remote sensing

MDPI

Article

Eddy Covariance CO_2 Flux Gap Filling for Long Data Gaps: A Novel Framework Based on Machine Learning and Time Series Decomposition

Dexiang Gao [1,2], Jingyu Yao [1,2], Shuting Yu [3], Yulong Ma [4], Lei Li [1,2] and Zhongming Gao [1,2,*]

[1] School of Atmospheric Sciences, Sun Yat-sen University, Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519082, China; gaodx6@mail2.sysu.edu.cn (D.G.); yaojy36@mail.sysu.edu.cn (J.Y.); lilei68@mail.sysu.edu.cn (L.L.)
[2] Key Laboratory of Tropical Atmosphere-Ocean System, Ministry of Education, Sun Yat-sen University, Zhuhai 519082, China
[3] School of Earth Sciences, Yangtze University, Wuhan 430100, China; 202001405@yangtzeu.edu.cn
[4] Guangdong-Hong Kong-Macau Greater Bay Area Weather Research Center for Monitoring Warning and Forecasting (Shenzhen Institute of Meteorological Innovation), Shenzhen 518040, China
* Correspondence: gaozhm3@mail.sysu.edu.cn

Abstract: Continuous long-term eddy covariance (EC) measurements of CO_2 fluxes (NEE) in a variety of terrestrial ecosystems are critical for investigating the impacts of climate change on ecosystem carbon cycling. However, due to a number of issues, approximately 30–60% of annual flux data obtained at EC flux sites around the world are reported as gaps. Given that the annual total NEE is mostly determined by variations in the NEE data with time scales longer than one day, we propose a novel framework to perform gap filling in NEE data based on machine learning (ML) and time series decomposition (TSD). The novel framework combines the advantages of ML models in predicting NEE with meteorological and environmental inputs and TSD methods in extracting the dominant varying trends in NEE time series. Using the NEE data from 25 AmeriFlux sites, the performance of the proposed framework is evaluated under four different artificial scenarios with gap lengths ranging in length from one hour to two months. The combined approach incorporating random forest and moving average (MA-RF) is observed to exhibit better performance than other approaches at filling NEE gaps in scenarios with different gap lengths. For the scenario with a gap length of seven days, the MA-RF improves the R^2 by 34% and reduces the root mean square error (RMSE) by 55%, respectively, compared to a traditional RF-based model. The improved performance of MA-RF is most likely due to the reduction in data variability and complexity of the variations in the extracted low-frequency NEE data. Our results indicate that the proposed MA-RF framework can provide improved gap filling for NEE time series. Such improved continuous NEE data can enhance the accuracy of estimations regarding the ecosystem carbon budget.

Keywords: CO_2 fluxes; long length gaps; moving average; empirical mode decomposition; random forest

Citation: Gao, D.; Yao, J.; Yu, S.; Ma, Y.; Li, L.; Gao, Z. Eddy Covariance CO_2 Flux Gap Filling for Long Data Gaps: A Novel Framework Based on Machine Learning and Time Series Decomposition. *Remote Sens.* **2023**, *15*, 2695. https://doi.org/10.3390/rs15102695

Academic Editors: Qiang Zhang, Yu Zhang, Ping Yue, Jun Wen, Zesu Yang and Yongli He

Received: 19 April 2023
Revised: 13 May 2023
Accepted: 16 May 2023
Published: 22 May 2023

1. Introduction

In recent decades, eddy covariance (EC) technology has advanced, and it has been used to perform continuous measurements of ecosystem–atmosphere exchanges of carbon, water, and energy around the world [1,2]. The data produced using EC technology is not only useful for estimating ecosystem-level annual carbon budgets; it can also help to establish parameterization relationships between CO_2 fluxes and meteorological and environmental variables [3]. Since the late 1990s, several regional networks of EC flux towers have been set up [4], the data from which have greatly improved our understanding of the temporal and spatial variations in net ecosystem exchange of CO_2 (NEE) [5,6]. To date, a few thousand EC towers have been established worldwide [7]. However, due to a

number of factors, including power outages, instrument malfunctions and maintenance, as well as data quality checking, there exist gaps, which account for approximately 30–60% of the half-hourly flux data at many EC sites [8–14]. These gaps have different temporal lengths, ranging from hours to days, or even to months for remote sites [15]. A robust gap-filling approach capable of filling such flux gaps is therefore urgently needed.

Accurate and robust NEE gap-filling methods are essential for quantifying the interannual variability in the carbon budget [12,13,16,17]. A variety of approaches and algorithms have been developed for performing gap filling in flux data, including nonlinear regressions [12], linear/multiple regressions [18], look-up tables [19–21], multiple imputation [22], etc. However, the performance of these approaches varies depending on the EC site, the time of the day, and the season of the year, and they are not able to achieve consistent results. Recently, marginal distribution sampling (MDS) [12,20] and machine learning (ML) [12,16,23] have become the standard approaches for NEE gap filling in the EC community [14]. However, there is still lack of robust methods for filling long data gaps.

Filling long data gaps is challenging for the most commonly used gap-filling approaches. Large degrees of uncertainty have been reported when using MDS to fill long gaps (e.g., 1–2 weeks) [24]. A variety of different ML algorithms, including artificial neural network (ANN), support vector machine (SVM), and random forest (RF), have been evaluated for their ability to fill flux gaps with different gap lengths and across different vegetation covers [23,25,26]. It is suggested that ML-based gap-filling models have the potential to fill long gaps, and RF generally outperforms both the other ML algorithms as well as the MDS method [16,23,25,27,28]. However, the overall performance of gap-filling models greatly depends on the type of ecosystem, climate, prediction target, and gap length [29].

Numerous studies have found that variations in NEE are affected by multiple environmental factors, including solar radiation or photosynthetic active radiation, air temperature, vapor pressure deficit, soil water content, water table depth, wind speed, and precipitation [30–34]. These factors are often used as the inputs for fitting or training gap-filling algorithms. For instance, the standard MDS model uses three meteorological parameters as the controlling factors for identifying similar conditions that would allow the NEE data to be used to fill the gaps [14]. Meanwhile, vegetation type, leaf area index and vegetation cover also affect CO_2 emissions [35], and these vegetation indices also change with weather and climate. The sensitivity of NEE variations to these factors has been shown to be site specific and time/season dependent, which limits the model performance for NEE gap filling [25,26]. For example, factors including needle drop, prescribed fire, wind sweep events, and morning venting of the canopy have been found to be responsible for the variability of NEE in a longleaf pine forest [36]. NEE gap-filling models usually have better performance for gaps occurring during the daytime period and in the growing season compared to those occurring in the nighttime and during the non-growing season [25]. Therefore, identifying how to better use the time and scale information embedded in the datasets could help to improve the performance of NEE gap-filling models.

In this study, we hypothesize that the observed NEE data are a product standing for the combined effects of different time and scale information of environmental and weather factors. After decomposing the NEE data into different time series corresponding to the effects of difference scales, we can then fill in the data separately. Based on this hypothesis, we propose a novel framework for performing gap filling in NEE data, combining the advantages of machine learning and time series decomposition. To examine the proposed framework, we select 25 AmeriFlux sites with different vegetation types and weather conditions. Two time series decomposition methods in combination with four ML algorithms are assessed with artificial gaps with different lengths. The objectives of this study are (1) to evaluate the performance of the novel framework compared to the commonly used approaches; and (2) to examine the reliability and accuracy of the novel framework across different sites with multiple vegetation types.

2. Data Sets and Study Sites

2.1. Eddy Covariance CO_2 Fluxes

The eddy covariance CO_2 fluxes used in this study were obtained from the AmeriFlux FLUXNET data product (https://ameriflux.lbl.gov/data/flux-data-products/, accessed on 17 September 2022). Eddy covariance technology consists of a three-dimensional sonic anemometer and a CO_2/H_2O gas analyzer, which measures longitudinal, lateral, and vertical wind velocities and scalars (CO_2, H_2O, and temperature) at a sample rate of 10–20 Hz. The high-frequency measurements are processed post field to estimate the net exchange in the scalars at the ecosystem level. The post-field processing procedures include: despiking and filtering for physically impossible values and abnormal diagnostic values of the instruments, applying coordinate rotation (double rotation or planar fit) to wind velocities, calculating raw fluxes, and correcting the raw fluxes for the effects of high- and low-pass filtering and air density fluctuations, respectively [37].

2.2. Ancillary Data for Gap-Filling Algorithms

Ancillary data used for the gap-filling algorithms include global radiation (Rg), air temperature (Tair), vapor pressure deficit (VPD), wind speed (WS), rain, soil temperature (Tsoil) and soil water content (SWC) obtained from the flux data product, and the normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI) from the Moderate-Resolution Imaging Spectroradiometer (MODIS). The meteorological data have an average interval of 30 min, and the gaps in these data are filled using the ERA5 Reanalysis global atmospheric product created by the European Centre for Medium-Range Weather Forecasts (ECMWF) with a spatial and temporal downscaling process [1]. The NDVI and EVI data around the flux tower locations are obtained from the MOD13Q1 version 6 data product (https://lpdaac.usgs.gov/products/mod13q1v006/, accessed on 13 October 2022) at 16 d temporal and 250 m spatial resolutions [38]. The 16 d NDVI and EVI data were resampled to 30 min using cubic spline interpolation.

2.3. Study Sites

We applied the proposed gap-filling framework to 25 AmeriFlux sites (Figure 1). Based on their land surface ecosystems, these sites can be categorized into five groups. US-AR1, US-AR2, US-Goo, US-IB2, US-Lin, US-SRG, and US-Var are predominantly composed of grasslands. US-ARM, US-CRT, US-GZ1, US-SZ2, and US-Tw2 are covered with croplands. US-Hn1, US-KS2, US-SRC, US-SRM, US-Sta, and US-Whs are shrubland sites. US-Me1, US-Me6, and US-NR1 are covered with evergreen coniferous forests, while US-Oho, US-UMd, and US-WCr are covered with deciduous broadleaved forest. US-Syv is a mixed forest site. At these sites, due to instrumental maintenance, power failures, data quality control, and other site-specific issues, approximately 46% of data, on average, are accounted for by high-quality CO_2 flux data, while the percentages corresponding to data gaps of different lengths vary from site to site (Figure 1b). Here, the high-quality NEE data refer to the data after quality checking and removing data obtained under low-turbulence conditions.

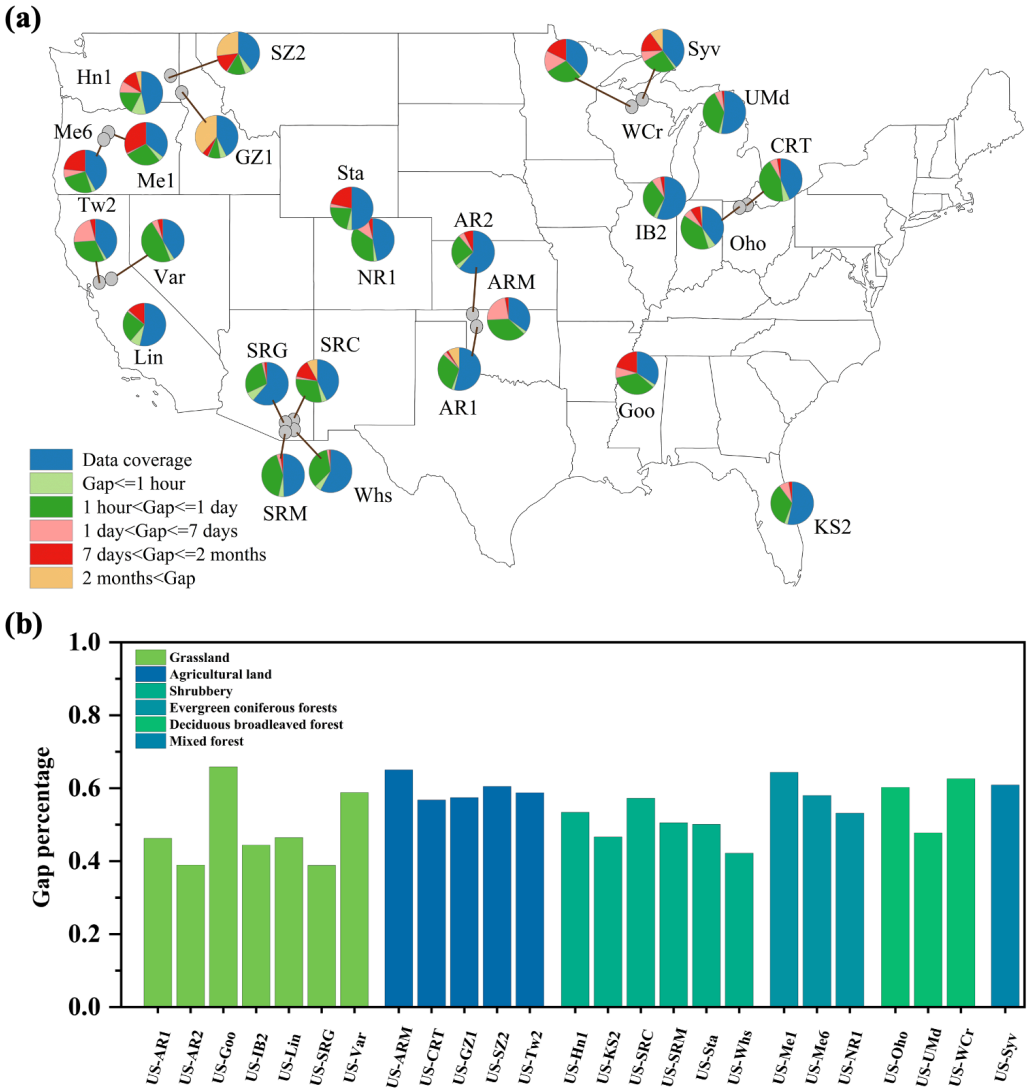

Figure 1. (**a**) Location of the 25 AmeriFlux sites and pie charts of percentages of high-quality data and data with gaps of different lengths at each site. The percentages of gaps are counted for five lengths (one hour, one day, one week, two months, and greater than two months), respectively. (**b**) Grouped histograms of all data gaps for different vegetation covers.

3. Methodology

3.1. Time Series Decomposition Approaches

The purpose of applying time series decomposition is to extract the dominant variations in the NEE time series that not only contribute greatly to the annual total NEE, but also determine the interannual variability of NEE (Figures A1 and A2 in Appendix A). In addition, the complexity of the trend term (i.e., nonlinear and non-stationary) is greatly reduced, which in turn is beneficial for ML model training. Here, we adopt moving average and empirical mode decomposition to decompose the pre-filled NEE data at each site. As stated above, the trend (T) and fluctuation (P) terms refer to the low- and high-frequency variations in the time series, respectively, and hence, a time series of NEE can be constructed as

$$X = T + P \tag{1}$$

3.1.1. Moving Average (MA)

By selecting an appropriate window, moving average can be applied to extract the trend term in NEE data. Considering the facts that (1) the impact factors for the half-hourly

variations in NEE are highly nonlinear, and (2) the balance between the surface available energy and the EC fluxes of sensible and latent heat is much better at the daily scale than at the half-hourly scale [39], we use 24 h as the length of the moving window. The fluctuation term is then determined by subtracting the trend term from the pre-filled NEE time series.

3.1.2. Empirical Mode Decomposition (EMD)

Compared to the moving average, the empirical mode decomposition is adaptive and suitable for the analysis of nonlinear and non-stationary signals, as well as linear and stationary signals. The EMD method assumes that a signal consists of a finite set of amplitude–frequency-modulated oscillatory components, each of which can be linear or nonlinear [40]. Using a sifting process, the oscillatory components can be extracted from the original data sequentially. For the EMD, the trend term is defined as the sum of certain oscillatory components, which represent the variations with scales longer than one day. Therefore, the trend terms determined by MA and EMD are comparable, although the EMD has the potential to decompose discontinuous data [41], suggesting that the impacts of pre-filled NEE data on the extracted trend term are constrained. The fluctuation term is calculated in the same way as for MA.

3.2. Machine Learning Approaches

To fill the gaps in the trend term of NEE variations, four machine learning (ML) approaches, including random forest (RF), extreme gradient boosting (XGboost), support vector regression (SVR), and back propagation (BP) neural network, are employed and evaluated. In the following subsections, we briefly introduce the characteristics and implementation of each ML algorithm. The optimal parameters for each of the four ML algorithms are listed in Appendix A, Table A1.

3.2.1. Random Forest (RF)

RF has commonly been used to perform gap filling in flux data [23,26]. The algorithm operates as an ensemble of many independent decision trees, each of which is trained independently on bootstrapped data [42]. The performance of the algorithm is influenced by multiple factors, including the number of decision trees, the maximum depth of the decision trees, the minimum number of samples required for each leaf node, the minimum number of samples required for the split nodes, etc. In this study, we optimize these parameters using grid search and five times cross-validation with the "scikit-learn" Python library.

3.2.2. EXtreme Gradient Boosting (XGboost)

XGBoost is an improved boost model based on the gradient boosting decision tree (GBDT), which has been widely used as an efficient GBDT framework to build decision trees sequentially [43]. The parameter optimization method is the same as RF, including the learning rate (0.01, 0.1 or 0.2), the minimum loss function descent value required for node splitting (0–0.5), the minimum sample weight sum in child nodes (1, 2, 5 or 10), and the proportion of features sampled when building the tree (0.6, 0.7, 0.8 or 0.9). In addition, the maximum depth of each tree and the number of decision trees are the same as in RF.

3.2.3. Support Vector Regression (SVR)

SVR was developed by Cortes and Vapnik [44], and it has now been applied to perform gap filling in flux data [23,25,26]. SVR is a supervised learning algorithm for predicting discrete values. The basic idea of SVR is to reduce the complexity of the algorithm by adding kernel functions to the SVR algorithm so that nonlinear regressions can be converted into linear regressions. The performance of the algorithm is greatly influenced by the type of kernel function. In this study, we optimize the parameters using a grid search, where the tuning parameters include the kernel function and the cost regularization parameter (C = 1, 10, 100 or 100).

3.2.4. Back Propagation (BP) Neural Network

The BP neural network was first proposed by Rumelhart et al. [45], and is a multi-layer feed-forward neural network whose main feature is that the signal is forward propagated while the error is backward propagated. The BP neural network model can be divided into three layers: input layer, hidden layer, and output layer. In this study, the structural design and parameter setting of the BP neural network are carried out using the "keras" Python library.

3.3. The Novel Framework Based on ML and Time Series Decomposition

To fill long gaps in a time series of CO_2 fluxes, we propose a novel framework based on machine learning and time series decomposition, the structure and inputs of which are summarized in Figure 2. The time series decomposition aims to extract the low-frequency variations in the NEE data, which largely correspond to changes in vegetation phenology. In order to decompose the NEE data, we first pre-fill the gaps in CO_2 fluxes using an RF model (i.e., the first-layer machine learning model), because most recent studies have shown that RF has the best performance at filling NEE gaps when compared to other gap-filling methods [23,25]. The RF model is trained using ancillary data at times when the learning target of the CO_2 fluxes is available in high quality. The trained RF model is then used to obtain the predicted NEE at times when the CO_2 fluxes are missing. This step is actually the same as in the traditional gap-filling framework when using ML algorithms.

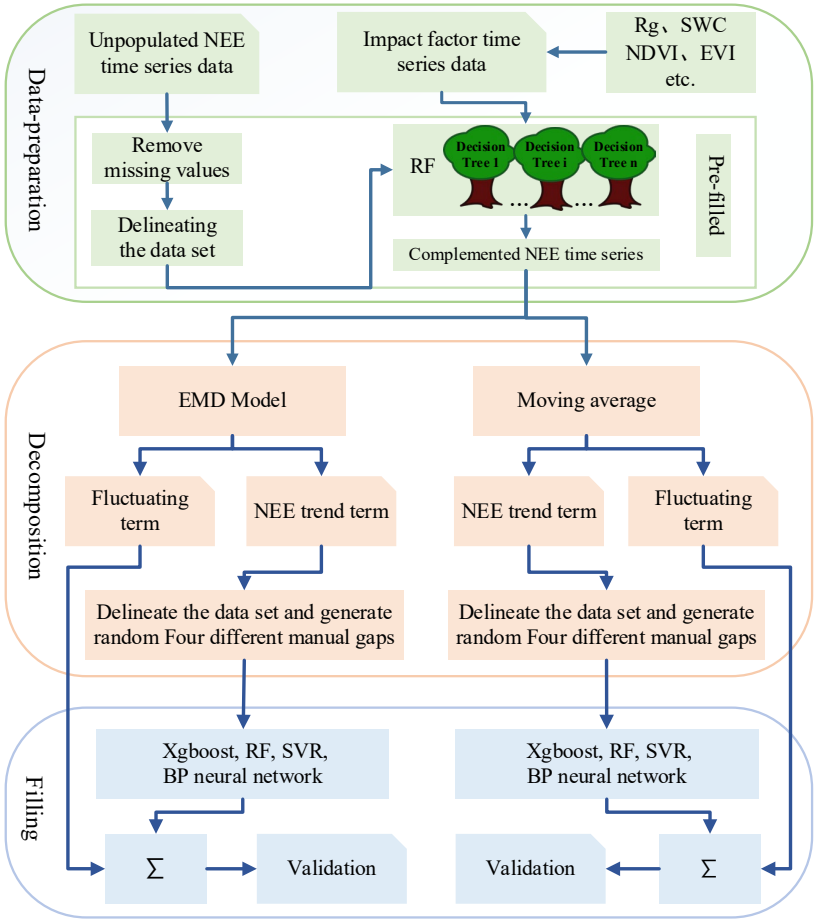

Figure 2. Flow chart for the proposed three-layer machine learning framework, consisting of (top panel) pre-filling NEE data using RF, (middle panel) extracting the dominant signal by decomposing the time series using the empirical mode decomposition (EMD) and moving average (MA) methods, and (bottom panel) re-filling the gaps in the extracted signal and reconstructing the NEE time series.

Second, we decompose the pre-filled NEE data into two time series representing high- and low-frequency variations (fluctuation and trend terms, hereafter), respectively, using the moving average (MA) and empirical mode decomposition (EMD) methods. The threshold used to separate the fluctuation and trend terms is determined by considering the contribution of the trend term to the annual total flux, as well as the distribution of the fluctuation terms. As shown in Figure 3, the histogram of the fluctuation term has a distribution more similar to a normal distribution, compared to that of the trend term, and thus the accumulated contribution of the fluctuation term to annual total NEE is constrained. In general, the trend term contributes more than 90% of the annual total NEE at each site, while the fluctuation term only accounts for a small portion of the annual total NEE. After separating the fluctuation and trend terms, the data in the trend term at times when the original NEE data are not available are removed, while the data in the fluctuation term are reserved for reconstructing the time series.

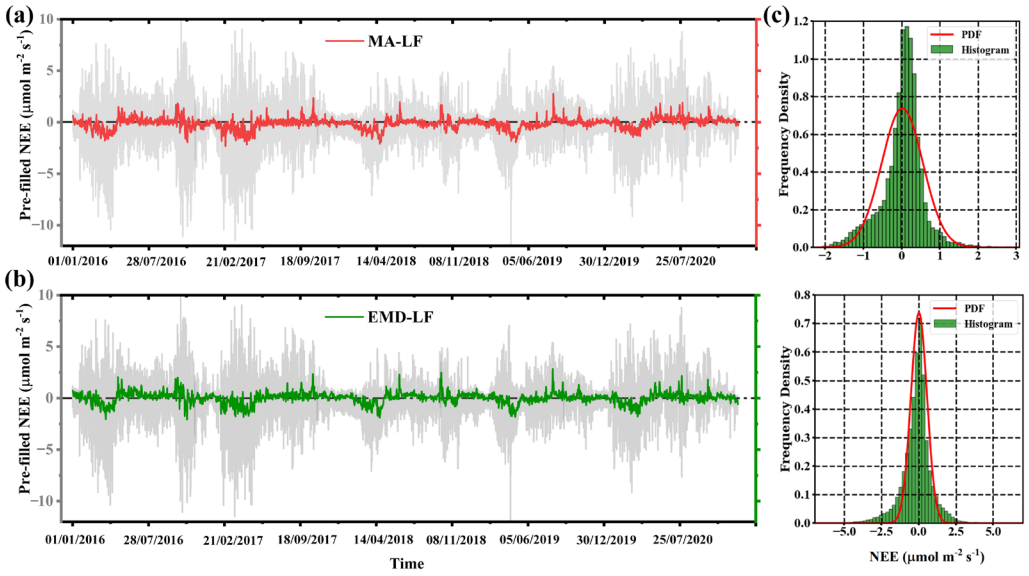

Figure 3. Time series decomposition using the (**a**) MA and (**b**) EMD methods, respectively, and (**c**) frequency histograms and PDF of the low- and high-frequency components after MA decomposition.

Using ancillary data and the extracted NEE trend term, the second-layer machine learning model is trained at times at which the learning target and inputs are available. The gaps in the trend term are obtained through the second-layer machine learning model. The time series of NEE is then reconstructed by adding the refilled trend term and the extracted fluctuation term together.

3.4. Model Evaluation

Gaps of different lengths occur non-randomly in the time series of CO_2 fluxes due to a variety of factors. In order to evaluate the performance of the gap-filling models, artificial gaps with four typical lengths are randomly generated in the original flux data, i.e., short gaps (1 h), medium gaps (1 day), long gaps (1 week), and very-long gaps (2 months). The total length of the artificial gaps accounts for approximately 10% of the total data length at each site. In addition, to eliminate the potential influence of gap location on the evaluation of model performance, each gap length scenario is permuted 10 times to create test sets with distinct artificial gaps, and for each test set, we also generate 10 training and validation sets to train and validate the ML model (Figure 4).

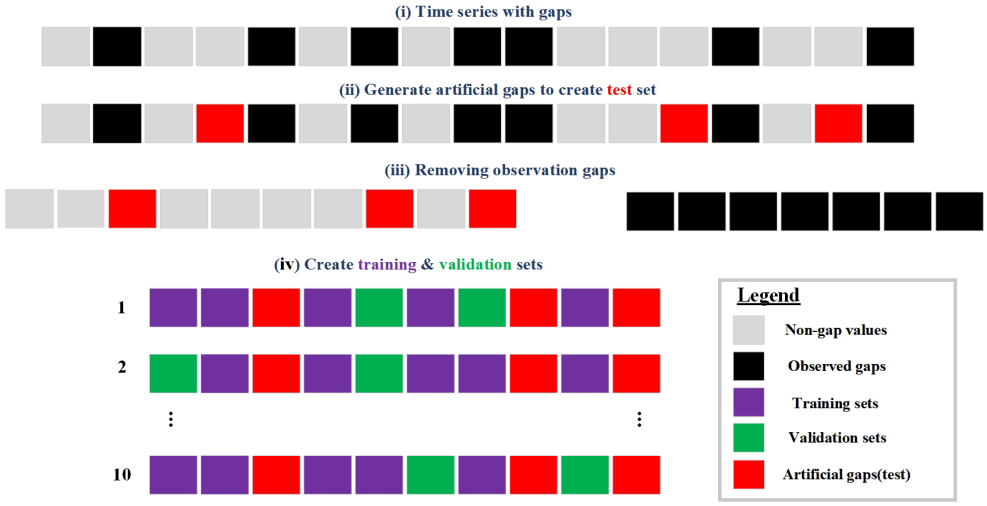

Figure 4. Procedures for the generation of the artificial gap test, training, and validation sets to evaluate model performance. Artificial gaps are introduced to (**i**) the original time series to (**ii**) create the test set, and then (**iii**) the observation gaps are removed, followed by (**iv**) the generation of ten training/validation sets. Ten test sets are created for each gap scenario, and the model performance is assessed by running each model on the test set.

With the artificial gaps, we evaluate the model performance by comparing the predicted and measured values for different gap lengths at each site. Four commonly used performance metrics are calculated, including the coefficient of determination (R^2), the root mean square error (RMSE), mean absolute error (MAE), and the bias error (Bias):

$$R^2 = \frac{\left(\sum (p-\bar{p})(m-\bar{m})\right)^2}{\sum (p-\bar{p})^2 \sum (m-\bar{m})^2} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum (m-p)^2}{n}} \quad (3)$$

$$MAE = \frac{\sum |p-m|}{n} \quad (4)$$

$$Bias = \frac{\sum p - \sum m}{n} \quad (5)$$

where m and p refer to the measured and predicted values, respectively, and the overbar denotes the mean value.

4. Results

4.1. Comparison of Model Performance between the Traditional and Proposed Gap-Filling Frameworks

Using the NEE data at five EC sites with different vegetation types, we first examine whether the proposed gap-filling framework improves the model performance compared to the traditional framework. Here, the traditional framework refers to an RF model similar to that trained for pre-filling the gaps. As shown in Figure 5, the proposed framework of either EMD or MA in combination with RF (i.e., EMD-RF or MA-RF) outperforms the gap-filling model with only RF, though the improvement in the proposed framework appears to be site dependent. At US-AR1, R^2 increases by approximately 0.4 from RF to EMD-RF or MA-RF, while at US-Oho, R^2 increases by around 0.1. On average, the traditional RF gap-filling framework has median R^2 and RMSE values of 0.78 and 2.58 μmol m^{-2} s^{-1}, respectively, at these five sites. The median R^2 values increase to 0.94 and 0.97 for the proposed framework

with EMD-RF and MA-RF, respectively, while the median values of RMSE decrease to 1.31 μmol m^{-2} s^{-1} and 1.04 μmol m^{-2} s^{-1} for EMD-RF and MA-RF, respectively. Note that, at US-GZ1, the large scatter of R^2 and RMSE for the gap-filling frameworks are mostly caused by the large fraction of very long data gaps and the changes in planted crops, as well as human activities. Therefore, the models have relatively low performance over croplands compared to over other vegetation types (Figure A3 in Appendix A). In addition, Figure 5 also suggests that when using MA to extract the dominant signal in NEE time series, the proposed framework has a slightly better performance than when using EMD.

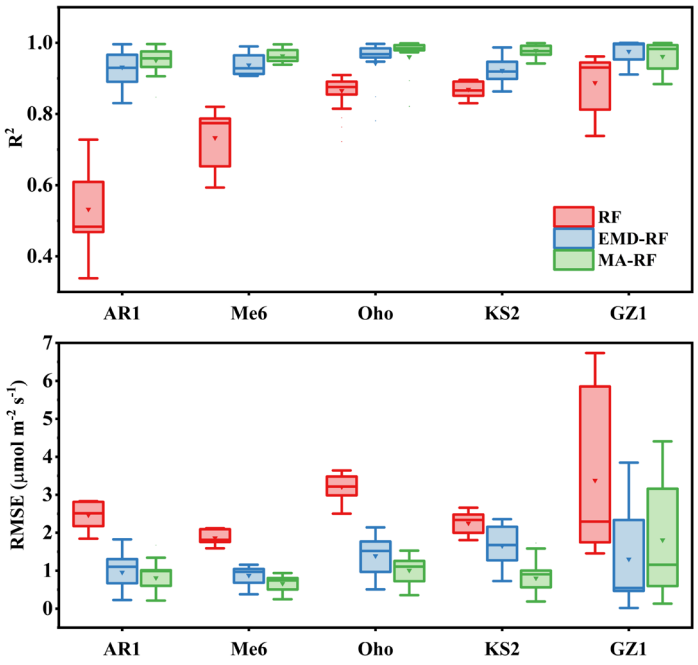

Figure 5. Comparison of model performance between the traditional and the proposed gap-filling frameworks at five EC sites with different vegetation types.

4.2. Comparison of the Model Performance of the Proposed Framework in Combination with Different ML Algorithms

To figure out which ML approach possesses the best performance when carrying out NEE gap filling using the proposed framework, we compare the performance of four commonly used ML approaches. After extracting the trend term of the NEE data from the 25 AmeriFlux sites, we train and evaluate the ML algorithms following the procedures explicated in Figures 2 and 4. As shown in Figure 6, the median performance of each ML algorithm degrades with increasing gap length for both EMD and MA combinations. Among the four ML algorithms, RF and XGboost have comparable performance for filling gaps of different lengths, and both outperform the SVR and BP neural network in each of the gap scenarios. For the short gap length scenario, all four ML algorithms have the highest R^2 and the lowest RMSE and MAE, as well as the lowest degree of scatter of bias. For the very long gap length scenario, both RF and XGboost have median values of R^2 of around 0.85 and 0.88 when combined with EMD and MA, respectively, while SVR and BP neural network have median R^2 values of around 0.78 and 0.82 when combined with EMD and MA, respectively. Compared to RF and XGboost, SVR and BP neural network display a larger decrease in R^2 and a larger increase in RMSE and MAE with increasing gap length. In addition, there are no significant differences between the performances of the ML algorithms when carrying NEE gap filling over different vegetation types. Overall, the ML algorithms combined with MA slightly outperform the ML algorithms combined with EMD.

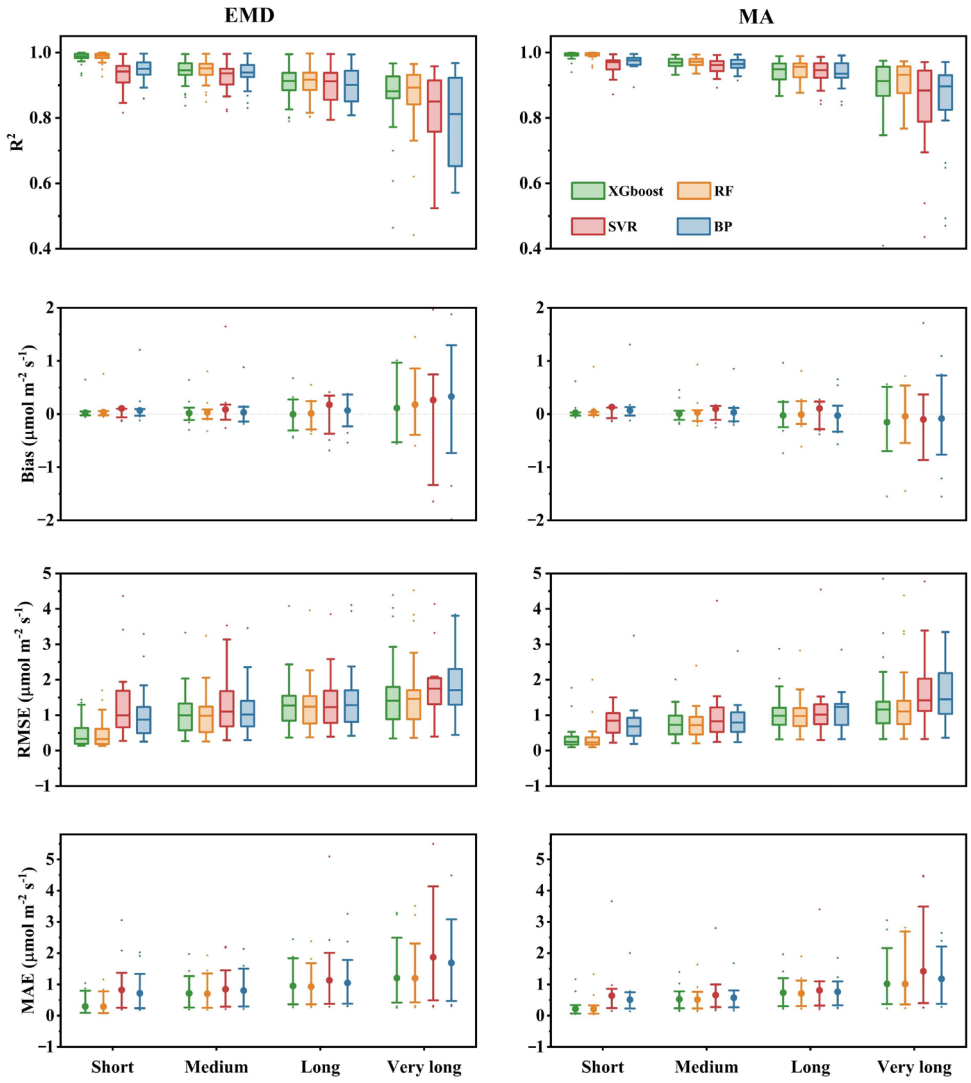

Figure 6. The median performance metrics of the proposed gap-filling framework for short, medium, long, and very long gap length scenarios, using two different time series decomposition methods (EMD and MA) in combination with four machine learning approaches (XGboost, RF, SVR, and BP neural network), respectively.

4.3. Relative Importance of the Input Variables

Figure 7 shows the relative importance of the input variables for the trend term of the NEE data at each site. Here, we use the absolute value of Pearson's correlation coefficient to characterize relative importance. Each row represents a driver and each column a site, and the color bar shows the absolute value of the correlation coefficient. For the trend term of the NEE data, EVI and NDVI are the most important variables at most of the sites, and the relative importance of the other variables varies among different sites, with precipitation being the least influential predictor. For the fluctuation term, neither EVI nor NDVI have any significant impacts on the variations of NEE, while VPD and rain are the most important variables. These results suggest that the variations in vegetation phenology play a critical role in the prediction of the trend term in ecosystem carbon flux.

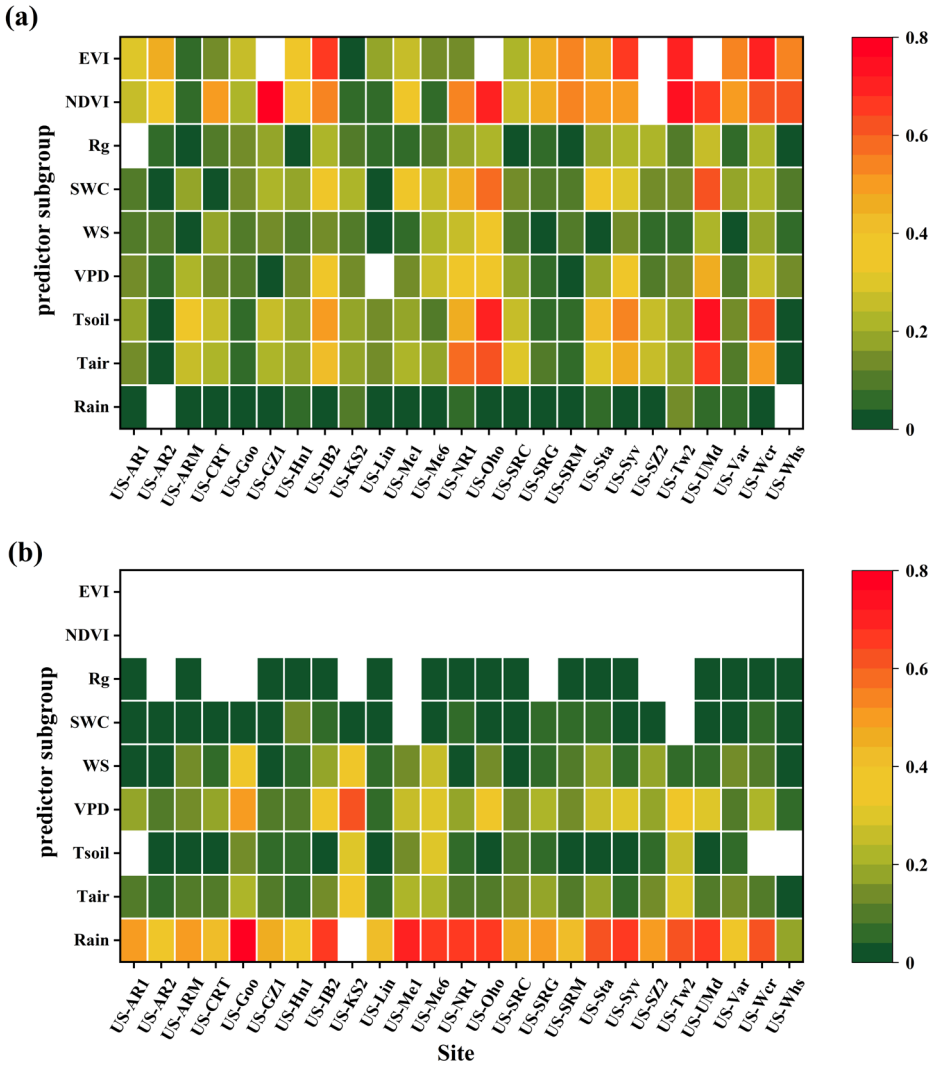

Figure 7. The relative importance of input variables for (**a**) the trend term and (**b**) the fluctuation term, respectively, at each site. White-colored spots refer to values with statistical significance lower than the 95% level.

4.4. Comparison with Results in a Peer's Study

In this subsection, we compare the performance of the proposed framework with the results of a recent study by Zhu et al. [28], in which an RF-based model with ten driving variables (RFR_{10}) outperformed other gap-filling models. The RFR_{10} was evaluated by filling artificial gaps of 24 h (24 H) and 7 days (7 D) in flux data at 16 AmeriFlux sites. Here, we select the same AmeriFlux sites and calculate the values of the same statistical metrics for EMD-RF and MA-RF with the two gap length scenarios. As shown in Figure 8, both EMD-RF and MA-RF outperform RFR_{10}, with larger values of R^2 and smaller values of RMSE. On average, for a gap length scenario of seven days, MA-RF has better performance, with an increase in R^2 by 0.24 (34%) and a decrease in RMSE by 1.30 μmol m^{-2} s^{-1} (55%), respectively, compared to RFR_{10}. Therefore, the proposed frameworks of EMD-RF and MA-RF outperform the RFR_{10}; however, due to the potential difference in the artificial gaps generated for testing, it is hard to quantify the improvement of the proposed framework for NEE gap filling compared to the RFR_{10}.

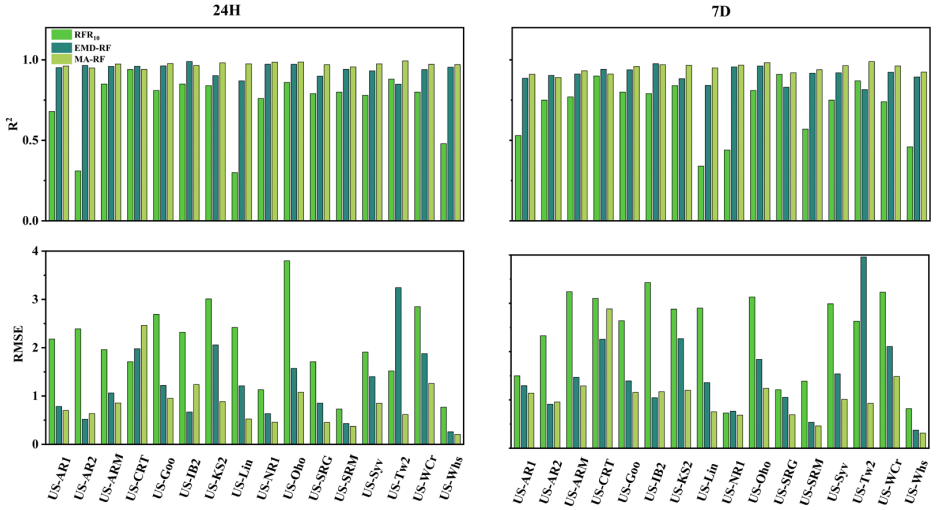

Figure 8. Comparison of model performance of the proposed framework with results of a peer's study for gap lengths of 24 h and 7 days, respectively.

4.5. Consistency of the Annual Total NEE Filled Using Different Methods

To assess the impacts of different gap-filling approaches on the annual total NEE, in this subsection, we compare the accumulated annual NEE using the gap-filled NEE data at these sites. As shown above, MA-RF outperforms the other models, and thus we use the NEE data filled using MA-RF (i.e., $NEE_{MA\text{-}RF}$) as a reference. The gap-filled NEE data [1] from the AmeriFlux FLUXNET data product ($NEE_{FLUXNET}$) are also obtained for comparison. As shown in Figure 9, the difference between the annual total $NEE_{MA\text{-}RF}$ and $NEE_{FLUXNET}$ is approximately 5%, with an R^2 of 0.96. The highest agreement is found between $NEE_{MA\text{-}RF}$ and $NEE_{MA\text{-}XGboost}$, with a linear regression coefficient of 0.99 and an R^2 of 0.99, respectively. Overall, the results suggest that the proposed framework of RF and XGboost in combination with MA or EMD can provide relatively consistent gap filling for NEE data. The comparison of the cumulative monthly NEE of MA-RF and FLUXNET in Figure A4 in Appendix A also shows a relatively good agreement, with constrained scattering.

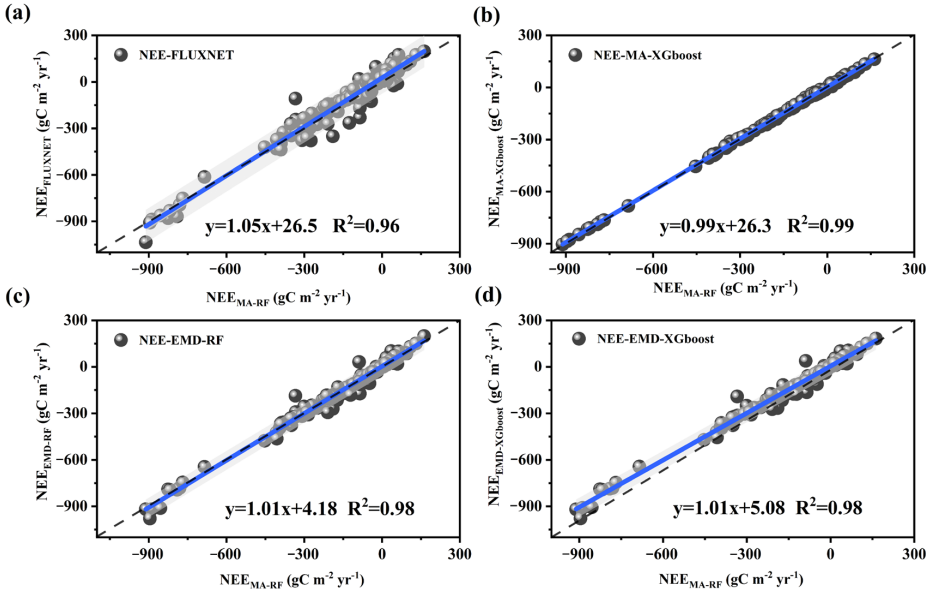

Figure 9. Comparison of the annual total NEE of the best gap-filling framework (MA-RF) with the other gap-filling approaches. (**a**) MA-RF and FLUXNET Comparison. (**b**) MA-RF and MA-XGboost Comparison. (**c**) MA-RF and EMD-RF Comparison. (**d**) MA-RF and EMD-XGboost Comparison.

5. Discussion

In this study, we proposed and evaluated a novel framework to perform gap filling in eddy covariance NEE data. The proposed framework is based on the concept that the observed variations in the NEE data stand for the combined effects of different environmental and weather factors, while the annual total NEE is largely determined by the low-frequency variations (the trend term) in the NEE data (Figure 3). The proposed framework can thus be summarized into three steps: (1) pre-fill the gaps using the RF algorithm (i.e., the first-layer ML model); (2) decompose the pre-filled NEE data to extract the trend term; and (3) refill the gaps in the trend term using the ML algorithm (i.e., the second-layer ML model) and reconstruct the time series by adding the refilled trend term and the extracted fluctuation term together (Figure 2). The commonly used RF-based gap-filling method is applied in the first step because different studies have shown that RF has relatively better performance at filling NEE gaps than other other ML algorithms [23,25,28]. The time series decomposition is performed using the MA and EMD methods, and on average, the trend term accounts for approximately 90% of the total annual fluxes, while the high-frequency term has a distribution similar to the normal distribution (Figures A1 and A2 in Appendix A).

The two time series decomposition methods, combined with four ML algorithms, are trained and evaluated with respect to their capacity to fill the gaps in the trend term. Using the NEE data from 25 AmeriFlux sites, our results suggest that MA combined with RF (MA-RF) outperforms the other combinations, as well as the traditional RF-based gap-filling model (Figures 5 and 6). The improved performance is most likely due to the reduced complexity of the NEE data in the trend term, which can ease the training and prediction of the ML algorithm. On the other hand, it implies that the trend term is generally able to capture the annual/seasonal variations in NEE. Therefore, an accurate prediction of the trend term can improve the accuracy of annual total NEE estimation.

The spatial complexity or variability of the targeted CO_2 flux and meteorological factors within the tower footprint [46,47] could be one of the reasons for the variable performance of gap-filling models at different sites. Huang and Hsieh [26] found that the ML algorithms showed better performance at forest and cropland sites than at grassland sites. Zhu et al. [28] also suggested that the performance of the gap-filling models varied as a function of ecosystem landcover classification. Figure A3 in Appendix A compares the performance of the model framework for different landcover types, with the RMSE for agricultural land showing relatively greater differences. This seems to be reasonably attributable to the fact that farmland crops often change, and are greatly affected by human activities (i.e., harvesting). Overall, MA-RF outperforms the other models and maintains a relatively good performance across different vegetation types. One plausible explanation could be that the variability of NEE is greatly reduced once the trend term has been extracted, resulting in improved performance among the gap-filling models.

The selection of inputs or driving variables for the ML algorithms can also influence the model performance. With an RF-based gap-filling model, Zhu et al. [28] found that the performance improved by 15% when using ten driving variables instead of three inputs. Yao et al. [25] found that Rg was the most important driving variable for the RF-based model. In this study, for the trend term, the vegetation phenology was the most important input, because the trend term represents the low-frequency variations in the NEE data. Overall, the results suggested that the proposed framework can indeed serve the purpose of improving the gap-filling performance, and that the trend term is better able to describe the annual–seasonal trend of NEE.

Although our results suggest that the model performance for filling gaps longer than 7 days is greatly improved, there is still a lot of room for improvement. For example, the results obtained using the EMD method were not consistent with those with the MA method at several sites, and therefore the performance of the model is degraded at individual sites (Figure 8). Additionally, considering the possible bias and distortion in the long gaps after pre-filling, future research should focus more on improving the performance when filling long gaps and the selection of driving variables.

6. Conclusions

In this study, a novel framework was proposed for gap filling in NEE data, and its performance was evaluated at multiple sites with different types of vegetation cover. The core of the research idea was the assumption that the annual/seasonal trend of NEE data dominates interannual and annual variations in NEE. Therefore, the pre-filled NEE data was decomposed into two time series representing the trend and fluctuation terms, respectively. Different time series decomposition methods and ML algorithms were combined to develop a robust model for gap filling in NEE data. The specific conclusions drawn are as follows: (1) the method of pre-filling followed by decomposition and re-filling reaps better gap-filling results; (2) among the two decomposition schemes, MA is slightly better than EMD decomposition; among the four algorithms, RF and XGboost exhibit higher performance than the SVR and BP neural networks; and overall, the framework using a moving average combined with RF has the best performance; (3) compared to a single-layer RF-based gap-filling model, the R^2 of the proposed framework with RF increased by 19%, and RMSE decreased by around 1.3 μmol m^{-2} s^{-1}. In addition, our results suggest that, for the trend term, the vegetation phenology is the most important predictor.

Author Contributions: Conceptualization, Z.G., D.G. and J.Y.; methodology, Z.G. and D.G.; software, D.G. and Z.G.; validation, D.G., Z.G., J.Y., S.Y., Y.M. and L.L.; formal analysis, D.G., Z.G., J.Y., Y.M. and L.L.; investigation, D.G., Z.G., J.Y., S.Y., Y.M. and L.L.; data curation, Z.G., D.G. and J.Y.; writing—original draft preparation, D.G. and Z.G.; writing—review and editing, Z.G., J.Y., S.Y., Y.M. and L.L.; visualization, D.G. and S.Y.; supervision, Z.G. and J.Y.; project administration, Z.G.; funding acquisition, L.L. and J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number U21A6001; the China National Postdoctoral Program for Innovative Talents, grant number BX20220366; and the China National Postdoctoral Program, grant number 2022M723576. The APC was funded by BX20220366 and 2022M723576.

Data Availability Statement: The original data used in this study can be downloaded from the websites provided in Section 2. Other processed data used in this study are available at https://doi.org/10.6084/m9.figshare.22598095.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Table A1. Grid search for the best parameters.

ML Algorithms	Best Parameters
RF	n_estimators = 1636, min_samples_split = 5, min_saples_leaf = 2, max_features = 0.5, max_depth = None, bootstrap = False, random_state = 0
XGboost	Subsample = 0.8, seed = 0, reg_lambda = 1, reg_alpha = 0, n_jobs = −1, n_estimtors = 3333, min_child_weight = 5, max_depth = 298, learning_rate = 0.01, gamma = 0.0, colsample_bytree = 0.7
SVR	kernel = 'rbf', gamma = 0.1, C = 100
BP neural network	input layer-intermediate layer-output layer = 120-10-1, activation function: sigmoid, trained 200 times

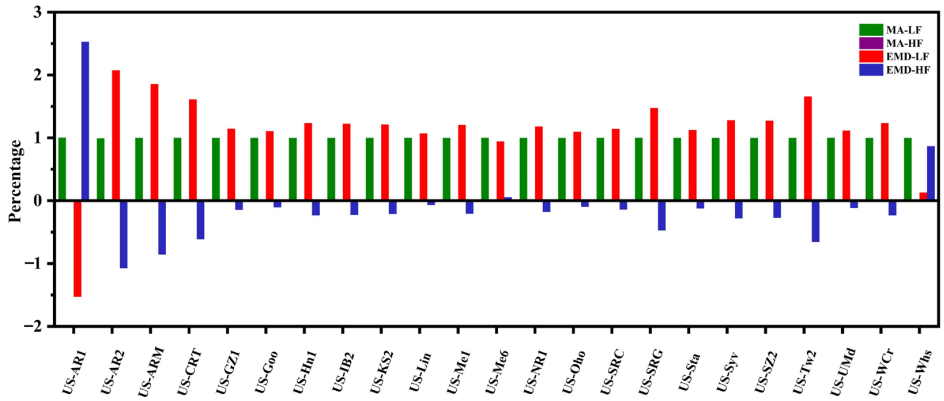

Figure A1. The ratio of the contribution of low- and high-frequency components to total NEE after MA and EMD decomposition.

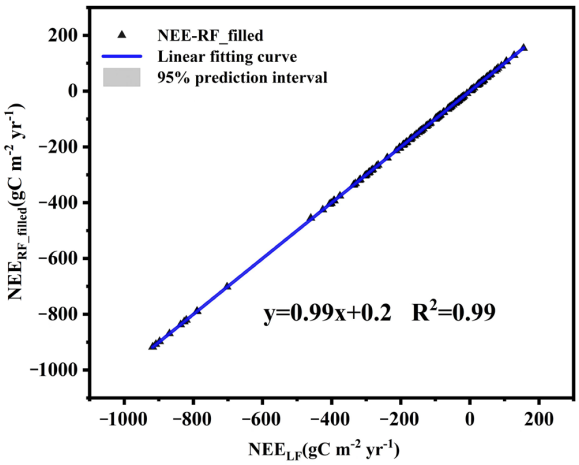

Figure A2. Comparison of MA decomposition trend terms and RF pre-filled annual total NEE.

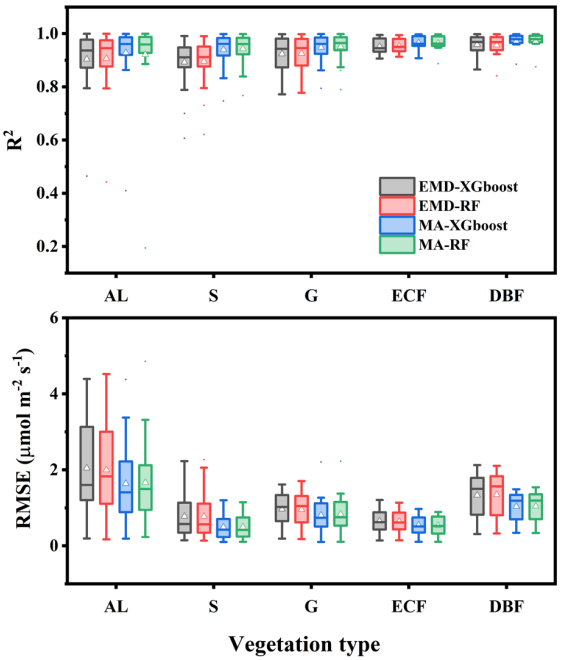

Figure A3. Comparison of the mean statistics of model performance for five different vegetation types. The five vegetation types are agricultural land (AL), shrubland (S), grassland (G), evergreen coniferous forest (ECF), and deciduous broadleaf forest (DBF).

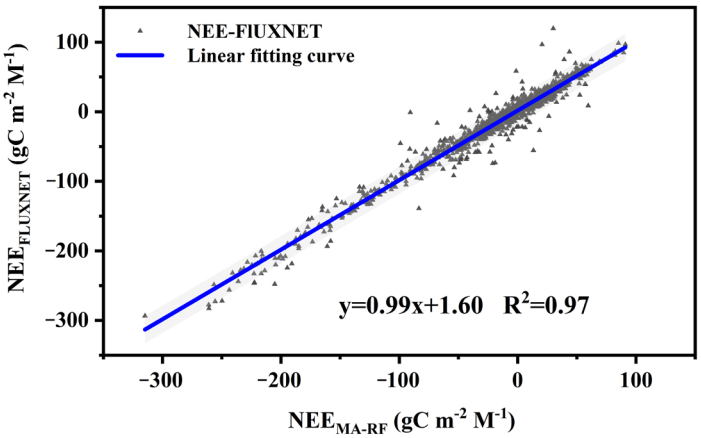

Figure A4. Comparison of monthly total NEE of the optimal gap-filling framework (MA-RF) using the FLUXNET gap-filling method.

References

1. Pastorello, G.; Trotta, C.; Canfora, E.; Chu, H.; Christianson, D.; Cheah, Y.W.; Poindexter, C.; Chen, J.; Elbashandy, A.; Humphrey, M.; et al. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Sci. Data* **2020**, *7*, 225. [CrossRef] [PubMed]
2. Twine, T.E.; Kustas, W.P.; Norman, J.M.; Cook, D.R.; Houser, P.R.; Meyers, T.P.; Prueger, J.H.; Starks, P.J.; Wesely, M.L. Correcting eddy-covariance flux underestimates over a grassland. *Agric. For. Meteorol.* **2000**, *103*, 279–300. [CrossRef]
3. Loescher, H.W.; Oberbauer, S.F.; Gholz, H.L.; Clark, D.B. Environmental controls on net ecosystem-level carbon exchange and productivity in a Central American tropical wet forest. *Glob. Chang. Biol.* **2003**, *9*, 396–412. [CrossRef]
4. Baldocchi, D.D. How eddy covariance flux measurements have contributed to our understanding of Global Change Biology. *Glob. Chang. Biol.* **2020**, *26*, 242–260. [CrossRef] [PubMed]
5. Goulden, M.L.; Munger, J.W.; Fan, S.-M.; Daube, B.C.; Wofsy, S.C. Exchange of Carbon Dioxide by a Deciduous Forest: Response to Interannual Climate Variability. *Science* **1996**, *271*, 1576–1578. [CrossRef]
6. Schimel, D.; Melillo, J.; Tian, H.; McGuire, A.D.; Kicklighter, D.; Kittel, T.; Rosenbloom, N.; Running, S.; Thornton, P.; Ojima, D.; et al. Contribution of increasing CO_2 and climate to carbon storage by ecosystems in the United States. *Science* **2000**, *287*, 2004–2006. [CrossRef]
7. Baldocchi, D.; Falge, E.; Gu, L.; Olson, R.; Hollinger, D.; Running, S.; Anthoni, P.; Bernhofer, C.; Davis, K.; Evans, R.; et al. FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem–Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities. *Bull. Am. Meteorol. Soc.* **2001**, *82*, 2415–2434. [CrossRef]
8. Dragoni, D.; Schmid, H.P.; Grimmond, C.S.B.; Loescher, H.W. Uncertainty of annual net ecosystem productivity estimated using eddy covariance flux measurements. *J. Geophys. Res.* **2007**, *112*. [CrossRef]
9. Falge, E.; Baldocchi, D.; Olson, R.; Anthoni, P.; Aubinet, M.; Bernhofer, C.; Burba, G.; Ceulemans, R.; Clement, R.; Dolman, H.; et al. Gap filling strategies for defensible annual sums of net ecosystem exchange. *Agric. For. Meteorol.* **2001**, *107*, 43–69. [CrossRef]
10. Lee, S.-C.; Christen, A.; Black, T.A.; Jassal, R.S.; Briegel, F.; Nesic, Z. Combining flux variance similarity partitioning with artificial neural networks to gap-fill measurements of net ecosystem production of a Pacific Northwest Douglas-fir stand. *Agric. For. Meteorol.* **2021**, *303*, 108382. [CrossRef]
11. Missik, J.E.C.; Liu, H.; Gao, Z.; Huang, M.; Chen, X.; Arntzen, E.; McFarland, D.P.; Ren, H.; Titzler, P.S.; Thomle, J.N.; et al. Groundwater—River Water Exchange Enhances Growing Season Evapotranspiration and Carbon Uptake in a Semiarid Riparian Ecosystem. *J. Geophys. Res. Biogeosci.* **2019**, *124*, 99–114. [CrossRef]
12. Moffat, A.M.; Papale, D.; Reichstein, M.; Hollinger, D.Y.; Richardson, A.D.; Barr, A.G.; Beckstein, C.; Braswell, B.H.; Churkina, G.; Desai, A.R.; et al. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agric. For. Meteorol.* **2007**, *147*, 209–232. [CrossRef]
13. Soloway, A.D.; Amiro, B.D.; Dunn, A.L.; Wofsy, S.C. Carbon neutral or a sink? Uncertainty caused by gap-filling long-term flux measurements for an old-growth boreal black spruce forest. *Agric. For. Meteorol.* **2017**, *233*, 110–121. [CrossRef]
14. Wutzler, T.; Lucas-Moffat, A.; Migliavacca, M.; Knauer, J.; Sickel, K.; Šigut, L.; Menzer, O.; Reichstein, M. Basic and extensible post-processing of eddy covariance flux data with REddyProc. *Biogeosciences* **2018**, *15*, 5015–5030. [CrossRef]
15. Whelan, A.; Mitchell, R.; Staudhammer, C.; Starr, G. Cyclic occurrence of fire and its role in carbon dynamics along an edaphic moisture gradient in longleaf pine ecosystems. *PLoS ONE* **2013**, *8*, e54045. [CrossRef]
16. Irvin, J.; Zhou, S.; McNicol, G.; Lu, F.; Liu, V.; Fluet-Chouinard, E.; Ouyang, Z.; Knox, S.H.; Lucas-Moffat, A.; Trotta, C.; et al. Gap-filling eddy covariance methane fluxes: Comparison of machine learning model predictions and uncertainties at FLUXNET-CH4 wetlands. *Agric. For. Meteorol.* **2021**, *308–309*, 108528. [CrossRef]

17. Wilson, K.; Goldstein, A.; Falge, E.; Aubinet, M.; Baldocchi, D.; Berbigier, P.; Bernhofer, C.; Ceulemans, R.; Dolman, H.; Field, C.; et al. Energy balance closure at FLUXNET sites. *Agric. For. Meteorol.* **2002**, *113*, 223–243. [CrossRef]
18. Boudhina, N.; Zitouna-Chebbi, R.; Mekki, I.; Jacob, F.; Ben Mechlia, N.; Masmoudi, M.; Prévot, L. Evaluating four gap-filling methods for eddy covariance measurements of evapotranspiration over hilly crop fields. *Geosci. Instrum. Methods Data Syst.* **2018**, *7*, 151–167. [CrossRef]
19. Falge, E.; Baldocchi, D.; Olson, R.; Anthoni, P.; Aubinet, M.; Bernhofer, C.; Burba, G.; Ceulemans, R.; Clement, R.; Dolman, H.; et al. Gap filling strategies for long term energy flux data sets. *Agric. For. Meteorol.* **2001**, *107*, 71–77. [CrossRef]
20. Reichstein, M.; Falge, E.; Baldocchi, D.; Papale, D.; Aubinet, M.; Berbigier, P.; Bernhofer, C.; Buchmann, N.; Gilmanov, T.; Granier, A.; et al. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm. *Glob. Chang. Biol.* **2005**, *11*, 1424–1439. [CrossRef]
21. Wang, E.; Smith, C.J.; Macdonald, B.C.T.; Hunt, J.R.; Xing, H.; Denmead, O.T.; Zeglin, S.; Zhao, Z.; Isaac, P. Making sense of cosmic-ray soil moisture measurements and eddy covariance data with regard to crop water use and field water balance. *Agric. Water Manag.* **2018**, *204*, 271–280. [CrossRef]
22. Hui, D.; Wan, S.; Su, B.; Katul, G.; Monson, R.; Luo, Y. Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. *Agric. For. Meteorol.* **2004**, *121*, 93–111. [CrossRef]
23. Kim, Y.; Johnson, M.S.; Knox, S.H.; Black, T.A.; Dalmagro, H.J.; Kang, M.; Kim, J.; Baldocchi, D. Gap-filling approaches for eddy covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal component analysis. *Glob. Chang. Biol.* **2020**, *26*, 1499–1518. [CrossRef] [PubMed]
24. Richardson, A.D.; Hollinger, D.Y. A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO_2 flux record. *Agric. For. Meteorol.* **2007**, *147*, 199–208. [CrossRef]
25. Yao, J.; Gao, Z.; Huang, J.; Liu, H.; Wang, G. Technical note: Uncertainties in eddy covariance CO_2 fluxes in a semiarid sagebrush ecosystem caused by gap-filling approaches. *Atmos. Chem. Phys.* **2021**, *21*, 15589–15603. [CrossRef]
26. Huang, H.; Hsieh, C. Gap-Filling of Surface Fluxes Using Machine Learning Algorithms in Various Ecosystems. *Water* **2020**, *12*, 3415. [CrossRef]
27. Mahabbati, A.; Beringer, J.; Leopold, M.; McHugh, I.; Cleverly, J.; Isaac, P.; Izady, A. A comparison of gap-filling algorithms for eddy covariance fluxes and their drivers. *Geosci. Instrum. Methods Data Syst.* **2021**, *10*, 123–140. [CrossRef]
28. Zhu, S.; Clement, R.; McCalmont, J.; Davies, C.A.; Hill, T. Stable gap-filling for longer eddy covariance data gaps: A globally validated machine-learning approach for carbon dioxide, water, and energy fluxes. *Agric. For. Meteorol.* **2022**, *314*, 108777. [CrossRef]
29. Cui, X.; Goff, T.; Cui, S.; Menefee, D.; Wu, Q.; Rajan, N.; Nair, S.; Phillips, N.; Walker, F. Predicting carbon and water vapor fluxes using machine learning and novel feature ranking algorithms. *Sci. Total Environ.* **2021**, *775*, 145130. [CrossRef]
30. Liu, X.; Xu, J.; Yang, S.; Zhang, J. Rice evapotranspiration at the field and canopy scales under water-saving irrigation. *Meteorol. Atmos. Phys.* **2017**, *130*, 227–240. [CrossRef]
31. Ouyang, Z.; Chen, J.; Becker, R.; Chu, H.; Xie, J.; Shao, C.; John, R. Disentangling the confounding effects of PAR and air temperature on net ecosystem exchange at multiple time scales. *Ecol. Complex.* **2014**, *19*, 46–58. [CrossRef]
32. Pahari, R.; Leclerc, M.Y.; Zhang, G.; Nahrawi, H.; Raymer, P. Carbon dynamics of a warm season turfgrass using the eddy-covariance technique. *Agric. Ecosyst. Environ.* **2018**, *251*, 11–25. [CrossRef]
33. Li, C.; Li, Z.; Zhang, F.; Lu, Y.; Duan, C.; Xu, Y. Seasonal dynamics of carbon dioxide and water fluxes in a rice-wheat rotation system in the Yangtze-Huaihe region of China. *Agric. Water Manag.* **2023**, *275*, 107992. [CrossRef]
34. Eckhardt, T.; Knoblauch, C.; Kutzbach, L.; Holl, D.; Simpson, G.; Abakumov, E.; Pfeiffer, E.-M. Partitioning net ecosystem exchange of CO_2 on the pedon scale in the Lena River Delta, Siberia. *Biogeosciences* **2019**, *16*, 1543–1562. [CrossRef]
35. Natali, S.M.; Watts, J.D.; Rogers, B.M.; Potter, S.; Ludwig, S.M.; Selbmann, A.-K.; Sullivan, P.F.; Abbott, B.W.; Arndt, K.A.; Birch, L.; et al. Large loss of CO2 in winter observed across the northern permafrost region. *Nat. Clim. Chang.* **2019**, *9*, 852–857. [CrossRef]
36. Whelan, A.; Starr, G.; Staudhammer, C.L.; Loescher, H.W.; Mitchell, R.J. Effects of drought and prescribed fire on energy exchange in longleaf pine ecosystems. *Ecosphere* **2015**, *6*, art128. [CrossRef]
37. Gao, Z.M.; Liu, H.P.; Missik, J.E.C.; Yao, J.Y.; Huang, M.Y.; Chen, X.Y.; Arntzen, E.; Mcfarland, D.P. Mechanistic links between underestimated CO2 fluxes and non-closure of the surface energy balance in a semi-arid sagebrush ecosystem. *Environ. Res. Lett.* **2019**, *14*, 044016. [CrossRef]
38. Chu, D.; Shen, H.; Guan, X.; Chen, J.M.; Li, X.; Li, J.; Zhang, L. Long time-series NDVI reconstruction in cloud-prone regions via spatio-temporal tensor completion. *Remote Sens. Environ.* **2021**, *264*, 112632. [CrossRef]
39. Foken, T. The energy balance closure problem: An overview. *Ecol. Appl.* **2008**, *18*, 1351–1367. [CrossRef]
40. Huang, N.E.; Wu, Z. A review on Hilbert-Huang transform: Method and its applications to geophysical studies. *Rev. Geophys.* **2008**, *46*. [CrossRef]
41. Barnhart, B.L.; Nandage, H.K.W.; Eichinger, W. Assessing Discontinuous Data Using Ensemble Empirical Mode Decomposition. *Adv. Adapt. Data Anal.* **2012**, *3*, 483–491. [CrossRef]
42. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
43. Chen, T.; Guestrin, C.J.A. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
44. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]

45. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
46. Chu, H.; Luo, X.; Ouyang, Z.; Chan, W.S.; Dengel, S.; Biraud, S.C.; Torn, M.S.; Metzger, S.; Kumar, J.; Arain, M.A.; et al. Representativeness of Eddy-Covariance flux footprints for areas surrounding AmeriFlux sites. *Agric. For. Meteorol.* **2021**, *301-302*, 108350. [CrossRef]
47. Stoy, P.C.; Mauder, M.; Foken, T.; Marcolla, B.; Boegh, E.; Ibrom, A.; Arain, M.A.; Arneth, A.; Aurela, M.; Bernhofer, C.; et al. A data-driven analysis of energy balance closure across FLUXNET research sites: The role of landscape scale heterogeneity. *Agric. For. Meteorol.* **2013**, *171–172*, 137–152. [CrossRef]