Available online at www.sciencedirect.com
SCIENCE DIRECT®
Agricultural and Forest Meteorology 121 (2004) 93–111
AGRICULTURAL AND FOREST METEOROLOGY
www.elsevier.com/locate/agrformet

Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations

Dafeng Hui [a,*], Shiqiang Wan [a], Bo Su [a], Gabriel Katul [b], Russell Monson [c], Yiqi Luo [a]

[a] *Department of Botany and Microbiology, University of Oklahoma, 770 Van Vleet Oval, Norman, OK 73019, USA*
[b] *Nicholas School of the Environment and Earth Sciences, Duke University, Durham, NC 27708, USA*
[c] *Department of Environmental, Population and Organismic Biology, University of Colorado, Boulder, CO 80309, USA*

Received 14 November 2002; received in revised form 20 June 2003; accepted 10 July 2003

Abstract

Missing data is a ubiquitous problem in evaluating long-term experimental measurements, such as those associated with the FluxNet project, due to the equipment failures, system maintenance, power-failure, and lightning strikes among other things. To estimate annual values of net ecosystem carbon exchange (NEE), latent heat flux (LE) and sensible heat flux (*H*), such gaps in the measured data must be filled or imputed. So far, no standardized method has been accepted and the imputation methods used are largely dependent on the researchers' choice. Here, we used multiple imputation (MI) to gap-fill the missing data for annual estimations of NEE, LE and *H* at three flux sites associated with the FluxNet effort. MI is a Monte Carlo technique in which the missing values are replaced by several simulated values. Each data set imputed is a complete one where the observed values are the same as those in the original data set; only the missing values are different. Thus, the normal statistical analysis (e.g. annual total calculation) can be applied to each data set separately. The results of each analysis can be recombined into one summary. We applied the MI method to eddy covariance measurements collected from Walker Branch Watershed (WBW) site (a deciduous forest), Duke site (a coniferous forest) and Niwot site (a subalpine forest). Results showed that annual estimations of NEE, LE and *H* by MI were comparable to other imputation methods but MI was much easier to apply because of readily available software and standard algorithms. Besides the normal statistical analyses, MI also provided confidence intervals for each estimated parameter. This confidence interval is most useful when assessing energy, water, and carbon balance closures at a given tower site. Significant differences in annual NEE, LE and *H* were found among years at the three AmeriFlux sites. NEE at the Niwot Ridge site was lower and LE and *H* were higher than at the other two sites. With the available software and realistic gap-filling capability, MI has the potential to become a standardized method to gap-fill eddy covariance flux data for annual estimations and to improve the analysis of uncertainties associated with annual estimations of NEE, LE and *H* from regional and global flux networks.

Keywords: Eddy covariance; Latent heat; Missing data; Multiple imputation; Net ecosystem carbon exchange; Regression analysis; Sensible heat

1. Introduction

Net ecosystem carbon exchange (NEE) between the atmosphere and biosphere is an important component in global carbon cycling. In order to understand the temporal and spatial variations of NEE, a worldwide

* Corresponding author. Tel.: +1-405-325-5003; fax: +1-405-325-7619.
E-mail address: dafeng@ou.edu (D. Hui).

0168-1923/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0168-1923(03)00158-8

network, FluxNet, equipped with eddy covariance flux towers, is operating to collect NEE, latent heat (LE) and sensible heat (*H*) fluxes from more than 140 sites around the world (Baldocchi et al., 2001). However, missing or rejected data in these measurements is a ubiquitous problem due to equipment failures (system/sensor breakdown), maintenance and calibration, spikes in the raw data, and physical and biological constraints (e.g. storms, hurricanes, and non-optimal wind directions). For example, about 22% of the total half-hourly daytime measurements were found to reflect gaps and rejected data at the Walker Branch Watershed (WBW) AmeriFlux site (Wilson and Baldocchi, 2001), and 35% at the Duke site (Katul et al., 2001), and less than 20% at the Niwot Ridge site (Monson et al., 2002). In general, about 17–50% of the observations in NEE are reported as missing or rejected at FluxNet sites (Falge et al., 2001a).

The gaps in observed data cause at least three problems: (1) difficulty in annual estimation of NEE, LE and *H*; (2) biased relationships between NEE, LE and *H* with climatic variables; and (3) low quality data for modeling validation. As most statistical methods, such as total calculation and regression analysis, can only handle complete data sets, observations with missing data for one or more variables should be ignored in the analysis (i.e. listwise deletion). Depending on the percentage of missing information, analytic power may be significantly reduced and the results may be biased (Little and Rubin, 2002; Allison, 2000).

To accurately calculate annual values of NEE and energy fluxes at FluxNet sites, gap-filling to account for the missing data is imperative. The commonly used methods for filling missing data include mean replacement (i.e. using mean of observed values to replace missing data), hot or cold dock (i.e. a randomly or systematically chosen value from an individual observation that has similar values on other variables), interpolation and extrapolation (i.e. an estimated value from other observations for the same variable), and regression analysis (i.e. the predicted value obtained by regressing the missing variable on other variables). For example, Greco and Baldocchi (1996) and Jarvis et al. (1997) have used modified mean replacement method (i.e. values of 15 days to replace the missing value in diurnal data variation). The regression analysis method has been widely used by other researchers (Goulden et al., 1996; Granier et al., 2000; Pilegaard et al., 2001; Grünwald and Bernhofer, 2000; Monson et al., 2002; Hui et al., 2003). The neural networks method has been proposed by Aubinet et al. (2000). In a comprehensive study, Falge et al. (2001a) compared three methods including mean diurnal variation (similar to mean replacement), look-up tables and nonlinear regression on the annual sum of NEE for 28 data sets from 18 FluxNet sites, and found that the differences in annual NEE estimation by different gap-filling methods ranged from -45 to $200\,\mathrm{g\,C\,m^{-2}}$ per year. Their study also emphasized the importance of the method of standardization during the data post-processing phase, so comparable data can be obtained to address intercomparisons across different ecosystems, climatic conditions, and multiple years (Falge et al., 2001a,b).

In this study, we applied a generic multiple imputation (MI) method to the data of eddy covariance measurements collected from the WBW site (a deciduous forest), the Duke Forest site (a coniferous forest) and the Niwot Ridge site (a coniferous forest). MI is a Monte Carlo technique in which the missing values are replaced by several simulated values (Rubin, 1987). Each data set imputed is a complete one where the observed values are the same as in the original data set, only the missing values are different. So the normal statistical analyses (e.g. annual total calculation, regression analysis) can be applied separately to each data set. The results of each analysis are then recombined into one summary. Compared with other methods, MI produces the mean estimate as well as a confidence internal of the mean. It has been successfully used in the social and behavioral sciences (King et al., 2001; Schafer and Graham, 2002), medical studies (Barnard and Meng, 1999), nursing research (Kneipp and McIntosh, 2001; Patrician, 2002), and public health research (Zhou et al., 2001). In this inter-comparison, we first describe the procedure of MI, then apply it to gap-fill the eddy covariance data for NEE, LE and *H* at three sites; we also compare the results with other methods.

2. Multiple imputation

MI is a general-purpose method for analyzing data sets with missing observations and is broadly applica-

ble to a variety of different types of data sets. It was proposed by Rubin (1977) and described in detail by Rubin (1987) and Schafer (1997). Briefly, three steps are involved in MI: imputation, analysis, and pooling. First, sets of plausible values for missing observations are created that reflect uncertainty about the imputation model. Each of these sets of plausible values is used to fill-in the missing values and create a complete data set. Second, each of these data sets is analyzed using normal statistical methods. Finally, the results are combined, which allows the uncertainty regarding the imputation to be taken into account (Horton and Lipsitz, 2001).

2.1. Imputation

Let Y be the $n \times p$ matrix of the complete data including p variables (e.g. PAR, T_A, T_S, ..., NEE, LE and H), which is not fully observed, and denote the observed part of Y by Y_{obs} and the missing part by Y_{mis}. Suppose that $Y = (Y_{obs}, Y_{mis})$ have a p-variate normal distribution with mean $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$ and covariance matrix $\Sigma = (\sigma_{jp})$. Imputation simulates Y_{mis} given Y_{obs}. MI generates m imputations, typically 3–5, for a given missing data point. MI data sets are simulated draws from a Bayesian predictive distribution of the missing data. To begin the imputation process, initial estimates of mean vector μ and covariance matrix Σ are needed which can be obtained by maximum likelihood estimation.

Maximum likelihood estimates of the mean vector and covariance matrix can be generated using the expectation and maximization (EM) algorithm. The EM algorithm is a technique that finds maximum likelihood estimates in parametric models for incomplete data (Dempster et al., 1977) and has been widely applied in genetic researches (e.g. Jiang and Hui, 1995; Jiang and Zeng, 1997; Hui et al., 1997) and many other studies (e.g. Dayan and Hinton, 1997; Barnard and Meng, 1999; Schafer and Graham, 2002; Carsob et al., 2002). It is an iterative procedure involving the following steps.

2.1.1. The expectation E-step

Given a set of parameter estimates θ, such as a mean vector μ and covariance matrix Σ for a multivariate normal distribution, the E-step calculates the condition expectation of the complete-data log likelihood, which can be expressed as

$$\ln L(\mu, \Sigma|Y_{obs}) = \sum_{l=1}^{g} \ln L_l(\mu, \Sigma|Y_{obs}) \tag{1}$$

where g is the number of groups with distinct missing patterns, $\ln L_l(\mu, \Sigma|Y_{obs})$ is the observed-data log likelihood from the lth group, and

$$\ln L_l(\mu, \Sigma|Y_{obs}) = -\frac{n_l}{2} \ln|\Sigma_l| - \frac{1}{2} \sum_i (y_{il} - \mu_l)' \Sigma_l^{-1} (y_{il} - \mu_l) \tag{2}$$

where n_l is the number of observations in the lth group, y_{il} is a vector of observed values corresponding to observed variables, μ_l is the corresponding mean vector, and Σ_l is the associated covariance matrix.

At the tth iteration of EM, let $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$ denote current estimates of parameters. The E-step of the algorithm consists in calculating

$$E\left(\sum_{i=1} y_{ij}|Y_{obs}, \theta^{(t)}\right) = \sum_{i=1} y_{ij}^{(t)}, \quad j = 1, 2, \ldots, p \tag{3}$$

and

$$E\left(\sum_{i=1} y_{ij} y_{ik}|Y_{obs}, \theta^{(t)}\right) = \sum_{i=1} (y_{ij}^{(t)} y_{ik}^{(t)} + c_{jki}^{(t)}), \quad j, k = 1, 2, \ldots, p \tag{4}$$

where

$$y_{ij}^{(t)} = \begin{cases} y_{ij} & \text{if } y_{ij} \text{ is observed} \\ E(y_{ij}|y_{obs,i}, \theta^{(t)}) & \text{if } y_{ij} \text{ is missing} \end{cases} \tag{5}$$

and

$$c_{jki}^{(t)} = \begin{cases} 0 & \text{if } y_{ij} \text{ or } y_{jk} \text{ is observed} \\ \mathrm{Cov}(y_{ij}, y_{jk}|y_{obs,i}, \theta^{(t)}) & \text{if } y_{ij} \text{ and } y_{jk} \text{ are missing} \end{cases} \tag{6}$$

Missing values y_{ij} are thus replaced by the conditional mean of y_{ij} given the set of observed values y_{obs} (Little and Rubin, 2002).

2.1.2. The maximization M-step

Given a complete-data, the *M*-step finds the parameter estimates to maximize the complete-data log likelihood from the *E*-step. The new estimates $\theta^{(t+1)}$ of the parameters are:

$$\mu_j^{(t+1)} = n^{-1}\sum_{i=1} y_{ij}^{(t)}, \quad j = 1, 2, \ldots, p \tag{7}$$

$$\sigma_{jk}^{(t+1)} = n^{-1}\sum_{i=1}[(y_{ij}^{(t)} - \mu_j^{(t+1)})(y_{ik}^{(t)} - \mu_k^{(t+1)}) + c_{jki}^{(t)}],$$
$$j, k = 1, 2, \ldots, p \tag{8}$$

The two steps are iterated until the iterations converge (i.e. the estimates barely change from one iteration to the next, e.g. less than a small number, 10^{-5}).

In the next step, a data augmentation algorithm, the Markov Chain Monte Carlo (MCMC), is used to generate the imputed data. MCMC uses the initial values obtained from the EM algorithm and constructs a Markov chain to simulate draws from the posterior distribution of $p(Y_{\text{mis}}|Y_{\text{obs}})$. This can be implemented using the imputation-posterior (IP) algorithm (Schafer, 1997), which is similar to EM. At the *t*th iteration, the steps can be defined as follows.

2.1.3. The imputation I-step

With the estimated mean vector and covariance matrix, the *I*-step simulates the missing values for each observation independently. That is, draw values for $Y_{\text{mis}}^{(t+1)}$ from $p(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$, a conditional distribution given observed variables Y_{obs}.

2.1.4. The posterior P-step

P-step simulates the posterior mean vector and covariance matrix from the complete data set, i.e. draws $\theta^{(t+1)}$ from $p(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)})$. These new estimates are then used in the next *I*-step. The two steps are iterated long enough for the results to be reliable for a multiply imputed data set. This creates a Markov chain $(\{Y^{(1)}, \theta^{(1)}\}, \{Y^{(2)}, \theta^{(2)}\}, \ldots, \{Y^{(t+1)}, \theta^{(t+1)}\}, \ldots)$ which converges in distribution to $p(Y_{\text{mis}}, \theta|Y_{\text{obs}})$. Assuming iterates converge to a stationary distribution, the goal is to simulate an approximately independent draw of the missing values from this distribution (see Appendix A for details in computation).

Imputations can be drawn from one Markov chain or multiple independent chains. Clearly *m* independent Markov chains are preferable, but the cost is running $m - 1$ additional MCMC simulations using the IP algorithm. The advantage of MCMC method is that it can handle arbitrary patterns of missing data and the downsides are: (1) it requires an assumption of multivariate normality; and (2) it is not readily intuitive and computationally expensive.

2.2. Analysis

With *m* imputed complete data sets, any chosen statistical analysis can be applied to each of them. In this study, we calculated the annual sum of NEE, LE and *H*, and their standard errors.

2.3. Pooling (combining results from multiply imputed data sets)

With *m* imputations, *m* different sets of the point and variance estimates for a parameter *Q* (i.e. annual sum here) can be computed (SAS Institute Inc., 2002; Fichman and Cummings, 2003). No matter which complete-data analysis is used, the process of combining results from multiple imputed data sets is essentially the same. Suppose $\hat{Q}_i$ and $\hat{U}_i$ are the point and variance estimates from the *i*th imputed data set, $i = 1, 2, \ldots, m$. Then the combined point estimate for *Q* from MI is the average of the *m* complete-data estimates:

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \tag{9}$$

Suppose $\bar{U}$ is the within-imputation variance, which is the average of the *m* complete-data estimates:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i \tag{10}$$

and *B* be the between-imputation variance:

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2 \tag{11}$$

Then the variance estimate associated with $\bar{Q}$ is the total variance:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B \tag{12}$$

The statistic $(Q - \bar{Q})T^{-(1/2)}$ is approximately distributed as t with ν_m degrees of freedom (Rubin, 1987), where $\nu_m = (m-1)[1 + (\bar{U}/(1+m^{-1})B)]^2$. A rough 95% confidence interval can be obtained as $\bar{Q} \pm t_{\nu_m, 0.05}\sqrt{T}$.

3. Site descriptions and data collection

As case studies, we applied MI to gap-fill the eddy covariance data collected from the Walker Branch Watershed, the Duke Forest and the Niwot Ridge AmeriFlux sites. General information about these sites is given briefly in the following paragraphs. For detailed information, see Baldocchi (1997), Baldocchi and Wilson (2001), Wilson and Baldocchi (2000, 2001), Katul et al. (1997, 1999) and Monson et al. (2002).

3.1. Walker Branch Watershed site

This site is located in the US Department of Energy reservation near Oak Ridge, TN (35°57′30″N, 84°17′15″W). The vegetation is temperate mixed broad-leaved forest, including species of oak (*Quercus* spp.), maple (*Acer* spp.), and tulip poplar (*Liriodendron tulipifera*). The site is in hilly terrain, and the upwind fetch of the forest extends several kilometers in all directions (Baldocchi et al., 2000). The forest is about 58 years old. Canopy height is 25–26 m and peak leaf area index is 4.9–6.0 $m^2\,m^{-2}$. Eddy covariance instruments have been operating continuously since August 1994. Wind speed (WS) and virtual temperature fluctuations were measured with a three-dimensional sonic anemometer (model SWS-211/3K, Applied Technology, Boulder, CO). Fluctuations in CO_2, and H_2O concentration were measured with an open path, infrared absorption gas analyzer. Temperature (T_A) and relative humidity (RH) were measured at 36.9 m with a temperature/humidity probe (HMP-35 A, Vaisala, Helsinki, Finland). Photosynthetically active radiation (PAR) was measured above the canopy with a quantum sensor (model LI-190S, Li-cor Inc., Lincoln, NE).

3.2. Duke Forest site

This site is located in Orange County, NC, USA (35°58′N, 79°05′W). The site consists of an even-aged loblolly pine (*Pinus taeda*) forest. Tree growth in the plantation is remarkably uniform, with a median height of 13 m, a mean diameter of about 15 cm and a peak leaf area index of about 3.5 $m^2\,m^{-2}$ (in 1996). Fluxes for CO_2, H_2O and sensible heat were measured using a Li-Cor 6262 gas analyzer together with a CSAT3 (Campbell Scientific) triaxial sonic anemometer. PAR was measured over the canopy using Li-190SZ (Li-Cor Ins., Lincoln, NE, USA). T_A and WS were measured using the CSAT3 anemometer at the canopy top. RH and VPD were measured using a Vaisala probe positioned at 2/3 the canopy height. T_S was measured via thermistors (Siemens GmbH, Nuernberg, Germany) at one point at 10–12 cm depth.

3.3. Niwot Ridge site

This site is located at Niwot Ridge, CO, USA (40°1′N, 105°32′W). The site consists of ∼97-year-old second growth subalpine forest, dominated by three conifers, Engelmann spruce (*Picea engelmanni*), lodgepole pine (*Pinus contorta*), and subalpine fir (*Abies lasiocarpa*). Canopy height is 11.4 m and maximum LAI is ∼4.2 $m^2\,m^{-2}$. WS was measured with a Campbell Scientific Inc. (model CSAT-3) sonic anemometer and CO_2 concentration was measured with a Licor Inc. (model 6262) closed-path infrared analyzer.

Data used in the analysis were mainly downloaded from the above three sites through links at the FluxNet website (http://ornl5.ornl.gov/ameriflux/Data/index.cfm, 10/1/2002). In order to compare with the methods used by Falge et al. (2001a,b), we also downloaded gap-filled data by Falge et al. (2001a,b) (http://public.ornl.gov/fluxnet/gapzips.cfm, 10/1/2002). When data sets contained both observed and gap-filled data, we deleted the gap-filled values, and flagged them as missing values. In total, 7 years of eddy covariance measurements including NEE, LE, H, PAR, T_A, T_S, RH, VPD, R_n, WS and u^* were obtained for the three sites. MI was conducted using SAS software (SAS Institute Inc., 2002). The analysis includes the following steps: (1) Using PROC MI to create the multiple data sets, each data set is a complete one as the missing values were imputed. Five imputations ($m = 5$) were created using multiple chains. We selected an EM algorithm (maximum number of iterations used in EM was set to 500) for initial value estimations and

MCMC for data augmentation. The number of burn-in iterations before the first imputation in each chain for MCMC method was set to 1000 and the number of iterations between imputations in a single chain was set to 500. (2) Using standard procedures, PROC MEANS and PROC CORR, to calculate mean and covariance of analyzed variables from these data sets (Hui and Jiang, 1996); (3) Using PROC MIANALYZE to combine the results from each data set analyzed in step 2, and get the means, standard errors and confidence intervals. The running time for each year was less than 3 min on a Dell Precision Workstation 340.

4. Results

4.1. Gaps in observed NEE, LE, H and climatic variables

Missing or rejected data in measured NEE, LE and H ranged from 9 to 60, 6 to 60, 6 to 34, for the Walker Branch, Duke Forest, and Niwot Ridge sites, respectively, among the studied years (Table 1). Missing information was 6–17% for NEE, LE and H at the Niwot Ridge site, remarkably less than that at the other two sites. Compared to ecosystem carbon and energy fluxes, values of climatic variables were well recorded. Only R_n and u^* had more missing data, ranging from 18–42% at the WBW site and 26–51% at the Duke site (Table 1). A small number of missing climatic variable values were imputed by the regression method before MI at the Niwot Ridge site (Monson et al., 2002).

4.2. Annual estimation of NEE, LE and H

Annual NEE estimated by MI at the WBW site was −842, −923, and −845 g C m^{-2} per year in 1995, 1996 and 1997, respectively (Table 2). The 95% confidence interval was −902 to −782 g C m^{-2} per year in 1995, −974 to −873 g C m^{-2} per year in 1996, and −900 to −790 g C m^{-2} per year in 1997. Annual NEE at the Duke site was −624 g C m^{-2} per year in 1998 and −743 g C m^{-2} per year in 1999. The corresponding 95% confidence interval

Table 2
Annual estimations of NEE, LE and H derived from gap-filled eddy covariance data by multiple imputation (MI) method

Site, year	Annual estimation	Standard error of annual estimation	95% confidence interval	
			Lower limit	Upper limit
NEE (g C m^{-2} per year)				
WBW, 1995	−842	29	−902	−782
WBW, 1996	−923	25	−974	−873
WBW, 1997	−845	28	−900	−790
Duke, 1998	−624	34	−702	−546
Duke, 1999	−744	20	−784	−703
Niwot, 1999	−162	11	−183	−142
Niwot, 2000	−123	9	−142	−105
LE (MJ m^{-2} per year)				
WBW, 1995	1336	19	1297	1374
WBW, 1996	1382	21	1340	1423
WBW, 1997	1538	22	1495	1582
Duke, 1998	1129	24	1075	1183
Duke, 1999	1254	19	1217	1290
Niwot, 1999	1622	16	1590	1653
Niwot, 2000	1685	16	1654	1716
H (MJ m^{-2} per year)				
WBW, 1995	967	21	926	1008
WBW, 1996	878	20	838	918
WBW, 1997	976	22	932	1020
Duke, 1998	745	27	685	806
Duke, 1999	907	20	869	946
Niwot, 1999	1311	35	1242	1380
Niwot, 2000	1320	35	1251	1388

Table 1
Percentages of missing or rejected NEE, LE and H, and climatic variables at WBW, Duke and Niwot sites

Site, year	NEE	LE	H	PAR	T_A	T_S	RH	VPD	R_n	WS	u^*
WBW, 1995	46.6	30.5	34.2	0	0.3	1.1	0	0	41.6	4.4	22.7
WBW, 1996	44.1	27.4	29.7	0	0	1.1	0	0	0	3.2	17.6
WBW, 1997	47.3	30.1	33.0	0	0	0.3	0	0	40.8	5.4	20.6
Duke, 1998	59.5	59.5	31.0	0	23.7	10.5	50.7	50.7	51.0	51.0	51.0
Duke, 1999	31.7	30.0	25.6	8.1	2.1	2.1	2.1	2.1	25.8	25.6	36.2
Niwot, 1999	16.9	12.6	12.6	0	0	49.6	0	0	0	0	0
Niwot, 2000	9.4	5.9	6.2	0	0	0	0	0	0	0	0

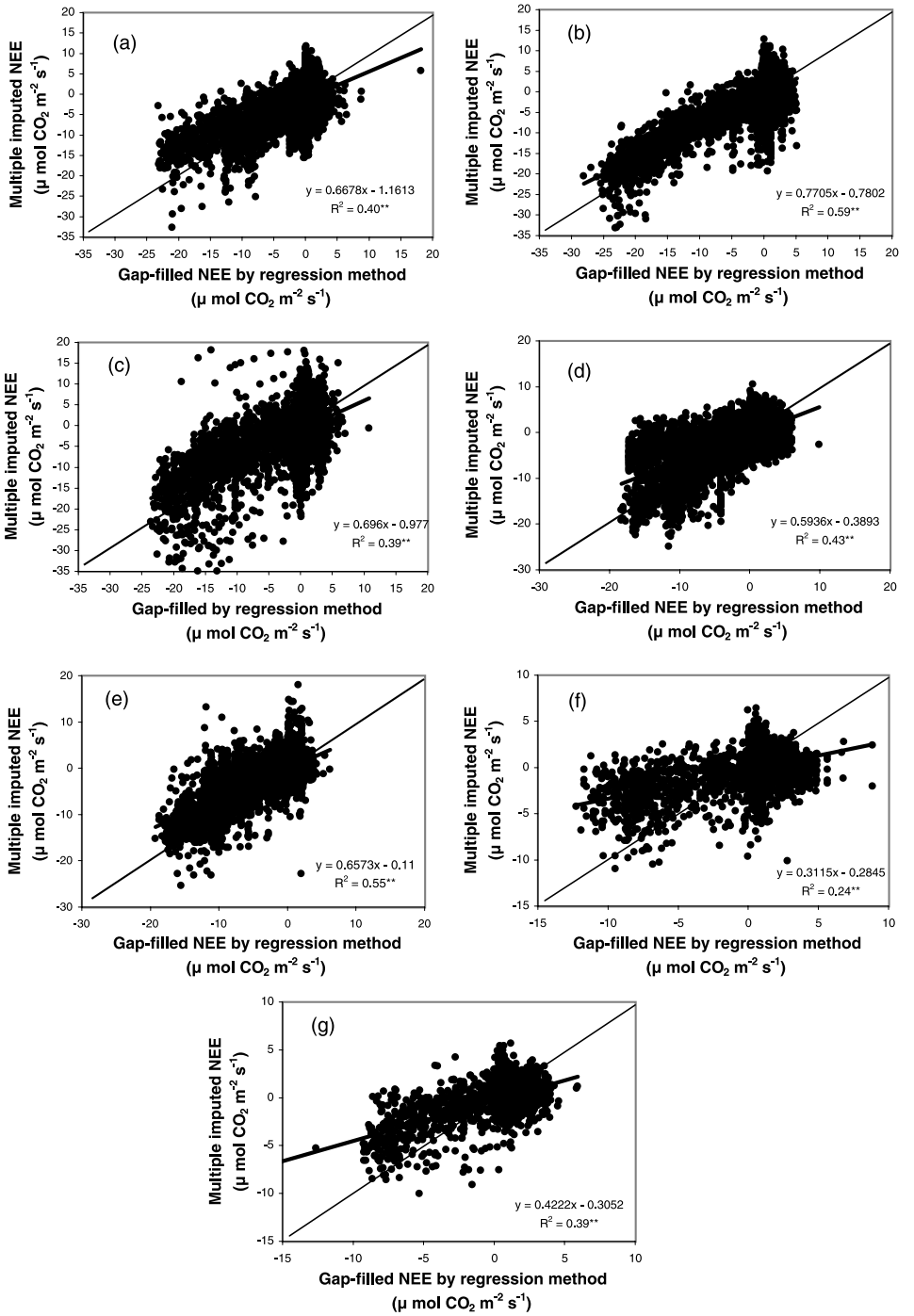

Fig. 1. Comparison of multiple imputed NEE (mean of five multiple imputed values) and imputed NEE by regression method at WBW site in 1995 (a), 1996 (b) and 1997 (c), at Duke Forest site in 1998 (d), in 1999 (e), at Niwot Ridge site in 1999 (f) and in 2000 (g). Imputed NEE by regression method in (a)–(e) are from Falge et al. (2001a) and in (f) and (g) are from Monson et al. (2002). (**) Represents significant at $\alpha = 0.01$ level.

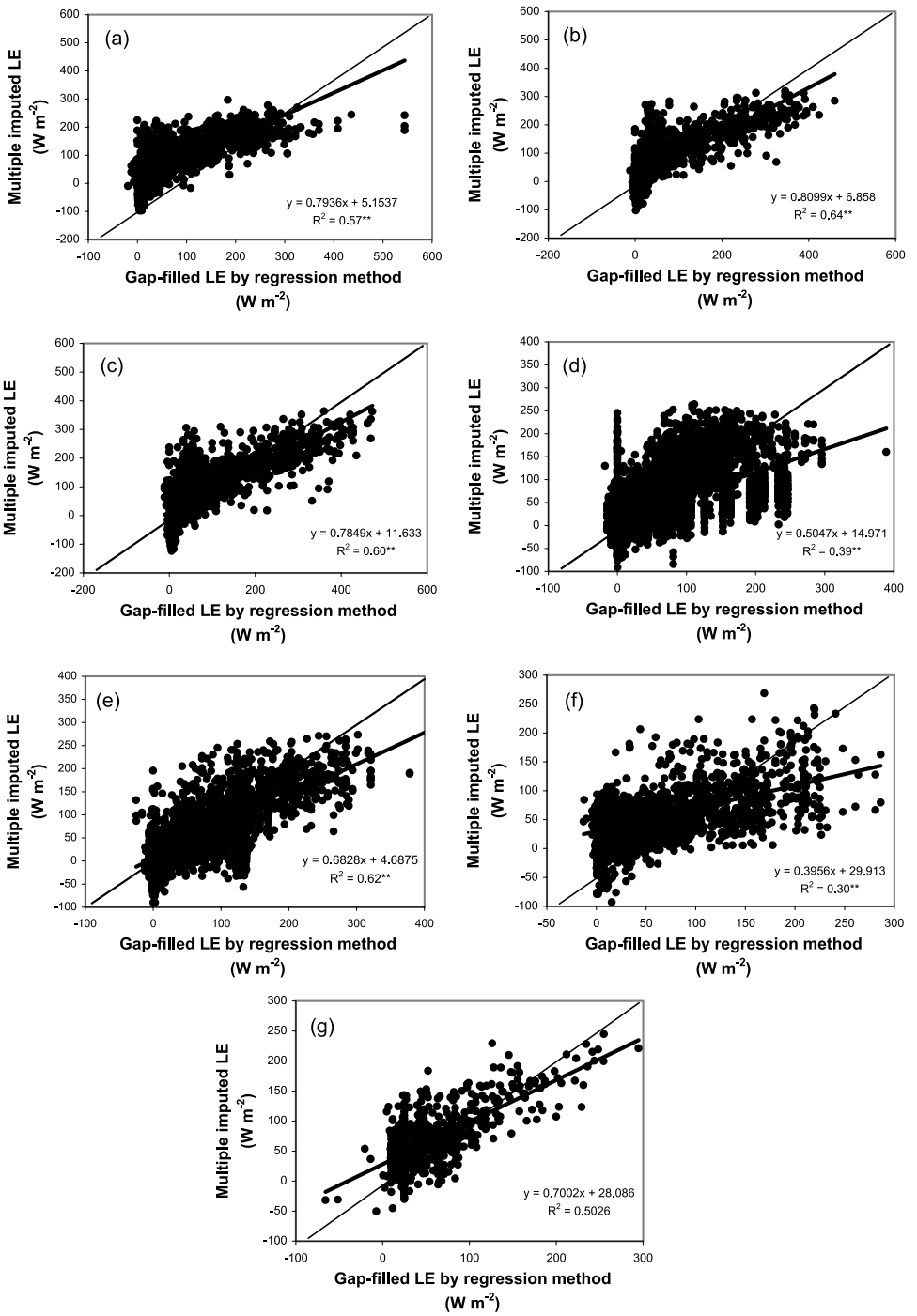

Fig. 2. Comparison of multiple imputed LE (mean of five multiple imputed values) and imputed LE by regression method at WBW site in 1995 (a), 1996 (b) and 1997 (c), at Duke Forest site in 1998 (d), in 1999 (e), at Niwot Ridge site in 1999 (f) and in 2000 (g). Imputed LE by regression methods in (a)–(e) are from Falge et al. (2001b) and in (f) and (g) are from Monson et al. (2002). (**) Represents significant at $\alpha = 0.01$ level.

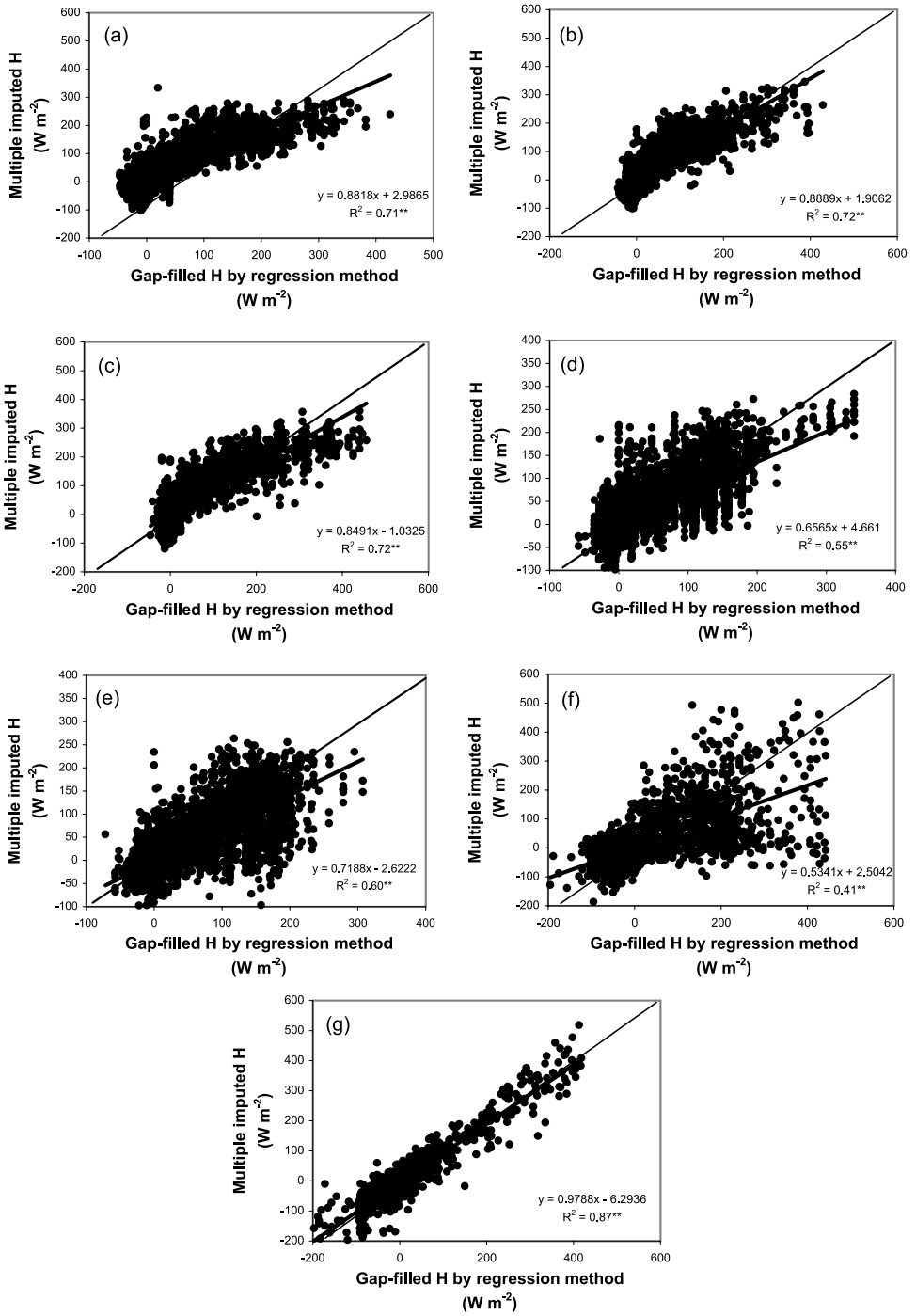

Fig. 3. Comparison of multiple imputed H (mean of five multiple imputed values) and imputed H by regression method at WBW site in 1995 (a), 1996 (b) and 1997 (c), at Duke Forest site in 1998 (d), in 1999 (e), at Niwot Ridge site in 1999 (f) and in 2000 (g). Imputed H by regression methods in (a)–(e) are from Falge et al. (2001b) and in (f) and (g) are from Monson et al. (2002). (∗∗) Represents significant at $\alpha = 0.01$ level.

was −702 to −546 g C m^{-2} per year in 1998 and −784 to −703 g C m^{-2} per year in 1999. Annual NEE at the Niwot Ridge site was −162 g C m^{-2} per year in 1999 and −123 g C m^{-2} per year in 2000. The corresponding 95% confidence interval was −183 to −142 g C m^{-2} per year in 1999 and −142 to −105 g C m^{-2} per year in 2000. The estimations varied significantly among years at these sites.

Annual LE estimated by MI at the WBW site was 1336, 1382, and 1538 MJ m^{-2} per year in 1995, 1996 and 1997, respectively (Table 2). The 95% confidence interval was 1297–1374 MJ m^{-2} per year in 1995, 1340–1423 MJ m^{-2} per year in 1996, and 1495–1582 MJ m^{-2} per year in 1997. Annual LE at the Duke Forest site was 1129 MJ m^{-2} per year in 1998 and 1254 MJ m^{-2} per year in 1999. Annual LE at the Niwot Ridge site was higher than other two sites, reaching 1622 and 1685 MJ m^{-2} per year in 1999 and 2000, respectively.

Similar to LE estimations, annual H estimated by MI ranged from 878 MJ m^{-2} per year in 1996 to 976 MJ m^{-2} per year in 1997 at the WBW site (Table 2). H at the Duke Forest site was close to that at the WBW site. Annual H at the Niwot Ridge site was also higher than that at the other two sites, reaching 1311 and 1320 MJ m^{-2} per year in 1999 and 2000, respectively.

4.3. Comparison of multiple imputed NEE, LE and H with gap-filled values by regression method and observed values

Missing NEE that imputed by MI was compared with gap-filled value by the regression method (Falge et al., 2001a; Monson et al., 2002). In general, there was a very significant correlation of NEE imputed by MI and regression method (Fig. 1). For most years, the determination coefficients were higher than 0.40. Even stronger correlation relationships were found for imputed LE by MI and regression method (Fig. 2) as well as for imputed H (Fig. 3).

NEE imputed by MI was consistent in magnitude and seasonality with observed data (figures not shown). The range of variation of multiple imputed NEE was similar to the observed values. To characterize the nature of multiple imputed data, we displayed the diurnal change of NEE, LE and H during the growing season in 1995 at the WBW site (Fig. 4). Data in other years at the sites showed similar trends. Estimated NEE, LE and H were consistent with the observed diurnal patterns, even when most of the data were missing for a day (e.g. days 197, 198 and 199). But in winter, while multiple imputed H fit the observed pattern, imputed NEE and LE did not show a clear diurnal pattern (Fig. 5). The observed NEE and LE in winter also did not show a clear diurnal pattern.

We also tested the goodness-of-filling of MI by deleting the observed NEE, LE and H for 2 weeks either in winter or in summer, then gap-filling these "missing" data using MI, and comparing the gap-filled values with observed ones. The goodness-of-filling was expressed as

$$R^2 = \frac{\sum(Y_{\text{obs}} - \bar{y})^2 - \sum(Y_{\text{obs}} - Y_{\text{imp}})^2}{\sum(Y_{\text{obs}} - \bar{y})^2}$$

where Y_{imp} is gap-filled value by MI. Multiple imputed NEE in summer in 2000 at Niwot Ridge site were similar to the observed data in most of the days (Fig. 6a). Only in a few days (e.g. day 156), multiple imputed NEE was smaller in magnitude than observed NEE. There was also good agreement in gap-filled LE and H with measurements for these 2 weeks at Niwot Ridge site (Fig. 6b and c). In the winter time, while multiple imputed H still followed the observed H well, NEE and LE were not consistent with observed values. However, due to the low winter value of NEE and LE in nature, the slightly mismatch did not have much influence on the total annual estimation on NEE and LE.

5. Discussion

5.1. Comparison of estimations by MI and other methods

By using the MI method to gap-fill the missing eddy covariance data, we estimated annual NEE and energy fluxes for 7 years at three AmeriFlux sites. The estimations, in general, were comparable to estimations by other gap-filling methods. For example, Falge et al. (2001a,b) estimated annual NEE using mean diurnal variation, look-up tables, and nonlinear regression methods, and total LE and H by mean diurnal variation and look-up tables for the same periods at the WBW and the Duke Forest sites. Their results showed that annual NEE estimation varied by different

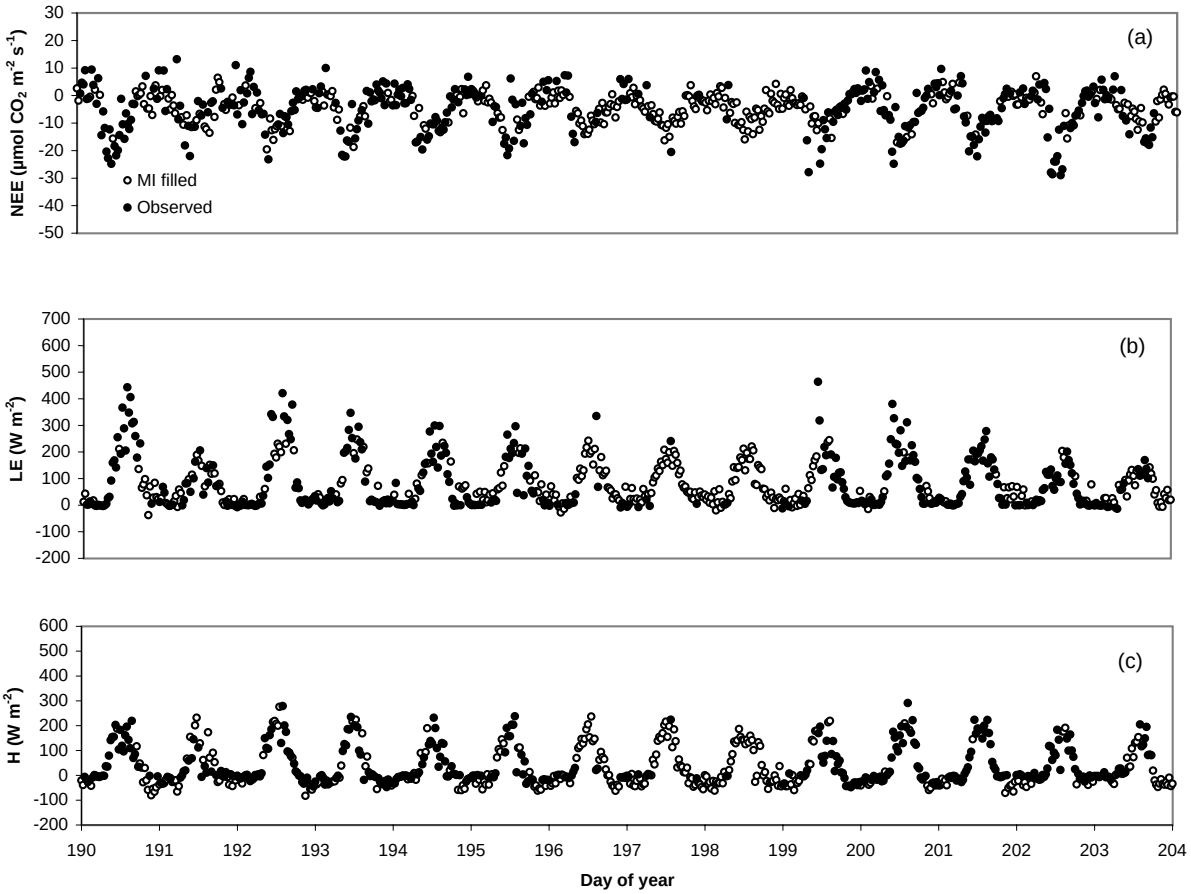

Fig. 4. Diurnal variation of observed (solid point) and imputed (empty point) of NEE (a), LE (b) and H (c) in the growth season in 1995 at WBW site.

gap-filling methods. Annual NEE in 1998 at the Duke Forest site ranged from −585 to −555 g C m^{-2} per year for u^*-corrected data by different methods and was −710 g C m^{-2} per year for not u^*-corrected data by regression method; our estimates showed annual NEE to be between −702 and −546 g C m^{-2} per year with 95% confidence. Annual H at the WBW site in 1995 estimated by Falge et al. (2001b) ranged from 954 to 1015 MJ m^{-2} per year, also mostly overlapped with our 95% confidence interval estimation. We also found some inconsistence in the estimations of NEE, LE and H with other methods. MI estimations of NEE at the WBW site in 1995 and 1996 were ∼30% higher in magnitude, and LE and H at the Duke Forest site in 1999 were ∼10% lower, in comparison to the estimations by Falge et al. (2001a,b).

In order to estimate the annual value of NEE, LE and H, missing data must be gap-filled before further analysis can be done. Current methods, such as mean replacement and regression analysis, work well under certain conditions. For example, mean diurnal variation, a mean replacement method, performs well when the gaps are not large; however, the window size may vary from site to site, making it difficult to compare data sets. Regression methods, based on either linear or nonlinear regression equations, may work well when predictors are strong, but the absence of sufficient variability can cause an underestimation of standard errors (Little and Rubin, 1989). Because regression methods are only able to produce the mean flux densities, the ranges of imputed data are smaller than observed ones. The high degree of scatter found

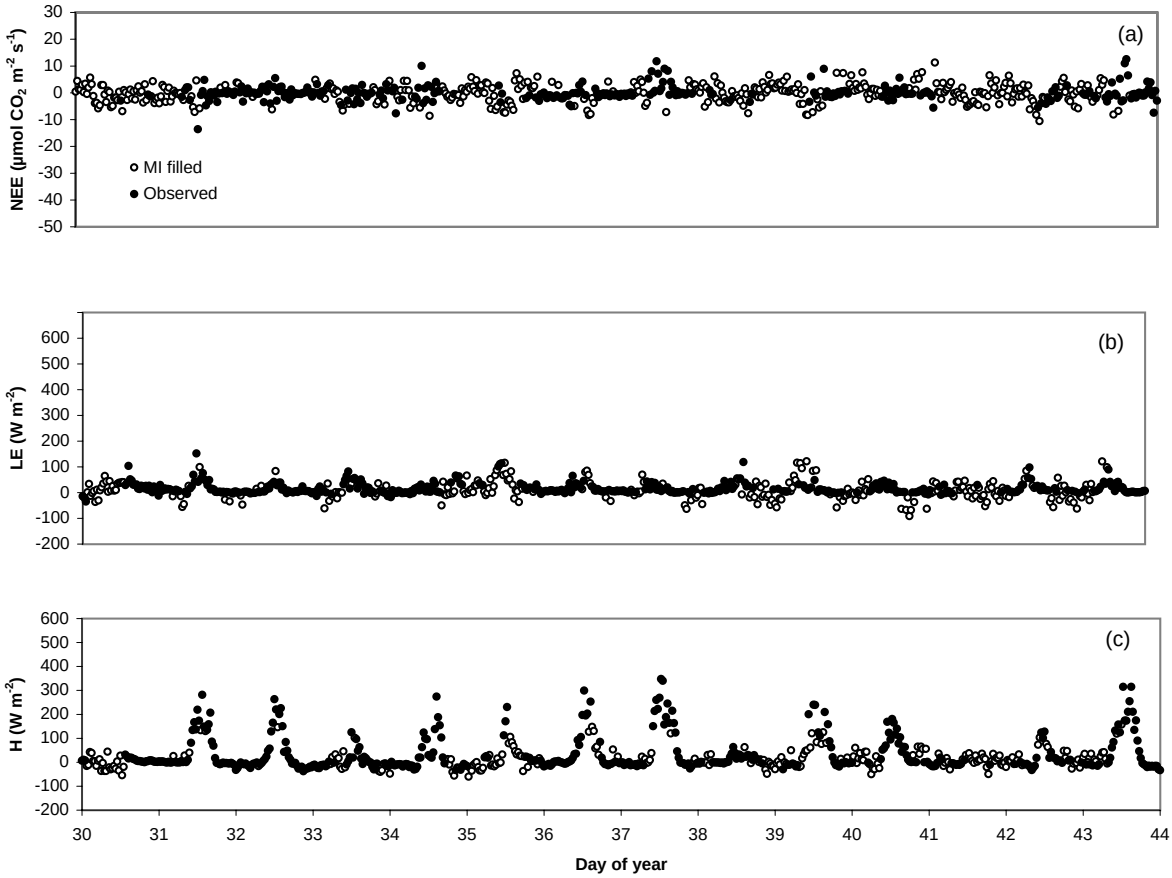

Fig. 5. Diurnal variation of observed (solid point) and imputed (empty point) of NEE (a), LE (b) and H (c) in winter in 1995 at WBW site.

in eddy covariance flux data also limits the application of regression method for data imputation.

Another approach is modeling approach (i.e. process-based gap-filling method). Theoretically, this method can be used to impute any missing data. It may be better than other gap-filling methods at certain sites (e.g. sites with many and large gaps). But in practice, that the model should be specifically configured for each FluxNet site makes it difficult to apply. That's one reason why so far, most of the imputation methods are empirical statistical methods. In addition, all these methods require specific computer programming. There appear to be clear advantages to the MI process.

Net ecosystem carbon exchange is underestimated by the eddy covariance approach during stable-night conditions because of CO_2 storage in the layer below the eddy flux system. To date, no general consensus has been found for correcting the fluxes, and considerable work in terms of methodology and underlying theory will be required to address the uncertainties associated with nighttime fluxes (Falge et al., 2001a). For the long-term budget, such as the daily and annual estimation of NEE, beneath-canopy CO_2 storage can theoretically be ignored (Aubinet et al., 2000; Falge et al., 2001a). However, if u^*-corrected, the annual NEE estimation by the regression method was on average 64 g C m^{-2} per year less than the not u^*-corrected data, for the 28 data set analyzed by Falge et al. (2001a). Data used in this MI exercise were not u^*-corrected. We may expect that estimated values for NEE will be lower than reported in this study, if u^*-corrected were used. Indeed, when we used u^*-corrected NEE data for the Niwot Ridge site, the annual estimation

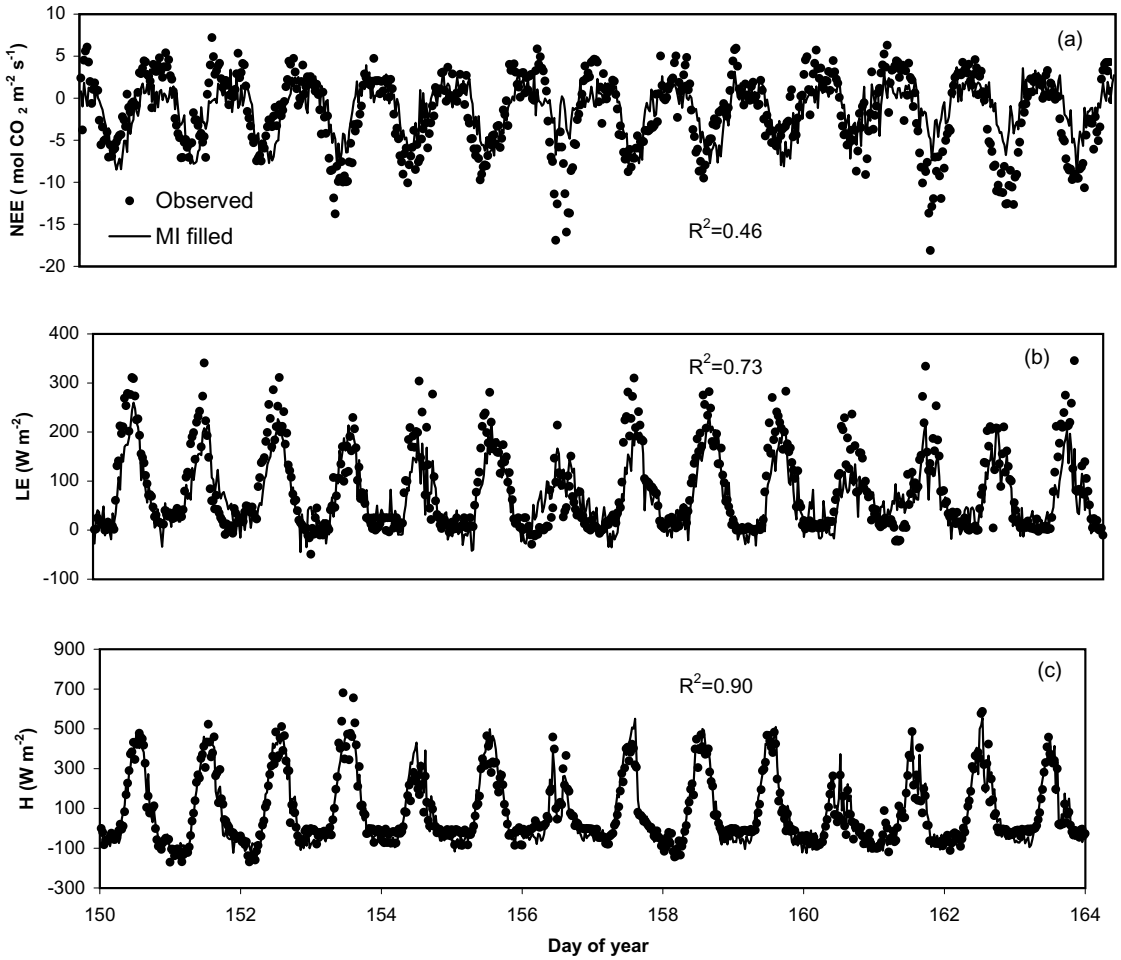

Fig. 6. Comparison of gap-filled NEE, LE and H (line) with measurements (solid point) in summer in 2000 at Niwot Ridge site.

of NEE by MI was $-94\,\mathrm{g\,C\,m^{-2}}$ per year, with a 95% confidence interval of -115 to $-73\,\mathrm{g\,C\,m^{-2}}$ per year in 1999, and $-51\,\mathrm{g\,C\,m^{-2}}$ per year with a 95% confidence interval of -70 to $-32\,\mathrm{g\,C\,m^{-2}}$ per year in 2000. These values are close to the reported values (80.5 and $57.6\,\mathrm{g\,C\,m^2}$ per year in 1999 and 2000, respectively) based on u^*-corrected data by Monson et al. (2002).

5.2. Seasonal and diurnal change of imputed NEE, LE and H

While in general MI data provided good estimates of annual sums and preserved the seasonal pattern of measurements, we noticed that some of the corrected data did not closely fit the observed pattern, especially in winter, when NEE is small in magnitude and the diurnal pattern is not always clear. This does not provide a significant source of error for the annual sum and may be improved by grouping data into those for the growing season and dormant season, and applying MI separately.

Preserving the relationships of NEE, LE and H with climatic variables such as PAR and temperature is another requirement that is often considered during gap filling. MI uses a multivariate normal distribution model and considers other variables, such as climatic variables, when imputing missing data. The missing

values for the climatic variables were imputed at the same time, so MI preserved the responses of NEE to these climatic variables as a whole. The data set we used for annual estimation was based on the observed values in 1 year. To preserve the short-term relationships of NEE, LE and *H* with climatic variables, we could apply MI to the short-term data set, for example, data collected from each month or grouped by growing season and dormant season, or separated into daytime and nighttime data.

5.3. Assumptions and constraints of MI

Like any statistical method, MI has assumptions (Schafer and Olsen, 1998). Understanding the assumptions underlying the MI method helps researchers evaluate robustness of imputation models and the appropriateness of interferences (Horton and Lipsitz, 2001). Those assumptions include data are missing at random (MAR), multivariate normal distributions, and proper model.

There are two types of randomness in missing data: MAR and missing completely at random (MCAR). MCAR assumes that the missing values are a random subsample of the entire data set. The assumption for MCAR is much restrictive and often unrealistic. The MI method does not require MCAR but MAR (Little and Rubin, 1989). Under the assumption of MAR, the probabilities of missing data may depend on data values that are observed but not ones that are missing. A simple example is that for a bivariate data set with one variable *X* that is always observed and a second variable *Y* that is sometimes missing. Under MAR, the probability that *Y* is missing for an observation may be related to the value of *X* but not to the value of *Y* itself. This applies that the statistical relationship of *Y* and *X* is on average no different for the observed data and missing data groups. MAR is the formal assumption that allows us to first estimate the relationships among variables from the observed data, and then use these relationships to obtain unbiased predictions of the missing values from the observed values. However, testing the MAR assumption remains a major statistical challenge (Schafer and Olsen, 1998; Allison, 2000; Horton and Lipsitz, 2001). Although it is difficult to test, we have good reason to assume the missing data in NEE, LE and *H* are MAR. The probability of missing values in NEE, for example, may be related to low u^*, but missing NEE values, may not be related to NEE itself, so the relationships developed from observed valid data can be applied to the missing values. Similarly, missing data under severe weather conditions can be imputed from the relationship indicated from the observed data under the normal weather conditions. Studies also showed that the assumption becomes more plausible as more variables are included in the imputation model (Schafer, 1997; van Buuren et al., 1999). In this study, we included many climatic variables in the model, so the assumption is more plausible. When the MAR assumption is tenable, MI provides less bias than other methods if the imputation model is correctly specified.

Second, the model used to generate the imputed values requires the variables to be multivariate normal distributed. This assumption is also required by many other imputation methods (Schafer and Olsen, 1998). As Schafer and Olsen (1998) pointed out, real data rarely conform to convenient models such as the multivariate normal. In most applications of MI, the model used to generate the imputations will at best be only approximately true. Indeed, Kolmogorov–Smirnov normal distribution test showed that observed NEE, LE and *H* and climatic variables are seldom truly normal distributed (results not shown). A number of simulation studies have demonstrated that MI is robust to violations of normality of the variables used in the analysis if the amount of missing information is not large (Ezzati-Rice et al., 1995; Schafer, 1997; Graham and Schafer, 1999). We examined the assumption of multivariate normal distribution by transforming data from Niwot Ridge site in 2000 and Duke Forest in 1999 using the Box–Cox transformation method provided in SAS software. Comparison of the results of transformed data with non-transformed data indicates the annual estimations only changed slightly. Annual NEE at Niwot Ridge site in 2000 changed from -123.2 ± 9.4 to $-121.6 \pm 9.4\,\mathrm{g\,C\,m^{-2}}$ per year. Annual NEE at Duke Forest in 1999 changed from 743.6 ± 20.3 to $746.2 \pm 19.7\,\mathrm{g\,C\,m^{-2}}$ per year. Similarly, annual values of LE and *H* were not affected by transformation. Our test confirmed that MI is robust to some departures to normality. So we used non-transformed results in this study. The goodness-of-fit test also showed that MI performed well on FluxNet data (Figs. 6 and 7).

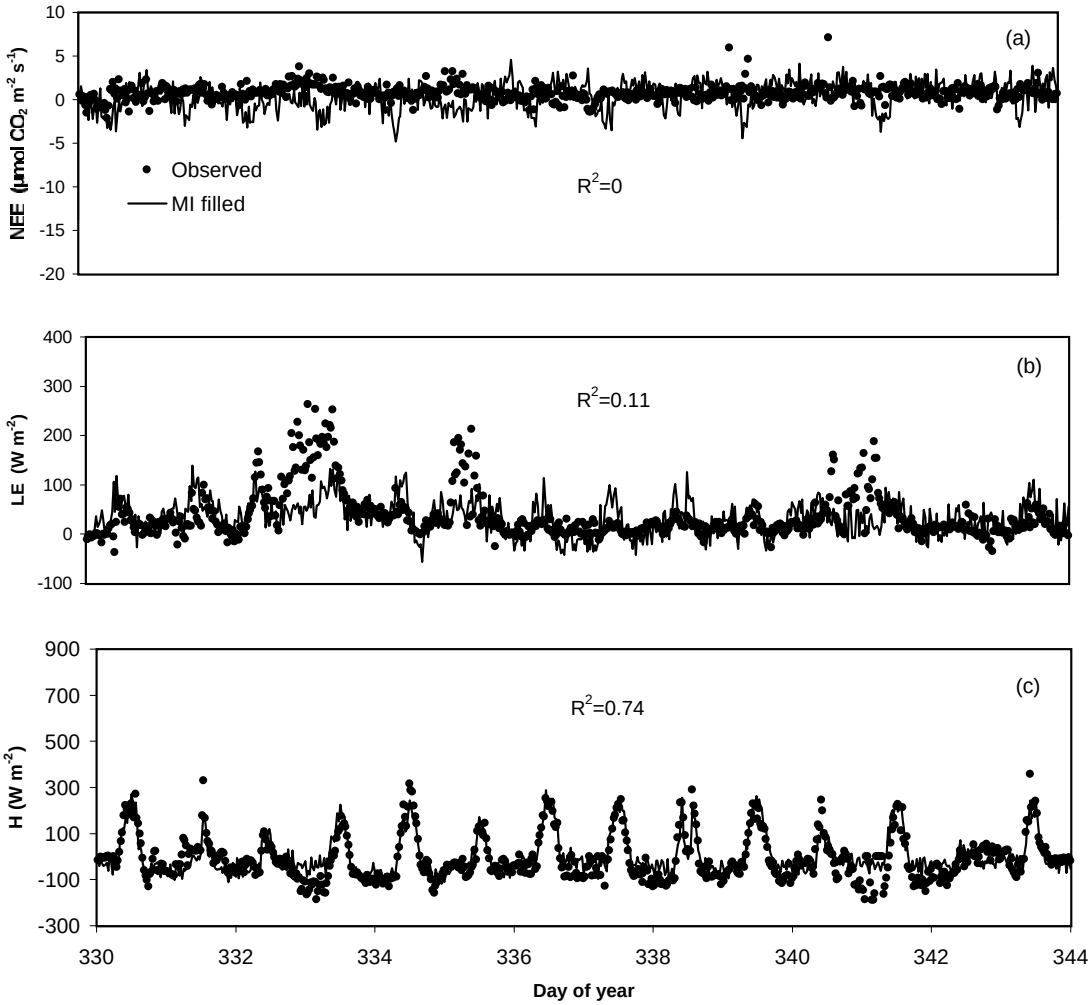

Fig. 7. Comparison of gap-filled NEE, LE and *H* (line) with measurements (solid point) in winter in 2000 at Niwot Ridge site.

Third, the model used for the analysis should match the model used in imputation and the algorithm used to generate imputed values should be correct; that is, it should accommodate the necessary variables and their associations. In this study, we included most influencing climatic variables in the imputation process, so the climatic variables can be included in the model for further analysis.

5.4. Advantages of MI

MI provides a general-purpose solution to statistical analysis with missing data and provides more valid estimates of statistical quantities (e.g. means, standard errors, regression coefficients) than other current practices (Fichman and Cummings, 2003). MI combines the well-known statistical advantages of EM and maximum likelihood with the ability of hot deck imputation to provide a raw data matrix to analyze. It also introduces statistical uncertainty into the model and uses that uncertainty to emulate the natural variability among observations one encounters in a complete database. MI then imputes actual data values to fill in the incomplete data points in the data matrix, just as hot deck imputation does.

Table 3
Fraction of missing information and efficiency of MI by using five imputations

Site, year	Fraction of missing information (%)			Efficiency of MI (%)		
	NEE	LE	H	NEE	LE	H
WBW, 1995	53.2	12.2	10.4	90.3	97.6	98.0
WBW, 1996	18.3	12.8	14.7	96.5	97.5	97.1
WBW, 1997	18.7	8.0	14.1	96.4	98.4	97.2
Duke, 1998	76.6	65.1	67.4	86.7	88.5	88.1
Duke, 1999	24.4	1.1	16.8	95.3	97.8	96.7
Niwot, 1999	9.5	2.9	6.0	98.1	99.4	98.8
Niwot, 2000	2.3	3.2	0.8	99.5	99.4	99.8

MI is highly efficient even for small values of m (Schafer, 1997). In many applications, just three to five imputations are sufficient to obtain excellent results. The relative efficiency of an estimate based on m imputations is approximately $(1+(\gamma/m))^{-1}$, where γ is the fraction of missing information for the variable being estimated, $\gamma = (r + 2/(\nu_m + 3))/r + 1$, and $r = ((1+m^{-1})B)/\bar{U}$. In this study, the fraction of missing information for NEE, LE and H ranged from 0.8 to 76.6%. With five imputations, the efficiency of the estimate ranged from 86.7 to 99.8% (Table 3).

Another advantage of MI is that the results obtained by individual researchers at different sites or among different years can be compared. Because repeated estimations are used, MI produces more reasonable estimates of standard errors than single imputation and current methods. This results in valid statistical inferences that properly reflect the uncertainty due to missing values. Given these objectives, MI becomes an attractive procedure for dealing with missing data issues (Fichman and Cummings, 2003).

Although MI was first proposed more than 20 years ago (Rubin, 1977), the method has remained largely unknown and unused by non-experts (Schafer and Olsen, 1998) due to the lack of computational tools. MI is especially new to ecologists, as there are few publications that formally address it. As there are many large-scale and long-term experiments (e.g. FluxNet, LTER, and FACE) that are operating, and more and more researchers are synthesizing meta-data (i.e. data collected from many individual studies in the literature) (e.g. Curtis, 1996; Wan et al., 2001), the issue of how to deal with missing observations, rather than simply deleting them, becomes important.

Before the advent of general purpose packages that support MI, the process of generating m imputed data sets, analyzing the results from each of the m data sets, and combining the results required specialized programming that was difficult to use. Complications in data handling and analysis have been greatly simplified by the existence of easy-to-use software packages such as SAS, SPSS, Norm, and S-plus (Schafer and Olsen, 1998; Hox, 1999; Horton and Lipsitz, 2001). With the availability of these techniques and useful software, MI using MCMC method can become part of the mainstream research practice of gap-filling missing data (Fichman and Cummings, 2003).

6. Conclusions

Prior to any analysis, researchers at FluxNet sites must examine their data sets for the amount and pattern of missing data and determine the best approach to handle them (Patrician, 2002). MI is one of the methods that need to be considered. By using MI to gap-fill the missing or rejected eddy covariance data, we estimated annual sums of NEE, LE and H and their confidence intervals based on the data collected from two coniferous forests and a deciduous forest for a combined 7 years. Results by MI were comparable to the estimations by other current methods. NEE, LE and H imputed by MI were also consistent with the seasonal and diurnal patterns in observed data. How to impute the missing data is not a trivial issue. While there are many other imputation methods available, we suggest that using MI as a standardized method for gap-filling eddy covariance measurements would improve the comparability of annual estimations of NEE, LE and H from regional and global flux networks, and provide a uniform as well as objective standard for evaluating uncertainties in annual sums.

Acknowledgements

This study was supported by the NSF/DOE/NASA/USDA/EPA/NOAA Interagency Program on Terrestrial Ecology and Global Change (TECO) by DOE under DE-FG03-99ER62800 to YL at the University of Oklahoma and the NIGEC SouthCentral Regional Center at Tulane University. Eddy-covariance

flux measurements were supported by the NIGEC Southeastern Regional Center at the University of Alabama, Tascaloosa (DOE Cooperative Agreement DE-FC03-90ER61010), the NIGEC SouthCentral Regional Center at Tulane University, and the Terrestrial Carbon Processes (TCP) program at DOE and the US Department of Energy (Terrestrial Carbon Program) and NASA/GEWEX. We thank Dr. Dennis Baldocchi and Dr. Kell Wilson for providing eddy-covariance flux measurements at WBW site.

Appendix A. MCMC method for missing data

The Markov Chain Monte Carlo is commonly used to generate pseudo-random draws from multidimensional and otherwise intractable probability distributions via Markov chains. In Bayesian inference, information about unknown parameters is expressed in the form of a posterior probability distribution. The posterior distribution is computed using Bayes' theorem

$$p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{\int p(y|\theta)\,p(\theta)\,\mathrm{d}\theta}$$

MCMC has been applied as a method for exploring posterior distribution in Bayesian. Through MCMC, one can simulate the entire joint posterior distribution of the unknown quantities and obtain simulation-based estimates of posterior parameters that are of interest. This appendix mainly followed the notation by SAS institute (2002) and Little and Rubin (2002).

When data set contains missing data, the observed-data posterior $p(\theta|Y_{\text{obs}})$ is intractable and cannot easily be simulated. But the complete-data posterior $p(\theta|Y_{\text{obs}}, Y_{\text{mis}})$ is much easier to simulate once Y_{obs} is augmented by a simulated value of the missing data Y_{mis}. Suppose that the data are from multivariate normal distribution and assume the conventional Jeffery's prior distribution for the mean and covariance matrix:

$$p(\mu, \Sigma) \propto |\Sigma|^{-(K+1)/2}$$

we present an iterative data augmentation algorithm for generating draws from the posterior distribution of $\theta = (\mu, \Sigma)$:

$$p(\mu, \Sigma|Y_{\text{obs}}) \propto |\Sigma|^{-(K+1)/2} L(\mu, \Sigma|Y_{\text{obs}})$$

A.1. The imputation I-step

Let $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$ and $Y^{(t)} = (Y_{\text{obs}}, Y_{\text{mis}}^{(t)})$ denote current draws of the parameters and gap-filled data matrix at iteration t. The I-step draws values for the missing data from the conditional distribution Y_{mis} given Y_{obs} with a given parameter θ

$$Y_{\text{mis}}^{(t+1)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$$

Since the observations of the data matrix Y are conditionally independent given θ, this is equivalent to drawing

$$y_{\text{mis},i}^{(t+1)} \sim p(y_{\text{mis},i}|y_{\text{obs},i}, \theta^{(t)}) \tag{A.1}$$

independently for $i = 1, 2, \ldots, n$. This distribution is multivariate normal with mean given by the linear regression of $y_{\text{mis},i}$ on $y_{\text{obs},i}$, evaluated at current draws $\theta^{(t)}$ of the parameters. The regression parameters and residual covariance matrix of this normal distribution is obtained computationally by sweeping on the augmented covariance matrix

$$\Sigma^{*(t)} = \begin{bmatrix} -1 & \mu^{(t)'} \\ \mu^{(t)} & \Sigma^{(t)} \end{bmatrix}$$

so that the observed variables are swept in and the missing variables are swept out. The draw $y_{\text{mis},i}^{(t+1)}$ is simply obtained by adding to the conditional mean in the E-step of EM algorithm, Eqs. (3) and (5), a normal draw with mean 0 and covariance matrix $\Sigma_{\text{mis},i.\text{obs},i}^{(t)}$.

A.2. The posterior step P-step

The P-step of data augmentation simulates the posterior population mean of vector μ and covariance matrix Σ from prior information for μ and Σ, and the complete data set. P-step draws

$$\theta^{(t+1)} \sim p(\theta|Y^{(t+1)})$$

where $Y^{(t+1)} = (Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)})$ is the gap-filled data from the I-step. The draw of $\theta^{(t+1)}$ can be carried out in two steps:

$$\begin{aligned} (\Sigma^{(t+1)}|Y^{(t+1)}) &\sim W^{-1}(n-1, (n-1)S^{(t+1)}) \\ (\mu^{(t+1)}|\Sigma^{(t+1)}, Y^{(t+1)}) &\sim N(\bar{y}^{(t+1)}, \Sigma^{(t+1)}/n) \end{aligned} \tag{A.2}$$

where $(\bar{y}^{(t+1)}, S^{(t+1)})$ is the sample mean and covariance matrix of Y from the gap-filled data $Y^{(t+1)}$,

$W^{-1}(n-1,(n-1)S)$ denotes the inverted Wishart distribution with $n-1$ degrees of freedom and scale matrix $(n-1)S$. The posteriors distribution of θ can be simulated directly using Eqs. (A.1) and (A.2), after a suitable burn-in period to achieve stationary draws.

References

Allison, P.D., 2000. Multiple imputation for missing data—a cautionary tale. Sociol. Methods Res. 28, 301–309.

Aubinet, M., Grelle, A., Ibrom, A., Rannik, U., Moncrieff, J., et al., 2000. Estimates of the annual net carbon and water exchange of forests: the EUROFLUX methodology. Adv. Ecol. Res. 30, 113–175.

Baldocchi, D.D., 1997. Measuring and modeling carbon dioxide and water vapor exchange over a temperate broad-leaved forest during the 1995 summer drought. Plant Cell and Environ. 20, 1108–1122.

Baldocchi, D.D., Wilson, K.B., 2001. Modeling CO_2 and water vapor exchange of temperate broadleaved forest across hourly to decadal time scales. Ecol. Model. 142, 155–184.

Baldocchi, D.D., Falge, E., Gu, L., Olson, R., Hollinger, D., et al., 2001. FluxNet: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bull. Am. Meteorol. Soc., 2415–2434.

Baldocchi, D.D., Finnigan, J., Wilson, K.B., Paw, U.K.T., Falge, E., 2000. On measuring net ecosystem carbon exchange over tall vegetation on complex terrain. Bound. Layer Meteorol. 96, 257–291.

Barnard, J., Meng, X.L., 1999. Applications of multiple imputation in medical studies: from AIDS as NHANES. Stat. Methods Med. Res. 8, 17–36.

Carsob, C., Belongie, S., Greenspan, H., Malik, J., 2002. Blobworld: Image segmentation using expectation-maximization and its application to image querying. IEEE Trans. Pattern Anal. Machine Intell. 24, 1026–1038.

Curtis, P.S., 1996. A meta-analysis of leaf gas exchange and N in trees grown under elevated carbon dioxide. Plant Cell Environ. 19, 127–137.

Dayan, P., Hinton, G.E., 1997. Using expectation-maximization for reinforcement learning. Neural Comput. 9, 271–278.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via EM algorithm. J. R. Stat. Soc. B: Methodol. 39, 1–38.

Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A., Rubin, D.B., Schafer, J.L., 1995. A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. In: Proceedings of the Annual Research Conference. Bureau of the Census, Washington, DC, pp. 257–266.

Falge, E., Baldocchi, D.D., Olson, R., Anthoni, P., Aubinet, M., et al., 2001a. Gap-filling strategies for defensible annual sums of net ecosystem exchange. Agric. Forest Meteorol. 107, 43–69.

Falge, E., Baldocchi, D.D., Olson, R., Anthoni, P., Aubinet, M., et al., 2001b. Gap-filling strategies for long-term energy flux data sets. Agric. Forest Meteorol. 107, 71–77.

Fichman, M., Cummings, J.N., 2003. Multiple imputation for missing data: Making the most of what you know. http://www.gsia.cmu.edu/andrew/mf4f/work/misspm.pdf. Org. Res. Methods, in press.

Goulden, M.L., Munger, J.W., Fan, S.M., Daube, B.C., Wofsy, S.C., 1996. Measurement of carbon storage by long-term eddy correlation: methods and a critical assessment of accuracy. Global Change Biol. 2, 169–182.

Granier, A., Ceschia, E., Damesin, C., Dufrêne, E., Epron, D., et al., 2000. The carbon balance of a young beech forest. Function. Ecol. 14, 312–325.

Graham, J.W., Schafer, J.L., 1999. On the performance of multiple imputation for multivariate data with small sample size. In: Hoyle, R. (Ed.), Statistical Strategies for Small Sample Research. Sage, Thousand Oaks, CA, pp. 1–29.

Greco, S., Baldocchi, D.D., 1996. Seasonal variations of CO_2 and water vapour exchange rates over a temperate deciduous forest. Global Change Biol. 2, 183–198.

Grünwald, Th., Bernhofer, Ch., 2000. Regression modelling used for data gap filling of carbon flux measurements. In: Ceulemans, R.J.M., Veroustraete, F., Gond, V., Van Rensbergen, J.B.H.F. (Eds.), Forest Modelling, Upscaling and Remote Sensing. SPB Academic Publishing, The Hague, The Netherlands, pp. 61–67.

Hui, D., Jiang, C., 1996. Practical SAS Usage. Beijing University of Aeronautics & Astronautics Press, Beijing, China.

Hui, D., Jiang, C., Mo, H., 1997. Comparison among mapping methods for detecting QTLs and estimating their effects. Acta Agron. Sinica 23, 129–136.

Hui, D., Luo, Y., Katul, G., 2003. Partitioning interannual variability in net ecosystem exchange between climatic variability and functional change. Tree Physiol. 23, 433–442.

Horton, N.J., Lipsitz, S.R., 2001. Statistical computing software reviews. Multiple imputation in practice: comparison of software packages for regression models with missing variables. Am. Stat. 55, 244–254.

Hox, J.J., 1999. A review of current software for handling missing data. Kwantitatieve Methoden 62, 123–138.

Jarvis, P.G., Massheder, J., Hale, D., Moncrieff, J., Rayment, M., Scott, S., 1997. Seasonal variation of carbon dioxide, water vapor and energy exchanges of a boreal black spruce forest. J. Geophys. Res. 102, 28953–28967.

Jiang, C., Hui, D., 1995. Joint mapping of genes for correlated quantitative traits. J. Jiangsu Agric. College 16, 1–7.

Jiang, C., Zeng, Z.-Z., 1997. Mapping quantitative loci with dominant and missing marks in various crosses from two inbred lines. Genetica 101, 47–58.

Katul, G., Oren, R., Ellsworth, D., Hsieh, C.I., Phillips, N., 1997. A Lagrangian dispersion model for predicting CO_2 sources, sinks, and fluxes in a uniform loblolly pine (*Pinus taeda* L.) stand. J. Geophys. Res. Atmos. 102, 9309–9321.

Katul, G., Hsieh, C.I., Bowling, D., et al., 1999. Spatial variability of turbulent fluxes in the roughness sublayer of an even-aged pine forest. Bound. Layer Meteorol. 93, 1–28.

Katul, G., Lai, C.T., Schafer, K., Vidakovic, B., Albertson, J., Ellsworth, D., Oren, R., 2001. Multiscale analysis of vegetation surface fluxes: from seconds to years. Adv. Water Resour. 24, 1119–1132.

King, G., Honaker, J., Joseph, A., Scheve, K., 2001. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. Am. Pol. Sci. Rev. 95, 49–69.

Kneipp, S.M., McIntosh, M., 2001. Handling missing data in nursing research with multiple imputation. Nurs. Res. 50, 384–389.

Little, R.J.A., Rubin, D.B., 1989. The analysis of social science data with missing values. Sociol. Methods Res. 18, 292–326.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data, second ed. Wiley, New York.

Monson, R.K., Turnipseed, A.A., Sparks, J.P., Harley, P.C., Scott-Denton, L.E., Sparks, K., Huxman, T.E., 2002. Carbon sequestration in a high-elevation, subalpine forest. Global Change Biol. 8, 459–478.

Patrician, P.A., 2002. Focus on research methods: multiple imputation for missing data. Res. Nurs. Health 25, 76–84.

Pilegaard, K., Hummelshoj, P., Jensen, N.O., Chen, Z., 2001. Two years of continuous CO_2 eddy flux measurements over a Danish beech forest. Agric. Forest Meteorol. 107, 29–41.

Rubin, D.B., 1977. Formalizing subjective notions about the effect of nonrespondents in sample surveys. J. Am. Stat. Assoc. 72, 538–543.

Rubin, D.B., 1987. Multiple Imputation for Nonresponses in Surveys. Wiley, New York.

SAS Institute Inc., 2002. Online version of SAS document version 9. SAS Institute Inc., Cary, NC. http://v9doc.sas.com/sasdoc.

Schafer, J.L., 1997. Analysis of Incompelete Multivariate Data. Chapman & Hall, New York.

Schafer, J.L., Graham, J.W., 2002. Missing data: our view of the state of the art. Psychol. Methods 7, 147–177.

Schafer, J.L., Olsen, M.K., 1998. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivariate Behav. Res. 33, 545–571.

van Buuren, S., Boshuizen, H.C., Knook, D.L., 1999. Multiple imputation of missing blood pressure covariates in survival analysis. Stat. Med. 18, 681–694.

Wan, S., Hui, D., Luo, Y., 2001. Fire effects on nitrogen pools and dynamics in terrestrial ecosystems: a meta-analysis. Ecol. Appl. 11, 1349–1365.

Wilson, K.B., Baldocchi, D.D., 2000. Seasonal and interannual variability of energy fluxes over a broadleaved temperate deciduous forests in North America. Agric. Forest Meteorol. 100, 1–18.

Wilson, K.B., Baldocchi, D.D., 2001. Comparing independent estimates of carbon dioxide exchange over 5 years at a deciduous forest in the southeastern United States. J. Geophys. Res. 106 (D24), 34167–34178.

Zhou, X.H., Eckert, G.J., Tierney, W.M., 2001. Multiple imputation in public health research. Stat. Med. 20, 1541–1549.