Statistical uncertainty of eddy covariance CO₂ fluxes inferred using a residual bootstrap approach

Huei-Jin Wang*, William J. Riley, William D. Collins

Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA
94720, USA

*** Correspondence:** Huei-Jin Wang, tel. +1 510 486 6397, fax +1 510 486 7897, e-mail: hjwang@lbl.gov

Abstract

High-frequency eddy-covariance measurements of net ecosystem CO₂ exchange (NEE) with the atmosphere are valuable resources for model parameterization, calibration, and validation. However, uncertainties in measured data, i.e., data gaps and inherent random errors, create problems for researchers attempting to quantify uncertainties in model projections of terrestrial ecosystem carbon cycling. Here, we demonstrate that a model-data fusion method (residual bootstrap) produces defensible annual NEE sums, through mimicking the behavior of random errors, filling missing values, and simulating gap-filling biases. This study estimated annual NEE sums for 53 site-years based on nine eddy-covariance tower sites in the USA, and found that our annual estimates were, in most cases, comparable in magnitude with those obtained from AmeriFlux gap-filled data. Additionally, compared to the AmeriFlux standardized gap-filling, our approach provides better NEE estimates for moderate to longer, and more frequent, data gaps. Annual accumulated uncertainties in NEE at the 95% confidence level were $\pm 30 \text{ gC m}^{-2} \text{ yr}^{-1}$ for evergreen needleleaf forests; $\pm 60 \text{ gC m}^{-2} \text{ yr}^{-1}$ for deciduous broadleaf forests; and $\pm 80 \text{ gC m}^{-2} \text{ yr}^{-1}$ for croplands. The residual bootstrap performed worst when gap length was greater than one month or data exclusion greater than 90% during the growing season, common to other gap-filling techniques. However, this study produced robust results for most site years

when monthly data coverage during the growing season is not extremely low. We therefore suggest that the inclusion of NEE uncertainty estimates and better estimation for moderate to longer, and more frequent, data gaps as provided by the residual bootstrap approach can be beneficial for ecosystem model evaluation.

Keywords

model-data fusion; multi-model ensembles; gap-filling comparison; long-term measurements; Monte Carlo; net ecosystem CO₂ exchange

1. Introduction

Despite progress in developing terrestrial ecosystem models over the past several decades, there is still very limited knowledge of the performance skills of these process models (Schwalm et al., 2010; Keenan et al., 2012; Luo et al., 2012). In order to evaluate and improve performance of terrestrial ecosystem models, more attention needs to be placed on validation against observations. Eddy covariance observations of ecosystem-atmosphere CO₂ exchanges are essential for evaluating dynamics of model predicted fluxes because these net ecosystem exchange (NEE) measurements are on a continuous basis of typical 30-minute averaging intervals (Falge et al., 2001; Baldocchi, 2003). However, recent studies have revealed that data uncertainty is a systematic cause of the low agreement between model predictions and observations (Schwalm et al., 2010; Dietze et al., 2011; Keenan et al., 2012). Many of the observational flux datasets used to develop and test ecosystem models are subject to systematic and random measurement errors, which weaken the quality of data and complicate model evaluation (Baldocchi, 2003; Hollinger and Richardson, 2005). Therefore, knowing the makeup of uncertainty in observed data is a prerequisite to quantifying the performance of ecosystem process models.

In the eddy covariance technique, uncertainties of flux measurements can be roughly categorized into systematic and random errors. Systematic errors often occur under stable, low-wind conditions at night due to insufficient turbulence mixing and are notoriously difficult to quantify (Lee, 1998; Loescher et al., 2006). The most common solution to these types of systematic errors is data filtering and data filling. Friction velocity (u^*) filtering has been developed to reject suspicious NEE measurements when u^* falls below a critical threshold (Gu et al., 2005; Barr et al., 2013b), and then data gaps created by u^* thresholds are filled using various gap-filling methods (Falge et al., 2001; Moffat et al., 2007). In addition, instrument failures and data quality controls (Foken and Wichura, 1996; Mahrt, 1998) result in further gaps in the data record. In general, data coverage over the course of a year is only ~65% (Falge et al., 2001). Consequently, these extensive, non-random data gaps are a major source of bias in estimating the magnitude of NEE integrals at various timescales, ranging from hours and years. Further, data gaps pose a challenge to quantitatively assess how well terrestrial ecosystem models simulate the processes governing NEE, i.e., gross primary production (GPP) and ecosystem respiration (RE). Because GPP and RE estimates rely only on a small amount of reliable nocturnal NEE measurements and are likely to be biased, they in turn complicate the ecosystem model validation (Reichstein et al., 2005; Desai et al., 2008).

Apart from data gaps, random errors are inherent in flux measurements at non-gap time

points. Random errors are stochastic and include turbulence sampling errors, statistical errors associated with time-varying flux footprints, and errors relevant to the measurement equipment, among others (Moncrieff et al., 1996). To characterize this type of data uncertainty, Hollinger and Richardson (2005) compared two adjacent tower measurement series in an evergreen needleleaf forest and found that random measurement errors were double-exponentially distributed with zero means and heteroscedastic variances. This heteroscedasticity depended on the flux magnitude, which varied in time, i.e., flux uncertainty was greater during the growing season than dormant season and greater in the daytime than nighttime. Therefore, these findings suggest that when not account for, this heteroscedastic random uncertainty has the potential to undermine model-measurement intercomparisons. Although Dietze et al. (2011) added artificial double-exponential errors to ecosystem synthetic data for the purpose of assessing model-measurement mismatch, it is unclear from the study of Hollinger and Richardson (2005) to what extent the application of the distribution parameter estimates is appropriate at other sites. In subsequent work, Richardson et al. (2008) used model residuals (mismatches between observed and modeled fluxes) directly to quantify the uncertainty distribution characteristics of a number of CarboEurope sites. However, their residuals did not reflect the nature of flux random errors (as could be inferred from Monte Carlo simulations) and was closely tied to an underlying model structure.

Due to a lack of two adjacent tower measurement series for most sites, Monte Carlo simulations, in conjunction with model residuals, have been used to resolve the problem of estimating uncertainty due to the random nature of any individual NEE observation. Also, when model residuals are resampled and added back to the model output, gap-free flux datasets can be constructed so that uncertainty in sums of flux estimates can be quantified at various timescales (Hagen et al., 2006; Stauch et al., 2008). Conceptually, this method requires a good model to give reasonable residuals, so that the resampled residuals reflect the behavior of the true measurement random errors even though residuals do not have mean zero (Hardle and Bowman, 1988). In this context, Hagen et al. (2006) and Stauch et al. (2008) used empirical models under the Monte Carlo framework. Although these empirical models are closely tuned to the data, their model parameters are tied to “non-gap-point” data and they in turn exert less capacity for extrapolation at gap points. The resampled residuals may therefore not reflect the behavior of random errors at gap points.

In this paper, rather than using empirical models, we used process models to separate residuals from NEE observations, for several reasons. First, process models contain useful prior functional constraints about ecosystem NEE fluxes and maintain mechanistic consistency in gap and non-gap predictions. Second, although process models exhibit persistent bias at certain times of year, they generally can adequately capture the diurnal cycle (Schwalm et al., 2010; Dietze et

al., 2011; Stoy et al., 2013). Because our approach does not require mean zero residuals, resampled residuals have the potential to mirror the behavior of measurement random errors. Third, the gulf between process-based and empirical approaches to predicting NEE fluxes may be bridged by the use of process model-data fusion. Because little agreement on model performance metrics exists to separate “good” and “bad” process models (Gleckler et al., 2008; Reichler and Kim, 2008; Luo et al., 2012), using multi-model ensemble means has been advocated because ensemble means generally provide more reliable information than any single model by alleviating individual model bias (Cantelaube and Terres, 2005; Thomson et al., 2006; Schwalm et al., 2010).

The goal of this study is to quantify data uncertainty, in association with random measurement errors and gap-filling errors, from eddy covariance measurements at nine sites spanning three vegetation types. We applied a Monte Carlo approach (residual bootstrap) to simulate multiple runs of gap-free NEE time series and hence estimate the mean NEE response at each point in time (pseudo data). To evaluate the degree to which process model errors confound random measurement errors, we differenced posterior residuals from eddy covariance observations and pseudo data in line with non-gap points. Having evaluated the confounded effect, we assessed the performance of residual bootstrap simulations at timescales longer than the measurement time intervals to ensure consistent error propagation. Finally, we inferred the

annual NEE sum with uncertainty limits, for the purpose of assessing the consequence of random errors and gap-filling errors in long-term measurements.

2. Materials and methods

2.1. Observed and modeled NEE Data

All eddy covariance data used were obtained from the AmeriFlux network (<http://public.ornl.gov/ameriflux/>). The obtained 30-minutue NEE values had been processed using a standardized protocol, including storage correction, spike removal, u^* filtering (Gu et al., 2005), and gap-filling using marginal distribution sampling (MDS; Reichstein et al., 2005) or artificial neural network (ANN; Papale and Valentini, 2003). The valid NEE observations (non-gap data) had data coverage ranging between 30% and 70% over the course of a year (Table 1).

Mean model ensemble (mean simulated value across all models) data were analyzed from 15 ecosystem models (Table 1): 13 models obtained through the NACP (North American Carbon Program) interim site synthesis model output (Barr et al., 2013a; Ricciuto et al., 2013), and two versions of the Community Land Model (Lawrence et al., 2011; Koven et al., 2013; Tang and Riley, 2013). Modeled NEE fluxes were model-specific runs using standardized meteorological data, soil types, and management history. Meteorological data, such as air temperature, precipitation, solar radiation, and humidity, were gap-filled using National Oceanic and

Atmospheric Administration (NOAA) meteorological station data and Daymet reanalysis products following Ricciuto et al. (2009). Locally observed values of soil texture and management history by model simulations were given by the AmeriFlux BADM templates (Law et al., 2008). All models were simulated at a 30- or 60-minute step using the standardized meteorological data as driving variables (http://nacp.ornl.gov/mast-dc/docs/Site_Synthesis_Protocol_v7.pdf).

Concerning the interannual variation in NEE provided by Monte Carlo simulations, we selected sites in the AmeriFlux network across the U.S. with at least five years of data collected between 2000 and 2007 and at least nine model outputs, with plant functional types that were represented by at least three sites. This resulted in nine eddy covariance sites spanning 53 site-years. Of these sites, three were characterized as evergreen needleleaf forest (US-Ho1, US-Me2, and US-NR1), three as deciduous broadleaf forest (US-Ha1, US-MMS, and US-UMB), and three as cropland (US-Ne1, US-Ne2, and US-Ne3). For the evergreen needleleaf sites, US-Ho1 and US-Me2 were temperate evergreen forest, and US-NR1 was subalpine conifer forest. Three cropland sites were subject to crop management: US-Ne1 was irrigated continuous maize site; US-Ne2 was irrigated maize-soybean rotation site; US-Ne3 was rainfed maize-soybean rotation site (Verma et al., 2005).

2.2. Residual bootstrap simulations: Gap-free NEE time series

The non-gap NEE observation $\hat{y}_{s,t}$ was assumed to be comprised of a true flux $y_{s,t}$ and a measurement error $\epsilon_{s,t}$, i.e., $\hat{y}_{s,t} = y_{s,t} + \epsilon_{s,t}$, for each site s at time point t . If a perfect model exactly predicted the true flux, i.e., $m_{s,t} = y_{s,t}$, where $m_{s,t}$ is the perfect model prediction, then an imperfect model might predict a flux $\hat{m}_{s,t}$ with some model error $\delta_{s,t}$, i.e., $\hat{m}_{s,t} = m_{s,t} + \delta_{s,t}$. Therefore, through applying multi-model ensemble means $\hat{m}_{s,t}$, we estimated residuals $\hat{\epsilon}_{s,t}$ from non-gap NEE observations $\hat{y}_{s,t}$ as

$$\hat{\epsilon}_{s,t} = \hat{y}_{s,t} - \hat{m}_{s,t} = (y_{s,t} + \epsilon_{s,t}) - (m_{s,t} + \delta_{s,t}) = \epsilon_{s,t} + \delta_{s,t} \quad [1].$$

Having specified $\hat{m}_{s,t}$ and $\hat{\epsilon}_{s,t}$ from the observed flux $\hat{y}_{s,t}$, the observations played no further role in residual bootstrap simulations. In our approach, the residuals $\hat{\epsilon}_{s,t}$ were sampled with replacement and added back to the ensemble mean time series $\hat{m}_{s,t}$ to create alternate realizations of true NEE time series. In this case, we simulated 1,000 alternate realizations, which were used to quantify NEE uncertainty through evaluation of statistics associated with a distribution of realizations. The details for residual bootstrap simulations are given below. For the i^{th} realization ($i = 1, 2, \dots, 1000$), a new observation $y_{i,s,t}^*$ was defined after resampled residual $\epsilon_{i,s,t}^*$ was added as

$$y_{i,s,t}^* = \hat{m}_{s,t} + \epsilon_{i,s,t}^* \quad [2]$$

where s identifies the site, and t' identifies any predicted time point, including non-gap and gap

time points. Further, in order to better retain the distributional characteristics of random measurement errors with time and the magnitude of flux, i.e., heteroscedasticity, we divided the entire yearly residual dataset into 2×3 individual bins accounting for the growing (May-October) and dormant (November-April) seasons with three categories of total incident solar radiation (defined as $x \leq 0$, $0 < x \leq 400$, and $x > 400 \text{ W m}^{-2}$, respectively). We sampled $\mathbf{e}_{i,s,r}^*$ from the correct bin of estimated residuals $\{\mathbf{e}_{s,r}\}_j$, where $j = 1, 2, \dots, 6$ denotes the specified bin. Then, through applying the distribution-free method (Iman and Conover, 1982), we permuted $\{\mathbf{e}_{i,s,r}^*\}_j$ based on the dependence of $\{\mathbf{e}_{s,r}\}_j$ on $\{\mathbf{m}_{s,r}\}_j$ to pair $\mathbf{e}_{i,s,r}^*$ and $\mathbf{m}_{s,r}$ for obtaining $\mathbf{y}_{i,s,r}^*$, which preserved a desired correlation structure between $\mathbf{e}_{s,r}$ and $\mathbf{m}_{s,r}$ (e.g., the data for the year of 2002: Table 2). Then the i^{th} realization $\{\mathbf{y}_{i,s,r}^*\}$ represented a potential complete NEE time series, and every time point in the time series had 1,000 bootstrap-predicted values from 1,000 realizations. The random deviation from the mean response of each NEE flux was estimated from these 1,000 bootstrap-predicted values (e.g., Table 3). The mean response at a point on time was also estimated as “a pseudo datum” in gap-free NEE time series.

2.3. Analysis of pseudo data

To evaluate the degree to which process-model structural errors confound random measurement errors, we differenced posterior residuals from eddy covariance observations and pseudo data in line with non-gap points, and grouped posterior residuals based on our 2×3

residual categories for each site-year. Then, we measured Kendall's tau and plotted histograms of residuals to assess whether any systematic pattern existed between pseudo data and posterior residuals. The Kendall's tau (τ) is a nonparametric statistical correlation coefficient (Higgins, 2004). If τ is close to zero, then posterior residuals should not contain significant predictive information (model structural errors) from pseudo data. In contrast, if $|\tau|$ is close to one, then pseudo data should contain influential model structural errors and could leak such explanatory information to posterior residuals. The histogram of residuals was used to ensure that the posterior residuals' properties were comparable to those of random errors, i.e., heteroscedastic variance and distributions with central peak at zero and heavy tails. Note that non-gap data were used here and only coincident pseudo data were taken into account.

2.4. Validation: Monthly and annual non-gap NEE means

To investigate the performance of residual bootstrap simulations at timescales longer than the measurement time intervals, we examined the accumulation of uncertainty in monthly and annual NEE measurements. In this analysis, the observed data were assumed to be the best representation of the true NEE fluxes available, although these observations contained uncertainty. For non-gap points, we averaged 30- or 60-minute flux values onto monthly and annual scales for the observed data and 1,000 realizations, respectively. Then we computed the correlation coefficient (R^2) and the root mean squared errors (RMSE) as measures of simulation

performance.

2.5. *Uncertainty in annual NEE estimates*

For the purpose of assessing the consequence of random errors and gap-filling-related errors in long-term measurements, we constructed 95% confidence intervals as a measure of the accuracy of annual NEE estimates, with error bounds in the form of ± 2 RMSE. In this analysis, the RMSE, which contains both variance and bias terms, could not be directly measured because its estimation requires both complete time series of NEE simulations and measurements. The variance term could be directly derived from the variability of the 1,000 realizations. In contrast, the annual bias was not a direct measure of the difference between bootstrap-mean and observed sums, due to the existence of gaps in the real dataset. To quantify this potential bias introduced by data exclusion, we directed our analysis on simulating the available 30- or 60-minute biases onto an annual scale. Because the operation of bootstrapping described in section 2.2 was identical between non-gap and gap points, biases at gap-filled points were expected to be analogous to those at non-gap points. The text below described the simulation for annual bias estimation:

- (1) At each non-gap point, the bias for the 1,000 realizations and the observation was estimated.

- (2) The yearly dataset of 30- or 60-minute biases was divided into 2×3 subsets, similar to

the bootstrap sampling scheme, accounting for the growing and dormant seasons with three categories of total incident solar radiation.

(3) A new dataset was created by resampling biases from each subset on a yearly basis (a total of 8,760 or 17,520 resampled biases).

(4) An annual bias was calculated by aggregating resampled biases from (3).

(5) Steps 3-4 were repeated 1,000 times.

(6) The quantiles for 95% bootstrap intervals were produced using the empirical distribution of the 1,000 annual biases. Then, the maximum of absolute values of these two quantiles was used to compute the RMSE.

3. Results

3.1. *Posterior residual analysis for pseudo data*

Posterior residuals and pseudo data were weakly correlated or uncorrelated, with pairwise Kendall's correlation coefficients close to zero $\tau \approx 0$ (e.g., the data in the year of 2002: Table 4).

Compared to the prior residuals (residuals derived from ensemble means), we noted that the deficiency existing in ensemble means was largely reduced across all sites under our Monte Carlo framework (Table 2 vs. Table 4). In the analysis of posterior residuals' properties, the grouped probability distributions had central peaks at zero and heavy tails, and hence the

appearance was more non-normal for all sites. Furthermore, for the evergreen forest sites, these non-normal distributions approximated a double exponential distribution type (e.g., the data of US-Ho1 in 2002: Fig. 1). Fig.1 also revealed that posterior residuals were roughly in proportion to the magnitude of the observed flux. For example, the residual variability was smaller during the dormant season than the growing season (Fig. 1a, 1b, 1c vs. Fig. 1d, 1e, 1f), and at night than in the daytime (Fig. 1a vs. Fig. 1b, 1c; Fig. 1d vs. Fig. 1e, 1f). In general, our posterior residuals' statistical properties were in agreement with the observed flux measurement uncertainty (Hollinger and Richardson, 2005), and pseudo data did not have evident structural errors confounded with random measurement errors.

3.2. Validation: Monthly and annual averaged non-gap fluxes

Our simulation approach validated reasonably well using the aggregated non-gap data. Monthly NEE flux averages were estimated by aggregating each of the 1,000 simulated non-gap time series to the monthly scale. The monthly simulated and observed flux averages were distributed along the 1:1 line with all $R^2 > 0.90$, and 80% of the month-sites had the $RMSE \leq 1 \mu\text{mol m}^{-2} \text{s}^{-1}$ (Fig. 2). At this scale, the mean difference between the simulated and observed flux averages (bias) exerted a much stronger influence on the RMSE value, primarily due to the unbalanced observation size between months across the sampling season, the growing season especially. For example, monthly data coverage of 32% (May), 0% (Jun), 8% (Jul), 0% (Aug),

14% (Sept), and 4% (Oct) applied to the growing season of US-Ha1 in 2005 caused biases in bootstrap NEE of -2.1, NA, 1.5, NA, 3.0, and 2.7 $\mu\text{mol m}^{-2} \text{s}^{-1}$, respectively. In contrast, at the annual scale, the expected value of the simulated flux average bias was an unbiased estimate of the observed, and 98% of the site-years had the RMSE $\leq 0.1 \mu\text{mol m}^{-2} \text{s}^{-1}$ on the order of 1-5% (Fig. 3). Therefore, our simulation method was robust and preserved the information of accumulated uncertainty.

3.3. *Uncertainty in annual NEE estimates*

Building on the validity of our monthly and annual results from non-gap-filled coincident NEE measurements, we sought to identify each ecosystem's NEE using 95% confidence intervals for annual sums (Fig. 4). The intervals of 95% uncertainty bounds for annual NEE sums at a site varied only slightly among years. Our uncertainty estimates in annual NEE sums at the 95% confidence level (± 2 RMSE) were roughly $\pm 30 \text{ gC m}^{-2} \text{yr}^{-1}$ ($\pm 10\%$) for evergreen needleleaf forests; $\pm 60 \text{ gC m}^{-2} \text{yr}^{-1}$ ($\pm 20\%$) for deciduous broadleaf forests; and $\pm 80 \text{ gC m}^{-2} \text{yr}^{-1}$ ($\pm 20\%$) for croplands (Table 5). Our results agreed with AmeriFlux gap-filled estimates about 70% of the time for MDS and 72% of the time for ANN, respectively (Fig. 4). Furthermore, our simulation method was able to characterize interannual patterns comparable to most of the MDS and ANN estimates. For all forest ecosystem sites other than US-NR1, the magnitude of interannual variability (1 SD) was roughly 20-30% of the mean flux ($50\text{-}120 \text{ gC m}^{-2} \text{yr}^{-1}$; Table 6). The

subalpine conifer forest ecosystem of US-NR1 approximated to a carbon equilibrium with an interannual variability of $20 \text{ gC m}^{-2} \text{ yr}^{-1}$. For croplands, interannual variability was $140\text{-}360 \text{ gC m}^{-2} \text{ yr}^{-1}$ (up to 100% of the mean flux) primarily due to crop management practices, e.g., crop rotation and water availability (Verma et al., 2005).

4. Discussion

4.1. *Appropriateness of bootstrap sampling scheme*

The process model-data fusion (residual bootstrap) method reflects the random nature of the flux observation from half-hourly to annual scales. We demonstrate that the impact of random errors on flux measurements could be reduced to a value within 1-5% because random errors partly compensate during time integration. At the 30- or 60-minute scale, eddy covariance measurements are frequently (about 95% of time) within the 95% confidence intervals of flux data, and most measurements have a relatively large proportion of random error with standard deviation of $> 1 \text{ } \mu\text{mol m}^{-2} \text{ s}^{-1}$ (Table 3). At the monthly scale, the random variability is reduced to a value of $\leq 1 \text{ } \mu\text{mol m}^{-2} \text{ s}^{-1}$ for most month-sites (Fig. 2). As for the annual scale, the effect of

random error becomes small relative to the half-hourly measurement (Fig. 3). The result of mimicking the behavior of random errors provides a solid foundation for implementing a process model-data fusion method in assessing the accumulated uncertainty in annual NEE sums (gap-free).

4.2. Error bounds on the annual NEE sum

The analysis presented here demonstrates a total random uncertainty at the 95% confidence intervals in (measured + filled) NEE of $\pm 30 \text{ gC m}^{-2} \text{ yr}^{-1}$ for evergreen needleleaf forests; $\pm 60 \text{ gC m}^{-2} \text{ yr}^{-1}$ for deciduous broadleaf forests; and $\pm 80 \text{ gC m}^{-2} \text{ yr}^{-1}$ for croplands. The different widths of the confidence intervals for the annual accumulated random uncertainty are relevant to the site-specific nature of random errors at the 30- or 60-minute level. For example, the crops sites have largest random errors at the high-frequency level; the deciduous forest sites second; and the evergreen forest sites third (Table 3). Our uncertainty estimates are generally comparable in magnitude to results from several other groups. Richardson and Hollinger (2005) reported 95% confidence limits in annual sums at the US-Ho1 evergreen forest of $\pm 25 \text{ gC m}^{-2} \text{ yr}^{-1}$ and Richardson and Hollinger (2007) $\pm 25\text{-}42 \text{ gC m}^{-2} \text{ yr}^{-1}$ during 1996-2004. Goulden et al. (1996) put 90% confidence limits of -80 to $+30 \text{ gC m}^{-2} \text{ yr}^{-1}$ on the annual NEE at the US-Ha1 deciduous

forest. Other studies at other sites also attempted to estimate uncertainty limits for the annual NEE as followings: $\pm 40 \text{ gC m}^{-2} \text{ yr}^{-1}$ for a deciduous broadleaf forest by Lee et al. (1999); and $\pm 170\text{-}180 \text{ gC m}^{-2} \text{ yr}^{-1}$ (Anthoni et al., 1999), $\pm 20\text{-}140 \text{ gC m}^{-2} \text{ yr}^{-1}$ (Griffis et al., 2003), and $\pm 30 \text{ gC m}^{-2} \text{ yr}^{-1}$ (Morgenstern et al., 2004) for several coniferous forests.

4.3. Potential gap-filling effects on annual NEE sums

Although the residual bootstrap method provides unbiased estimates for non-gap annual averages (Fig. 3), the accuracy of its application to gap filling depends on the pre-treatment of the data used for filling gaps, particularly when the choice of the u^* threshold filter is made to remove nighttime NEE measurement deficits. Thus, annual NEE estimates might suffer from gap filling effects as NEE time series have fractional data exclusions. In this study, potential biases introduced by data gaps are considered, and the effect on annual sums is $0\pm 10\text{-}40 \text{ gC m}^{-2} \text{ yr}^{-1}$ at the 95% confidence level, up to two times larger than the random errors (Table 5). This estimate is comparable in magnitude to the u^* -threshold-related uncertainty in annual NEE ($-18 \pm 49 \text{ gC m}^{-2} \text{ yr}^{-1}$: median $\pm$ inter-quartile range) reported by Barr et al. (2013b), and the effects of gap filling on annual sums ($\pm 25 \text{ gC m}^{-2} \text{ yr}^{-1}$: ± 2 SD) associated with the AmeriFlux standardized gap-filling methods (Moffat et al., 2007).

Although our estimates of annual sums generally are consistent with those obtained from the AmeriFlux gap-filled data, most years of US-Ha1, especially, do not have compatible annual sums from residual bootstrap and AmeriFlux standardized gap filling (MDS and ANN), compared to the other site-years (Fig. 4). A likely explanation is that different filling strategies work differently in response to long gaps, which generally degrade the performance of gap filling (Richardson and Hollinger, 2007). Although the MDS and ANN methods make use of meteorological data to fill gaps, their filled values only depend on the available NEE data (Papale and Valentini, 2003; Reichstein et al., 2005). When datasets contain long or many gaps, the parameterization of these filling algorithms would shift towards a local condition and may produce large bias in the annual sum (Falge et al., 2001; Moffat et al., 2007), such as occurred at US-Ha1 in 2003. As for the residual bootstrap approach, instead of the available data, residuals combine with ensemble means to fill gaps, which enhances the capacity for extrapolation. However, when long gaps prevail and result in insufficient residual sample size, the residual bootstrap would also give large bias in the annual sum. For example, examination of the observed data for the US-Ha1 in 2005 shows that the growing-season months of Jun-Oct have extreme fractional data exclusions (>90%) and the impact on the annual sum is large compared to the other years. On the other hand, the annual NEE at this site-year made by the MDS and ANN might also not be reliable due to extremely low observed NEE density. Overall we contend

that annual sums of NEE we estimate here are confirmed by those obtained from the AmeriFlux standardized gap-filling methods. When annual sums of NEE obtained from different gap-filling methods are not compatible with each other, we recommend further data examination (e.g., fractional data exclusion for nighttime/daytime and growing/dormant season) to ensure the applicability of each gap-filling strategy.

5. Summary and conclusions

This study estimated annual NEE sums for 53 site-years based on nine eddy-covariance tower sites in the USA. We used a model-data fusion method to bracket the range of likely annual NEE sums, through mimicking the behavior of random errors, filling missing values, and simulating potential biases introduced by gap filling. This method shows good performance of R^2 and RMSE and good annual sum reliability, which can be recommended as a tool for estimating uncertainty in eddy-covariance data. Uncertainties in annual NEE sums are $\pm 30 \text{ gC m}^{-2} \text{ yr}^{-1}$ ($\pm 10\%$) for evergreen needleleaf forests; $\pm 60 \text{ gC m}^{-2} \text{ yr}^{-1}$ ($\pm 20\%$) for deciduous broadleaf forests;

and $\pm 80 \text{ gC m}^{-2} \text{ yr}^{-1}$ ($\pm 20\%$) for croplands at the 95% confidence level. Our results indicate that the uncertainty due to gap-filling is greater than the random measurement uncertainty. However, we caution that long gaps or extreme fractional data exclusions pose an additional challenge in assessing the gap-filling effect on annual sums, which is a common problem to all gap-filling techniques and difficult to quantify due to its non-random distribution.

Acknowledgements

This research was supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 as part of their Earth System Modeling and Regional and Global Climate Modeling (RGCM) Programs.

References

- Anthoni, P.M., Law, B.E. and Unsworth, M.H., 1999. Carbon and water vapor exchange of an open-canopied ponderosa pine ecosystem. *Agricultural and Forest Meteorology*, 95(3): 151-168.
- Baldocchi, D.D., 2003. Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change Biology*, 9(4): 479-492.
- Barr, A.G. et al., 2013a. NACP Site: Tower Meteorology, Flux Observations with Uncertainty, and Ancillary Data. Data set. Available on-line [<http://daac.ornl.gov>] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA <http://dx.doi.org/10.3334/ORNLDAAAC/1178>.
- Barr, A.G. et al., 2013b. Use of change-point detection for friction-velocity threshold evaluation in eddy-covariance studies. *Agricultural and Forest Meteorology*, 171: 31-45.
- Cantelaube, P. and Terres, J.M., 2005. Seasonal weather forecasts for crop yield modelling in Europe. *Tellus Series a-Dynamic Meteorology and Oceanography*, 57(3): 476-487.
- Desai, A.R. et al., 2008. Cross-site evaluation of eddy covariance GPP and RE decomposition

- techniques. *Agricultural and Forest Meteorology*, 148(6-7): 821-838.
- Dietze, M.C. et al., 2011. Characterizing the performance of ecosystem models across time scales: A spectral analysis of the North American Carbon Program site-level synthesis. *Journal of Geophysical Research-Biogeosciences*, 116.
- Falge, E. et al., 2001. Gap filling strategies for defensible annual sums of net ecosystem exchange. *Agricultural and Forest Meteorology*, 107(1): 43-69.
- Foken, T. and Wichura, B., 1996. Tools for quality assessment of surface-based flux measurements. *Agricultural and Forest Meteorology*, 78(1-2): 83-105.
- Gleckler, P.J., Taylor, K.E. and Doutriaux, C., 2008. Performance metrics for climate models. *Journal of Geophysical Research-Atmospheres*, 113(D6).
- Goulden, M.L., Munger, J.W., Fan, S.M., Daube, B.C. and Wofsy, S.C., 1996. Measurements of carbon sequestration by long-term eddy covariance: Methods and a critical evaluation of accuracy. *Global Change Biology*, 2(3): 169-182.
- Griffis, T.J. et al., 2003. Ecophysiological controls on the carbon balances of three southern boreal forests. *Agricultural and Forest Meteorology*, 117(1-2): 53-71.
- Gu, L.H. et al., 2005. Objective threshold determination for nighttime eddy flux filtering. *Agricultural and Forest Meteorology*, 128(3-4): 179-197.
- Hagen, S.C. et al., 2006. Statistical uncertainty of eddy flux-based estimates of gross ecosystem

- carbon exchange at Howland Forest, Maine. *Journal of Geophysical Research-Atmospheres*, 111(D8).
- Hardle, W. and Bowman, A.W., 1988. Bootstrapping in nonparametric regression - Local adaptive smoothing and confidence bands. *Journal of the American Statistical Association*, 83(401): 102-110.
- Higgins, J.J., 2004. *Introduction to Modern Nonparametric Statistics*. Brooks/Cole - Thomson Learning, Pacific Grove, California, 366 pp.
- Hollinger, D.Y. and Richardson, A.D., 2005. Uncertainty in eddy covariance measurements and its application to physiological models. *Tree Physiology*, 25(7): 873-885.
- Iman, R.L. and Conover, W.J., 1982. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics*, 11(3): 311-334.
- Keenan, T.F. et al., 2012. Terrestrial biosphere model performance for inter-annual variability of land-atmosphere CO₂ exchange. *Global Change Biology*, 18(6): 1971-1987.
- Koven, C.D. et al., 2013. The effect of vertically-resolved soil biogeochemistry and alternate soil C and N models on C dynamics of CLM4. *Biogeosciences Discuss.*, 10: 7201-7256.
- Law, B. et al., 2008. *Terrestrial carbon observations: protocols for vegetation sampling and data submission*, Rome, Italy.
- Lawrence, D.M. et al., 2011. Parameterization improvements and functional and structural

advances in version 4 of the Community Land Model. *Journal of Advances in Modeling Earth Systems*, 3.

Lee, X.H., 1998. On micrometeorological observations of surface-air exchange over tall vegetation. *Agricultural and Forest Meteorology*, 91(1-2): 39-49.

Lee, X.H., Fuentes, J.D., Staebler, R.M. and Neumann, H.H., 1999. Long-term observation of the atmospheric exchange of CO₂ with a temperate deciduous forest in southern Ontario, Canada. *Journal of Geophysical Research-Atmospheres*, 104(D13): 15975-15984.

Loescher, H.W. et al., 2006. Uncertainties in, and interpretation of, carbon flux estimates using the eddy covariance technique. *Journal of Geophysical Research-Atmospheres*, 111(D21).

Luo, Y.Q. et al., 2012. A framework for benchmarking land models. *Biogeosciences*, 9(10): 3857-3874.

Mahrt, L., 1998. Flux sampling errors for aircraft and towers. *Journal of Atmospheric and Oceanic Technology*, 15(2): 416-429.

Moffat, A.M. et al., 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agricultural and Forest Meteorology*, 147(3-4): 209-232.

Moncrieff, J.B., Malhi, Y. and Leuning, R., 1996. The propagation of errors in long-term measurements of land-atmosphere fluxes of carbon and water. *Global Change Biology*, 2(3): 231-240.

- Morgenstern, K. et al., 2004. Sensitivity and uncertainty of the carbon balance of a Pacific Northwest Douglas-fir forest during an El Nino La Nina cycle. *Agricultural and Forest Meteorology*, 123(3-4): 201-219.
- Papale, D. and Valentini, A., 2003. A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Global Change Biology*, 9(4): 525-535.
- Reichler, T. and Kim, J., 2008. How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, 89(3): 303-311.
- Reichstein, M. et al., 2005. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global Change Biology*, 11(9): 1424-1439.
- Ricciuto, D.M. et al., 2013. Terrestrial Biosphere Model and Aggregated Flux Data in Standard Format. Data set. Available on-line [<http://daac.ornl.gov>] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA.
<http://dx.doi.org/10.3334/ORNLDAAAC/1183>.
- Ricciuto, D.M. et al., 2009. How uncertainty in gap-filled meteorological input forcing at eddy covariance sites impacts modeled carbon and energy flux. American Geophysical Union, Fall Meeting 2009: Abstract B54A-03.

Richardson, A.D. and Hollinger, D.Y., 2005. Statistical modeling of ecosystem respiration using eddy covariance data: Maximum likelihood parameter estimation, and Monte Carlo simulation of model and parameter uncertainty, applied to three simple models. *Agricultural and Forest Meteorology*, 131(3-4): 191-208.

Richardson, A.D. and Hollinger, D.Y., 2007. A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO₂ flux record. *Agricultural and Forest Meteorology*, 147(3-4): 199-208.

Richardson, A.D. et al., 2008. Statistical properties of random CO₂ flux measurement uncertainty inferred from model residuals. *Agricultural and Forest Meteorology*, 148(1): 38-50.

Schwalm, C.R. et al., 2010. A model-data intercomparison of CO₂ exchange across North America: Results from the North American Carbon Program site synthesis. *Journal of Geophysical Research-Biogeosciences*, 115.

Stauch, V.J., Jarvis, A.J. and Schulz, K., 2008. Estimation of net carbon exchange using eddy covariance CO₂ flux observations and a stochastic model. *Journal of Geophysical Research-Atmospheres*, 113(D3).

Stoy, P.C. et al., 2013. Evaluating the agreement between measurements and models of net ecosystem exchange at different times and time scales using wavelet coherence: an example

using data from the North American Carbon Program Site-Level Interim Synthesis.

Biogeosciences Discuss., 10: 3039-3077.

Tang, J.Y. and Riley, W.J., 2013. A new top boundary condition for modeling surface diffusive exchange of a generic volatile tracer: theoretical analysis and application to soil evaporation.

Hydrology and Earth System Sciences, 17(2): 873-893.

Thomson, M.C. et al., 2006. Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature, 439(7076): 576-579.

Verma, S.B. et al., 2005. Annual carbon dioxide exchange in irrigated and rainfed maize-based agroecosystems. Agricultural and Forest Meteorology, 131(1-2): 77-96.

Figure Caption

Fig. 1 Histograms of residuals from pseudo data at US-Ho1 in 2002, in the association with total incident solar radiation (defined as $x \leq 0$, $0 < x \leq 400$, and $x > 400 \text{ W m}^{-2}$, respectively) in the dormant (Nov-Apr) and growing (May-Oct) seasons: (a) $x \leq 0$ in the dormant season, (b) $0 < x \leq 400$ in the dormant season, (c) $x > 400 \text{ W m}^{-2}$ in the dormant season, (d) $x \leq 0$ in the growing season, (e) $0 < x \leq 400$ in the growing season, and (f) $x > 400 \text{ W m}^{-2}$ in the growing season.

Fig. 2 Comparisons of monthly flux averages at non-gap points between simulated and measured fluxes: (a1) US-Ho1; (a2) US-Me2; (a3) US-NR1; (b1) US-Ha1; (b2) US-MMS; (b3) US-UMB; (c1) US-Ne1; (c2) US-Ne2; (c3) US-Ne3. Error bars represent $\pm \text{RMSE}$.

Fig. 3 Comparisons of annual flux averages at non-gap points between simulated and measured fluxes for 53 site-years. Error bars represent $\pm \text{RMSE}$.

Fig. 4 Estimates of annual NEE sums based on bootstrap results and AmeriFlux gap-filled data (marginal distribution sampling and artificial neural network): (a1) US-Ho1; (a2) US-Me2; (a3) US-NR1; (b1) US-Ha1; (b2) US-MMS; (b3) US-UMB; (c1) US-Ne1; (c2)

US-Ne2; (c3) US-Ne3. The gray area represents the 95% confidence interval for the annual sum.

Table 1 Sites and models used for the uncertainty analysis.

Site ID	Biome ^a	Years ^b	Nongap ^c (%)	Models															
				ArgoIBIS	CanIBIS	CLMCN	CLMBGC	CNCLASS	ECOSYS	ED2	ISOLSM	LoTECDA	ORCHIDEE	SiB	SiBCASA	SiBcrop	SSiB2	TECO	
US-Ho1	ENF	00-04	56.2		x	x	x	x	x		x	x	x	x	x		x	x	
US-Me2	ENF	02-07	44.2		x	x	x	x	x		x	x	x	x	x		x	x	
US-NR1	ENF	00-07	56.4		x	x	x	x	x			x	x	x	x		x	x	
US-Ha1	DBF	00-06	28.8		x	x	x	x	x		x	x	x	x	x		x	x	
US-MMS	DBF	00-06	59.2		x	x	x	x	x			x	x	x	x		x	x	
US-UMB	DBF	00-06	41.2		x	x	x	x	x		x	x	x	x	x		x	x	
US-Ne1	CRO	01-06	67.1	x	x			x	x	x			x		x	x	x	x	
US-Ne2	CRO	01-06	62.6	x				x	x	x			x		x	x	x	x	
US-Ne3	CRO	01-06	67.9	x	x			x	x	x		x	x		x	x	x	x	

^a Biome codes: ENF, evergreen needleleaf forest; DBF, deciduous broadleaf forest; CRO, cropland

^b Data used for US-Ne1, US-Ne2, and US-Ne3: Jun 2001 – May 2006.

^c Data coverage over the course of a year on average

Table 2 Kendall's tau (τ) coefficients for diagnosing model structure in relationship between ensemble means and the corresponding residuals.

Site ID	Dormant Season (Nov-Apr) in 2002			Growing Season (May-Oct) in 2002		
	$x^* \leq 0$	$0 < x \leq 400$	$x > 400$	$x \leq 0$	$0 < x \leq 400$	$x > 400$
US-Ho1	0.008	-0.288	-0.206	-0.045	0.198	0.334
US-Me2	-0.149	-0.615	-0.625	-0.154	-0.112	-0.033
US-NR1	-0.228	-0.504	-0.391	0.195	0.067	0.204
US-Ha1	-0.207	-0.321	0.211	0.225	0.238	0.198
US-MMS	-0.001	0.100	0.324	0.318	0.198	0.220
US-UMB	-0.182	-0.549	-0.399	-0.051	-0.150	-0.013
US-Ne1	-0.350	-0.549	-0.619	0.028	-0.506	0.122
US-Ne2	-0.362	-0.419	-0.517	-0.052	-0.202	0.401
US-Ne3	-0.224	-0.499	-0.355	-0.150	-0.198	0.413

* total incident solar radiation as $x \leq 0$, $0 < x \leq 400$, and $x > 400$ W m⁻²

Table 3 Mean standard deviation of simulated random errors for a NEE flux at a point in time.

Site ID	Dormant Season (Nov-Apr) in 2002			Growing Season (May-Oct) in 2002		
	$x^* \leq 0$	$0 < x \leq 400$	$x > 400$	$x \leq 0$	$0 < x \leq 400$	$x > 400$
US-Ho1	0.56	1.21	2.13	2.86	3.09	3.68
US-Me2	0.93	1.65	2.48	1.55	2.25	2.86
US-NR1	0.66	0.86	1.29	1.32	2.16	2.81
US-Ha1	1.41	1.56	2.02	2.86	3.86	6.32
US-MMS	0.80	1.12	1.38	1.76	4.35	5.79
US-UMB	0.64	0.91	1.09	1.70	3.29	5.90
US-Ne1	0.79	1.43	1.65	3.91	4.78	11.97
US-Ne2	0.71	0.90	1.04	3.21	3.21	5.44
US-Ne3	1.11	1.03	1.17	2.52	3.18	4.56

* total incident solar radiation as $x \leq 0$, $0 < x \leq 400$, and $x > 400$ W m⁻²

Table 4 Kendall's tau (τ) coefficients for diagnosing model structure in relationship between pseudo data (bootstrap mean estimates) and the corresponding residuals.

Site ID	Dormant Season (Nov-Apr) in 2002			Growing Season (May-Oct) in 2002		
	$x^* \leq 0$	$0 < x \leq 400$	$x > 400$	$x \leq 0$	$0 < x \leq 400$	$x > 400$
US-Ho1	0.001	-0.015	-0.043	0.015	-0.012	0.000
US-Me2	-0.011	0.044	-0.176	0.021	0.000	0.000
US-NR1	-0.122	-0.066	-0.027	0.018	-0.020	0.016
US-Ha1	-0.011	-0.107	0.002	-0.066	-0.014	0.000
US-MMS	0.000	-0.008	0.004	-0.013	-0.008	0.004
US-UMB	-0.024	-0.039	-0.109	0.003	-0.012	0.000
US-Ne1	-0.011	0.044	-0.176	0.021	0.000	0.000
US-Ne2	-0.014	-0.220	-0.225	-0.021	-0.065	-0.036
US-Ne3	-0.063	-0.039	-0.101	-0.022	-0.051	0.014

* total incident solar radiation as $x \leq 0$, $0 < x \leq 400$, and $x > 400$ W m⁻²

Table 5 Annual uncertainties in NEE sums reflected by

RMSE, which incorporates both random error (1 SD) and

gap-filling-related error (bias: 95% confidence intervals).

Site ID	Annual uncertainty (gC m ⁻² yr ⁻¹)		
	SD	Bias	RMSE
US-Ho1	7	0 ±14	15
US-Me2	7	0 ±12	14
US-NR1	5	0 ±10	11
US-Ha1	14	0 ±24	28
US-MMS	17	1 ±33	37
US-UMB	11	0 ±24	27
US-Ne1	22	1 ±41	46
US-Ne2	18	0 ±32	37
US-Ne3	18	0 ±31	36

Table 6 Interannual variability (1 SD) in annual NEE

sums. Mean of annual sums (Mean) and mean annual error

(RMSE) in annual sums are also given for reference.

Site ID	Annual NEE sums (gC m ⁻² yr ⁻¹)		
	Mean	SD	RMSE
US-Ho1	-314	57	15
US-Me2	-519	118	14
US-NR1	-23	19	11
US-Ha1	-315	107	28
US-MMS	-378	72	37
US-UMB	-213	49	27
US-Ne1	-411	144	46
US-Ne2	-352	360	37
US-Ne3	-371	322	36

PDF

— Residuals from pseudo data - - - - Double-exponential distribution

Simulated nongap NEE monthly mean ($\mu\text{mol m}^{-2} \text{s}^{-1}$)

Measured nongap NEE monthly mean ($\mu\text{mol m}^{-2} \text{s}^{-1}$)

Simulated nongap NEE annual mean

($\mu\text{mol m}^{-2} \text{s}^{-1}$)

0
-1
-2
-3
-4

-4

-3

-2

-1

0

Measured nongap NEE annual mean

($\mu\text{mol m}^{-2} \text{s}^{-1}$)

— Residual bootstrap

⋯●⋯ Marginal distribution sampling

-△- Artificial neural network

Annual NEE sums ($\text{gC m}^{-2} \text{yr}^{-1}$)

Year