



OPEN

A widely-used eddy covariance gap-filling method creates systematic bias in carbon balance estimates

Henriikka Vekuri¹✉, Juha-Pekka Tuovinen¹, Liisa Kulmala¹, Dario Papale^{2,3}, Pasi Kolari⁴, Mika Aurela¹, Tuomas Laurila¹, Jari Liski¹ & Annalea Lohila^{1,4}

Climate change mitigation requires, besides reductions in greenhouse gas emissions, actions to increase carbon sinks in terrestrial ecosystems. A key measurement method for quantifying such sinks and calibrating models is the eddy covariance technique, but it requires imputation, or gap-filling, of missing data for determination of annual carbon balances of ecosystems. Previous comparisons of gap-filling methods have concluded that commonly used methods, such as marginal distribution sampling (MDS), do not have a significant impact on the carbon balance estimate. By analyzing an extensive, global data set, we show that MDS causes significant carbon balance errors for northern (latitude > 60°) sites. MDS systematically overestimates the carbon dioxide (CO₂) emissions of carbon sources and underestimates the CO₂ sequestration of carbon sinks. We also reveal reasons for these biases and show how a machine learning method called extreme gradient boosting or a modified implementation of MDS can be used to substantially reduce the northern site bias.

Climate change is one of the most severe challenges faced by humankind. In addition to limiting greenhouse gas emissions from fossil fuels and land use, it is necessary to find effective ways of sequestering carbon (C), most notably carbon dioxide (CO₂), already present in the atmosphere. Natural climate solutions, such as climate-smart agriculture, afforestation, reforestation and peatland restoration, are considered the most feasible means for this^{1,2}. To make these solutions credible for climate policy and carbon markets, reliable verification of carbon sequestration is necessary³. The verification also involves the micrometeorological eddy covariance (EC) technique, a key method to directly measure the CO₂ fluxes between ecosystems and the atmosphere⁴. The popularity of this method is manifest in the FLUXNET network, which has had over 900 EC sites worldwide registered over the years⁵. In principle, EC provides continuous data on the short-term net ecosystem exchange (NEE) of CO₂ with the atmosphere, which can be integrated temporally to determine the related carbon balance of an ecosystem. Even though EC measurements can be run continuously, in practice there are gaps in the collected data, for example due to technical failures and, most importantly, due to the need for filtering of the data collected in atmospheric conditions compromising the validity of the EC technique. For example, in the global FLUXNET2015 data set, with 1532 site-years of data⁶, on average 68% of the half-hourly CO₂ fluxes are missing⁷. Even if the site-years that have gaps longer than two months are excluded, the mean data coverage is 40%. Only 50 site-years have a coverage higher than 60% and only 5 site-years a coverage higher than 70%.

Various methods have been used to impute, or gap-fill, missing data, with methods ranging from simple linear interpolation and mean diurnal variation to more complex methods such as artificial neural networks (ANN). In a comparison of 15 CO₂ flux gap-filling methods, it was concluded that the effect of gap-filling is modest on the annual C balance and that the accuracy of the best-performing methods, which proved to be nonlinear regression, look-up table, marginal distribution sampling (MDS), a semi-parametric model and ANN, is already reaching the noise limit of measurements⁸. However, this comparison only included forest sites from a 20° latitudinal range. Other comparisons have missed the gap-filling methods most commonly used nowadays⁹, namely MDS and machine learning-based approaches, or focused on long gaps^{7,10,11}. Despite these shortcomings, different machine learning methods and especially MDS have become the standard methods for gap-filling

¹Finnish Meteorological Institute, 00101 Helsinki, Finland. ²DIBAF University of Tuscia, 01100 Viterbo, Italy. ³Euro-Mediterranean Center on Climate Change CMCC IAFES, 01100 Viterbo, Italy. ⁴Institute for Atmospheric and Earth System Research, Physics, University of Helsinki, 00014 Helsinki, Finland. ✉email: henriikka.vekuri@fmi.fi

EC data. Notably, MDS is used for gap-filling the standardized open-access NEE data provided by FLUXNET⁶ and the European research infrastructure Integrated Carbon Observation System (ICOS). MDS is also implemented within the free gap-filling tool REddyProc¹² and as part of ToviTM, a commercial software for EC data post-processing¹³. However, knowledge on the performance of different gap-filling methods, especially MDS, is lacking for data from northern high-latitude (latitude > 60°) sites. In northern ecosystems, growing seasons are short and the amount of solar radiation, a key environmental driver in CO₂ exchange, is distributed very unevenly throughout the year. Therefore, the amount of potentially available nighttime data during the short northern growing season is low even before filtering for data quality.

Here, we investigated the performance of two methods to gap-fill CO₂ flux time series with the aim to reveal if the gap-filling-induced uncertainties limit our abilities to verify carbon sequestration estimates in northern latitudes. The gap-filling methods considered were MDS and a machine learning method called eXtreme gradient boosting (XGBoost). XGBoost was chosen from a wide range of machine learning methods because decision tree-based methods have performed well in previous studies on gap-filling CO₂ and methane flux data^{7,10,14}. First, we compared the performance of MDS and XGBoost on the global FLUXNET2015 data set including annual time series of half-hourly data from the northern hemisphere that had at least a 20% temporal CO₂ flux data coverage. Next, we focused on the special case of northern ecosystems by investigating gap-filling performance for ten northern sites. We addressed a basic gap-filling task involving a maximum of three-day data gaps and the use of only readily available environmental drivers: air temperature (T_{air}), incoming shortwave radiation (SWR) and atmospheric vapour pressure deficit (VPD). More specifically, we wanted to know if the gap-filling methods (1) result in similar C balance estimates and (2) perform equally well in northern ecosystems as elsewhere. We did this by inserting artificial gaps into real and synthetic data sets, gap-filling the artificial gaps, and evaluating the performance of both methods. We adopted the mean flux bias as the main performance metric, as it directly translates to the accuracy of an annual balance. Another key metric that was used was the balance error, which is the sum of flux biases over missing data. This was calculated using synthetic data.

Results

Skewed radiation distribution causes a positive bias at northern sites when using MDS. In this work, the micrometeorological sign convention is adopted, which means that a positive CO₂ flux denotes a flux from the ecosystem to the atmosphere and a negative flux indicates uptake from the atmosphere into the ecosystem. Gap-filling of artificial gaps inserted into 882 site-years of data in the FLUXNET2015 data set revealed that with MDS there was a clear positive bias in the gap-filled fluxes during daytime (SWR ≥ 20 W m⁻²) in high latitudes (> 50°N, N = 105) (Fig. 1a; for statistical tests, see Supplementary Table S1). During nighttime (SWR < 20 W m⁻²), there was a negative, but much smaller bias (Fig. 1b) leading to a positive total flux bias (Fig. 1c). A positive flux bias indicates that either emission was overestimated or uptake was underestimated, while a negative bias indicates the opposite. When XGBoost was used for gap-filling, some positive and negative flux biases were observed (Fig. 1d,e), but the magnitude of these biases was small compared to the daytime bias

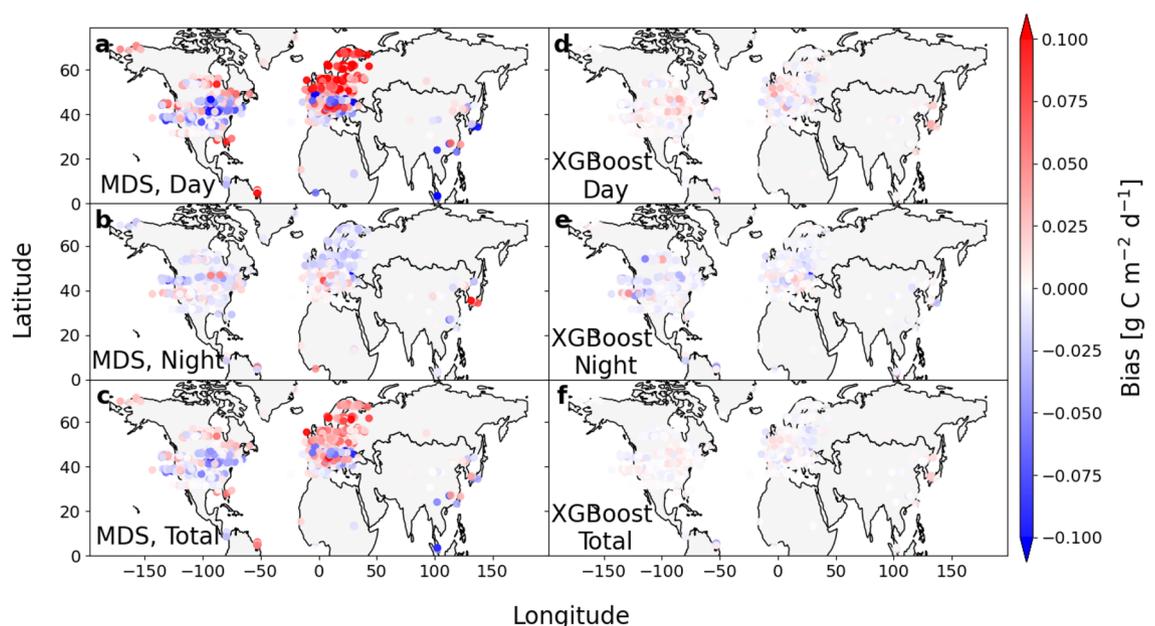


Figure 1. MDS causes a clear positive bias in the gap-filled NEE at northern latitudes. (a,b) Daytime, (c,d) nighttime and (e,f) total bias in the CO₂ flux data gap-filled using marginal distribution sampling (MDS) and extreme gradient boosting (XGBoost). The data cover 882 site-years from 141 sites from the global FLUXNET2015 data set. Outputs of each method were compared to measured data labelled as artificial gaps. The total, daytime and nighttime mean flux bias is plotted for each site-year.

of MDS. The total flux bias with XGBoost was insignificant or very small at all latitudes (Fig. 1f and Supplementary Table S1).

The MDS method is based on the covariation between NEE and meteorological variables and primarily gap-fills a missing NEE observation by the mean of the available NEE values measured under similar conditions. The similarity of meteorological conditions is determined based on sampling tolerances, or maximum accepted deviations from the actual conditions, specified for each predictor variable (for details see Methods). The positive daytime bias at the northern latitudes (Fig. 2a) resulted from a very skewed radiation distribution (Fig. 2b and Supplementary Fig. S1), causing more data to be sampled from the lower range of the radiation distribution (Fig. 2c). An underestimated radiation level corresponds to underestimated photosynthetic uptake (negative flux component) and thus to overestimated NEE (Fig. 2d). Examples using measured data are shown in Supplementary Fig. 2a,b. The imbalance of daytime and nighttime flux bias was obvious at latitudes 50°–70°, affecting 133 site-years of data which corresponds to 15% of the site-years in the selected subset of the FLUXNET2015 data set. At latitudes 30°–50° there was also a substantial number of site-years with a negative flux bias, indicating that the biases at those latitudes were caused by other reasons. Because the positive daytime flux bias with MDS was especially evident at latitudes 60°–70°, we investigated more thoroughly data from ten sites located within this zone (Table 1). These data were acquired from the ICOS Warm Winter 2020 EC flux product, since there were

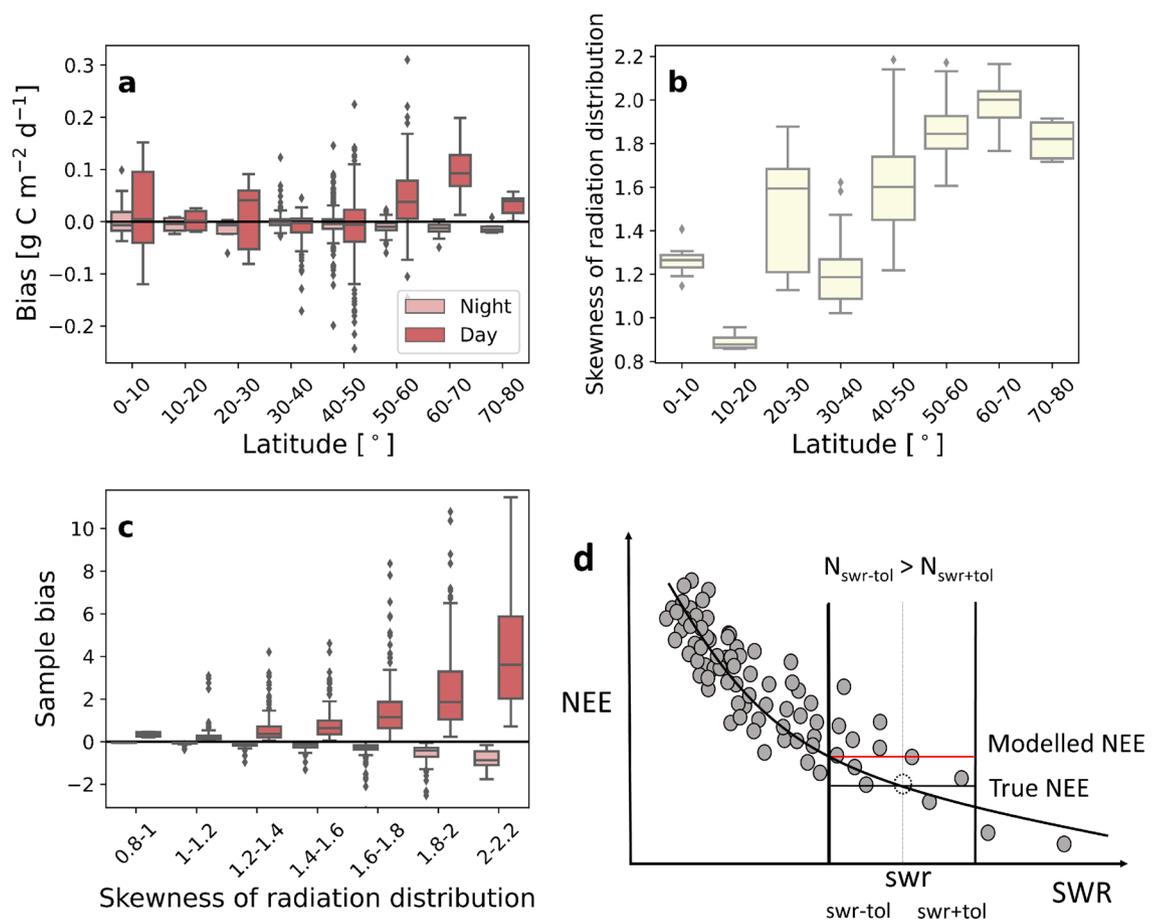


Figure 2. The radiation distribution is skewed at northern latitudes which results in a sample bias and a positive flux bias. (a) Daytime and nighttime CO₂ flux bias of MDS at different latitudes, (b) skewness of the distribution of incoming shortwave radiation at different latitudes, (c) sample bias during daytime and nighttime at different skewness values of the radiation distribution and (d) a schematic figure illustrating how the positive bias is produced. The black curve shows an example of the true dependence of NEE on SWR (a short period of daytime NEE during the growing season). Sample bias (c) indicates how many more data points are sampled with lower than higher radiation values compared to the measured value when selecting the NEE data that are averaged to impute a missing observation. If the number of measurements (N) within the tolerance interval below measured swr ($\text{swr-tol} \dots \text{swr}$) is higher than the number of measurements within the tolerance interval above it ($\text{swr} \dots \text{swr+tol}$), there is a positive sample bias, and the modelled NEE at certain swr is likely to be overestimated. The sample size ranged from 2 to 507 with a mean of 29. The data cover 882 site-years. All boxplots illustrate the median (solid line), quartiles (box) and rest of the distribution, excluding outliers, which are plotted individually. Outliers were defined as points that were outside 1.5 times the interquartile range. Sample biases over 12 ($N = 6$) and under -3 ($N = 4$) are not shown for clarity.

more site-years with good data coverage available for the northern latitudes¹⁵. The ICOS data product is fully compatible with the FLUXNET2015 data set.

Gap-filling of synthetic data reveals C balance errors with MDS. To assess what implications the detected bias has on the annually accumulated CO₂ fluxes, i.e. the estimates of annual C balances, we generated synthetic full time series corresponding to the CO₂ fluxes observed at the ten northern sites (Supplementary Figs. S3, S4 and Supplementary Table S2). We then introduced realistic, both in length and timing, artificial gaps (30%, 50% and 70% of data) into the synthetic data sets and compared the C balances derived from the gap-filled NEE time series to the true balances. This was done for two reasons. In the first part of the study, we assessed the flux bias based on the available measurement time series that were compromised by data gaps. However, most gaps fall into nighttime, meaning that the majority of the data used to assess the bias were daytime fluxes. Secondly, we wanted to know what is the combined effect of realistic gap lengths, timing and biases, which depend on the time of the day, on estimating the C balance.

We found that MDS systematically overestimated the annual C balance in all except one case (site SE-Nor with 70% gaps) (Fig. 3 and Supplementary Table S3). The site-specific median balance error with MDS ranged from 2–10 g C m⁻² y⁻¹ at the 30% gap level to 3–17 g C m⁻² y⁻¹ at the 70% gap level. For 5 out of the 10 data sets, balance errors exceeded 30 g C m⁻² y⁻¹ and the largest balance error was 42 g C m⁻² y⁻¹. Noteworthy is that the annual absolute balance error was similar for sites with synthetic balances ranging from tens (FI-Qvd, FI-Let) to hundreds (FI-Hyy, SE-Ros) of grams of C m⁻² y⁻¹. Also, when more gaps fell into the nighttime than daytime data, the annual error of MDS decreased, because the larger number of underestimated NEE values during nighttime compensated for the smaller number of overestimated NEE during daytime. In four out of thirty cases, there was a significant error in the balance calculated with XGBoost, but the median balance errors were only -4 to 2 g C m⁻² y⁻¹.

C balance errors can be decreased by using a modified version of MDS. To test if MDS could be improved by modifying the method and to further verify that the positive flux bias was caused by the biased radiation sampling, we investigated the model performance with three alterations using the synthetic data set FI-Let, for which MDS overestimated the most. We either (1) narrowed the two tolerance settings defining the SWR limits of each data sample, (2) used only one radiation tolerance, or (3) for daytime data, calculated separately the mean NEE of low- and high-SWR subsamples and then averaged these means.

When the SWR tolerances were different for lower and higher SWR levels, the positive flux bias during daytime was larger than the negative nighttime bias, causing a positive total bias and thus an overestimation of the C balance (Fig. 4a,c and Supplementary Tables S4 and S5). When the SWR classes were narrowed, the errors were smaller compared to the original implementation, but the root mean squared error (RMSE) of modelled NEE increased (Fig. 4b and Supplementary Table S6). When a common SWR tolerance of 25 W m⁻² was used, the error of the gap-filled C balance was insignificant when 30% or 50% of data was missing and negative when 70% of data was missing (Fig. 4a and Supplementary Table S4). Also in this case RMSE increased (Fig. 4c and Supplementary Table S6). Averaging the low- and high-SWR subsamples separately and taking their mean decreased the positive daytime flux bias and the error of the annual balance without affecting the RMSE (Fig. 4a–c and

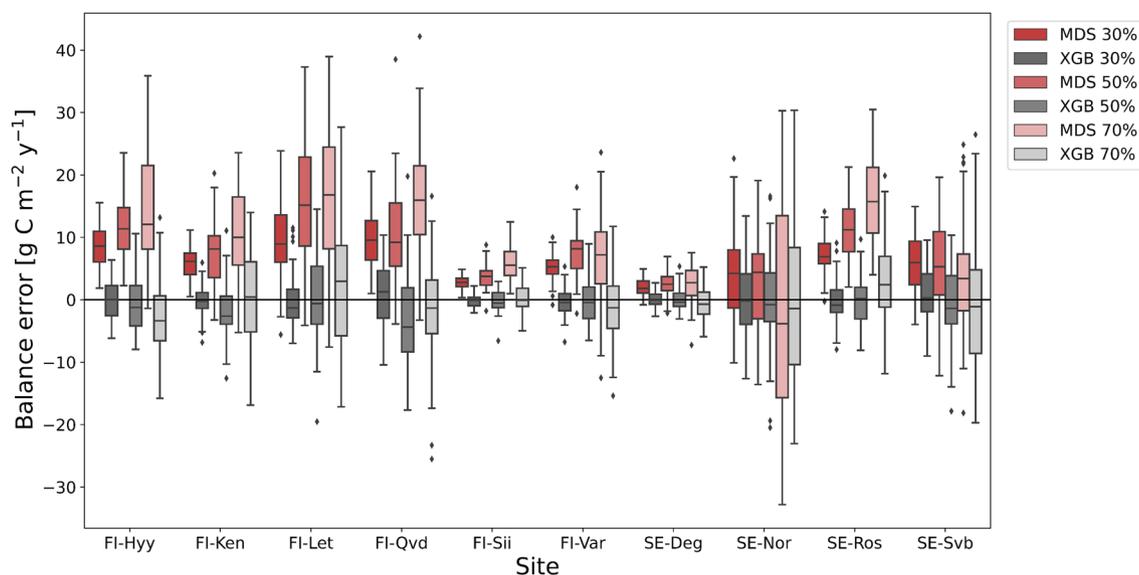


Figure 3. Gap-filling with MDS results in C balance errors. Errors of the gap-filled annual C balance for different gap percentages. A synthetic data set was generated five times for each site (Table 1), and each data coverage class was tested 10 times for each data set. Random gaps of 0.5 h to 3 days in length were generated based on the original gap distribution taking also into account the timing of gaps. Positive balance errors mean that the gap-filled C balance estimate indicates too large a source or too small a sink.

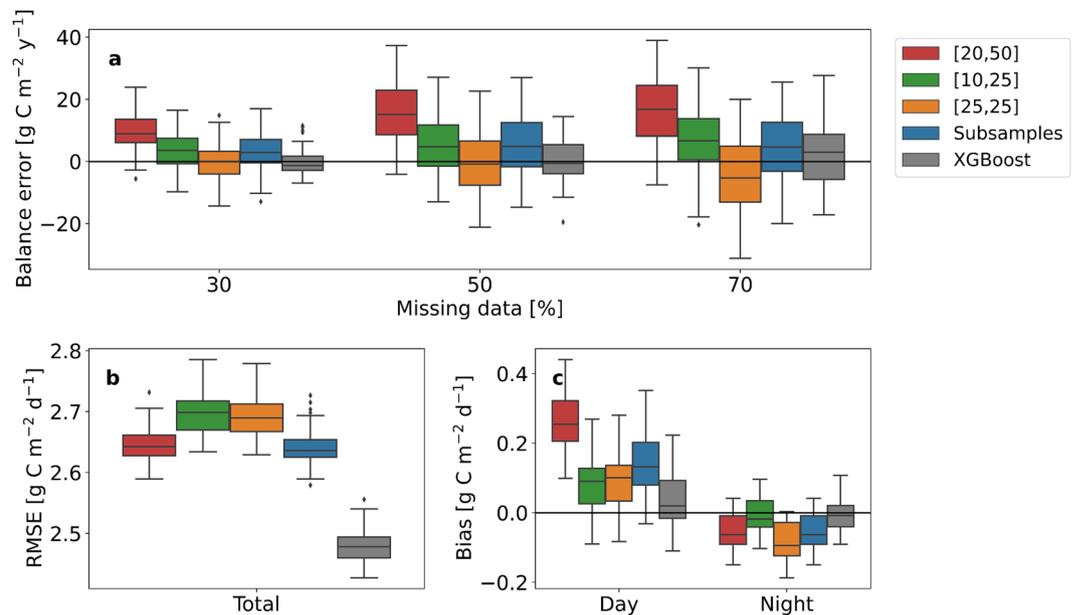


Figure 4. A modified version of MDS can be used to decrease C balance errors. Performance of different implementations of the MDS gap-filling method and XGBoost for the synthetic data set FI-Let. [20,50], [10,25] and [25,25] (W m^{-2}) indicate the sampling tolerances used for SWR. ‘Subsamples’ refers to averaging of the mean NEEs of the low- and high-SWR subsamples of data during daytime. [20,50] is the standard implementation of MDS. **(a)** Annual error, **(b)** root mean squared error (RMSE) and **(c)** daytime and nighttime bias of gap-filled synthetic CO_2 flux data. For evaluating errors, each data coverage class was tested 10 times, and RMSE and biases were calculated for data with 70% artificial gaps.

Supplementary Tables 4, 5). However, the balance error was not eliminated completely and an even lower RMSE was obtained by using XGBoost. Finally, we gap-filled all the synthetic data sets using the implementation of MDS with subsample averaging and found that only in one case (SE-Nor with 70% missing data) the magnitude of the annual error was larger than when using the original implementation of MDS (Supplementary Fig. S5 and Supplementary Table S7). In all other cases the error was either decreased or eliminated completely.

Discussion

Artificial differences among site-specific C balances can be generated by gap-filling, especially if methods with a tendency for both a negative and positive bias are used for different sites. This is a reason why flux networks have standardized their data processing. However, even when gap-filling with the same method, data imputation can lead to significant differences among sites because the biases are site-specific and dependent on data coverage. Previously, the random uncertainties at an annual timescale have been estimated to be $10\text{--}40 \text{ g C m}^{-2} \text{ y}^{-1}$ ^{16,17}, uncertainties related to gap-filling $10\text{--}30 \text{ g C m}^{-2} \text{ y}^{-1}$ ^{18,16} and the total uncertainty of the C balance at nearly ideal sites less than $\pm 50 \text{ g C m}^{-2} \text{ y}^{-1}$ ¹⁴. The magnitude of the annual errors found in this study is comparable to the estimated uncertainties, meaning that in some cases the true balance might not be captured by the confidence interval of the original estimate. We showed that a machine learning algorithm, here the extreme gradient boosting technique, can substantially reduce the gap-filling error.

In principle, any method that correctly accounts for the covariation between NEE and its meteorological drivers, such as other machine learning methods and regression models, could reduce the gap-filling error. The key problem with MDS in its original implementation is that the mean flux within a certain radiation interval is a biased estimate due to the favouring of lower radiation levels and therefore higher NEE. A different but similar problem could be observed in regression models if the assumed shape of the response of NEE to radiation is incorrect. The advantage of data-driven machine learning methods is that they do not make any, or make fewer, assumptions on the response function.

Even though standard methods aid in making comparisons, there is still a need to improve these methods, in particular considering specific environmental conditions like those at northern sites. It should also be noted that the amount of gap-filled, i.e. modelled, data is typically at least 50% and, therefore, it is crucial that the modelled data are unbiased. We showed that the original implementation of MDS is a sub-optimal method for gap-filling data from northern sites where the radiation distribution is very skewed. Other methods, such as XGBoost or the proposed modified MDS, should be considered in the standardized processing pipeline of FLUXNET, at least for the northern high-latitude (latitude $> 60^\circ$) sites. The results could hold also for other sites, such as southern high-latitude sites, and if applied to different drivers.

While the absolute errors found in this work may not seem large, systematic errors should be eliminated where possible. With increasing interest in the C sequestration potential of ecosystems, it is noteworthy that C balances of high-latitude ecosystems are generally small. The observed mean annual NEE has been reported to

Site id	Year	Name	Country	Latitude	Longitude	Site type	Data coverage
FI-Hyy	2015	Hyytiala	Finland	61.84741	24.29477	Evergreen needleleaf forests	0.53
FI-Ken	2020	Kenttarova	Finland	67.98721	24.24301	Evergreen needleleaf forests	0.30
FI-Let	2012	Lettosuo	Finland	60.64183	23.95952	Evergreen needleleaf forests	0.50
FI-Sii	2020	Siikaneva	Finland	61.83265	24.19285	Permanent wetlands	0.55
FI-Var	2018	Varrio	Finland	67.7549	29.61	Evergreen needleleaf forests	0.46
FI-Qvd	2020	Qvidja	Finland	60.29550	22.39281	Agricultural grassland	0.40
SE-Deg	2006	Degero	Sweden	64.18203	19.55654	Permanent wetlands	0.63
SE-Nor	2020	Norunda	Sweden	60.0865	17.479504	Evergreen needleleaf forests	0.53
SE-Svb	2020	Svartberget	Sweden	64.25611	19.7745	Evergreen needleleaf forests	0.40
SE-Ros	2018	Rosinedal-3	Sweden	64.1725	19.738	Evergreen needleleaf forests	0.39

Table 1. Site information for ten northern sites adopted for detailed gap-filling analysis.

be -17 g C m^{-2} in the high-latitude boreal and tundra biomes ($> 45^\circ$) which cover an area of $20.6 \times 10^6 \text{ km}^2$ ¹⁸. Therefore, the systematic gap-filling errors discovered here can have a significant relative impact on the C balance estimates of northern ecosystems, with implications for the verification of C sequestration.

Methods

Data. The FLUXNET2015 data set and the data acquired from an ICOS data product (Warm Winter 2020 ecosystem eddy covariance flux product for 73 stations in FLUXNET) consist of open-access eddy covariance CO_2 flux data and supporting measurements that have been processed in a standardized manner^{6,15}. The FLUXNET data have been collected from 206 sites distributed around the globe. From the FLUXNET2015 data set, we used all site-years from the northern hemisphere that had at least 20% annual coverage (Supplementary Table S8). We utilized measured NEE by selecting those values in the data product NEE_VUT_REF where the quality flag NEE_VUT_REF_QC was zero. The selected environmental drivers were shortwave radiation (SW_IN_F), atmospheric vapour pressure deficit (VPD_F_MDS) and air temperature (TA_F_MDS). Gaps in the drivers were filled according to the FLUXNET data processing protocol⁶. From the ICOS data product we used data from all northern ($> 60^\circ$) sites that had at least 30% annual coverage and T_{air} , VPD, SWR and soil temperature available. For each site we selected the site-year that had the highest data coverage (Table 1).

Gap-filling methods. Two gap-filling methods were used in this work: extreme gradient boosting and marginal distribution sampling.

The extreme gradient boosting algorithm is based on parallel boosted decision trees. The 'xgboost' Python package¹⁹ was used to apply this method. Hyperparameters, which control for the subsample ratio of columns when constructing each tree (0.4, 0.6, 0.8, 1), the maximum depth of a tree (3, 5, 10, 15), minimum number of samples required to create a new node in a tree (2, 5, 10) and the fraction of observations that is randomly sampled for each tree (0.65, 0.75, 1) were determined for the FLUXNET2015 data based on grid search using ten randomly selected data sets and by selecting the mode of each hyperparameter. For the northern sites, the same hyperparameters were optimized for each synthetic data set using the original artificial gap-free data and grid search. For learning rate we used the default value of 0.1. The squared error was used as the loss function. Since MDS uses a moving data window to resolve temporal correlations in the flux data, XGBoost was augmented by two cyclical functions for both month and time of day, and a linear description of time as additional drivers:

$$t_1 = \sin\left((\text{month} - 1) \times \frac{2\pi}{12}\right) \quad (1)$$

$$t_2 = \cos\left((\text{month} - 1) \times \frac{2\pi}{12}\right) \quad (2)$$

$$t_3 = \sin\left(\text{hour} \times \frac{2\pi}{24}\right) \quad (3)$$

$$t_4 = \cos\left(\text{hour} \times \frac{2\pi}{24}\right) \quad (4)$$

$$t_5 = i \quad (5)$$

where i is the number of half hours from the beginning of the year.

In MDS, gaps are filled with the average of fluxes measured under similar meteorological conditions using a moving window for data sampling, or with the mean diurnal course (MDC) if a sufficiently large sample could not

be found. The default meteorological drivers and their tolerances, or accepted deviations from actual conditions, are SWR, with a tolerance of 20 W m^{-2} for $\text{SWR} \leq 50 \text{ W m}^{-2}$ and 50 W m^{-2} for $\text{SWR} > 50 \text{ W m}^{-2}$, T_{air} with a tolerance of 2.5 K and VPD with a tolerance of 5 hPa. If T_{air} or VPD is missing, only SWR is used. If none of the meteorological drivers are available, the gaps are filled with MDC. The specific sampling procedure is described in Wutzler et al.¹². Noteworthy is that also the REdDyProc tool uses SWR tolerances of 20 W m^{-2} and 50 W m^{-2} (<https://github.com/bgctw/REddyProc/tree/1.1.3>) even though a single tolerance approach is reported.

In addition to the standard implementation of MDS, we tested modified versions of it. First, we only altered the radiation sampling limits. We tested using two different tolerances as in the original implementation, but lowered the SWR tolerances to 10 and 25 W m^{-2} . We also tested using only one tolerance at all radiation levels and used an SWR tolerance of 25 W m^{-2} for this. Finally, to better account for the sample bias, for daytime data, we first calculated the average NEE separately for the subsamples of data with higher and lower SWR than the present SWR and then averaged these two NEE values. The MDS variants were implemented using a C code by Papale et al.²⁰.

Synthetic data. An artificial neural network (ANN) was used to generate synthetic data sets with 100% coverage that could be used for comparing the gap-filling methods. The ANN used here was a sequential model with four hidden layers, with 16 nodes in the first hidden layer and 32 in the other layers. The activation functions that were used were linear, hyperbolic tangent (tanh) and rectified linear activation (relu), and the structure of the network was linear-tanh-tanh-relu-linear. Mean squared error was used as the loss function. The ANN was implemented using the Keras library²¹. Air temperature, soil temperature, SWR and VPD were used as predictors for the neural network. For each site, we used all the available measured data to train the ANN and after modelling for all 30-min periods of one year, noise was added to the modelled 30-min NEE. This was done by binning the model residuals based on the season (winter months 1–4 and 11–12, and summer months 5–10), time of day (nighttime with $\text{SWR} \leq 20 \text{ W m}^{-2}$ and daytime with $\text{SWR} > 20 \text{ W m}^{-2}$), and air temperature (five bins of equal size). After binning the data, a residual was randomly selected from the correct bin and added to each 30-min NEE. The whole procedure was repeated five times for each site to obtain 50 different synthetic data sets.

In practice, the synthetic data represent a case in which NEE is driven by T_{air} , SWR, VPD and soil temperature. We assumed that there is noise in the data and that it is of different magnitude during the active summer months and winter, during daytime and night and in different temperatures. When gap-filling the data with MDS and XGBoost, we assumed that they should be able to solve the covariation between NEE and T_{air} , SWR and VPD. We also assumed that the average errors of an unbiased gap-filling method would be zero.

Artificial gaps. Of the FLUXNET2015 data, we sampled all available measured data to calculate the mean bias for each site-year. For MDS we labelled one half hour at a time as an artificial gap since it is easy to implement in a moving window. For XGBoost we labelled one percent of the measured data as artificial gaps at a time since training a different model for each half hour was not reasonable due to computational time constraints. However, the results for XGBoost represent a conservative estimate since training a separate model for each individual half hour would improve rather than worsen the results.

For the synthetic data sets, the artificial gaps were drawn from the original gap distributions, taking into account both gap length and timing. More specifically, the gaps in the original time series were labelled by their length and starting time, and the artificial gaps were randomly chosen from the original gaps and inserted into the synthetic data sets until 30%, 50% or 70% coverage was achieved. When inserting the artificial gaps into the synthetic data, the starting time of each gap was retained. The artificial gaps were not allowed to overlap. For each data coverage we generated ten different gap sequences that were used for all synthetic data sets.

Performance measures. Comparisons between gap-filled and measured values were carried out using mean bias, RMSE and the error of the annual gap-filled C balance. In this paper, we defined the C balance as the C balance measured with EC excluding the potential impact of harvests, fertilization, leaching and lateral transport of C.

Mean bias and RMSE were defined as:

$$\text{Bias} = \frac{1}{N} \sum_{n=1}^N (\text{NEE}_{\text{gap},i} - \text{NEE}_{\text{meas},i}) \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^N (\text{NEE}_{\text{gap},i} - \text{NEE}_{\text{meas},i})^2}{N}} \quad (7)$$

where N equals the amount of artificial gaps, and $\text{NEE}_{\text{gap},i}$ indicates a gap-filled and $\text{NEE}_{\text{meas},i}$ a measured value.

The error of the gap-filled C balance was calculated as the difference between the gap-filled and true (synthetic) C balances:

$$\text{Balance error} = \sum_{n=1}^N \text{NEE}_{\text{gap},i} - \sum_{n=1}^N \text{NEE}_{\text{synth},i} \quad (8)$$

where N equals the number of 30-min NEE in the data and $\text{NEE}_{\text{synth},i}$ a synthetic NEE value.

Statistical tests. To determine whether the mean biases of half hourly NEE and the errors of the gap-filled carbon balances differed from zero, we used the non-parametric Wilcoxon signed-rank test. Normality of the data was first evaluated using the Shapiro–Wilk test. Pairwise comparisons were done using Conover’s test and Holm’s method to adjust p-values. All statistical tests were performed in Python using the SciPy package²².

Data availability

The FLUXNET2015 data set is available at <http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/>. The ICOS Warm Winter 2020 ecosystem eddy covariance flux product for 73 stations in FLUXNET-Archive format-release 2022-1 is available at <http://www.icos-cp.eu/data-products/2G60-ZHAK>.

Code availability

Code for making synthetic data and gap-filling using XGBoost is available at https://github.com/hvekuri/co2_gapfilling.

Received: 21 November 2022; Accepted: 25 January 2023

Published online: 31 January 2023

References

1. Paustian, K. *et al.* Climate-smart soils. *Nature* **532**, 49–57 (2016).
2. Griscom, B. W. *et al.* Natural climate solutions. *Proc. Natl. Acad. Sci.* **114**, 11645–11650 (2017).
3. Smith, P. *et al.* How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. *Glob. Change Biol.* **26**, 219–241 (2020).
4. Baldocchi, D. D. Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: Past, present and future. *Glob. Change Biol.* **9**, 479–492 (2003).
5. Baldocchi, D. D. How eddy covariance flux measurements have contributed to our understanding of Global Change Biology. *Glob. Change Biol.* **26**, 242–260 (2020).
6. Pastorello, G. *et al.* The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Sci. Data* **7**, 1–27 (2020).
7. 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.* **314**, 108777 (2022).
8. Moffat, A. M. *et al.* Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agric. For. Meteorol.* **147**, 209–232 (2007).
9. Falge, E. *et al.* Gap filling strategies for defensible annual sums of net ecosystem exchange. *Agric. For. Meteorol.* **107**, 43–69 (2001).
10. Mahabbati, A. *et al.* A comparison of gap-filling algorithms for eddy covariance fluxes and their drivers. *Geosci. Instrum. Methods Data Syst.* **10**, 123–140 (2021).
11. Kim, Y. *et al.* 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**, 1499–1518 (2020).
12. Wutzler, T. *et al.* Basic and extensible post-processing of eddy covariance flux data with REddyProc. *Biogeosciences* **15**, 5015–5030 (2018).
13. LI-COR Biosciences, USA. <https://www.licor.com/env/support/Tovi/manuals.html>.
14. Irvin, J. *et al.* Gap-filling eddy covariance methane fluxes: Comparison of machine learning model predictions and uncertainties at FLUXNET-CH4 wetlands. *Agric. For. Meteorol.* **308**, 108528 (2021).
15. Warm Winter 2020 Team, & ICOS Ecosystem Thematic Centre. (2022). Warm Winter 2020 ecosystem eddy covariance flux product for 73 stations in FLUXNET-Archive format-release 2022-1 (version 1.0). icos carbon portal. <https://doi.org/10.18160/2g60-zhak> (2022).
16. Richardson, A. D. & Hollinger, D. Y. A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO₂ flux record. *Agric. For. Meteorol.* **147**, 199–208 (2007).
17. Liu, M. *et al.* Uncertainty analysis of CO₂ flux components in subtropical evergreen coniferous plantation. *Sci. China Ser. D Earth Sci.* **52**, 257–268 (2009).
18. Virkkala, A.-M. *et al.* Statistical upscaling of ecosystem CO₂ fluxes across the terrestrial tundra and boreal domain: Regional patterns and uncertainties. *Glob. Change Biol.* **27**, 4040–4059 (2021).
19. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).
20. Papale, D. *et al.* Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: Algorithms and uncertainty estimation. *Biogeosciences* **3**, 571–583 (2006).
21. Chollet, F. *et al.* Keras. <https://keras.io> (2015).
22. Virtanen, P. *et al.* Fundamental algorithms for scientific computing in Python. *SciPy 1.0. Nat. Methods* **17**, 261–272 (2020).

Acknowledgements

This research was funded by Maj and Tor Nessling foundation (grant no. 202000391), the Strategic Research Center at the Academy of Finland (grant nos. 327214, 335204), the Academy of Finland Flagship funding (grant no. 337549) and the Academy of Finland (grant nos. 328310, 325549). DP acknowledges the support of the CoCO₂ H2020 European Project (GA 958927) and the Open Earth Monitor Horizon Europe Project (GA 101059548). MA, TL and AL acknowledge the Finnish Ministry of Transport and Communications through ICOS Finland. This work used eddy covariance data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The FLUXNET eddy covariance data processing and harmonization was carried out by the ICOS Ecosystem Thematic Center, AmeriFlux Management Project and Fluxdata project of FLUXNET, with the support of CDIAC, and the OzFlux, ChinaFlux and AsiaFlux offices. All the northern sites used in this study except FI-Qvd and SE-Ros are part of the ICOS European Research Infrastructure and their data distributed by the ICOS Carbon Portal.

Author contributions

H.V., J-P.T., and A.L. designed the study. H.V. performed the data analyses. D.P. provided the code for MDS and developed the modified version of MDS. All authors were involved in the discussion of the analysis, data interpretation, writing, and editing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-28827-2>.

Correspondence and requests for materials should be addressed to H.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023, corrected publication 2023