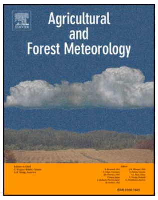
Agricultural and Forest Meteorology 353 (2024) 110038

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet



Toward spectrally truthful models for gap-filling soil respiration and methane fluxes. A case study in coastal forested wetlands in North Carolina

Bhaskar Mitra[a,b,*], Kevan Minick[c], Michael Gavazzi[d], Prajaya Prajapati[e], Maricar Aguilos[f], Guofang Miao[g], Jean-Christophe Domec[c,h], Steve G. McNulty[d], Ge Sun[d], John S. King[f], Asko Noormets[a]

[a] Department of Ecology and Conservation Biology, Texas A&M University College Station, TX, United States
[b] Information and Computational Sciences Department, The James Hutton Institute, Aberdeen, UK
[c] Nicholas School of the Environment, Duke University, Durham, NC, USA
[d] Eastern Forest Environmental Threat Assessment Center, USDA Forest Service, Research Triangle Park, North Carolina, USA
[e] Campbell Scientific, Logan, UT, United States
[f] Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, North Carolina, USA
[g] School of Geographical Sciences, Fujian Normal University, Fuzhou, Fujian Province, China
[h] Bordeaux Sciences Agro, UMR INRA-ISPA 1391, Gradignan 33195, France

ARTICLE INFO

Keywords:
Gap filling
Machine learning
Methane
Soil respiration
Wavelet filtering

ABSTRACT

Soil respiration (*Rs*) and methane (F_{CH4}) fluxes are two important metrics of ecosystem metabolism. An accurate estimate of the budget of these two greenhouse gases is critical to understanding their response to climate and land-use changes. Reconstructing continuous time series of gappy chamber *Rs* and eddy-covariance derived F_{CH4} measurements is usually done based on correlative relationships of these fluxes with environmental variables. However, current approaches do not account for the fact that different environmental drivers affect the carbon fluxes at different temporal scales. Here we propose a novel gapfilling technique that accounts for the specific spectral frequencies at which each of the environmental variables covaries with *Rs* and F_{CH4} - photosynthetically active radiation at diel scale, soil temperature at synoptic scale, and soil moisture, water table depth and atmospheric pressure at synoptic and seasonal scale. The method was applied on two operational loblolly pine plantations of different ages and a mixed hardwood forested wetland on the lower coastal plain of North Carolina. The time series of these environmental drivers were reconstructed using wavelet decomposition and a Daubechies wavelet filter. Further, to consider the joint influence of the environmental drivers, parametric (elastic net regression, support vector machine, gradient boost and artificial neural network), and nonparametric (Bayesian) statistical models were chosen, and compared the results with Q_{10} and Marginal Distribution Sampling (*MDS*) outputs. In all cases, the algorithms were trained on 70 % of the data and validated with the remaining data. Spectral-filtered models did not significantly differ from those driven by unfiltered data with respect to *Rs* and F_{CH4} predictions. While all the spectrally driven algorithms achieved high predictive accuracy against Q_{10}, the increase in model fit compared to *MDS* was minimal. Spectral data filtering modestly improves model accuracy, shedding light on complex environmental and biological factors affecting greenhouse gas flux variability.

1. Introduction

Accurate future climate projections require the predictive capability of land surface biogeochemistry, including the carbon cycle (Jones et al., 2003; Meissner et al., 2021). Soil respiration (*Rs*), the second largest carbon flux to the atmosphere, is currently represented by both statistical and process-based approaches in ecosystem carbon cycle models (Lloyd and Taylor 1994; Turetsky et al., 2014; Parton et al., 1987; Li et al., 1992). The former incorporate either a simple exponential response to temperature function (Arrhenius, 1889; Jones et al., 2003) or more detailed data-oriented approach that integrates various biotic and abiotic drivers, including soil properties, temperature, moisture and carbon substrate availability (Bond-Lamberty & Thomson,

* Corresponding author at: Information and Computational Sciences Department, The James Hutton Institute, Aberdeen, Scotland, UK.
E-mail address: Bhaskar.Mitra@hutton.ac.uk (B. Mitra).

https://doi.org/10.1016/j.agrformet.2024.110038
Received 2 July 2023; Received in revised form 11 April 2024; Accepted 30 April 2024

List of symbols

S_4	approximation coefficient
P (kPa)	atmospheric pressure
T_a (°C)	air temperature
P_a (kPa)	atmospheric pressure
$p(x_i\|y_j)$	conditional probabilities
d1, d2, d3, d4, d5, d6, d7, d8, d9, d10, d11, d12	detailed coefficients
LE (W m^{-2})	ecosystem-scale latent energy
$H(x)$	entropy of variable x
$H(y)$	entropy of variable y
$J(x,y)$	joint entropy of x and y
$p(x_i,y_j)$	joint probabilities
CH_4 (umol m^{-2} s^{-1})	methane flux
F_{CH4} (umol m^{-2} s^{-1})	net ecosystem CH_4 exchange
NEE_{CO2} (umol m^{-2} s^{-1})	net ecosystem CO_2 exchange
p	probability distribution of measured data
q	probability distribution of modeled data
J_s, (g m^{-2} s^{-1})	sap flow velocity
T_{s5} (°C)	soil temperature at 5 cm
WTD (cm)	water table depth

2010; Warner et al., 2019; Tang et al., 2020). The latest process-based models include microbial processes interacting with soil microenvironment, roots and one another, aiming for improved representation of temporal dynamics and future projection capability compared to statistical models (Wieder et al., 2015; Sulman et al., 2017; Robertson et al., 2019). Yet, the improvement in large-scale and long-term projections due to the incorporation of microbial processes remains to be elucidated (Wieder et al., 2019).

As we recently demonstrated, the interactive effects of different biotic and abiotic influences on *Rs* can statistically be separated in the temporal domain (Mitra et al., 2019). It emerged that the causal influence of environmental factors affected the synoptic and seasonal dynamics of *Rs*, whereas diurnal dynamics appeared controlled by plant carbohydrate status. Notably, in a follow-up analysis, we also found that methane flux (F_{CH4}), which is exclusively a microbial process, exhibited similar co-spectral signatures with environmental drivers as did *Rs* (Villa et al., 2019, Mitra et al., 2020). Earlier studies have also observed temporal offsets between F_{CH4} and its presumed drivers (Chang et al., 2021; Sturtevant et al., 2016; Sturtevant et al., 2012; Vizza et al., 2017). In their 2021 meta-analysis, Knox et al. identified the primary factors influencing F_{CH4}, noting that soil and air temperature, water table depth, and atmospheric pressure exhibit variations across diel, synoptic, and seasonal timescales. Many F_{CH4} simulation studies do not consider the temporal domain of environmental factors, even though Kim et al., (2020) explored the impacts of time-lagged controls on F_{CH4} gap filling and considered net ecosystem exchange as a proxy of carbon substrate related to vegetation processes. Resolving these knowledge gaps is critical, especially the coupling between plant and soil processes, as it is relevant for both future projection models, as well as filling inevitable gaps in observation time series (Aubinet et al., 2012).

The statistical evidence of the significance of plant-microbe interactions for both CO_2 and CH_4 emissions is significant in that it is currently not explicitly considered in the ecosystem and Earth system models and is not guaranteed to naturally arise from the inclusion of microbial processes in the above-cited models. It is possible that this omission contributes to the large remaining uncertainties about *Rs* and F_{CH4} trajectories to future climate projections in global land surface models (Anav et al. 2013; Ciais et al., 2014; Friedlingstein et al. 2006; Hoffman et al. 2014).

Moving from partitioning the spectral signatures of different effects to practical application is the next logical step. The statistical nature of our original spectral analysis makes it suitable for gapfilling discontinuous *Rs* and F_{CH4} data when the time series of the drivers are available. Current community standard gap-filling models – different temperature response functions (Lloyd and Taylor, 1994; Fang and Moncrieff, 2001), marginal distribution sampling (Reichstein et al. 2005) and artificial neural networks (Moffat et al. 2010) – all fall in the statistical category. They have been rigorously evaluated and they tend to perform reliably in all biomes (Falge et al. 2001; Moffat et al. 2007; Pastorello et al. 2020). The basic approach has been to select meteorological variables (e.g., air temperature, soil moisture, incoming shortwave radiation, wind speed) or flux variables (e.g: sensible heat and latent energy) that are consistently measured along with the *GHGs*' and then use them as forcing parameters, even though they may not have a causal relationship with the fluxes (Reichstein et al., 2019).

To that end, we propose a novel gap-filling and simulation procedure for estimating *Rs* and F_{CH4} fluxes by building on the emerging conceptual understanding of scale-dependence of the controls of the two greenhouse gases (Mitra et al., 2019, 2020) by accounting for the main periodic components of individual drivers and maintaining the spectral information of individual variables and their relationships. The objective of this research was to systematically evaluate the effectiveness of simulating *Rs* and F_{CH4} with spectrally filtered input data and evaluate model performance against two community standards driven by measured inputs: i) Marginal Distribution Sampling (*MDS*; Reichstein et al. 2005; Wutzler et al. 2018; Pastorello et al. 2020) and (ii) simple Q_{10} model (Lloyd & Taylor 1994). Secondarily, we evaluated the performance of an ensemble of machine learning models (Elastic Net regression, Support Vector Machine, XGBoost, Random Forest, and ANN) and Bayesian workflows that are agnostic to data structure and normality and do not make a-priori assumptions about the linearity of the relationships among independent and dependent variables. To our knowledge, the incorporation of spectral information in the simulation of biogeochemical processes remains scarce (but see Bakhshian and Romanak, 2021) and can potentially elevate the predictive capabilities of *Rs* and F_{CH4} simulation by moving beyond ubiquitous correlation techniques and emphasizing causal relations.

2. Methods

2.1. Site description

The data used for this study were collected at three Ameriflux eddy covariance sites in North Carolina, USA: *Rs* data at US-NC1 and US-NC2 (Noormets, 2018; Noormets et al., 2022a), and ecosystem net CH_4 exchange (F_{CH4}) at US-NC4 (Noormets et al., 2022b). The vegetation, soil and climate characteristics, and instrumentation have been described in detail elsewhere (Aguilos et al. 2020, 2021). To summarize, the main salient features are as follows. The sites are in the lower coastal plain in North Carolina (35.8°, -76.7°). The climate is subtropical maritime, with a mean annual temperature of 17 °C and mean annual precipitation of 1300 mm. The soils are deep histosols with a shallow water table (<1.5 m). The US-NC1 and US-NC2 sites are operational loblolly pine (*Pinus teada* L.) plantations, managed on a 25-year rotation, planted at 1300 trees per acre, thinned at around age 10, and fertilized both at the establishment and after thinning. The sites are drained with a network of ditches (100 m spacing), keeping the forest from flooding after harvest. US-NC4 is a natural bottomland forested wetland (*Nyssa-Taxodium* type) within the Alligator River National Wildlife Refuge in Dare County on the Albermarle-Pamlico peninsula, with similar soil and climate conditions as the other sites, except for minimal draining and lower elevation, resulting in approximately 8-month hydroperiod (Miao et al., 2017). Eddy covariance and chamber flux measurements were conducted from 2005-2012 at US-NC1, 2005-present at US-NC2, and 2009-present at US-NC4. The data used in the current study spanned years 2009–2011 in US-NC1 (tree age 5–7 years), 2010–2014 in US-NC2 (tree age 18-22 years), and 2013–2016 in US-NC4 (average tree age ~100 years). Data selection was determined by data coverage and continuity (e.g., instrument availability), interfering management activities (e.g., thinning at US-NC2), and weather events (e.g., hurricanes and storms at US-NC1 and US-NC4; Aguilos et al., 2020, 2021).

2.2. Measurements

Rs was measured continuously by the eddy covariance instrument towers of US-NC1 and US-NC2 on polyvinyl chloride collars using an automated soil CO_2 efflux chamber (LI 8100-104, LICOR Inc). The spatial representativeness of the continuous single-chamber measurements was confirmed against stand-wide survey measurements (5 plots with 5 collars each, within 300 m of the instrument tower; Noormets et al., 2012). At each tower, and adjacent to the *Rs* collars, soil temperature (T_S; CS107 Campbell Scientific (CSI), USA) and moisture (*q*; CS616, CSI) were measured at 0–5 and 0–30 cm (integrated) depth, respectively. All sites were also instrumented for eddy covariance measurements of CO_2 and H_2O fluxes (LI-7500; Licor, USA) and a CSAT-3 (CSI; US-NC1 and US-NC2) or WindMaster (Gill, UK; US-NC4)), as well as for air temperature (*T*a; HMP45AC, Vaisala, Finland), photosynthetically active radiation (*PAR*; PARLITE, Kipp & Zonen, Netherlands), and precipitation (P; TE525MM, Texas Electronics, USA). US-NC4 also included an open-path analyzer for F_{CH4} (LI-7700, Licor) measurements. Data processing of the 10 Hz data using Eddypro (v 6.1.0; LICOR) and post-processing of the 30 min fluxes has been documented elsewhere (Miao et al., 2017; Mitra et al., 2020). All fluxes, as well as meteorological data, were computed at a 30-minute resolution (e.g., Aguilos et al., 2020; 2021). The three years chosen from each site (US-NC1: 2009, 2010, 2011; US-NC2: 2010, 2013, 2014; US-NC4: 2013, 2015, 2016) had gaps in the meteorological time series which represented less than 30 % of the total data collected. While short gaps (~ hours) were filled using linear interpolation, longer gaps were filled using the *MDS* approach (Falge et al., 2001).

2.3. Wavelet denoising

Wavelet denoising incorporated wavelet transformation, whereby a wavelet function was partitioned to multiple sub-signals by adjusting the scaling and shifting parameters of the mother wavelet function. Wavelets are mathematical functions with a wave shape that partitions a time series into different frequency components which are then used to localize a given function in both position and scale. Sub-signals with two different types of coefficients were produced: detailed and approximation (*S*). The former is a high-frequency (or noise or rapidly changing) component of the original signal, while the latter represents slow changing (or residual or trend) aspect of the original signal. The signal decomposition process was repeated multiple times to generate numerous high-resolution components (Drago and Boxall, 2002; Chou, 2007; Naley et al., 2012, 2013; Joshi et al., 2016; Liu and Menzel, 2016).

For this study, the non-decimated discrete wavelet transformation with the Daubechies orthonormal compactly supported wavelet (Length = 8) basis function acted as low and high pass filters. Daubechies (db) wavelet is advantageous as it has compact support and orthogonal properties (Vonesch et al., 2007; Nalley et al., 2012; 2013), thereby allowing it to localize signals (Liu and Menzel, 2016). The Daubechies orthonormal compactly supported wavelet of length 8 was chosen to provide a trade-off between time and frequency localization. A lower value would have provided poorer frequency localization, while longer filters provide better frequency localization but may have poorer time localization. The environmental time-series signals were decomposed by a factor of 2 (dyadic degradation) into several detailed hierarchical levels as shown below (Mallat 1989; Daubechies 1990; Daubechies 1992). The detailed (d1 – d12) wavelet vectors represented different time scales, which can broadly be classified under four broad categories:

- Hourly scale: d1 (0.5 h), d2 (1 h), d3 (2 h)
- Diel scale: d4 (4 h), d5 (8 h), d6 (16 h), d7 (1.33 days)
- Synoptic scale: d8 (2.66 days), d9 (5.33 days), d10 (10.67 days), d11 (21.33 days)
- Seasonal scale: d12 (42.67 days).

In the studies by Mitra et al. (2019), (2020), distinct environmental drivers were identified for soil respiration (*Rs*) and methane flux (F_{CH4}). Both *Rs* and F_{CH4} demonstrated statistically significant diurnal cospectral peaks correlating with Photosynthetically Active Radiation (*PAR*), suggesting a link with carbohydrate availability via Gross Primary Productivity (*GPP*). Additionally, they exhibited synoptic and seasonal peaks associated with various factors: soil temperature (*Ts*) influenced both *Rs* and F_{CH4}; atmospheric pressure (*P)*, water table depth (*WTD)* and soil moisture (*q*) specifically impacted F_{CH4} and soil moisture also affected *Rs*.

The time series of the environmental drivers of *Rs* and F_{CH4} were therefore reconstructed with only the dominant and statistically significant wavelet coefficients and added to the *S* coefficient (Fig. 1, Supp. Figs. S6, S7, Liu et al., 2019). The other wavelet coefficients were set at 0. The reconstructed time series, expressed as variance in the selected energetic wavelet coefficients, can be expressed as follows (Renaud et al., 2005):

$$PAR_r \ = d4 + d5 + d6 + d7 + \ S \tag{1}$$

$$T_{sr} \ = d8 + d9 + d10 + d11 + \ S \tag{2}$$

$$\theta_r \ = d8 + d9 + d10 + d11 + d12 + S \tag{3}$$

$$WTD_r \ = d8 + d9 + d10 + d11 + d12 + S \tag{4}$$

$$P_r \ = d8 + d9 + d10 + d11 + d12 + S \tag{5}$$

The multiresolution wavelet partition of the environmental drivers to demonstrate the signal decomposition process has been highlighted in Supplementary Figs. S1–S5. Due to space limitations, results from every site and year are not shown. From a statistical perspective, retaining the dominant frequencies is akin to denoising or band-pass filtering. Thus, these meteorological sub-series generated by wavelet transformation were used as input variables for *Rs* (T_{sr}, PAR_r, q $_r$*)* and F_{CH4} (T_{sr}, PAR_r, P_r and WTD_r*)* and hybridized with machine learning, *MDS* and Bayesian algorithms for improved prediction of the two greenhouse gases.

2.4. Gapfilling algorithms

The original ("measured") and decomposed ("spectrally filtered") signal components were then used as drivers to gap-fill R_S and F_{CH4} using different algorithms. Wherever the decomposed signals (T_{sr}, PAR_r, q $_r$, P_r and WTD_r*)* have been used as input, the corresponding gap-filled model has been referred to as a spectral model, and the approach is referred to as a scalar-based approach. Gap-filled model using unfiltered measured data (T_s, *PAR, q, P* and *WTD)* has been referred as measured-data based approach. The different algorithms analyzed in this study have been summarized below.

2.4.1. Q_{10} model

A nonlinear regression approach was used to gap-fill the half-hourly measurements of a carbon flux (Fx = either R_S or F_{CH4}) using a Q_{10} function with T_s as shown below:

$$F_x \ = \ R_{ref} Q_{10}^{(T_s \ -10)/10} \tag{6}$$

The model parameters were determined using the Ordinary Least Squares (*OLS*) algorithm, and the optimum parameters were chosen to minimize the Residual Sum of Squares (*RSS*).

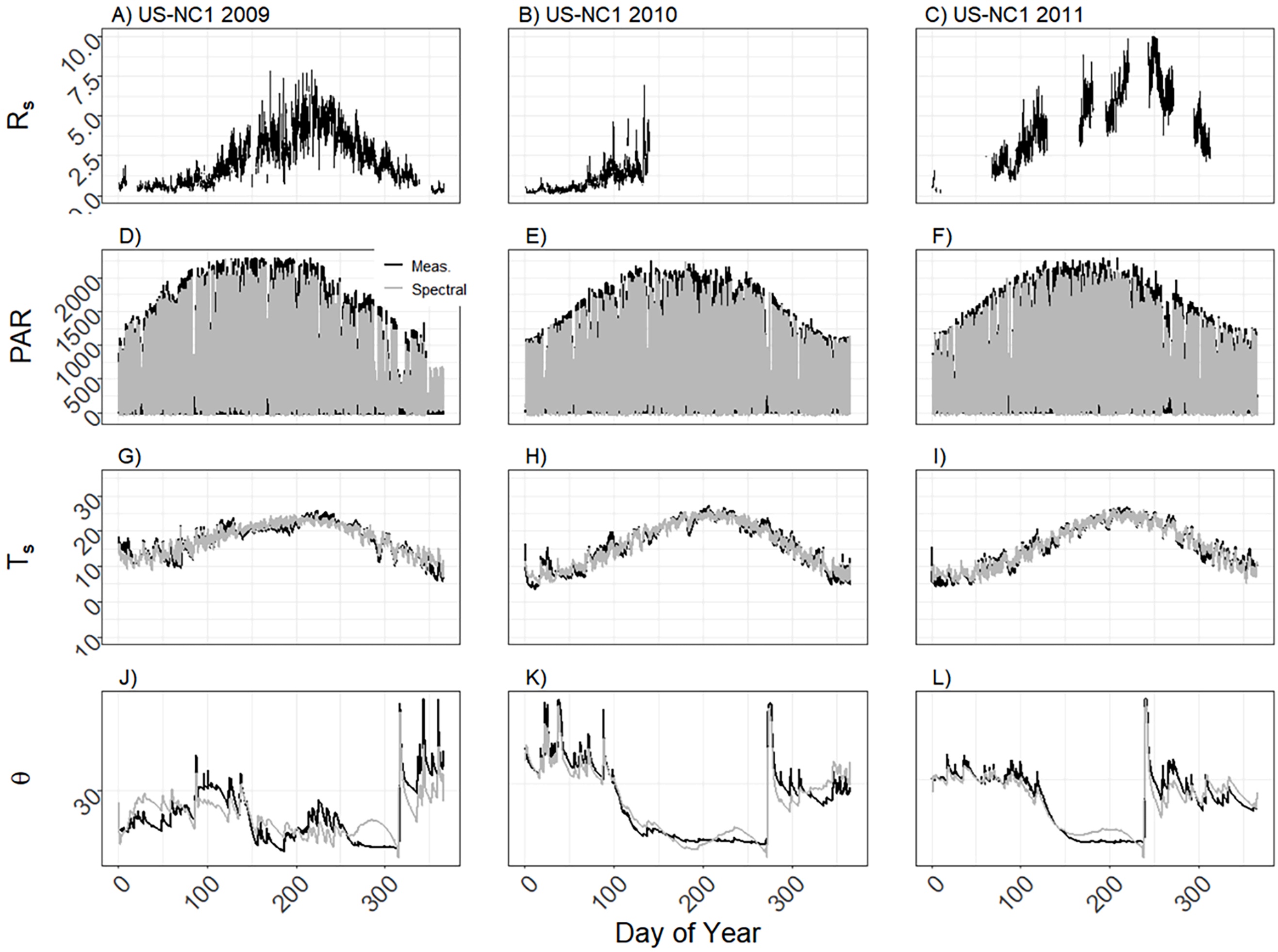


Fig. 1. 30 min. temporal sequences of original soil respiration (*R*s, umol m^{-2}s^{-1}, A–C) as well as original (Meas.) and spectrally reconstructed (Spectral) time series of Photosynthetically Active Radiation (PAR, umol m^{-2} s^{-1}, D–F), soil temperature (Ts, °C, G–I) and soil moisture (θ, %, J–L) for US-NC1 for three different years.

2.4.2. *Marginal distribution sampling (MDS)*

MDS is based on the look-up table and mean diurnal curve approach. Spectral (only for *R*s) and non-spectral approaches were incorporated, whereby any single gap in the flux (*R*s) were replaced by average flux values within a certain window frame and based on similar meteorological conditions (REddyProc; Reichstein et al., 2005; Wutzler et al., 2018).

2.4.3. *Elastic net regression*

A penalized multivariate variant of Ordinary Least Squares regression (*OLS*), elastic net regression is ideal when the input data set have high multicollinearity between predictor variables. Mathematically, the cost function of elastic net regression can be expressed as follows:

$$\left(RSS+\,1\left[(1-\mathrm{a})\left(\mathrm{S}|\mathrm{b}j|^{2}\right)\Big/2\;+\;\mathrm{a}*(\mathrm{S}|\mathrm{b}j|)\right]\;\Big/\mathrm{N}\right) \tag{7}$$

Where β*j* refers to coefficients, λ is the shrinkage parameter used to balance the fit of data and magnitude of coefficients, *N* is the number of observations, and *RSS* is the residual sum of squares. To make the estimated parameters close to the "true" parameters, improve model accuracy, and account for multicollinearity, a mixing parameter α is incorporated (Zou and Hasie, 2005).

2.4.4. *Support vector machine (SVM)*

Ideal for nonlinear systems, the regression version of Support Vector Machine (SVM) is a nonparametric machine learning algorithm that defines boundaries in a high dimensional sub-space using a hyperplane (Smola and Schölkopf, 2004; Chapelle et al., 2002; Birzhandi et al. 2019). A line is used to separate classes for two-dimensional space, while a plane is used for three-dimensional space. The perpendicular distance of observations from the hyperplane is calculated, and the minimum distance value is identified as the margin. The hyperplane with the maximum value of the margin is chosen. The points which fall within the margin are known as support vectors. A misclassification budget (*B*) is established, whereby the algorithm limits the sum of distances of the points on the wrong side of the margin. The hyperparameter, *c* (μ (1/*B*)), is chosen so that it will both prevent overfitting of the measured data and prevent too many misclassified observations. The radial kernel was used to establish nonlinear boundaries between two observations. SVM is scale sensitive, and hence all datasets were scaled to have a mean of 0 and a standard deviation of 1.

2.4.5. *Ensemble methods*

The basic building block of ensemble methods are decision trees. Decision trees conduct a recursive partition of the space defined by the independent variables to smaller regions. Features are partitioned in such a way that the residual sum of squares (*RSS*) improves the most. Subsequent splits are not performed on the entire set but on the prior split. To control the split from leading to unnecessary branches, one needs to prune the tree to an optimal size to prevent low bias and high variance. Based on this decision tree algorithm, two ensemble methods were adopted: Extreme Gradient Boosting (*XGBoost*) and Random Forest (*RF*).

The *XGBoost* algorithm uses the predictions from a set of "weak" learners to create a strong "learner", ensuring no overfitting. In addition, the objective function for *XGBoost* incorporates predictive and regularization terms and incorporates optimal computational efficiency by automatically allowing parallel calculations (Friedman, 2001; Chen et al., 2015). The development of multiple decision trees using the bagging ensemble methodology is the random forest (*RF*) algorithm (Breiman, 2001). First, data is randomly partitioned into the training and testing component at a fixed rate. Next, the model is initialized by generating many trees (parameter = *ntree*) using the bootstrap approach, which then leads to the development of a large number of decision subtrees. Those datasets (approx.: 1/3rd) which are not included in this step are referred to as "out-of-bag". Then, only a subset (parameter = *mtry*) of the predictor variables is selected randomly and tested on the training dataset. Finally, the optimal model developed from the training dataset is evaluated on the test dataset.

2.4.6. *Artificial neural networks (ANN)*

Compared to linear regression, *ANN* is an alternative nonlinear parametric procedure that will have less bias and low values in the estimation errors (Dayhoff 1990; Ripley 1996; Bishop 1995; Diamantopoulo 2005). *ANN* mimics human brain processing and is a parallel distributed processor consisting of multiple components, including a set of connected nodes (artificial neurons) that form the input layer, multiple hidden layers, output layer, weights schemes, thresholds, and transfer activation functions. *ANN*s assume no prior assumption about the relationship between independent and dependent variables and are capable of quantifying complex nonlinear processes. However, the above-mentioned relation remains "hidden" within the neural architecture, thereby not allowing the relationship to be expressed mathematically. Other advantages of *ANN* include high fault tolerance that would ensure the reliability of the system's input-output relation, efficient parallel interconnectedness, and efficient storage and processing of information using a distributed system. The trained *ANN* was a multilayered perceptron with a backpropagation training algorithm for this study. Three different values of the hidden layers were experimented with in this study (Supp. Table 1). Training and testing of *ANN* were implemented in *R* using the "*Neuralnet*" package (Günther and Fritsch, 2010).

2.4.7. *Heterogeneous ensemble methods*

A simple weighted linear combination of *Elastic net regression, SVM, XGBoost, RF,* and *ANN* were integrated to form an ensemble model. The outputs from the ensemble model were combined using a linear greedy stacking optimization strategy to generate the final output. The optimization strategy selects the most optimal output at each time step. The ensemble model was implemented in R using the '*caretEnsemble*' package (Deane Mayer and Knowles, 2019).

2.4.8. *Bayesian regression*

A Bayesian framework (Bayes, 1763) with the underlying assumption that errors are independent and normally distributed was used to quantify R_S and F_{CH4}. Probabilistically, the single-level regression model can be expressed as follows:

$$R_s = \beta_o + \beta_1 * PAR_r + \beta_2 * T_{sr} + \beta_3 * \theta_r \tag{8}$$

$$F_{CH4} = \beta_o + \beta_1 * PAR_r + \beta_2 * T_{sr} + \beta_3 * WTD_r + \beta_4 * P_r \tag{9}$$

The parameters β_o, β_1, β_2, β_3, β_4, are the intercept and regression coefficients corresponding to the decomposed signal components of the associated environmental drivers. With the addition of the error term (ϵ_i), the likelihood function for Rs and F_{CH4} can be expressed as follows:

$$R_{si} \sim (\beta_o + \mathrm{X}_i * \beta + \epsilon_i) \tag{10}$$

$$F_{CH4i} \sim (\beta_o + \mathrm{X}_i * \beta + \epsilon_i) \tag{11}$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{12}$$

R_{si} (& F_{CH4i}) are the elements temporally corresponding to the decomposed environmental drivers X_i (PAR_r, T_{sr}, θ_r, WTD_r, P_r) and errors ϵ_i are normally distributed with mean 0 and variance σ^2 . Eqs. (8)–(12) were also adopted for the unfiltered measured data. Weakly informative priors corresponding to the parameters were used to prevent overfitting (Muth et al., 2018). The priors and likelihood functions were combined to form the joint posterior distributions, sampled using Markov Chain Monte Carlo (MCMC) simulations. The weakly informative prior adopted in this paper has the advantage of reduced uncertainty in the posterior distribution of the parameters and reduces the high computational cost of running MCMC simulations (Muth et al., 2018).

Four MCMC chains were run with 5000 iterations each, with the first 2000 iterations discarded as they comprised the "warm-up" period. Chain convergence was analyzed visually by analyzing the trace plots (Kass et al., 1998) and calculating the scale reduction factor (Gelman and Rubin, 1992). Convergence was determined when the scale reduction factor value was less than 1.2 (Muth et al., 2018) and the sample size was well above 1000 (Muth et al., 2018). Model fit was analyzed using a graphical posterior predictive model check, whereby the measured data (R_s and F_{CH4}) were compared against the histogram of the corresponding posterior predictive distribution. Posterior estimates of β and ϵ_i were summarized by the 95 % credible interval (C.I) (Supp. Tables 3–5). Simulating draws from the posterior predictive distribution were used to generate predictions for the missing observations and the corresponding 95 % C.I. The Bayesian regression model was implemented with the *R* package "*rstanarm*" (Goodrich et al., 2020). The *rstanarm* package incorporates prior predictive rescaling, whereby the prior scales of the intercept, coefficients, and the error are internally rescaled by the standard deviation of the errors, thereby adjusting the posterior range. Thus, the posterior range of the predictors would appear to be different from the absolute values of the predictor variables (Supp. Tables 3–5).

2.5. *Hyperparameters*

Another key aspect is hyperparameters or tuning parameters, which express the complexity of the different machine-learning algorithms (Elastic net regression, SVM, XGBoost, RF, and ANN). The hyperparameters for each algorithm were not fixed. Still, they were given a range of values (Supp. Table 1), and hyperparameter values were determined using grid search (hyperparameter optimization technique) in conjunction with a 10-fold cross-validation sampling strategy. The training dataset was randomly partitioned into nine training and one validation dataset in this approach. The model was trained on the training component of the dataset and validated against the validation component of the dataset. This operation was repeated ten times. The average of the ten individual scores from the 10-fold cross-validation estimate was used to generate the best hyperparameters. Those hyperparameters were then used to evaluate the unseen test (30 %) data set.

2.6. *Model evaluation*

Quantifying the predictive performance of the different machine learning algorithms is a crucial component of the different gap-filling strategies. However, applying the model's predictive performance on the same set of data that it has been analyzed upon incurs the risk of overfitting. That risk is minimized if the model is evaluated on an independent dataset. This approach was implemented by randomly partitioning the dataset (Rs & F_{CH4}) into training and testing classes. We used 70 % of the partitioned data for training, while the remaining 30 % was used to evaluate the performance of the gap-filling algorithms.

Model performance was evaluated by their root mean square error (*RMSE*), mean absolute percentage error (*MAPE*), mean absolute error

(*MAE*), and coefficient of determination (R^2). These four indicators were combined into a synthesis index (SI) to provide a comprehensive measure of the model performance (Chou et al., 2014). *RMSE* was first normalized with the other nine *RMSE* values generated during the 10-fold cross-validation process to generate *SI*. Then, a similar operation was implemented for *MAPE* and *MAE*. As the direction of R^2 is opposite to that of the other three measured errors, $1\text{-}R^2$ was used instead (personal communication, Dr. Chou). Finally, *SI* was quantified as shown below:

$$SI = \frac{1}{n}\sum_{i=1}^{n}\frac{(P_i - P_{min,i})}{(P_{max,i} - P_{min,i})} \tag{13}$$

P_i refers to the ith performance measure, and n is the number of performance indexes. High model performance is marked by low *SI*.

Forecast skill (*FS*) was computed as shown below to evaluate the performance of the scalar-based gap-filling algorithms against three baseline models, including standard measured-data driven machine learning, Q_{10} and *MDS* approaches (Eq. (6)) (Chu et al., 2015):

$$FS = 1 - \frac{RMSE_{model}}{RMSE_{bm}} \tag{14}$$

$RMSE_{\text{model}}$ and $RMSE_{\text{bm}}$ refer to the *RMSE* value calculated over the entire dataset for the scalar-data driven algorithm and the baseline model, respectively. Positive and negative *FS* values indicated better and worse performance of the scalar-based gap-filling algorithms compared to the three baseline approaches. The difference between the mean *SI* for the training and testing component quantifies the stability of the four algorithms and is summarized in Table S2. An increase in mean SI when the algorithm was tested on the unknown dataset implies a decrease in prediction accuracy. A t-test was conducted for each algorithm, corresponding to each site and year to check for statistically significant difference between the training and testing component.

2.7. .Uncertainty analysis and prediction

Artificial gaps were incorporated within the *Rs* and F_{CH4} time series to evaluate the impact of gaps on the performance of the machine learning algorithms. These gap lengths represented a spectrum, including short (~ few hours), medium (< 2 days), and long (> 2 days) gaps (Moffat et al., 2007). Overall, 10 % of additional gaps were created and, in many cases, the new gaps overlapped with the existing gaps. These gaps were simulated ten times. The above-mentioned gap-filling operation was repeated for each of the 10-time series for each site year.

3. Results

The spectrally filtered (reconstructed) time series followed most of the measured values as expected, similar to a running mean (Figs. S1–S5, panels A-L). Only the intermediate scale fluctuations in soil moisture (Fig. S4), and less so in water table depth (Fig. S5), exhibited occasional divergence. In all of the time series, the approximation coefficient (S; Figs. S1–S5, panel M) was the slowest component (i.e., the trend) of the time series.

The efficacy of spectral filtering on the performance metrics of the individual machine learning algorithms compared to model run on measured data has been delineated in Table 1. Within the context of the US-NC1 site, the enhancement in model performance, quantified by the positive and negative alterations in R^2 and RMSE values, was most pronounced for the elasticnet regression model, whereas no significant changes were observed for any algorithms at the US-NC2 site. In the case of F_{CH4} fluxes at the US-NC4 location, the elasticnet regression (average $\Delta R^2 = 8.3$ %) and Artificial Neural Network (ANN) (average $\Delta R^2 = 13.36$ %) algorithms exhibited the most substantial improvement. While MDS_Spectral incorporated spectral filtering (Fig. S8), the enhancement in model fit over traditional MDS was minimal (Fig. 2)

Table 1
Summary of the efficacy of iterative enhancements of various machine learning algorithm, quantified by ΔR2 and ΔRMSE, which represent the percentage improvements in the coefficient of determination (R2) and root mean square error (RMSE), respectively, of the spectral version against the measured data driven approach. The flux of interest was *Rs* (umol m^{-2} s-1) in US-NC1 and US-NC2, and F_{CH4} (umol m^{-2} s-1) in US-NC4 (C).

	US-NC1						US-NC2						US-NC4					
	2009		2010		2011		2010		2013		2014		2013		2015		2016	
Algorithms	ΔR^2	ΔRMSE	ΔR^2	ΔRMSE	ΔR^2	ΔRMSE	ΔR^2	ΔRMSE	ΔR^2	ΔRMSE	ΔR^2	ΔRMSE	ΔR^2	ΔRMSE	ΔR^2	ΔRMSE	ΔR^2	ΔRMSE
Elastic Net	3.01	-7.2	1.87	-3.19	-0.49	2.9	-0.37	3.18	-1	8.34	-0.35	2.15	16.81	-1.3	4.2	-1.4	3.9	-1.9
SVM	-1.92	11.27	2.11	-4.12	-0.15	0.21	-0.33	4.07	-0.95	11.29	-0.69	5.92	-2.91	0.62	3.82	-1.4	0.85	-1.04
XGBoost	-1.38	8.8	3.45	-14.51	0.52	-8.5	-0.34	5.12	-1	13.9	-0.65	7.05	2.5	-1.29	-0.17	0.12	1.35	-1.27
Random Forest	-0.58	9.39	1.47	-22.1	0.41	-17.5	0.14	-7.3	0.03	-1.5	0.31	-11.7	1.47	-4.3	0.54	-1.45	0.69	-1.98
ANN	-1.27	6.1	3.26	-8.53	-1.67	16.3	-0.34	3.82	-0.89	9.74	-0.83	6.67	30.4	-2.7	8.1	-3.6	1.6	-1.1

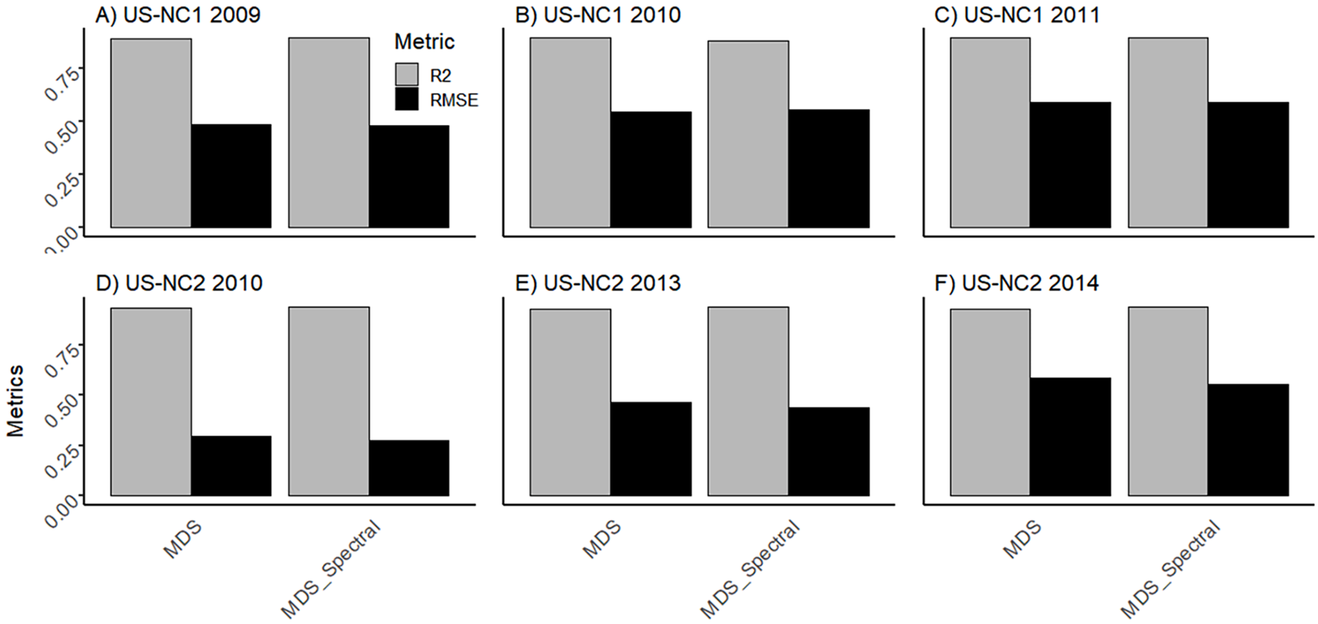


Fig. 2. Bar plot summarizing statistical model performance (quantified by R^2, RMSE) of soil respiration (*Rs*) gapfilling using two different variants of the Marginal Distribution Sampling algorithm, employing two distinct sets of environmental inputs for model calibration: measured data (MDS) and spectrally filtered data (MDS_Spectral).

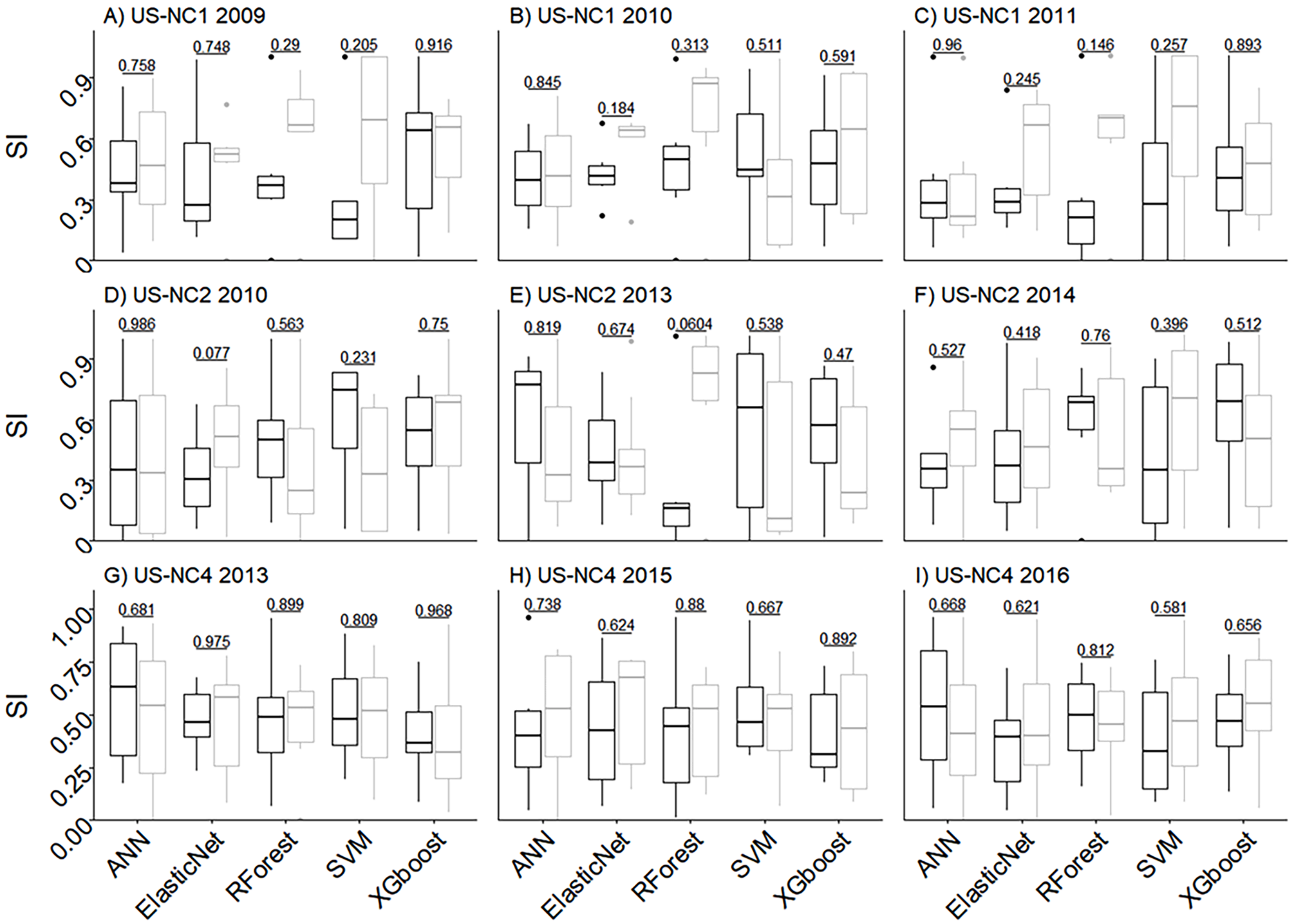


Fig. 3. Box plots of the Synthesis Index (SI) used to evaluate the performance of various machine learning algorithms for training (black bars) and testing (grey bars) phases of the dataset. Panels A-C depict SI distributions for soil respiration (*Rs*, umol $m^{-2}s^{-1}$) at US-NC1 across three different years, panels D-F show SI for *Rs* at US-NC2, and panels G-I illustrate SI for methane flux (F_{CH4}, umol $m^{-2}s^{-1}$) at US-NC4. Each panel provides a comparison of algorithm performance within a specific site-year context. Notably, p-values from t-tests comparing algorithm performances are annotated above the respective box plots for each site-year combination.

Fig. 3 illustrates variability in robustness, consistency, and relative performance across different site-years for the different machine learning algorithms when evaluated across training and unknown test datasets. All the models utilized spectrally filtered (PAR_r, T_{sr}, θ_r, WTD_r, P_r) environmental drivers as input. Notably, the *SVM* algorithm and *RForest* algorithm demonstrated a broad range of performance efficiency for the same site and years when predicting the unknown test dataset (Fig 3A–C). Similarly, across US-NC2 and US-NC4, divergences in the mean stability analysis value for various machine learning algorithms across multiple years presented no discernable pattern (Fig. 3, Supp Table 2). Across all three sites, no algorithm showed a statistically significant difference between training and testing datasets ($p < 0.05$).

Fig. 4 provides a comprehensive evaluation of the efficiency of various spectrally driven algorithms by pooling F-Score (*FS*) values across different sites and years, facilitating a broad assessment of their performance relative to the measured-data driven *MDS* and Q_{10} approaches. On average, the *RForest* algorithm exhibited superior performance compared to both *MDS* and Q_{10} algorithms across all sites, as evidenced by its consistently positive average *FS* scores. Except for *ElasticNet* across US-NC2, all the spectrally driven algorithms outperformed the Q_{10} algorithm. As indicated by the average negative *FS*-score, *ElasticNet* and *SVM* algorithms were consistently inferior to the *MDS* algorithm. When *MDS* was run in the spectral mode for *R*s, the benefit (as measured by positive *FS* score) was minimal.

The spectrally driven individual machine learning algorithms (Figs S9 A–E), as well as MDS (Fig. S9 G) and the original data driven *MDS* (Fig. S9 F) and Q_{10} (Fig. S9 H) algorithms exhibited a saturation response pattern, wherein the models were unable to capture the variability of high or low observed values. Of these, *RForest*, followed by *MDS* and *MDS_Spectral* demonstrated the least bias with regard to saturating effect, as evident by the tighter clustering of data points around the line of perfect fit. The Q_{10} algorithm demonstrated the highest level of bias and low accuracy (lower R^2 values).

Fig. 5 provides an overall assessment of four different algorithms: *MDS*, Q_{10}, spectral (Spectral) and original (Meas.) data driven ensemble model. Across all sites, spectrally driven ensemble model ranked highest in terms of accuracy (R^2) and efficiency (RMSE), followed by the ensemble measured data driven approach with the Q_{10} algorithm being the least efficient. Between the ensemble Spectral and Meas. approach, the difference in terms of model improvement was minimal (e.x: less than 2 % across US-NC1).

Histograms of the posterior predictive distribution of the trained dataset (*R*s, F_{CH4}) simulated by their corresponding likelihood functions (Eqs. (10), (11)) are summarized in Fig. S10. The histograms are a visualization tool to demonstrate the accuracy of the parameter space and the likelihood function. In all cases, the histograms plotted with a zero-valued baseline demonstrated symmetric unimodal distribution with no outliers. Three aspects (mean, median, and standard deviation (SD)) of the distributional features of the dataset have been summarized and compared with the mean statistics of the observed values. In all cases, the mean of the observed data was in the middle of the distribution (Fig. S10).

Bayesian simulations of *R*s, F_{CH4} generated by their corresponding likelihood functions have been summarized in Fig. 6. For the three different sites and different years, 70–90 % of the raw data (*R*s, F_{CH4}) were within the 95 % credible interval (C.I). The low range of the statistically significant coefficients (lack of zero in the 95 % credible interval) of the different parameters identified in the likelihood functions have been summarized in the supplementary tables (Supp. Tables 3–5) and reflect the reduction in model uncertainty. Spatial filtering also did not lead to discernable difference in the bayesian prediction of the greenhouse gas against measured data driven *R*s (Figs. 6, S11).

4. Discussion

The research spanned across different land types (managed and unmanaged) having different soil biogeochemical conditions, ecosystems that are carbon-rich yet notably underrepresented in major monitoring networks such as Ameriflux and the North American Carbon Program (Aguilos et al., 2020, 2021). Positioned at the nexus of terrestrial and aquatic environments, these wetlands are facing a sharp decline due to climatic shifts, sea level rise, extreme events, and direct human activities

Table 2
Synopsis of Machine Learning Algorithms: Advantages, limitations, and performance insights for key algorithms within this study are tabulated, alongside pertinent literature references.

Algorithm	Pros	Cons	Notes / Specific Findings	Refs.
Random Forest	• Ensemble of weak learners to strong learners. • Trains several decision trees in parallel • Unique trees reduce variance • Robust to noise and sample size. • Less dependent on hyperparameter tuning	N/A	• Superior performance in various studies • Effective for methane fluxes (Figs. 3, S9, Supp. Table S6)	• Misra and Li (2020) • Williams et al. (2020) • Misra and Wu (2020)
SVM	• Potentially be improved with advanced techniques like coordinate descent, genetic algorithms, particle swarm optimization	• Sensitive to noise and outliers. • Highly sensitive to hyperparameters. • - Not all kernel functions analysed	• Less effective compared to Random Forest in certain years (Figures S9, Supp. Table S6)	• Bai-dong (2010) • Tsirikoglou et al. (2017) • Ito and Nakano (2003) • Zhao et al. (2015) • Zhao et al. (2016) • Wang et al. (2012)
XGBoost	• Mathematical similarities to Random Forest	• Performance not always on par with Random Forest. • Grid-based hyperparameter optimization may not be optimal	• Performance issues with methane fluxes	• Chen and Guestrin (2016)
Elastic Net Regression	• Assumes linearity	• Outperformed by other algorithms, especially with methane fluxes	• Linear assumption may limit its applicability (Fig. S9, Supp. Table 6)	• De Clercq et al. (2020)
ANN	• Efficient in modeling non-linear and complex relationships	• Performance dependent on the amount of training data • Back propagation algorithm susceptible to local minima	• Requires sufficient training data for optimal performance. • Issues with local minima (Fig. S9, Supp. Table 6)	• Kala et al. (2009) • Chevalier et al. (2011)

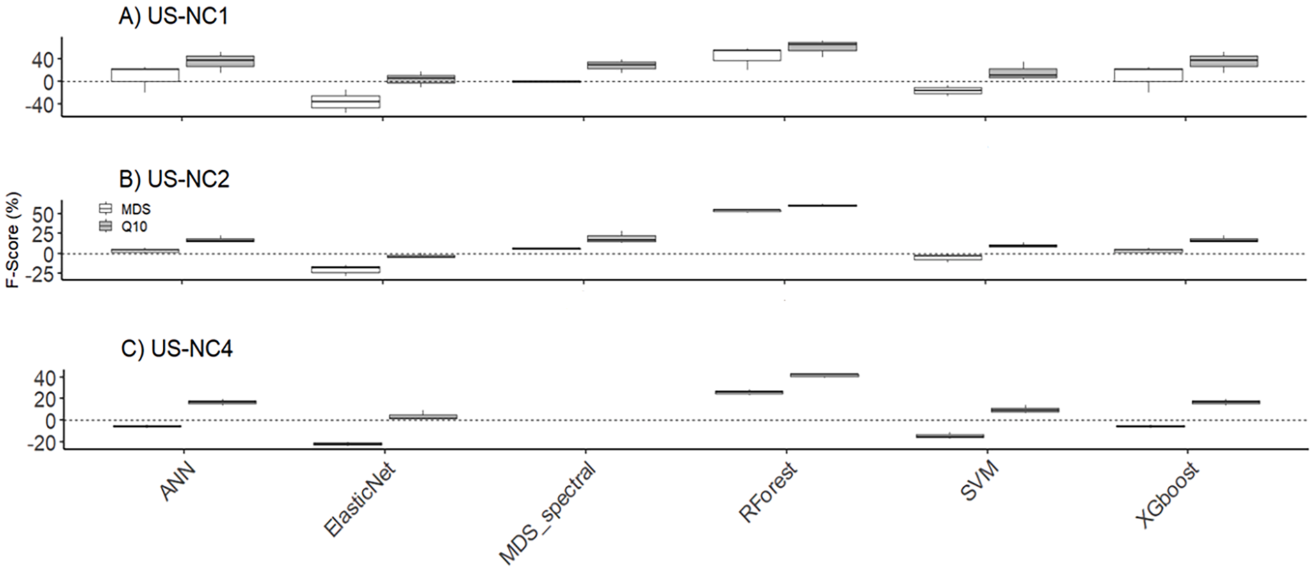


Fig. 4. Comparative analysis of the change in model effectiveness for predicting soil respiration (*Rs*, umol m^{-2}s^{-1}) and methane flux (F_{CH4}, umol m^{-2}s^{-1}) using individual machine learning algorithms and the spectrally-filtered data driven *MDS* algorithm (MDS_Spectral), benchmarked against the traditional data-driven *MDS* and Q_{10} models. Panels (A) US-NC1 and (B) US-NC2 illustrate the performance variations for *Rs*, while panel (C) US-NC4 depicts the changes for F_{CH4}.

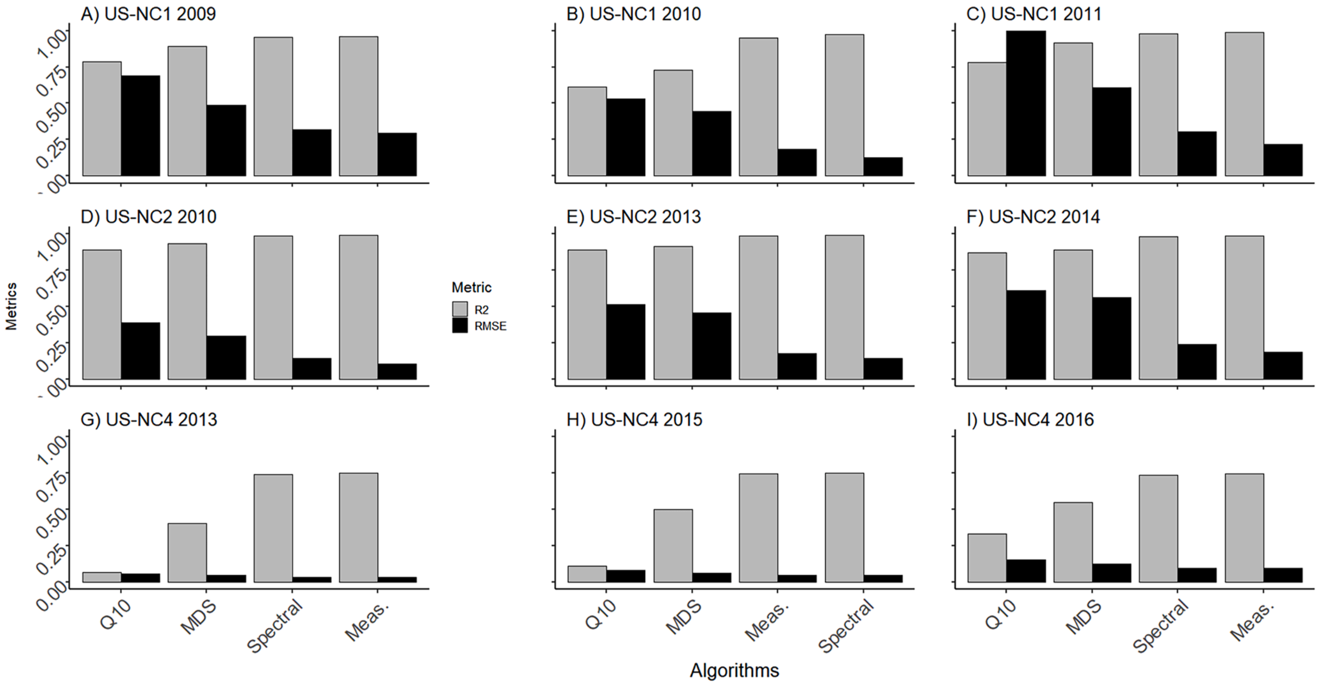


Fig. 5. Evaluation of four gap-filling methodologies for their accuracy in estimating gas fluxes, quantified by the correlation coefficient (R^2) and root mean square error (RMSE). Soil respiration (*Rs*, umol m^{-2}s^{-1}) at locations US-NC1 and US-NC2 are analyzed in Panels A-C and D-F, respectively, while methane flux (F_{CH4}), umol m^{-2}s^{-1} at US-NC4 is presented in Panels G–I.

such as land conversion and drainage (Raabe et al., 2016). Our selected research locations typify the forest varieties found within the lower coastal plains of the Southeastern United States. Furthermore, our research has attempted to address a critical theme in the understanding of the global carbon cycle that seeks to move beyond the standard correlative statistical simulation of two important biochemical processes (*Rs*, F_{CH4}) and seek to develop a parsimonious causal understanding of how temperature and other biological and non-biological factors overlap in a non-linear asynchronous fashion to induce the two greenhouse fluxes (Vargas et al., 2010; Vargas et al., 2011; Sturtevant et al., 2016).

4.1. Scale, spectral filtering and gapfilling

Scale is "*…the fundamental conceptual problem in ecology, if not all of science*" (Levin 1992) and is key to understanding patterns and processes (Bissonette 2017; Hernandez, 2020). To that end, partitioning the variability of the two fluxes at different scales can help to provide a more improved understanding of the role of different drivers in driving the temporal patterns in *Rs* and F_{CH4}. Multiresolution analysis using wavelets, as depicted in Figs. S1–S5 and described by Eqs. (1)–(5), offered a distinctive perspective by segmenting information into discrete packages (d1-d12), facilitating an understanding of the temporal scale

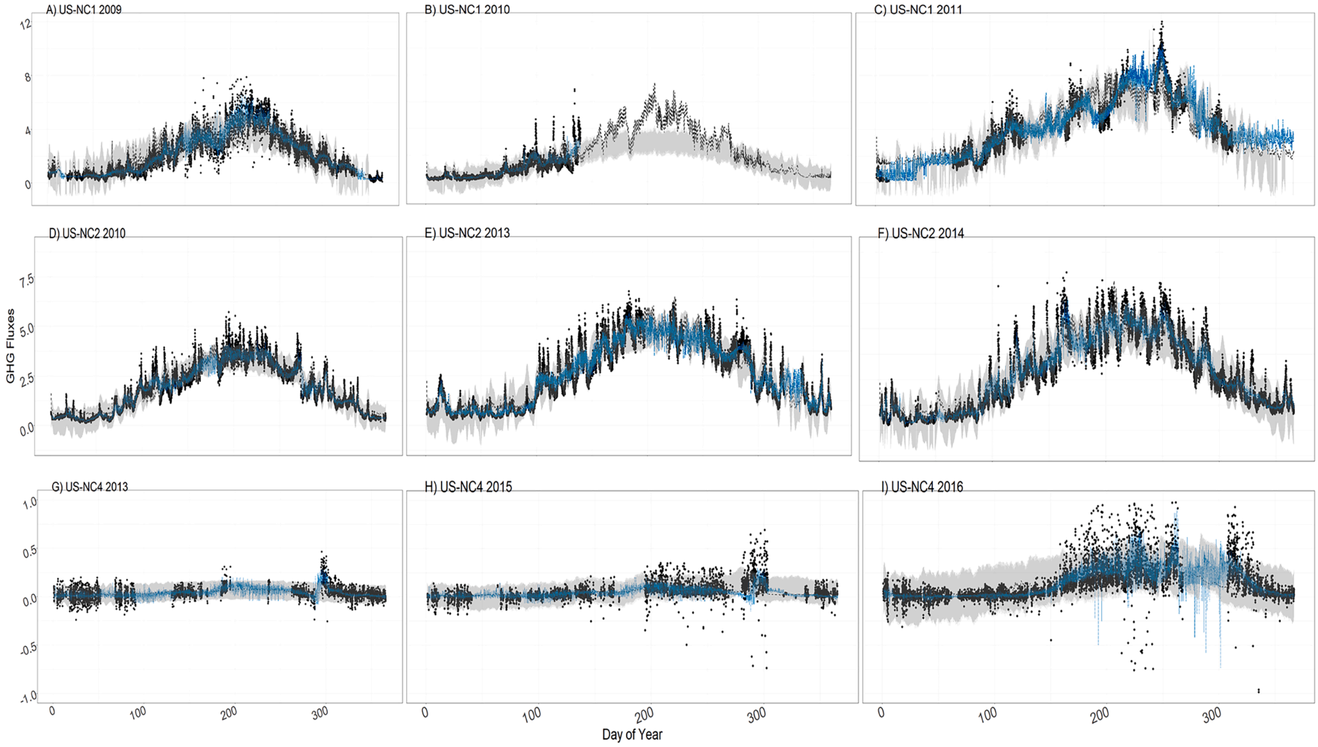


Fig. 6. Bayesian simulation of spectrally-filtered data driven *Rs* (umol m^{-2} s^{-1}) (US-NC1 & US-NC2) and *NEE*$_{CH4}$ (umol m^{-2} s^{-1}) (US-NC4) across different years. Solid grey lines represent the 95 % credible interval (C.I), measured values are represented by circle symbols, black and blue lines represent the gap-filling performance of the measured-data driven *Q*$_{10}$ *and MDS* approach.

responses of *Rs* (and F_{CH4}). These distinct information represent plant, microbial and abiotic processes that causally induced variation in *Rs* and F_{CH4} fluxes (Mitra et al., 2019; Mitra et al., 2020). By a close-up of finely detailed resolution (d1-d2, Eq. (1), Supp Fig. S1), one was able to link the rapid energetic frequencies of R_S and F_{CH4} emission to plant activity, modulated primarily by photosynthesis (Baldocchi et al., 2001) while at the diel scale (d3–d6, Eq. (1), Supp Fig S1), diurnal variation of the meteorological drivers modulated the fluxes. Synoptic scales (d7-d12, Eqs. (2)–(5), Supp Figs S2–S5) established the importance of atmospheric pressure, soil moisture, and water table dynamics as well as superimposed seasonal changes in plant physiological status in controlling the slow-moving waveforms of the two greenhouse fluxes. Taking into account the diverse and intricate enzymatic processes among various taxa that underpin the multi-scale variability of *Rs* and F_{CH4}, this aspect represents a novel contribution of this study.

The employment of spectrally filtered data in the gap-filling of *Rs* and F_{CH4} fluxes, as opposed to an unfiltered measured data approach, resulted in only minimal variance among the different algorithms (Table 1, Figs. 2, 6 and S11). Nevertheless, applying spectrally filtered data aligns with the overarching principle that incorporating established biochemical insights alongside diverse machine learning techniques can be a highly effective strategy for the estimation of carbon budgets (Liu et al., 2024). Therefore, the study decisively utilized spectrally filtered data in advancing the gap-filling process for *Rs* and F_{CH4} fluxes, reinforcing the premise that such a methodological approach harmonizes with the nuanced impacts exerted by each environmental and biological factor on the variability observed in these two pivotal greenhouse gas fluxes.

The fusion of wavelet decomposition with machine learning algorithms for prediction adopted in this study has already been documented across various fields (Renaud et al., 2003; Kim and Valdes, 2003; Belayneh et al., 2014; Belayneh et al., 2016; Tiwari and Chatterjee, 2010, 2011). Traditional spectral tools like wiener filtering, Kalman filtering, and Fourier transform are limited to linear systems and fail against non-stationary events (Sang et al., 2009). Daubechies' wavelet transformation in this paper overcomes the above-mentioned limitation (Daubechies 1990; Daubechies 1992).

4.2. Model comparison

Disparities in stability analysis (Fig. 3, Supp. Table 2) for the same algorithms across sites and years indicate a lack of model generalization and suggests all models (Fig. 4) may have inherent limitations in capturing the full range of variability in the data across different environmental or temporal conditions. These variations (Figs. 3, 4, S9, Supp. Table 2) can be attributed to multiple factors, including overfitting and generalization, stochastic variance, sensitivity to hyperparameters and data gaps, algorithm complexity, and ease of interpretability. Table 2 offers a nuanced synthesis of the confluence of factors impacting the outputs of the various machine learning algorithms evaluated within the context of this study.

After evaluating multiple models, a heterogeneous ensemble approach that combined *Elastic net regression, SVM, XGBoost, RF,* and *ANN* was adopted (*Spectral*, Fig. 5) as such a strategy can compensate for the structural deficiencies of each technique along with the errors induced by hyperparameter optimization and noise in data (Weng and Huang, 2021; Yin et al., 2021). Simply put, the performance gain from the heterogeneous ensemble approach is linked to model diversity (Sagi and Lokach 2018). The machine learning heterogeneous model ensemble (*Spectral*), applied to the spectrally filtered driver inputs, outperformed both *Q*$_{10}$ and *MDS* model estimates (Fig. 5). However, much of this difference originated from the performance of the *RForest* model that best captured the variability in *the* two greenhouse fluxes (Fig. S9, Supp. Table 6).

The effective model fit of the *MDS* in both the scalar and measured data based approaches (Fig. S9, Supp. Table 6) can be attributed to its non-parametric approach based on diurnal interpolation, and look-up tables (Reichstein et al., 2005; Wutzler et al., 2018). The former was especially effective in the current study as there was strong diurnal variation in both *Rs* (Mitra et al., 2019) and F_{CH4} (Mitra et al., 2020). However, that may not be true for methane fluxes or other trace gases across all sites or within a biome on account of strong variability (Lucas-Moffat et al., 2022). The look-up table exploits the strong correlation between the fluxes and environmental parameters. However, the current FLUXNET standard gapfilling MDS model uses only three predictor variables (Kim et al., 2020). Hence, the spectral version for *MDS* was run only for *R*s and not for F_{CH4}, thereby limiting its ability to capture the responses in many-dimensional parameter space. More importantly, gaps longer than 12 days cannot be filled with the *MDS* approach (Moffat et al., 2007). The exponential Q_{10} model was adopted in this paper as it remains a commonly used function to fit greenhouse emission flux with temperature (Lloyd and Taylor, 1994). Furthermore, it obscures the compounding effects of moisture, temperature, and carbon substrate and does not consider the temporal domain of influence for each drivers (Ryan and Law, 2005; Davidson et al., 2006), a solution to which remains the core theme of this paper.

Inference from Bayesian statistics remains fully a function of the observed data (Hackenberger 2019). The information used to simulate *R*s and F_{CH4} was consistent with the information contained in the trained data set (Fig. S10). Thus, the choice of the parameter space (Eqs. (1)–(5)) and the likelihood function (Eqs. (10)–(12)) was appropriate. Furthermore, weakly informative prior used in this study not only disallows any risk of Type I (rejection of null hypothesis when it is true) and Type M (exaggeration of statistically significant effects) errors (Lemoine 2019) but also helps to stabilize computation (Goodrich et al., 2020). Use of Bayesian (Figs 6, S11) along with different machine learning algorithms to predict the two greenhouse gases may not be redundant. Looking into the future and against the backdrop of extreme climate events, data distribution of environmental drivers may not mimic the current trend. New information on extreme events can easily be assimilated into the Bayesian framework as prior information. However, for machine learning models, it will need to be retrained with any new data in order to ensure that the training set is similar to the real-world scenario.

4.3. *Implications*

From the perspective of developing a robust framework for predictive modeling, there has been increased emphasis on identifying drivers that have a causal relationship with the GHG fluxes (Yuan et al., 2022). The principle of causality can be incorporated into other sites (Pearl, 2019; Reichstein et al., 2019), promising to improve gap-filling performance across different biomes. Appropriate coupling of driving factors to *R*s and F_{CH4} outputs at relevant time scales may provide a missing constraint to current ecosystem carbon cycle models and thus improve their projection capabilities under novel environmental conditions. In this study, we have demonstrated one approach to incorporating spectral information into *R*s and F_{CH4} gap-filling models. While the numerical improvement in model fit statistics was small (at least with the current data and models evaluated), there was a trend for greater improvement in parametric models that implicitly assume the universality of causality at all time scales. More nuanced non-parametric models (especially *MDS*) showed less improvement with spectrally filtered input data (Figs. 2, S8), presumably because the time scales and covariance with other factors is already implicitly accounted for in such structures. More importantly, though, the explicit consideration of timescales of variability represents the mechanisms more realistically, or spectrally truthful, approach. As shown here, spectral filtering can be applied to many different model structures and would be an easy modification to make in functional biogeochemical models, as well. We propose that current gap-filling tools be updated to include such an option.

Although the proposed approach is statistical in nature, it is consistent with other recent developments in soil carbon dynamics and ecosystem carbon cycle research. The significance of newly assimilated carbohydrates in controlling both root and soil microbial processes has been highlighted in recent ecosystem models (Wieder et al., 2015; Robertson et al., 2019). These studies, both directly and indirectly, underscore the importance of carbohydrate substrates for microbial activity, thus supporting the overarching narrative of this research. Similarly, the surplus carbon hypothesis (Prescott et al., 2020) highlights the significance of new carbohydrate availability to katabolic processes on short timescales, while also allowing for temporary temporal decoupling between assimilatory and respiratory processes (Noormets et al., 2021).

5. Conclusions

In this study, gap-filling soil respiration and methane fluxes using temporally filtered time series variables within a machine learning framework slightly enhanced the model's accuracy compared to the *MDS* method. The enhancement over the Q_{10} model was more significant. Although using spectrally filtered data over unfiltered data did not markedly alter the two greenhouse gas predictions, we conclude that adopting a scale-based methodology for gap-filling is beneficial and could be applied across different biomes. The combination of wavelet filtering with a machine learning framework offers the advantage of 'generalizability' to new environments, facilitating adaptability across scales and time.

CRediT authorship contribution statement

Bhaskar Mitra: Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Kevan Minick:** Writing – review & editing, Writing – original draft. **Michael Gavazzi:** Writing – original draft, Data curation. **Prajaya Prajapati:** Data curation. **Maricar Aguilos:** Writing – original draft, Data curation. **Guofang Miao:** Writing – original draft, Data curation. **Jean-Christophe Domec:** Writing – review & editing, Writing – original draft. **Steve G. McNulty:** Writing – review & editing, Writing – original draft, Funding acquisition. **Ge Sun:** Writing – original draft, Funding acquisition. **John S. King:** Writing – original draft, Investigation, Funding acquisition. **Asko Noormets:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Code is available on Github website for public access. Data is available on Ameriflux repository.

Meteorological and methane flux data used for this study can be downloaded from the Ameriflux database (doi:10.18140/FLX/16696; 10.17190/AMF/1246083). Codes implemented in this paper is available for download from Github (Mitra 2024; https://zenodo.org/records/10418521). Soil respiration data is available from the authors on reasonable request.

Acknowledgments

Funding for this AmeriFlux core site was provided by the U.S. Department of Energy's Office of Science, US DOE National Institute of Climatic Change Research (08-SC-NICCR-1072), the USDA Forest Service Eastern Forest Environmental Threat Assessment Center (08-JV11330147-038, 13-JV-11330110-081), the US DOE Terrestrial

Ecosystems Program (11-DE-SC0006700) and the National Science Foundation (NSF-IOS 1754893). We thank Dennis Stewart, Scott Lanier and the staff of Alligator River National Wildlife Refuge for site access and logistical support.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.agrformet.2024.110038.

References

Aguilos, M., Mitra, B., Noormets, A., Minick, K., Prajapati, P., Gavazzi, M., Sun, G., McNulty, S., Li, X., Domec, J.C., Miao, G., King, J., 2020. Long-term carbon flux and balance in managed and natural coastal forested wetlands of the Southeastern USA. Agric. For. Meteorol. 288 https://doi.org/10.1016/j.agrformet.2020.108022.

Aguilos, M., Sun, G., Noormets, A., Domec, J.C., McNulty, S., Gavazzi, M., Minick, K., Mitra, B., Prajapati, P., Yang, Y., King, J., 2021. Effects of land-use change and drought on decadal evapotranspiration and water balance of natural and managed forested wetlands along the southeastern US lower coastal plain. Agric. For. Meteorol. 303, 108381 https://doi.org/10.1016/j.agrformet.2021.108381, 2021ISSN 0168-1923.

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R., Zhu, Z., 2013. Evaluating the land and ocean components of the global carbon cycle in the CMIP5 earth system models. J. Clim. 26 (18), 6801–6843. https://doi.org/10.1175/JCLI-D-12-00417.1.

Arrhenius, S., 1889. Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren. Z. Phys. Chem. 4, 226–248.

Aubinet, M., Vesala, T., Papale, D., 2012. Eddy covariance: a Practical Guide to Measurement and Data Analysis. Springer Science & Business Media.

Bai-dong, J., 2010. Support vector machines based on reducing noise. Support Vector Mach. Based Reducing Noise 351–354.

Bakhshian, S., Romanak, K., 2021. DeepSense: a physics-guided deep learning paradigm for anomaly detection in soil gas data at geologic CO2 storage sites. Environ. Sci. Technol. 55 (22), 15531–15541. https://doi.org/10.1021/acs.est.1c04048. Epub 2021 Oct 25. PMID: 34694136.

Baldocchi, D., Falge, E., Wilson, K., 2001. A spectral analysis of biosphere-atmosphere trace gas flux densities andmeteorological variables across hour to multi-year time scales. Agric. For. Meteorol. 107 (1), 1–27. https://doi.org/10.1016/S0168-1923(00)00228-8.

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances, by the late Rev. Mr. Bayes, F. R. S., communicated by Mr. Price, in a letter to John Canton. A. M. F. R. S. Philos. Trans. 53, 370–418.

Belayneh, A., Adamowski, J., Khalil, B., Ozga-Zielinski, B., 2014. Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models. J. Hydrol. 508, 418–429. https://doi.org/10.1016/j.jhydrol.2013.10.052 (Amst).

Belayneh, A., Adamowski, J., Khalil, B., Quilty, J., 2016. Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. Atmos. Res. 172, 37–47. https://doi.org/10.1016/j.atmosres.2015.12.017.

Birzhandi, P., Kim, K.T., Lee, B., Youn, H.Y., 2019. Reduction of training data using parallel hyperplane for support vector machine. Appl. Artif. Intell. 33 (6), 497–516. https://doi.org/10.1080/08839514.2019.1583449.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Bissonette, J.A., 2017. Avoiding the scale sampling problem: a consilient solution. J. Wildl. Manag. 81 (2), 192–205. https://doi.org/10.1002/jwmg.21187.

Bond-Lamberty, B., Thomson, A., 2010. A global database of soil respiration data. Biogeosciences 7, 1915–1926. https://doi.org/10.5194/bg-7-1915-2010.

Breiman, L., 2001. Random Forests. Machine Learning, 45, pp. 5–32. https://doi.org/10.1023/A:1010933404324.

Chang, K.Y., Riley, W.J., Knox, S.H., Jackson, R.B., McNicol, G., Poulter, B., Aurela, M., Baldocchi, D., Bansal, S., Bohrer, G., Campbell, D.I., Cescatti, A., Chu, H., Delwiche, K.B., Desai, A., Euskirchen, E., Friborg, T., Goeckede, M., Helbig, M., Hemes, K.S., Kang, M., Keenan, T., King, J.S., Krauss, K.W., Lohila, A., Mammarella, I., Mitra, B., Miyata, A., Nilsson, M.B., Noormets, A., Oechel, W., Papale, D., Peichl, M., Reba, M.L., Rinne, J., Runkle, B.R.K., Ryu, Y., Sachs, T., Schäfer, K.V.R., Schmid, H.P., Shurpali, N., Sonnentag, O., Tang, A.C.I., Torn, M.S., Trotta, C., Tuittila, E.S., Ueyama, M., Vargas, R., Vesala, T., Windham-Myers, L., Zhang, Z., Zona, D., 2021. Substantial hysteresis in temperature sensitivity of global wetland methane emissions. Nat. Commun. 12, 2266. https://doi.org/10.1038/s41467-021-22452-1.

Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S., 2002. Choosing multiple parameters for support vector machines. Mach. Learn. 46 (1–3), 131–159. https://doi.org/10.1023/A:1012450327387.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD, pp. 785–794.

Chen, J.L., Li, G.S., Xiao, B.B., Wen, Z.F., Lv, M.Q., Chen, C.D., Jiang, Y., Wang, X.X., Wu, S.J., 2015. Assessing the transferability of support vector machine model for estimation of global solar radiation from air temperature. Energy Convers. Manag.e 89, 318–329. https://doi.org/10.1016/j.enconman.2014.10.004.

Chevalier, R.F., Hoogenboom, G., McClendon, R.W., et al., 2011. Support vector regression with reduced training sets for air temperature prediction: a comparison with artificial neural networks. Neural Comput. Appl. 20, 151–159. https://doi.org/10.1007/s00521-010-0363-y.

Chou, J.S., Tsai, C.F., Pham, A.D., Lu, Y.H., 2014. Machine learning in concrete strength simulations: Multi-nation data analytics. Constr. Build. Mater. 73, 771–780. https://doi.org/10.1016/j.conbuildmat.2014.09.054.

Chou, C.M., 2007. Applying multi-resolution analysis to differential hydrological grey models with dual series. J. Hydrol. 332 (1–2), 174–186. https://doi.org/10.1016/j.jhydrol.2006.06.031 (Amst).

Chu, Y., Urquhart, B., Gohari, S.M.I., Pedro, H.T.C., Kleissl, J., Coimbra, C.F.M., 2015. Short-term reforecasting of power output from a 48 MWe solar PV plant. Sol. Energy 112, 68–77. https://doi.org/10.1016/j.solener.2014.11.017.

Ciais, P., Dolman, A.J., Bombelli, A., Duren, R., Peregon, A., Rayner, P.J., Miller, C., Gobron, N., Kinderman, G., Marland, G., Gruber, N., Chevallier, F., Andres, R.J., Balsamo, G., Bopp, L., Bréon, F.-M., Broquet, G., Dargaville, R., Battin, T.J., Borges, A., Bovensmann, H., Buchwitz, M., Butler, J., Canadell, J.G., Cook, R.B., DeFries, R., Engelen, R., Gurney, K.R., Heinze, C., Heimann, M., Held, A., Henry, M., Law, B., Luyssaert, S., Miller, J., Moriyama, T., Moulin, C., Myneni, R.B., Nussli, C., Obersteiner, M., Ojima, D., Pan, Y., Paris, J.-D., Piao, S.L., Poulter, B., Plummer, S., Quegan, S., Raymond, P., Reichstein, M., Rivier, L., Sabine, C., Schimel, D., Tarasova, O., Valentini, R., Wang, R., van der Werf, G., Wickland, D., Williams, M., Zehner, C., 2014. Current systematic carbon-cycle observations and the need for implementing a policy-relevant carbon observing system. Biogeosciences 11 (13), 3547–3602. https://doi.org/10.5194/bg-11-3547-2014.

Daubechies, I., 1990. The wavelet transform, time-frequency localization and signal analysis. IEEe Trans. Inf. Theory 36 (5), 961–1005. https://doi.org/10.1109/18.57199.

Daubechies, I., 1992. Ten Lectures on Wavelets. SIAM, Philadelphia, PA, USA, p. 357.

Davidson, E.A., Janssens, I.A., Luo, Y.Q., 2006. On the variability of respiration in terrestrial ecosystems: moving beyond Q(10). Glob. Change Biol. 12 (2), 154–164. https://doi.org/10.1111/j.1365-2486.2005.01065.x.

Dayhoff, J.E., 1990. Neural Network Architectures. Van Nostrand Reinhold, New York.

De Clercq, D., Wen, Z., Fei, F., Caicedo, L., Yuan, K., Shang, R., 2020. Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion. Sci. Total Environ. 712, 134574 https://doi.org/10.1016/j.scitotenv.2019.134574.

Deane-Mayer, Z.A., Knowles, J.E., 2019. R Package Caret. Ensemble, Version 2.0.1. https://cran.r-project.org/package=caretEnsemble.

Diamantopoulou, M.J., 2005. Artificial neural networks as an alternative tool in pine bark volume estimation. Comput. Electron. Agric. 48 (3), 235–244. https://doi.org/10.1016/j.compag.2005.04.002.

Drago, A.F., Boxall, S.R., 2002. Use of the wavelet transform on hydro-meteorological data. In: Proceedings of the Physics and Chemistry of the Earth, 27(26th General Assembly of the European-Geophysical-Society), pp. 1387–1399. https://doi.org/10.1016/S1474-7065(02)00076-1.

Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., Burba, G., Ceulemans, R., Clement, R., Dolman, H., Granier, A., Gross, P., Grünwald, T., Hollinger, D., Jensen, N.-O., Katul, G., Keronen, P., Kowalski, A., Lai, C.T., Law, B.E., Meyers, T., Moncrieff, J., Moors, E., Munger, J.W., Pilegaard, K., Rannik, Ü., Rebmann, C., Suyker, A., Tenhunen, J., Tu, K., Verma, S., Vesala, T., Wilson, K., Wofsy, S., 2001. Gap filling strategies for defensible annual sums of net ecosystem exchange. Agric. For. Meteorol. 107 (1), 43–69. https://doi.org/10.1016/S0168-1923(00)00225-2.

Fang, C., Moncrieff, J.B., 2001. The dependence of soil CO2 efflux on temperature. Soil Biol. Biochem. 33, 155–165.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., Von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H.D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K.-G., Schnur, R., Strassmann, K., Weaver, A.J., Yoshikawa, C., Zeng, N., 2006. Climate-carbon cycle feedback analysis: results from the (CMIP)-M-4 model intercomparison. J. Clim. 19 (14), 3337–3353. https://doi.org/10.1175/JCLI3800.1.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Stat. 29 (5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7 (4), 457–472. https://doi.org/10.1214/ss/1177011136. November.

Goodrich, B., Gabry, J., Ali, I., Brilleman, S., 2020. rstanarm: Bayesian applied regression modeling via Stan. R package, version 2.17.4. https://mc-stan.org/rstanarm.

Guenther, F., Fritsch, S., 2010. neuralnet: training of neural networks. R J. 2 (1), 30–38.

Hackenberger, B.K., 2019. Bayes or not Bayes, is this the question? Croat. Med. J. 60 (1), 50–52. https://doi.org/10.3325/cmj.2019.60.50.

Hernandez, F., 2020. The importance of scale in ecological studies. J. Wildl. Manag. 84 (8), 1427–1434.

Hoffman, F.M., Randerson, J.T., Arora, V.K., Bao, Q., Cadule, P., Ji, D., Jones, C.D., Kawamiya, M., Khatiwala, S., Lindsay, K., Obata, A., Shevliakova, E., Six, K.D., Tjiputra, J.F., Volodin, E.M., Wu, T., 2014. Causes and implications of persistent atmospheric carbon dioxide biases in earth system models. J. Geophys. Res. Biogeosciences 119 (2), 141–162. https://doi.org/10.1002/2013JG002381.

Ito, K., Nakano, R., 2003. Optimizing support vector regression hyperparameters based on cross-validation. In: Proceedings of the International Joint Conference on Neural Networks, 2003, 3, pp. 2077–2082. https://doi.org/10.1109/IJCNN.2003.1223728.

Jones, C.D., Cox, P., Huntingford, C., 2003. Uncertainty in climate-carbon-cycle projections associated with the sensitivity of soil respiration to temperature. Tellus B 55 (6), 642–648. https://doi.org/10.1034/j.1600-0889.2003.01440.x. -Chemical and Physical Meteorology.

Joshi, N., Gupta, D., Suryavanshi, S., Adamowski, J., Madramootoo, C.A., 2016. Analysis of trends and dominant periodicities in drought variables in India: A wavelet transform-based approach. Atmos. Res. 182, 200–220. https://doi.org/10.1016/j.atmosres.2016.07.030.

Kala, R., Shulkla, A., Tiwari, R., 2009. Fuzzy neuro systems for machine learning for large data sets. In: Proceedings of the 2009 IEEE International Advance Computing Conference, pp. 541–545. https://doi.org/10.1109/IADCC.2009.4809069.

Kass, R.E., Carlin, B.P., Gelman, A., Neal, RM., 1998. Markov chain Monte Carlo in practice: a roundtable discussion. Am. Stat. 52 (2), 93–100. https://doi.org/10.2307/2685466.

Kim, T.W., Valdes, J.B., 2003. Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. J. Hydrol. Eng. 8 (6), 319–328. https://doi.org/10.1061/(ASCE)1084-0699(2003)8:6(319).

Kim, Y., Johnson, M.S., Knox, S.H., Black, T.A., Dalmagro, H.J., Kang, M., Kim, J., Baldocchi, D., 2020. Gap-filling approaches for eddy covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal component analysis. Glob. Change Biol. 26 (3), 1499–1518. https://doi.org/10.1111/gcb.14845.

Lemoine, N.P., 2019. Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. Oikos 128, 912–928. https://doi.org/10.1111/oik.05985.

Levin, S.A., 1992. The problem of pattern and scale in ecology. Ecology 73, 1943–1967.

Li, C., Frolking, S., Frolking, T.A., 1992. A model of nitrous oxide evolution from soil driven by rainfall events: 1. Model structure and sensitivity. J. Geophys. Res. Atmos. 97 (D9), 9759–9776.

Liu, Z.Y., Menzel, L., 2016. Identifying long-term variations in vegetation and climatic variables and their scale-dependent relationships: A case study in Southwest Germany. Glob. Planet. Change 147, 54–66. https://doi.org/10.1016/j.gloplacha.2016.10.019.

Liu, S.Y., Huang, S.Z., Xie, Y.Y., Huang, Q., Wang, H., Leng, G.Y., 2019. Assessing the non-stationarity of low flows and their scale-dependent relationships with climate and human forcing. Sci. Total Environ. 687, 244–256. https://doi.org/10.1016/j.scitotenv.2019.06.025.

Liu, L., Zhou, W., Guan, K., et al., 2024. Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. Nat. Commun. 15, 357. https://doi.org/10.1038/s41467-023-43860-5.

Lloyd, J., Taylor, J.A., 1994. On the temperature dependence of soil respiration. Funct. Ecol. 8 (3), 315–323. https://doi.org/10.2307/2389824.

Lucas-Moffat, A.M., Schrader, F., Herbst, M., Brümmer, C., 2022. Multiple gap-filling for eddy covariance datasets. Agric. For. Meteorol. 325, 109114 https://doi.org/10.1016/j.agrformet.2022.109114.

Mallat, 1989. A theory for multiresolution signal decomposition - the waveletrepresentation. IEEe Trans. Pattern. Anal. Mach. Intell. 11 (7), 674–693. https://doi.org/10.1109/34.192463.

Meissner, K.J., Brook, E., Finkelstein, S.A., Rae, J., 2021. Carbon cycle dynamics during episodes of rapid climate change. Environmental Research Letters 16 (4). https://doi.org/10.1088/1748-9326/abeade.

Miao, G., Noormets, A., Domec, J.C., Fuentes, M., Trettin, C.C., Sun, G., McNulty, S.G., King, J.S., 2017. Hydrology and microtopography control carbon dynamics in wetlands: Implications in partitioning ecosystem respiration in a coastal plain forested wetland. Agric. For. Meteorol. 247, 343–355. https://doi.org/10.1016/j.agrformet.2017.08.022.

Misra, S., Li, H., Misra, S., Li, H., He, J., 2020. Noninvasive fracture characterization based on the classification of sonic wave travel times. Machine Learning for Subsurface Characterization. Gulf Professional Publishing, pp. 243–287. https://doi.org/10.1016/B978-0-12-817736-5.00009-0.

Misra, S., Wu, Y., Misra, S., Li, H., He, J., 2020. Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. Machine Learning for Subsurface Characterization. Gulf Professional Publishing, pp. 289–314. https://doi.org/10.1016/B978-0-12-817736-5.00010-7.

Mitra, B., Miao, G., Minick, K., McNulty, S.G., Sun, G., Gavazzi, M., King, J.S., Noormets, A., 2019. Disentangling the effects of temperature, moisture, and substrate availability on soil CO2 efflux. J. Geophys. Res. Biogeosciences 124 (7), 2060–2075. https://doi.org/10.1029/2019JG005148.

Mitra, B., Minick, K., Miao, G., Domec, J-C., McNulty, S.G., Sun, G., King, J.S, Noormets, A., 2020. Spectral evidence for multiple drivers of methane fluxes from a coastal forested wetland in North Carolina. Agric. For. Meteorol. 291, 108062 https://doi.org/10.1016/j.agrformet.2020.108062.

Moffat, A.M., Papale, D., Reichstein, M., Barr, A.G., Beckstein, C., Braswell, B.H., Churkina, G., Desai, A.R., Falge, E., Gove, J.H., Heimann, M., Hollinger, D.Y., Hui, D., Jarvis, A.J., Kattge, J., Noormets, A., Richardson, A.D., Stauch, V.J., 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agric. For. Meteorol. 147, 209–232.

Moffat, A.M., Beckstein, C., Churkina, G., Mund, M., Heimann, M., 2010. Characterization of ecosystem responses to climatic controls using artificial neural networks. Glob. Change Biol. 16 (10), 2737–2749. https://doi.org/10.1111/j.1365-2486.2010.02171.x.

Muth, C., Oravecz, Z., Gabry, J., 2018. User-friendly Bayesian regression modeling: A tutorial with rstanarm and shinystan. Quant. Methods Psychol. 14 (2), 99–119. https://doi.org/10.20982/tqmp.14.2.p099.

Nalley, D., Adamowski, J., Khalil, B., 2012. Using discrete wavelet transforms to analyze trends in streamflow and precipitation in Quebec and Ontario (1954-2008). J. Hydrol. 475, 204–228. https://doi.org/10.1016/j.jhydrol.2012.09.049 (Amst).

Nalley, D., Adamowski, J., Khalil, B., Ozga-Zielinski, B., 2013. Trend detection in surface air temperature in Ontario and Quebec, Canada during 1967-2006 using the discrete wavelet transform. Atmos. Res. 132, 375–398. https://doi.org/10.1016/j.atmosres.2013.06.011.

Noormets, A., McNulty, S.G., Domec, J.C., Gavazzi, M., Sun, G., King, J.S., 2012. The role of harvest residue in rotation cycle carbon balance in loblolly pine plantations. Respiration partitioning approach. Glob. Change Biol. 18 (10), 3186–3201. https://doi.org/10.1111/j.1365-2486.2012.02776.x.

Noormets, A., Bracho, R., Ward, E., Seiler, J., Strahm, B., In, W., McElligott, K., Domec, J.-C., Gonzalez-Benecke, C., Jokela, E.J., Markewitz, D., Meek, C., Miao, G., McNulty, S.G., King, J.S., Samuelson, L., Sun, G., Teskey, R., Vogel, J., Will, R., Yang, J., Martin, T.A., 2021. Heterotrophic respiration and the divergence of productivity and carbon sequestration. Geophys. Res. Lett. 48 https://doi.org/10.1029/2020GL092366 e2020GL092366.

Noormets, A., Mitra, B., Sun, G., Miao, G., King, J., Minick, K., Yang, L., Aguilos, M., Gavazzi, M., Prajapati, P., McNulty, S., 2022a. AmeriFlux BASE US-NC2 NC_Loblolly Plantation, Ver. 10-5. AmeriFlux AMP. https://doi.org/10.17190/AMF/1246083 (Dataset).

Noormets, A., Yang, L., Aguilos, M., Mitra, B., Prajapati, P., King, J., 2022b. AmeriFlux BASE US-NC4 NC_AlligatorRiver, Ver. 5-5. AmeriFlux AMP. https://doi.org/10.17190/AMF/1480314 (Dataset).

Noormets, A., 2018. AmeriFlux BASE US-NC1 NC_Clearcut, Ver. 3-5. AmeriFlux AMP. https://doi.org/10.17190/AMF/1246082 (Dataset).

Parton, W.J., Schimel, D.S., Cole, C.V., Ojima, D.S., 1987. Analysis of factors controlling soil organic matter levels in Great Plains grasslands. Soil Sci. Soc. Am. J. 51 (5), 1173–1179.

Pastorello, G., Trotta, C., Canfora, E., et al., 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddycovariance data. Sci. Data 7 (1). https://doi.org/10.1038/s41597-020-0534-3.

Pearl, J. (2019). The limitations of opaque learning machines. Chapter 2 in John Brockman (Ed.), Possible Minds: 25 Ways of Looking at AI, Penguin Press, 2019. (pp. 13-19).

Prescott, C.E., Grayston, S.J., Helmisaari, H., Kaštovská, E., Körner, C., Lambers, H., Meier, I.C., Millard, P., Ostonen, I., 2020. Surplus carbon drives allocation and plant–soil interactions. Trends Ecol. Evol. 35, 1110–1118.

Raabe, E.A., Stumpf, R.P., 2016. Expansion of tidal marsh in response to sea-level rise: gulf coast of Florida, USA. Estuaries Coasts 39, 145–157. https://doi.org/10.1007/s12237-015-9974-y.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Grunwald, T., Havrankova, K., Ilvesniemi, H., Janous, D., Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers, T., Miglietta, F., Ourcival, J.M., Pumpanen, J., Rambal, S., Rotenberg, E., Sanz, M., Tenhunen, J., Seufert, G., Vaccari, F., Vesala, T., Yakir, D., Valentini, R., 2005. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. Glob. Change Biol. 11 (9), 1424–1439. https://doi.org/10.1111/j.1365-2486.2005.001002.x.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566 (7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1.

Renaud, O., Starck, J.L., Murtagh, F., 2003. Prediction based on a multiscale decomposition. Int. J. Wavelets. Multiresolut. Inf. Process. [Online]. Available: http://www.unige.ch/fapse/PSY/persons/profana/renaud/papers/PredictMultiscale.pdf.

Renaud, O., Stark, J.L., Murtagh, F., 2005. Wavelet–based combined signal filtering and prediction. IEEE Trans. Syst. Man Cybern. 35 (6), 1241–1251.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press.

Robertson, A.D., Paustian, K., Ogle, S., Wallenstein, M.D., Lugato, E., Cotrufo, M.F., 2019. Unifying soil organic matter formation and persistence frameworks: the MEMS model. Biogeosciences. 16 (6), 1225–1248. https://doi.org/10.5194/bg-16-1225-2019.

Ryan, M., Law, B., 2005. Interpreting, measuring, and modeling soil respiration. Biogeochemistry 73, 3–27. https://doi.org/10.1007/s10533-004-5167-7.

Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8 (4), e1249.

Sang, Y.F., Wang, D., Wu, J.C., Zhu, Q.P., Wang, L., 2009. Entropy-based wavelet denoising method for time series analysis. Entropy 11 (4), 1123–1147. https://doi.org/10.3390/e11041123.

Smola, A.J., Scholkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

Sturtevant, C.S., Oechel, W.C., Zona, D., Kim, Y., Emerson, C.E., 2012. Soil moisture control over autumn season methane flux, arctic coastal plain of Alaska. Biogeosciences 9 (4), 1423–1440.

Sturtevant, C., Ruddell, B.L., Knox, S.H., Verfaillie, J., Matthes, J.H., Oikawa, P.Y., Baldocchi, D., 2016. Identifying scale-emergent, nonlinear, asynchronous processes of wetland methane exchange. J. Geophys. Res. Biogeosciences 121 (1), 188–204. https://doi.org/10.1002/2015JG003054.

Sulman, B.N., Brzostek, E.R., Medici, C., Shevliakova, E., Menge, D.N.L., Phillips, R.P., 2017. Feedbacks between plant N demand and rhizosphere priming depend on type of mycorrhizal association. Ecol. Lett. 20 (8), 1043–1053. https://doi.org/10.1111/ele.12802.

Tang, X., Fan, S., Du, M., Zhang, W., Gao, S., Liu, S., et al., 2020. Spatial and temporal patterns of global soil heterotrophic respiration in terrestrial ecosystems. Earth Syst. Sci. Data 12, 1037–1051. https://doi.org/10.5194/essd-12-1037-2020.

Tiwari, M.K., Chatterjee, C., 2010. Development of an accurate and reliable hourly flood forecasting model using wavelet-bootstrap-ANN (WBANN) hybrid approach. J. Hydrol. 394 (3–4), 458–470. https://doi.org/10.1016/j.jhydrol.2010.10.001 (Amst).

Tiwari, M.K., Chatterjee, C., 2011. A new wavelet-bootstrap-ANN hybrid model for daily discharge forecasting. J. Hydroinformatics 13 (3), 500–519. https://doi.org/10.2166/hydro.2010.142.

Tsirikoglou, P., Abraham, S., Contino, F., Lacor, C., Ghorbaniasl, G., 2017. A hyperparameters selection technique for support vector regression models. Appl. Soft. Comput. 61, 139–148. https://doi.org/10.1016/j.asoc.2017.07.017.

Turetsky, M.R., Kotowska, A., Bubier, J., Dise, N.B., Crill, P., Hornibrook, E.R.C., Minkkinen, K., Moore, T.R., Myers-Smith, I.H., Nykanen, H., Olefeldt, D., Rinne, J., Saarnio, S., Shurpali, N., Tuittila, E.S., Waddington, J.M., White, J.R., Wickland, K. P., Wilmking, M., 2014. A synthesis of methane emissions from 71 northern, temperate, and subtropical wetlands. Glob. Change Biol. 20 (7), 2183–2197. https://doi.org/10.1111/gcb.12580.

Vargas, R., Detto, M., Baldocchi, D.D., Allen, M.F., 2010. Multi-scale analysis of temporal variability of soil CO2 production as influenced by weather and vegetation. Glob. Change Biol. 16 (5), 1589–1605. https://doi.org/10.1111/j.1365-2486.2009.02111.x.

Vargas, R., Carbone, M.S., Reichstein, M., Baldocchi, D.D., 2011. Frontiers and challenges in soil respiration research: from measurements to model-data integration. Biogeochemistry 102 (1–3), 1–13. https://doi.org/10.1007/s10533-010-9462-1.

Villa, J.A., Ju, Y., Vines, C., Rey-Sanchez, C., Morin, T.H., Wrighton, K.C., Bohrer, G., 2019. Relationships between methane and carbon dioxidefluxes in a temperate cattail-dominated freshwater wetland. J. Geophys. Res. Biogeosciences 124, 2076–2089. https://doi.org/10.1029/2019JG005167.

Vizza, C., West, W.E., Jones, S.E., Hart, J.A., Lamberti, G.A., 2017. Regulators of coastal wetland methane production and responses to simulated global change. Biogeosciences 14, 431–446. https://doi.org/10.5194/bg-14-431-2017.

Vonesch, C., Blu, T., Unser, M., 2007. Generalized daubechies wavelet families. IEEE Trans. Signal Process. 55 (9), 4415–4429. https://doi.org/10.1109/TSP.2007.896255.

Wang, J., Li, L., Niu, D., Tan, Z., 2012. An annual load forecasting model based on support vector regression with differential evolution algorithm. Appl. Energy 94, 65–70. https://doi.org/10.1016/j.apenergy.2012.01.010.

Warner, D.L., Bond-Lamberty, B., Jian, J., Stell, E., Vargas, R., 2019. Spatial predictions and associated uncertainty of annual soil respiration at the global scale. Global. Biogeochem. Cycles 33, 1733–1745. https://doi.org/10.1029/2019gb006264.

Weng, C.H., Huang, C.K., 2021. A hybrid machine learning model for credit approval. Appl. Artif. Intell. 35 (15), 1439–1465. https://doi.org/10.1080/08839514.2021.1982475.

Wieder, W.R., Grandy, A.S., Kallenbach, C.M., Taylor, P.G., Bonan, G.B., 2015. Representing life in the Earth system with soil microbial functional traits in the MIMICS model. Geosci. Model. Dev. 8 (6), 1789–1808. https://doi.org/10.5194/gmd-8-1789-2015.

Wieder, W.R., Lawrence, D.M., Fisher, R.A., Bonan, G.B., Cheng, S.J., Goodale, C.L., et al., 2019. Beyond static benchmarking: Using experimental manipulations to evaluate land model assumptions. Global. Biogeochem. Cycles 33, 1289–1309. https://doi.org/10.1029/2018GB006141.

Williams, B., Halloin, C., Löbel, W., Finklea, F., Lipke, E., Zweigerdt, R., Cremaschi, S., Pierucci, S., Manenti, F., Bozzano, G.L., Manca, D., 2020. Data-driven model development for cardiomyocyte production experimental failure prediction. Comput. Aided Chem. Eng. 48, 1639–1644. https://doi.org/10.1016/B978-0-12-823377-1.50274-3. ISBN 9780128233771.

Wutzler, T., et al., 2018. Basic and extensible post-processing of eddy covariance flux data with Reddyproc. Biogeosciences 15 (16), 5015–5030. https://doi.org/10.5194/bg-15-5015-2018.

Yin, J., Medellín-Azuara, J., Escriva-Bou, A., Liu, Z., 2021. Bayesian machine learning ensemble approach to quantify model uncertainty in predicting groundwater storage change. Sci. Total Environ. 769, 144715 https://doi.org/10.1016/j.scitotenv.2020.144715.

Yuan, K., Zhu, Q., Li, F., Riley, W.J., Torn, M., Chu, H., McNicol, G., Chen, M., Knox, S., Delwiche, K., Wu, H., Baldocchi, D., Ma, H., Desai, A.R., Chen, J., Sachs, T., Ueyama, M., Sonnentag, O., Helbig, M., Tuittila, E.S., Jurasinski, G., Koebsch, F., Campbell, D., Schmid, H.P., Lohila, A., Goeckede, M., Nilsson, M.B., Friborg, T., Jansen, J., Zona, D., Euskirchen, E., Krauss, K., Bohrer, G., Jin, Z., Liu, L., Iwata, H., Goodrich, J., Jackson, R., 2022. Causality Guided Machine Learning Model on Wetland CH4 Emissions Across Global Wetlands. SSRN. https://ssrn.com/abstract=4034619.

Zhao, W., Tao, T., Zio, E., 2015. System reliability prediction by support vector regression with analytic selection and genetic algorithm parameters selection. Appl. Soft. Comput. 30, 792–802. https://doi.org/10.1016/j.asoc.2015.02.026.

Zhao, W., Tao, T., Zio, E., Wang, W., 2016. A novel hybrid method of parameters tuning in support vector regression for reliability prediction: particle swarm optimization combined with analytical selection. IEEE Trans. Reliab. 65 (3), 1393–1405. https://doi.org/10.1109/TR.2016.2515581.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. B Stat. Methodol. 67, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.