; optimization solver = auto; tolerance =0.001; L2 regularization (L1 ratio =0.0); Fit intercept = true
Stochastic Gradient Descent (SGDR)	Alpha= 0.001; Loss function = episilon insensitive; learning rate = invscaling; L2 regularization; L1 ratio = 0.15; Fit intercept = true
Machine Learning	
Adaptive Boosting (ADA)	Number of estimators=1000; learning rate =0.01; loss function = linear base estimator is DT with depth =500; splitter =best
Decision Tree (DT)	Maximum depth =500; splitter =best; splitting criterion = mse
K Nearest Neighbour (KNN)	Number of nearest neighbours assessed K = 8; distance metric = Minkowski with p = 2 (Euclidian); neighbour selection algorithm = auto
Multi-layer Perceptron (MLP)	3 hidden layers with 100, 50 and 25 neurons; activation fn. = tanh; Solver = adam; alpha=0.0001; Learning rate = adaptive with initial learning rate = 0.001
Random Forest (RF)	Number of estimators = 1000; maximum depth = 100; Splitting criterion = mse
Support Vector Classifier (SVC)	Kernel = rbf; For CA-LP1: C = 30; gamma = 0.08 For US-Mpj: C = 15; gamma = 1.0
Extreme Gradient Boosting (XGB)	Number of estimators=1500; Maximum depth =4; eta = 0.01; Subsample = 0.7; Columns sampled per tree = 0.8

adequate leading to high prediction errors. K-fold cross validation provides a statistical method for establishing the suitability of specific data splits for the dataset considered. The K-fold-cross-validation method works by randomly dividing a dataset into K-equal-sized subsets, with the value of K typically varying from about 4 to 15, depending on the number of data records available. Sequentially, one of the K subsets serves as the validation subset, while the other K-1 subsets are used to train the model. That sequence is repeated until, one at a time, all the K subsets are evaluated as the validation subset. The results are then compiled to assess the mean and standard deviation of their errors. It is typically worthwhile repeating that process several times with different random subdivisions to improve the statistical confidence in the error results. In this study, the SciKit Learn K-fold validation routine [79] is employed, and the results provided and interpreted in Section 4, which justify the use of data record splits of 90% training : 10% validation for detailed analysis of the datasets compiled for this study.

3.3 Statistical Measures of Prediction Performance Assessed

Prediction errors associated with the MLR and ML models applied in this study are evaluated in terms of three widely used statistical measures. These are:

Root Mean Squared Error (RMSE)

$$RMSE = \left[\frac{1}{m}\sum_{i=1}^{m}((cNEE_i) - (pNEE_i))^2\right]^{\frac{1}{2}} \quad (11)$$

$cNEE_i$ = measured *NEE* value based on eddy covariance measurements (i^{th} data record), and $pNNN_i$ = predicted *NEE* value derived from the regression or ML methods for data record i, while m = quantity of distinct data records evaluated.

Mean Absolute Error (MAE)

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|cNEE_i - pNEE_i| \quad (12)$$

In this study, both MAE and RMSE values presented relate to the daily units (gC m^{-2} d^{-1}) in which *NEE* values are calculated and presented for AmeriFlux sites. Consequently, when interpreting MAE and RMSE values it is useful to consider them in terms of the range of *NEE* values reported for specific sites.

Coefficient of Determination (R^2)

$$R^2 = \left\{\frac{\sum_{i=1}^{m}(cNEE_i - \overline{cNEE})(pNEE_i - \overline{pNEE})}{\sqrt{\sum_{i=1}^{m}(cNEE_i - \overline{cNEE})^2}\sqrt{\sum_{i=1}^{m}(pNEE_i - \overline{pNEE})^2}}\right\}^2 \quad (13)$$

$\overline{cNEE}$ and $\overline{pNEE}$ are the arithmetic means of the *cNEE* and *pNEE* variable distributions. The R^2 value derived varies between 0 to 1.

3.4 Pre-processing of Weekly Data Records

Some data records recorded at each AmeriFlux site are missing values for certain weeks of specific variables. There are some gaps in the data variable distributions recorded. Such missing values may not have been collected due to equipment issues, or, during the FLUXNET2015 processing pipeline of the recorded data may have been identified as invalid or unreliable. For this study, any data record with missing weekly values of the normally recorded variables values was removed from the dataset compiled for evaluation by the regression and ML models. This resulted in 323 data records with 20 independent variables being compiled for site CA-LP1 (Table 1) and 624 data records with 16 independent variables being compiled for site US-Mpj (Table 2).

The values of all variables in the compiled datasets were normalized to the scale range –1 to +1 by implementing Equation (14) for each variables distribution.

$$X_i^* = 2*[(X_i - Xmin)/(Xmax - Xmin)] - 1 \quad (14)$$

$X_i = i^{th}$ data record in variable X distribution;
$Xmin$ = minimum value in distribution X;
$Xmax$ = maximum in distribution X; and,
X_i^* = normalized value of i^{th} record of variable X.

Normalization is necessary to avoid scaling biases introduced to the prediction methods by variable values extending over different scale ranges.

Figure 4 provides a workflow diagram of the *NEE* prediction and variable-influence detection methodology applied to the two woodland AmeriFlux datasets considered. It applies a number of regression and ML algorithms to predict assess the *NEE* distributions using a large suite of measured environmental influences.

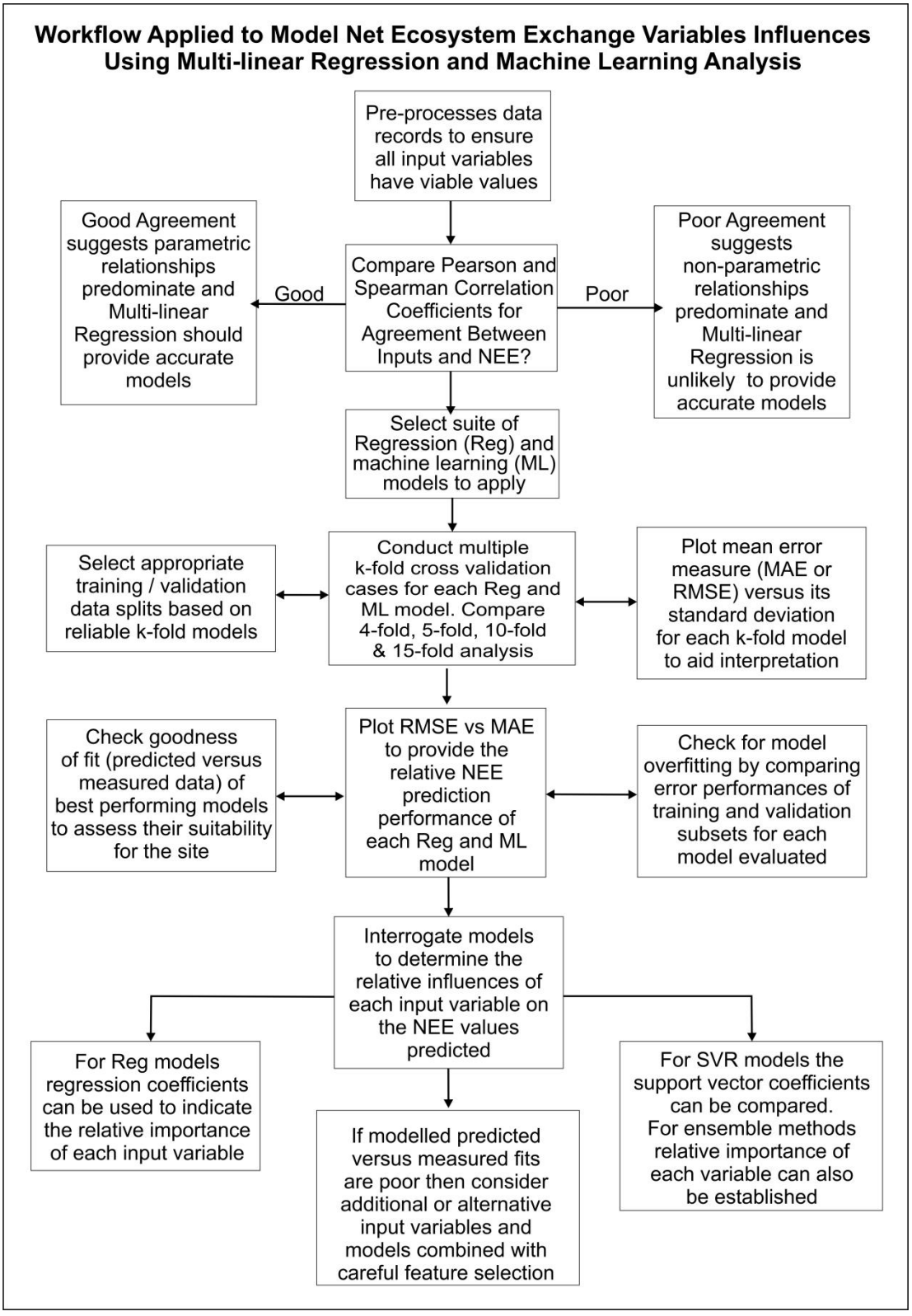

Figure 4. Workflow diagram describing the methodology adopted in this study for comparing regression and machine learning models to prediction *NEE* values from multiple influencing variables.

4. Results

4.1 Regression Versus Machine Learning *NEE* Predictions

The results of applying four regression and seven ML algorithms to the weekly, multi-year datasets of the two woodland AmeriFlux sites (CA-LP1/US-Mpj) confirm the superior *NEE* predictions generated by most of the ML models. This is illustrated in Figures 5 and 6, displaying the results for the best performing regression and ML models for each site.

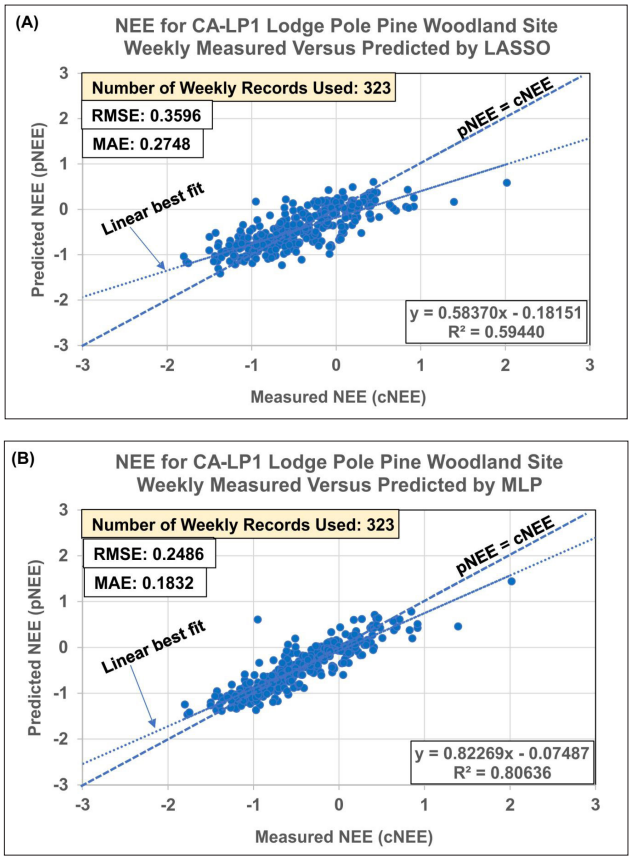

Figure 5. *NEE* predictions for site CA-LP1: (A) multi-linear regression LASSO model; and (B) Multi-layered Perceptron (MLP) model. *NEE* units are gC m^{-2} d^{-1}. In the equations for the best fit straight lines (y = mx + c), generated by linear regression and shown in the lower right corner of each graphic, y = p*NEE*, x = c*NEE* and the numerical values refer to coefficient m and constant c. The higher the value of c, the further that best-fit line deviates from passing through the origin of the graph.

In Figure 5, it is clear that MAE and RMSE are substantially lower, and R^2 substantially higher, for the MLP model than the LASSO model for site CA-LP1. Moreover, predicted *NEE* vs measured *NEE* are linearly arranged to more closely follow a *pNEE = cNEE* relationship for the SVR model (Figure 5B), making it a more credible solution.

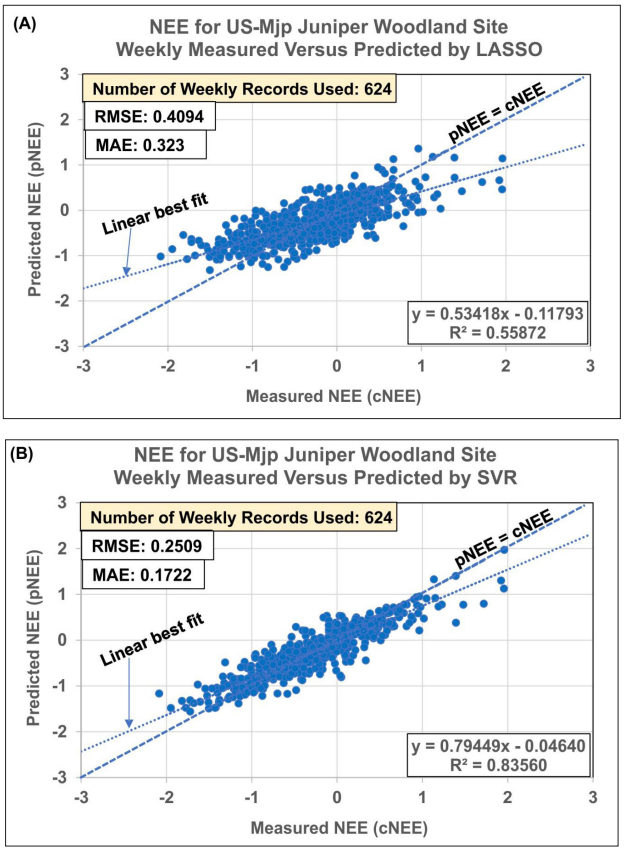

Figure 6. *NEE* predictions for site US-Mpj: (A) multi-linear regression LASSO model; and (B) Support Vector Regression (SVR) model. *NEE* units are gC m^{-2} d^{-1}. In the equations for the best fit straight lines (y = mx + c), generated by linear regression and shown in the lower right corner of each graphic, y = p*NEE*, x = c*NEE* and the numerical values refer to coefficient m and constant c. The higher the value of c, the further that best-fit line deviates from passing through the origin of the graph.

Figure 6 shows a similar outcome for site US-Mpj. The results for the SVR model (Figure 6B) are substantially superior to the regression model displayed (Figure 6B) and more closely approximate a *pNEE = cNEE* relationship.

4.2 K-fold Cross Validation Analysis of *NEE* Prediction Models

The K-fold cross validation technique provides confidence in the reliability of both MLR and ML models evaluated to predict *NEE* from the influencing variables for the two woodland sites considered. 4-fold, 5-fold, 10-fold and 15-fold cross validation analysis (as described in Section 3.2) was performed, and repeated three times, for each model and the mean MAE and standard deviation of the MAE for each model are displayed in Tables 4 and 5. These results are compared in Figures 7 and 8.

Table 4. K-fold cross validation results for *NEE* prediction analysis applying four regression and seven machine learning models to data for the lodge pole pine CA-LP1 woodland site. Best performing MLP model highlighted in bold type. StDev = standard deviation.

K-Fold Cross Validation *NEE* Prediction Errors (MAE) for Woodland Site CA-LP1								
	4-Fold (12 Cases)		**5-Fold (15 Cases)**		**10-Fold (30 Cases)**		**15-Fold (45 Cases)**	
	Mean MAE	**StDev MAE**	**Mean MAE**	**StDev MAE**	**Mean MAE**	**StDev MAE**	**Mean MAE**	**StDev MAE**
Regression								
LR	0.2987	0.0309	0.2958	0.0331	0.2911	0.0430	0.2912	0.0568
LASSO	0.2963	0.0305	0.2946	0.0303	0.2922	0.0422	0.2919	0.0556
RIDGE	0.2949	0.0325	0.2970	0.0305	0.2908	0.0429	0.2908	0.0565
SGDR	0.3139	0.0299	0.3151	0.0266	0.3130	0.0414	0.3100	0.0557
Machine Learning								
ADA	0.2979	0.0354	0.2973	0.0371	0.2957	0.0554	0.2940	0.0655
DT	0.4141	0.0292	0.3974	0.0334	0.3809	0.0656	0.4019	0.0853
KNN	0.2974	0.0333	0.2940	0.0296	0.2910	0.0480	0.2911	0.0574
MLP	0.2725	0.0306	0.2692	0.0250	**0.2566**	0.0425	0.2582	0.0460
RF	0.2943	0.0373	0.2961	0.0386	0.2908	0.0554	0.2877	0.0634
SVR	0.2715	0.0366	0.2642	0.0327	0.2612	0.0448	0.2575	0.0482
XGB	0.2834	0.0363	0.2830	0.0353	0.2757	0.0535	0.2745	0.0577

Table 5. K-fold cross validation results for *NEE* prediction analysis applying four regression and seven machine learning models to data for the juniper US-Mpj woodland site. Best performing SVR model highlighted in bold type. StDev = standard deviation.

K-Fold Cross Validation *NEE* Prediction Errors (MAE) for Woodland Site US-Mpj								
	4-Fold (12 Cases)		**5-Fold (15 Cases)**		**10-Fold (30 Cases)**		**15-Fold (45 Cases)**	
	Mean MAE	**StDev MAE**	**Mean MAE**	**StDev MAE**	**Mean MAE**	**StDev MAE**	**Mean MAE**	**StDev MAE**
Regression								
LR	0.3333	0.0213	0.3321	0.0266	0.3326	0.0309	0.3326	0.0427
LASSO	0.3314	0.0220	0.3308	0.0281	0.3310	0.0326	0.3305	0.0442
RIDGE	0.3328	0.0213	0.3317	0.0267	0.3323	0.0311	0.3323	0.0428
SGDR	0.3559	0.0181	0.3527	0.0280	0.3517	0.0356	0.3505	0.0451
Machine Learning								
ADA	0.2564	0.0177	0.2541	0.0249	0.2504	0.0295	0.2468	0.0359
DT	0.3593	0.0192	0.3516	0.0234	0.3613	0.0358	0.3595	0.0430
KNN	0.2621	0.0137	0.2629	0.0235	0.2599	0.0282	0.2587	0.0339
MLP	0.2460	0.0151	0.2374	0.0207	0.2370	0.0260	0.2364	0.0305
RF	0.2525	0.0202	0.2501	0.0257	0.2476	0.0280	0.2473	0.0336
SVR	0.2271	0.0144	0.2252	0.0215	**0.2194**	0.0282	0.2203	0.0318
XGB	0.2424	0.0176	0.2370	0.0227	0.2332	0.0262	0.2336	0.0328

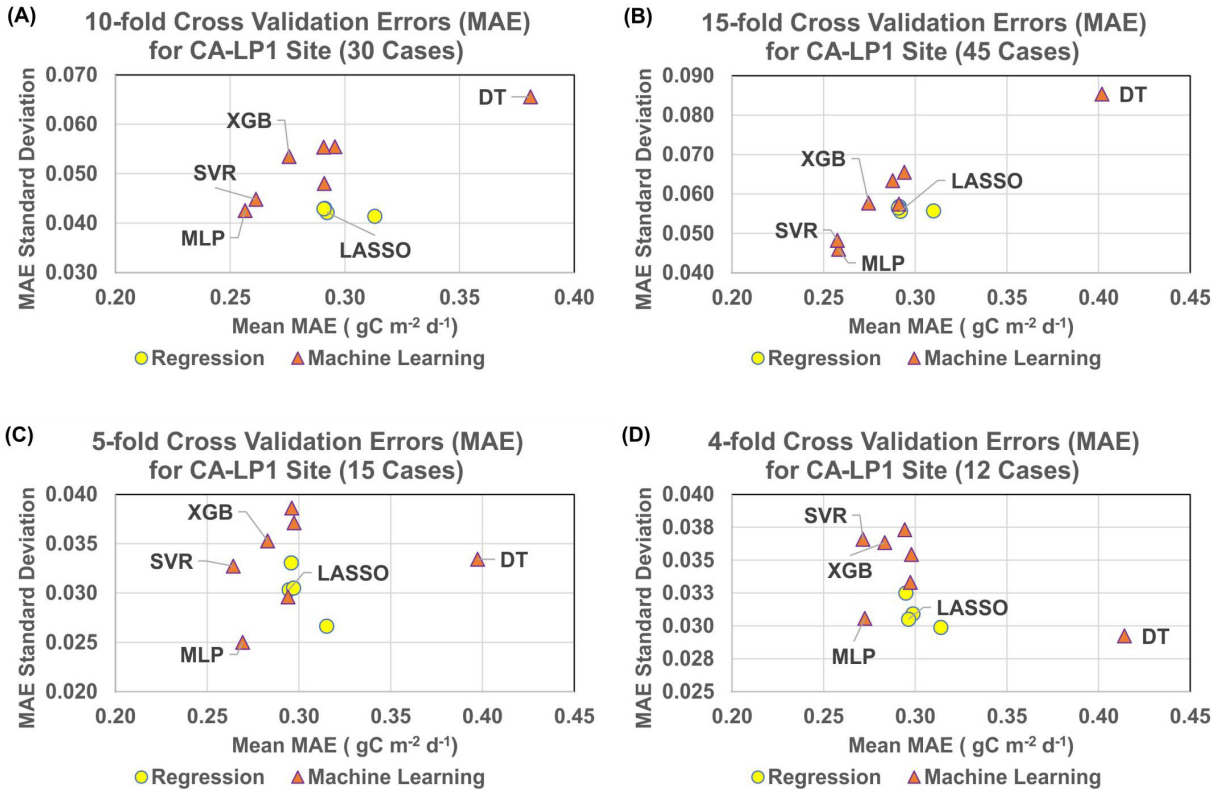

Figure 7. K-fold cross validation analysis of model *NEE* predictions for site CA-LP1: (A) 10-fold (splits 90% training: 10% validation); (B) 15-fold (splits 93.33%:6.67%); (C) 5-fold (splits 80%:20%); and, (D) 4-fold (splits75%:25%).

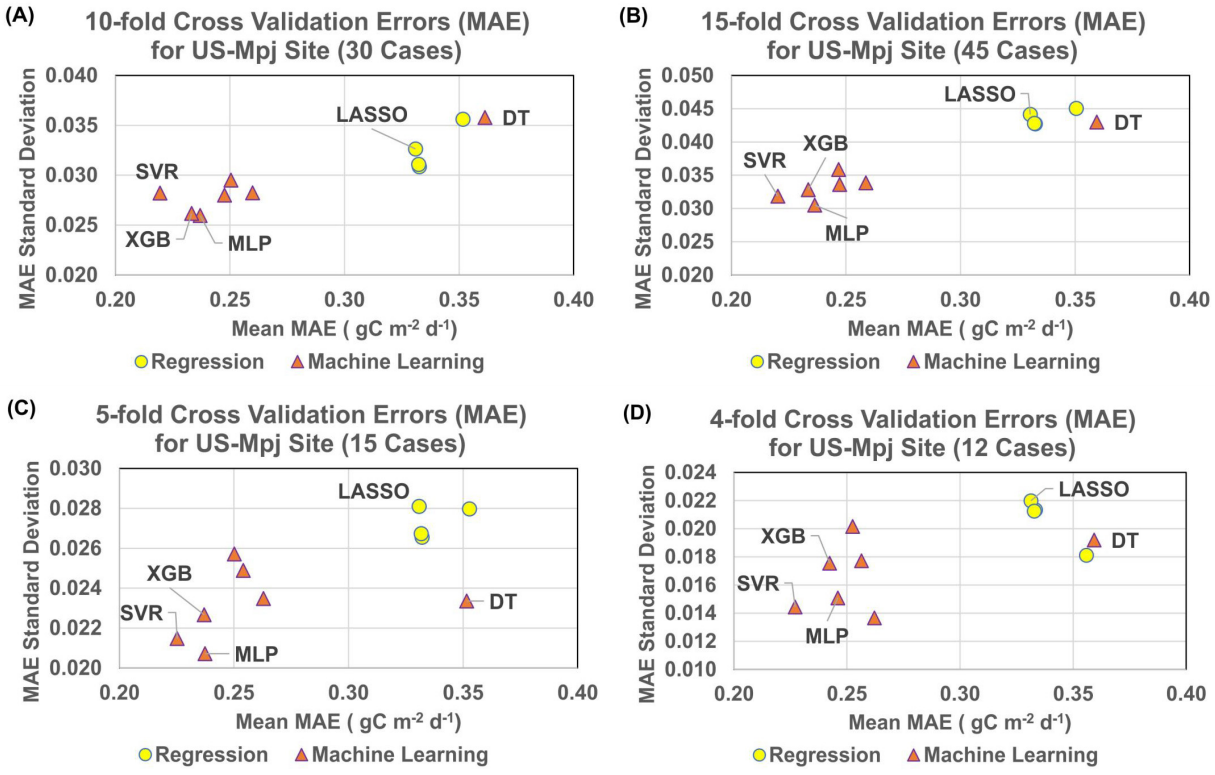

Figure 8. K-fold cross validation analysis of model *NEE* predictions for site US-Mpj: (A) 10-fold (splits 90% training: 10% validation); (B) 15-fold (splits 93.33%:6.67%); (C) 5-fold (splits 80%:20%); and, (D) 4-fold (splits75%:25%).

The 4-fold cross validation randomly splits the data records into four subsets and alternately one of those subsets is assigned for model validation and the other three sets are used to train the regression or ML model. This results in four cases being evaluated with each validation subset involving 25 percent of the available data records. Repeating the 4-fold process three times, each with distinct random selections, results for twelve cases are generated and used to calculate the MAE mean and standard deviation for the models. In a similar way, the 15-fold cross validation randomly splits the data records into fifteen subsets and alternately one of those subsets is assigned for model validation and the other fourteen sets are used to train the regression or ML model. This results in fifteen cases being evaluated with each validation subset involving 6.7 percent of the available data records. Repeating the 15-fold process three times, each with distinct random selections, results for forty-five cases are generated and used to calculate the MAE mean and standard deviation for the models. In a similar way, the 5-fold and 10-fold processes repeated three times results in fifteen and thirty cases being generated, respectively.

The rigorous k-fold process and the low MAE means and standard deviations it generates for all the models evaluated confirms the robustness of those models for all four splits considered. Even with only 6.7% of the data records assigned to the validation subset results in reproducible results with relatively low MAE standard deviations. For both woodland sites the 10-fold process generally yields the lowest MAE mean for the models considered. Moreover, the models display similar relative *NEE* prediction accuracies for each k-fold process applied (Figures 7 and 8).

For the CA-LP1 site, the MLP model provides the lowest MAE mean for the 10-fold process, followed by the SVR and XGB models (Figure 7). However, for the 4-fold, 5-fold and 15-fold processes the MLP and SVR models generate almost identical means for the CA-LP1 site. The better performing regression models (all but SGDR) show similar MAE means for all k-fold processes to the ADA, RF, KNN models. Nevertheless, the DT model generates the poorest *NEE* predictions for the CA-LP1 site with substantially higher mean MAE values than the other regression and ML models evaluated for each k-fold process.

The K-fold results for the US-Mpj site (Figure 8) are similar to those of the CA-LP1 site but with some distinctive features. SVR, XGB and MLP are the best performing models, in that order, for all k-fold processes. However, for US-Mpj, there is a substantial gap in performance between the six best ML models and the regression models. The poor-performing DT model generates predictions that are similar but slightly worse than the SGDR model for this site. The LASSO regression model provides a slightly better *NEE* prediction performance than the other three regression models for both sites. The LASSO, LR and ridge models provide quite similar *NEE* prediction performances that outperform the other regression model considered (SGDR) for both sites. The k-fold analysis result lead to the conclusions that the SVR, XGB and MLP are the best performing prediction models with the datasets from the two sites evaluated.

4.3 Training and Validation Subset Performances

Further consideration of the *NEE* prediction results of specific training and validation subsets provides further insight to the performance of the regression and ML models applied to woodland sites CA-LP1 and US-Mpj. These results are displayed in Tables 4 and 5 and Figure 9 for a representative case (90% training subset: 10% validation subset random split). The use of the 90%:10% split is justified based upon the results of the 10-fold cross validation analysis (Tables 4 and 5). The execution times of each model are also listed in Tables 6 and 7.

For the regression models, the prediction performances of the training subset and the trained model applied to the complete dataset (100% of the data records) are similar for both sites (Tables 6 and 7). Overfitting is clearly not an issue for the regression models. The prediction performance for the validation subset is slightly worse than for the training subset in the case of site CA-LP1, but slightly better for site US-Mpj. Such variations are consistent with what should be expected from the random selections of the validation and training subsets considered.

For the ML models, it is apparent that all but the SVR and MLP models involve a degree of overfitting. The100% prediction accuracy achieved with the training subset by the ADA, DT, KNN models, compared with much lower accuracies for those models applied to the validation subset, are clear indications of overfitting (Tables 6 and 7). The RF and XGB models also show similar but less extreme prediction relationships between their training subset and validation subset results, indicative of a degree of overfitting associated with those models. Despite that overfitting the XGB model is still able to provide prediction performances for both sites that rival the best performing SVR and MLP models. The trends shown for the 10% validation subset in Figure 9 are consistent with the k-fold analysis (Figures 7 and 8) for both sites.

Table 6. Training and validation subset *NEE* prediction performances of MLR and ML models for the CA-LP1 woodland site.

***NEE* Forecasting Accuracy for Training and Validation Analysis Applied to the Full Dataset for Site CA-LP1**										
	Example Training Subset (90% of Data Records)			**Example Validation Subset (10% of Data Records)**			**Example Full Dataset (100% of Data Records)**			
Model	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**Ex Time**
Regression										
LR	0.6225	0.3430	0.2606	0.4241	0.4613	0.3652	0.6002	0.3569	0.2713	4.5
LASSO	0.6149	0.3464	0.2647	0.4293	0.4592	0.3630	0.5944	0.3596	0.2748	4.6
Ridge	0.6217	0.3433	0.2611	0.4276	0.4599	0.3651	0.6000	0.3570	0.2718	4.3
SGDR	0.5040	0.3932	0.2945	0.3039	0.5072	0.4074	0.4819	0.4063	0.3061	4.5
Machine Learning										
ADA	1.0000	0.0000	0.0000	0.3238	0.4999	0.3760	0.9199	0.1598	0.0384	71.1
DT	1.0000	0.0000	0.0000	-0.1579	0.6311	0.5171	0.8723	0.2017	0.0528	5.3
KNN	1.0000	0.0000	0.0000	0.3741	0.4809	0.3863	0.9258	0.1537	0.0395	4.4
MLP	0.8425	0.2216	0.1697	0.5303	0.4166	0.3024	0.8064	0.2486	0.1832	64.3
RF	0.9408	0.1358	0.1013	0.3224	0.5004	0.3704	0.8677	0.2053	0.1288	60.2
SVR	0.8249	0.2336	0.1443	0.5561	0.4050	0.3056	0.7936	0.2564	0.1608	4.5
XGB	0.9950	0.0394	0.0303	0.4200	0.4629	0.3425	0.9269	0.1526	0.0622	12.5
Notes: (1) RMSE and MAE are expressed in terms of the measured weekly NEE range for the site: -2.47935 to 2.01894 gC m^{-2} d^{-1} (2) Execution times (Ex Time) are expressed in seconds and include 10-fold cross validation.										

Table 7. Training and validation subset *NEE* prediction performance results for the regression and ML models applied to the US-Mpj woodland site.

***NEE* Forecasting Accuracy for Training and Validation Analysis Applied to the Full Dataset for Site US-Mpj**										
	Example Training Subset (90% of Data Records)			**Example Validation Subset (10% of Data Records)**			**Example Full Dataset (100% of Data Records)**			
Model	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**Ex Time**
Regression										
LR	0.5510	0.4140	0.3287	0.6399	0.3506	0.2682	0.5595	0.4080	0.3226	4.4
LASSO	0.5465	0.4161	0.3294	0.6540	0.3437	0.2657	0.5577	0.4094	0.3230	4.4
Ridge	0.5508	0.4141	0.3289	0.6417	0.3497	0.2683	0.5595	0.4080	0.3228	4.3
SGDR	0.4659	0.4515	0.3516	0.5885	0.3748	0.2866	0.4775	0.4444	0.3450	4.4
Machine Learning										
ADA	1.0000	0.0000	0.0000	0.7618	0.2852	0.2213	0.9783	0.0906	0.0223	119.7
DT	1.0000	0.0000	0.0000	0.5170	0.4061	0.3011	0.9559	0.1290	0.0304	5.4
KNN	1.0000	0.0000	0.0000	0.7346	0.3010	0.2387	0.9758	0.0956	0.0241	10.0
MLP	0.8559	0.2346	0.1860	0.7762	0.2764	0.2079	0.8487	0.2391	0.1882	53.0
RF	0.9624	0.1198	0.0922	0.7635	0.2842	0.2165	0.9443	0.1451	0.1048	40.7
SVR	0.8376	0.2490	0.1694	0.7908	0.2672	0.1966	0.8356	0.2509	0.1722	2.6
XGB	0.9824	0.0821	0.0635	0.7835	0.2718	0.2019	0.9642	0.1163	0.0775	3.9
Notes: (1) RMSE and MAE are expressed in terms of the measured weekly NEE range for the site: -2.44429 to 1.95933 gC m^{-2} d^{-1} (2) Execution times (Ex Time) are expressed in seconds and include 10-fold cross validation.										

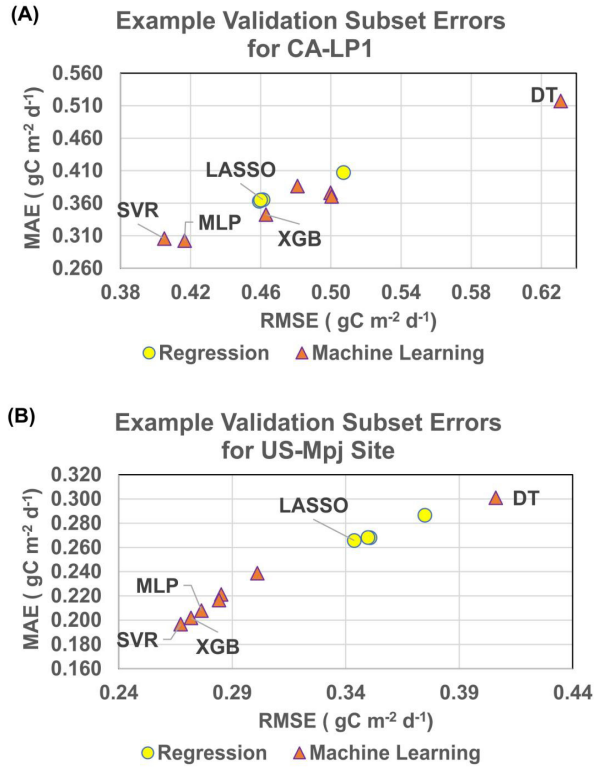

Figure 9. Validation subset (representative 10% split of the dataset) prediction performances compared for the regression and ML models applied to: (A) CA-LP1 site data; and, (B) US-Mpj site data.

The regression models are all executed rapidly (about 4.5 seconds). Whereas the SVR model rivals or improves upon the MLR models for execution speeds, the MLP, RP and ADA models involve more computational seconds (Tables 6 and 7). All the models applied to data compiled for site CA-LP1 evaluate 20 variables and 323 data records (Table 1). All the models applied to data compiled for site US-Mpj evaluate 16 variables and 624 data records (Table 2).

4.4 Input Variable Importance Derived from Model Solutions

MLR models all generate their solutions transparently by routinely revealing the regression coefficients associated with each solution they generate. The SVR models can reveal their support vector coefficients for each solution generated, which are also useful for estimating feature influences on their solutions. Also, the DT models and the tree-ensemble ML models (ADA, RF and XGB) can also provide the relative influences of each input variable in the solutions they generate. DT model estimates of feature importance are derived from the relative contributions each attribute-split point (node) makes to improving the selected prediction-performance metric, for example, the Gini index of concentration [80,81]. To establish that the node contributions need to be weighted by the quantity of values associated with each node. For the ensemble-tree models feature-importance estimates calculated in that way for each tree involved need to be averaged. Such variable influence information cannot be easily extracted from the KNN and MLP solutions.

It is useful to compare variable influence information for the two woodland sites considered for the regression and ML models for which it can be readily extracted. Figures 10 and 11 make such a comparison in the form of bar charts; one for the four regression models plus SVR, and one for five ML models (SVR, XGB, RF, ADA, DT) for each site. It is apparent that both regression and ML models are influenced quite differently by the input variables at the two woodland sites. However, in the case of almost all models, all the input variables are assigned a weight, implying that they do exert some influence, albeit very small in the case of some models.

The three best-performing regression models (LASSO, LR and Ridge) applied to the CA-LP1 site are generally in agreement with respect to the respective weights given to input variables (Figure 10A). They assign most weight to TA followed by SWIN, NetRad, PPFDIN, TS and LE. The LR and Ridge models assign more weight to SWOUT and PPFDOUT than the LASSO model. The LASSO model assigns slightly more weight to TA, SWIN, NetRad, PPFDIN, TS, LE and H than the LR and Ridge models. On the other hand, the SGDR model assigns quite different weights to the variables in establishing its somewhat poorer solution, in particular, assigning much higher weights to SWIN, LWIN, PF, LWOUT and H and much lower weights to TA, NetRad and LE than the other three regression model.

In comparison with the best-performing regression models, SVR (Figure 10A) gives more weight to LWOUT and TA and less weight to SWIN, CO_2, LE and H, otherwise its variable weightings are quite similar to those regression models. On the other hand, SVR shows quite different priorities in its input variable weightings than the other ML models evaluated for sites CA-LP1 (Figure 10B).

The DT model stands out in Figure 10B in that it assigns almost 40% of its weightings to the input variable SWout, twice as much as other ML models. This probably explains the poorer *NEE* prediction performance of the DT model in that it converges too quickly to a solution skewed towards the variations of just one variable. The three tree-ensemble models (XGB, RF and ADA; Figure 10B) show generally similar input variable weightings

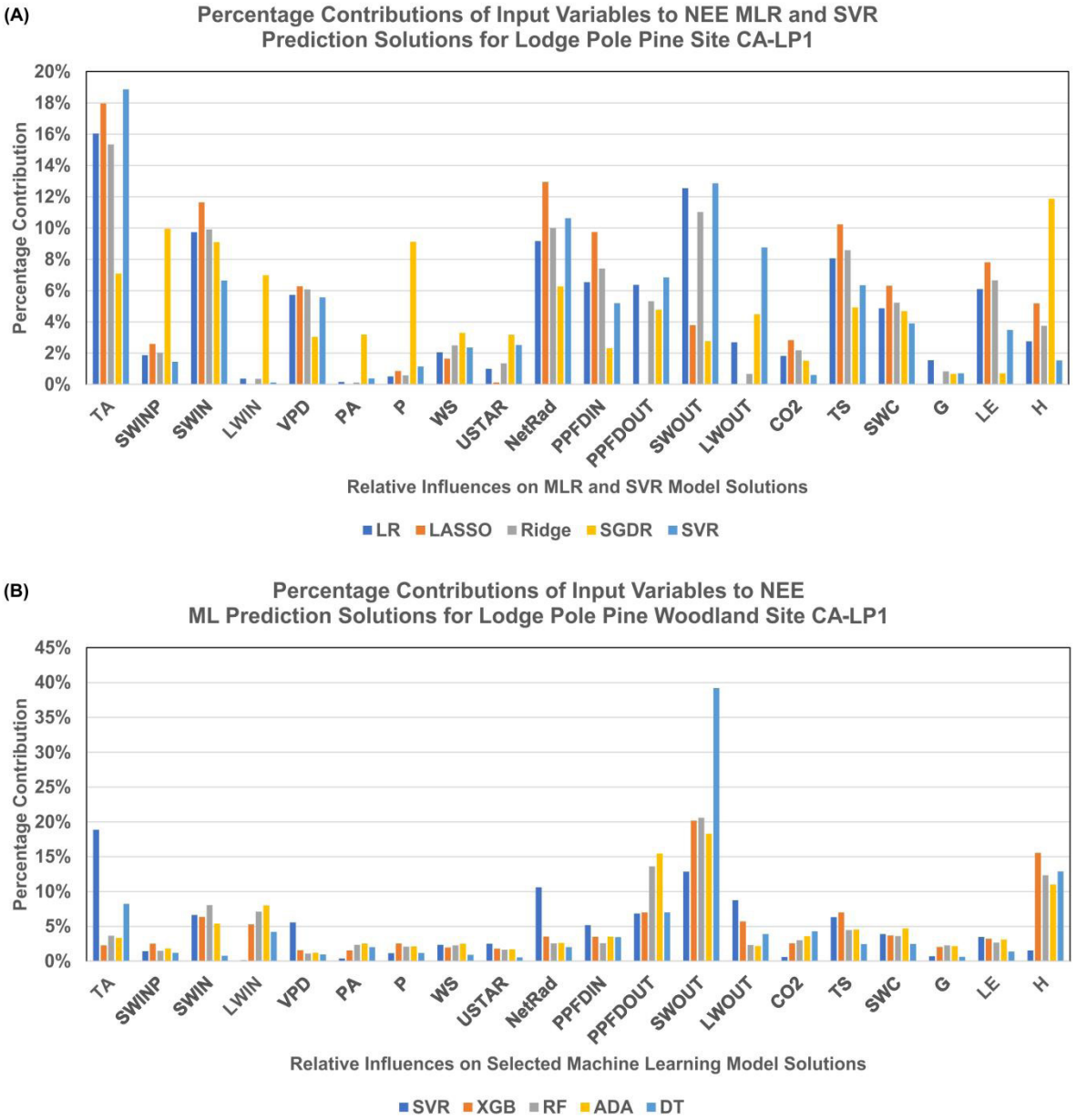

Figure 10. Relative importance of the twenty influencing variables on the *NEE* predictions for woodland site CA-LP1 of: (A) regression models; and (B) those ML models from which variable influences can be extracted.

giving most priority to variables SWout, H and PPFDout. Although XGB assigns slightly more weight to H, TSF and LWout, and less weight to PPFDout than the other tree-ensemble models. SVR is distinctive from the other ML models in that it assigns more weight to TA, VPDF, NetRad and LWout but less weight to LWinF, SWout and H.

The three best-performing regression models (LASSO, LR and Ridge) applied to the US-Mpj site are, as they are for the CA-LP1 site, generally in agreement with respect to the respective weights given to input variables (Figure 11A). However, distinctively from the CA-LP1 site, those models assign most weight to VPD followed by SWIN, LWIN, TA, SWOUT and P, in that order. LASSO gives more slightly more weight to VPD SWIN, LWIN, TA and SWOUT than the LR and Ridge models. LASSO also assigns no weight to NetRad, PPFDIN, LWOUT and H, in contrast to the other regression models. Nevertheless, the SGDR model, distinctive to the other regression models, assigns substantially higher weights to LWIN, SWOUT, CO_2, LE and H and a much lower weight to SWIN and almost no weight to TA.

In comparison with the best-performing regression models, SVR (Figure 11A) gives more weight to PPFDIN and slightly more weight to TA but lower weights

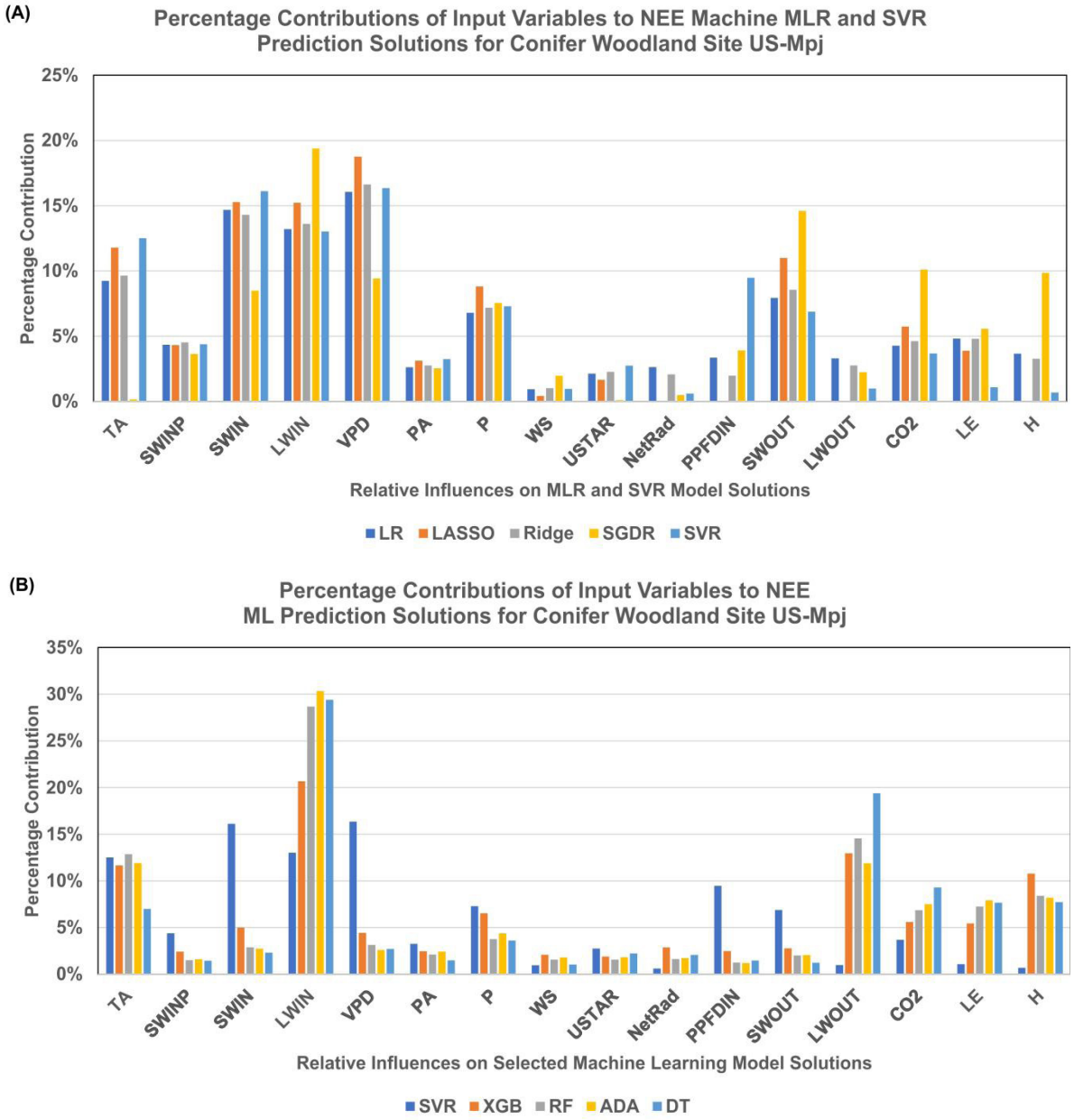

Figure 11. Relative importance of the sixteen influencing variables on the *NEE* predictions for woodland site US-Mpj of: (A) regression models; and (B) those ML models from which variable influences can be extracted.

to SWOUT, LWOUT, LE and H. On the other hand, SVR shows quite different priorities in its input variable weightings to the other ML models evaluated for site US-Mpj (Figure 11B), as was the case for site CA-LP1.

For site US-Mpj the DT shows general agreement with the tree-ensemble models in the weights it assigns the input variables, with slightly higher weights assigned to LWOUT and CO_2 and slightly lower weights assigned to TA and SWOUT (Figure 11B). The three tree-ensemble models (XGB, RF and ADA; Figure 11B) assign substantially higher weights to LWIN than the other variables, with LWOUT, TA, H, LE and CO_2 also being assigned relatively high weights. The XGB model assigns a somewhat lower weight to LWIN and slightly higher weights to H, P and SWIN than the other tree-ensemble models. SVR is distinctive from the other ML models applied to site US-Mpj in that it assigns higher weights to SWIN, VPD, PPDFIN, SWOUT and SWINP but less weight to LWOUT, H, LE and CO_2.

As the SVR model provides the most accurate *NEE* predictions of the models applied to data from both woodland sites considered, from a variable-influence perspective, its weightings should be given the greatest consideration. They suggest that:

- ***For site CA-LP1***: variables TA, SWOUT, NetRad, LWOUT, SWIN and VPD, in that order, are the most influential, whereas variables LWIN, H, G, CO_2, PA and P have very little influence on the SVR solution.
- ***For site US-Mpj***: variables VPD, SWIN, LWIN, TA, PPFDIN, SWOUT and P, in that order, are the most influential, whereas variables H, LE, LWOUT, NetRad and WS have very little influence on the SVR solution. See Figure 1 for variable abbreviation definitions.

Although there are some commonalities regarding variable influences on the SVR solutions generated for the two sites (i.e., high influences by TA, VPD and SWIN and very low influence by H), overall there are substantial differences. In particular, site CA-LP1 assigns relatively high weights to LWOUT and NetRad, whereas site US-Mpj assigns very low weight to those two variables.

5. Discussion: Significance for Ecosystem Assessments

Two evergreen conifer woodland sites from the AmeriFlux dataset incorporating eddy covariance recordings and prepared to comply with FLUXNET2015 protocols provide useful insights regarding variable influences at those sites. The relative degree of importance assigned to a suite of measured variables in the prediction of net ecosystem exchange (*NEE*) weekly-averaged trends over multiple years can be meaningfully ascertained by evaluating four multi-linear-regression (MLR) and seven machine-learning (ML) models. The findings of this study indicate that *NEE* can be predicted with confidence and a relatively high degree of accuracy by the SVR, MLP and XGB models (Figures 7 and 8) considering a large suite of recorded input variables (20 variables for site CA-LP1; 16 variables for site US-Mpj). That result implies that the variables being recorded are sufficient to explain the *NEE* trends observed over multiple years at those two sites. It is an important piece of information to ascertain for all sites recording eddy covariance datasets [28].

Unlike the two woodland sites studied, for some woodland ecosystem sites being monitored to FLUXNET standards, the full suite of recorded variables are not able to reliably predict the observed *NEE* trends over several years. For instance, the same methodology applied in this study has been applied to AmeriFlux sites MX-Tes, a dry temperate deciduous forest in Mexico [82,83], and PE-QFR, a tropical palm swamp in Peru [84] without being able to generate highly accurate predictions with predicted *NEE* versus measured *NEE* values failing to approximate *pNEE=cNEE* relationships. In such situations, it implies that the recorded input variables are insufficient and that there are additional factor(s) influencing *NEE* that are not being recorded. This situation is also the case for many cropland sites for which variables adequately assessing the impacts of tillage and harvests are typically not being recorded [28,85].

In some seasonally dry ecosystems, carbon fluxes can demonstrate time-lag effects, caused by variations in timing and precipitation levels during the wet season, resulting in fluctuating carbon flux responses during subsequent growing seasons [86]. In such situations short-term, daily or weekly, variations in environmental variables are inadequate predictors on their own of *NEE*, as it is being partially affected by certain environmental influence from several months earlier.

With the confidence that an adequate set of input variables are being recorded for the two evergreen-conifer sites studied (CA-LP1/US-Mpj), assessments of variable influences on MLR regression and ML prediction models can be considered a worthwhile exercise. For sites where the models generate poor or moderate *NEE* predictions, with predicted *NEE* versus measured *NEE* trends deviating substantially from *pNEE=cNEE*, that is not the case, as at least some key influential variables are not being taken into consideration.

The results of this study highlight for the two woodland sites considered that multi-linear regression models are only able to generate moderately accurate solutions with the input variables available. The fact that several ML models can generate much more reliable *NEE* predictions for these sites indicates that there is some degree of non-linearity between at least some of the input variables and *NEE*. The MLR models are unable to capture such relationships as these methods rely upon linear relationships between dependent and independent variables. This finding has implications for other woodland sites, particularly those with established conifer stands. Further studies are required to determine whether that is a general feature of all woodland site including deciduous woodlands. Moreover, the role of the understory [87], tree-stand densities [12] and periodic disturbances [88] making substantial *NEE* contributions to the non-linearity of relationships between *NEE* and its influencing variables, although considered likely, requires further investigation.

It is not a surprise that the relative influence of the input variables on the best *NEE* prediction solutions is quite different for the two woodland sites modelled in this study. Although both are conifer woodlands, the sites involve different evergreen tree species and are located in quite different geographic locations (latitudes) and climatic zones. However, the high relative influences of variables TA, VPD and SWIN and very low influence by H at both

sites, associated with the best *NEE* prediction models, are features worth assessing at other woodland sites. A knowledge of the key influencing variables at specific sites from which *NEE* can be predicted is valuable, as it can assist in the understanding the likely impacts of climate change on a site in the future.

Such variable-influence information can be used to establish baseline parameters with which to assess the influences of other dynamic factor potentially at play over time at most woodland sites. There are a substantial number of dynamic factors that could potentially influence woodland sites over the medium and long term. For example climate change and extreme weather events (droughts and wild fires in particular), anthropogenic disturbances (periodic tree harvesting) various kinds of insect and microbial infestations, introduction of invasive understory species, altering stand density and tree species planting mix over time. Understanding how such changes might alter the balance of influence between independent variables and *NEE* could assist in woodland management decisions and the selection of certain protection strategies.

Notwithstanding the importance of knowledge pertaining to the most influential factors at specific woodland sites, the fact that almost all the input variables measured at both sites considered exert some influence on the most accurate *NEE* prediction model solutions suggests that none of these variables should be ignored or disregarded via feature selection processes. It may be tempting to filter out some of the less influential input variables to construct simpler and quicker to implement prediction models involving just a few key variables. However, considering the dynamic factors impacting woodland sites over time, just described, it is possible that variables exerting low influence on *NEE* at prevailing woodland conditions could become much more significant when certain disturbances are introduced.

6. Conclusions

The lodge-pole-pine site (CA-LP1; British Columbia, Canada) evaluated involves twenty environmental variables, and the pinon-juniper site (US-Mpj; New Mexico, U.S.A.) evaluated involves sixteen environmental variables. Weekly-averaged data for these variables recorded over several years can reliably predict net ecosystem exchange (*NEE*) at these two sites. Close similarities in the Pearson (R) vs Spearman (p) correlation coefficients of these variables with weekly *NEE* distributions at both sites indicate that their relationships with *NEE* are at least approximately parametric.

4-fold, 5-fold, 10-fold and 15-fold cross validation analysis of eleven prediction models applied to each site show reliable and reproducible *NEE* prediction performances with relatively low mean and standard deviation mean absolute errors (MAE). Machine learning (ML) models, support vector regression (SVR), extreme gradient boosting (XGB) and multilayer perceptron (MLP), provide substantially more accurate *NEE* weekly predictions than the multi-linear regression (MLR) models evaluated. Predicted *NEE* (*pNEE*) versus measured *NEE* (*cNEE*) distributions at both sites follow *pNEE = cNEE* trends passing approximately through the origin of cross plots of those two variable, whereas they do not for MLR models.

At site CA-LP1 variables air temperature, shortwave radiation outgoing, net radiation, longwave radiation outgoing, shortwave radiation incoming and vapor pressure deficit, in that order, are the most influential, whereas variables longwave radiation incoming, sensible heat, soil heat flux, carbon dioxide in wet air, atmospheric pressure and precipitation have very little influence on *NEE* predictions in the best-performing SVR solution. At site US-Mpj variables vapor pressure deficit, shortwave radiation incoming, longwave radiation incoming, air temperature, photosynthetic photon flux density incoming, shortwave radiation outgoing and precipitation, in that order, are the most influential, whereas variables sensible heat, latent heat, longwave radiation outgoing, net radiation and wind speed have very little influence on *NEE* predictions in the best-performing the SVR solution. At both sites variables air temperature, vapor pressure deficit and shortwave radiation incoming exert high influence on *NEE* predictions, whereas variable sensible heat exerts very low influence.

Comparing Pearson and Spearman correlation coefficients between influential variables and *NEE*, and the *NEE* prediction solutions of a suite of MLR and ML models provides valuable insight to the relative importance of variables in determining weekly *NEE* trends at specific woodland sites.

Conflicts 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 those operating and monitoring the AmeriFlux eddy covariance measurement sites for their meticulous work.

References

[1] Baldocchi, D.D., Hicks, B.B., Meyers, T.P., 1988. Measuring biosphere-atmosphere exchanges of biologically related gases with micrometeorological 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., 2003. Fluxes of carbon, water and energy of European forests. Ecological Studies. pp. 270.
DOI: https://doi.org/10.1007/978-3-662-05171-9

[4] Goulden, M.L., Munger, W., Fan, S.M., Daube, B.C., Wofsy, S.C., 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., Kondo, H., Nishimura, N., 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., Field, C.B., Hall., F.G., 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., Abraha, M., Desai, A.R., Bohrer, G., Arain, M.A., 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] Holtmann, A., Huth, A., Pohl, F., Rebmann, C., Fischer, R., 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

[13] Falge, E., Aubinet, M., Bakwin, P., Baldocchi, D., Berbigier, P., Bernhofer, C., Black, T., et al., 2005. FLUXNET Marconi conference gap-filled flux and meteorology data, 1992-2000. https://catalog.data.gov/dataset/fluxnet-marconi-conference-gap-filled-flux-and-meteorology-data-1992-2000 (Accessed 20th March 2022)

[14] Neog, P., Kumar, A., Srivastava, A.K., Chakravarty, N.V.K., 2005. Estimation and application of Bowen ratio fluxes over crop surfaces - an overview. Journal of Agricultural Physics. 5(1), 36-45.

[15] Yuan, W., Liu, S., Zhou, G., Zhou, G., Tieszen, L.L., Baldocchi, D., Bernhofer, C., 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

[16] Ge, S., Smith, R.G., Jacovides, C.P., Kramer, M.G., Carruthers, R.I., 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

[17] 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

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

[19] Niu, B., He, Y., Zhang, X., Fu, G., Shi, P., Du, M., Zhang, Y., Zong, N., 2016. Tower-based validation and improvement of MODIS gross primary production in an alpine swamp meadow on the Tibetan Pla-

teau. Remote Sensing. 8(7), 592.
DOI: https://doi.org/10.3390/rs8070592

[20] Xu, C., Qu, J.J., Hao, X., Zhu, Z., Gutenberg, L., 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

[21] Zhou, X., Wang, X., Tong, L., Zhang, H., Lu, F., Zheng, F., Hou, P., Song, W., Ouyang, Z., 2012. Soil warming effect on net ecosystem exchange of carbon dioxide during the transition from winter carbon source to spring carbon sink in a temperate urban lawn. Journal of Environmental Sciences (China). 24(12), 2104-2112.
DOI: http://www.ncbi.nlm.nih.gov/pubmed/23534206

[22] Valentini, R., Matteucci, G., Dolman, A.J., Schulze, E.D., Rebmann, C., Moors, E.J., Granier, A., 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

[23] Zhu, S., Clement, R., McCalmont, J., Davies, C.A., Hill, T., 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

[24] Rödig, E., Huth, A., Bohn, F., Rebmann, C., Cuntz, M., 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

[25] 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

[26] Churkina, G., Schimel, D., Braswell, B.H., Xiao, X., 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

[27] Mendes, K.R., Suany Campos, S., da Silva, L.L., Mutti, P.R., Ferreira, R.R., Medeiros, S.S., 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

[28] 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

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

[30] Abbasian, H., Solgia, E., Hosseini, S.M., Kia, H., 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

[31] 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

[32] AmeriFlux, 2022. AmeriFlux Management Project. https://ameriflux.lbl.gov/about/ameriflux-management-project/ (Accessed 20th March 2022)

[33] FLUXNET, 2022. International network of eddy covariance measurement sites. https://fluxnet.org/ (Accessed 20th March 2022)

[34] Kirschbaum, M.U., Mueller, R., 2001. Net ecosystem exchange: workshop proceedings, cooperative research centre for greenhouse accounting. pp. 136. https://www.kirschbaum.id.au/NEE_Workshop_Proceedings.pdf (Accessed 20th March 2022)

[35] Reichstein, M., Falge, E.M., Baldocchi, D.D., Papale, D., Aubinet, M., Berbigier, P., et al., 2005. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. Global Change Biology. 11, 1424-1439.
DOI: https://doi.org/10.1111/j.1365-2486.2005.001002.x

[36] Luyssaert, S., Reichstein, M., Schulze, E-D., Janssens, A., Law, B.E., Papale, 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

[37] Fei, X., Jin, Y., Zhang, Y., Sha, L., Liu, Y., Song, Q., Zhou, W., Liang, N., Yu, G., Zhang, L., Zhou, R., Li, J., Zhang, S., Li, P., 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

[38] 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. Bulletin of the American Meteorological Society. 82(82), 2415-2434.

[39] 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

[40] Ameriflux, 2022. Flux/met data processing pipeline overview. https://ameriflux.lbl.gov/data/data-processing-pipelines/ (Accessed 20th March 2022).

[41] Ameriflux, 2022. Data variable descriptions for the FLUXNET product. https://ameriflux.lbl.gov/data/aboutdata/data-variables/ (Accessed 20th March 2022).

[42] Black, T.A., 2021. AmeriFlux FLUXNET-1F CA-LP1 British Columbia - Mountain pine beetle-attacked lodgepole pine stand. AmeriFlux AMP, (Dataset).
DOI: https://doi.org/10.17190/AMF/1832155

[43] Brown, M., Black, T.A., Nesic, Z., Foord, V.N., Spittlehouse, D.L., Fredeen, A.L., Grant, N.J., Burton, P.J., Trofymow, J.A., 2010. Impact of mountain pine beetle on the net ecosystem production of lodgepole pine stands in British Columbia. Agricultural & Forest Meteorology. 150(2), 254-264.
DOI: https://doi.org/10.1016/j.agrformet.2009.11.008

[44] Litvak, M., 2021. AmeriFlux FLUXNET-1F US-Mpj Mountainair Pinyon-Juniper Woodland. AmeriFlux AMP, (Dataset).
DOI: https://doi.org/10.17190/AMF/1832161

[45] Morillas, L., Pangle, R.E., Maurer, G.E., Pockman, W.T., McDowell, N., Huang, C., Krofcheck, D.J., Fox, A.M., Sinsabaugh, R.L., Rahn, T.A., Litvak, M.E., 2017. Tree mortality decreases water availability and ecosystem resilience to drought in piñon-juniper woodlands in the southwestern U.S. Journal of Geophysical Research: Biogeosciences. 122(12), 3343-3361.
DOI: https://doi.org/10.1002/2017JG004095

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

[47] 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

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

[49] Boddy, R., Smith, G., 2009. Statistical Methods in Practice: For scientists and technologists. Chichester, U.K.: Wiley. pp. 95-96.

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

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

[52] 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

[53] Harrell, F.E., 2015. Regression Modeling Strategies. Second Edition. Springer, Switzerland. pp. 582.
DOI: https://doi.org/10.1007/978-3-319-19425-7

[54] Goldberger, A.S., 1964. Classical linear regression. Econometric Theory. New York: John Wiley & Sons. pp. 158.

[55] Stigler, S.M., 1981. Gauss and the Invention of Least Squares. Annals of Statistics. 9(3), 465-474.
DOI: https://doi.org/10.1214/aos/1176345451

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

[57] SciKit Learn, 2022. Linear models. https://scikit-learn.org/stable/modules/linear_model.html (Accessed 20th March 2022).

[58] SciKit Learn, 2022. Supervised and unsupervised machine learning models in Python. 2022a. https://scikit-learn.org/stable/ (Accessed 20th March 2022).

[59] 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

[60] 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

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

[62] Debeljak, M., Džeroski, S., 2011. Decision trees in ecological modelling. Modelling complex ecological dynamics. Springer, Berlin, Heidelberg. pp. 197-209.

[63] Fix, E., Hodges Jr., J.L., 1951. Discriminatory analysis, nonparametric discrimination: consistency properties. Technical Report, USAF School of Aviation Medicine.

[64] Fu, Y., He, H.S., Hawbaker, T.J., Henne, P.D., Zhu, Z., Larsen, D.R., 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

[65] Rosenblatt, F., 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Cornell Aeronautical Laboratory. Psychological Review. 65(6), 386-408.
DOI: https://doi.org/10.1037/h0042519

[66] Eshel, G., Dayalu, A., Wofsy, S.C.C., Munger, J.W., Tziperman, E., 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

[67] Safa, B., Arkebauer, T.J., Zhu, Q., Suyker, A., Irmak, S., 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

[68] 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

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

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

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

[72] IIlie, I., Dittrich, P., Carvalhais. N., Jung, M., Heinemeyer, A., Migliavacca, M., Morison, J.I.L., Sippel, S., Subke, J.A., Wilkinson, M., Mahecha, M.D., 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

[73] Li, Z., Chen, C., Nevins, A., Pirtle, T., Cui, S., 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

[74] Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In Krishnapuram, Balaji; Shah, Mohak; Smola, Alexander J.; Aggarwal, Charu C.; Shen, Dou; Rastogi, Rajeev (eds.). Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA. ACM. pp. 785-794.
DOI: https://doi.org/10.1145/2939672.2939785

[75] Yan, S., Wu, L., Zhang, F., Zou, Y., Wu, Y., 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

[76] Liu, J., Zuo, Y., Wang, N., Yuan, F., Zhu, X., Zhang, L., Zhang, J., Sun, Y., Guo, Z., Guo, Y., Song, X., Song, C., Xu, X.F., 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

[77] SciKit Learn. 2022 GridSearchCV: Exhaustive search over specified parameter values for an estimator in Python. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (Accessed 20th March 2022).

[78] SciKit Learn. 2022 Bayesian optimization of hyperparameters in Python.. https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html (Accessed 20th March 2022).

[79] SciKit Learn. 2022 Cross-validation: evaluating estimator performance. https://scikit-learn.org/stable/modules/cross_validation.html (Accessed 20th March 2022).

[80] Gini, C., 1997. Concentration and dependency ratios (published 1909 in Italian). English translation in Rivista di Politica Economica. 87, 769-778.

[81] Guillermina, J., 1979. On Gini's Mean Difference and Gini's Index of Concentration. American Socio-

logical Review. 44(5), 867-870.
DOI: https://doi.org/10.2307/2094535

[82] Verduzco, V.S., Garatuza-Payán, J., Yépez, E.A., Watts, C.J., Rodríguez, J.C., Robles-Morua, A., Vivoni, E.R., 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

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

[84] Griffis, T., Roman, D., Wood, J., Deventer, J., Fachin, L., Rengifo, J., Del Castillo, D., Lilleskov, E., Kolka, R., Chimner, R., del Aguila-Pasquel, J., Wayson, C., Hergoualc'h, K., Baker, J., Cadillo-Quiroz, H., Ricciuto, D., 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

[85] Schulze, E.D., Valentini, R., Bouriaud, O., 2021. The role of net ecosystem productivity and of inventories in climate change research: the need for net ecosystem productivity with harvest (NEPH). Forest Ecosystems. 8, 15.
DOI: https://doi.org/10.1186/s40663-021-00294-z

[86] Cable, J., Ogle, K., Barron-Gafford, G., Bentley, L., Cable, W., Scott, R., Williams, D., Huxman, T., 2013. Antecedent conditions influence soil respiration differences in shrub and grass patches. Ecosystems. 16, 1230-1247.

[87] Wiesner, S., Staudhammer, C.L., Javaheri, C.L., Kevin Hiers, J.K., Boring, L.R., Mitchell, R.J., Starr, G., 2019. The role of understory phenology and productivity in the carbon dynamics of longleaf pine savannas. Ecosphere. 10(4), e02675.
DOI: https://doi.org/10.1002/ecs2.2675

[88] Matusick, G., Hudson, S.J., Garrett, C.Z., Samuelson, L.J., Kent, J.D., Addington, R.N., Parker, J.M., 2020. Frequently burned loblolly-shortleaf pine forest in the southeastern United States lacks the stability of longleaf pine forest. Ecosphere. 11(2), e03055.
DOI: https://doi.org/10.1002/ecs2.3055