ISEIS

Journal of Environmental Informatics 34(2) 68-87 (2019)

Journal of Environmental Informatics

www.iseis.org/jei

A Multiple Imputation Strategy for Eddy Covariance Data

D. Vitale[1*], M. Bilancia[2], and D. Papale[1]

[1]*Department for Innovation in Biological, Agro-food and Forest Systems (DIBAF), University of Tuscia, via San Camillo de Lellis, 01100 Viterbo, Italy*

[2]*Ionian Department of Law, Economics and Environment, University of Bari Aldo Moro, Via Lago Maggiore angolo Via Ancona, 74121 Taranto, Italy*

Received 23 February 2017; revised 14 September 2017; accepted 25 October 2017; published online 18 June 2018

ABSTRACT. Half-hourly time series of net ecosystem exchange (NEE) of CO_2, latent heat flux (LE) and sensible heat flux (H) measured through the micro-meteorological eddy covariance (EC) technique are noisy and show a high percentage of missing data. By using EC measurements that are part of the FLUXNET2015 dataset, we evaluate the performance of a multiple imputation (MI) strategy based on an efficient computational strategy introduced in Honaker and King (2010), combining the classic Expectation-Maximization (EM) algorithm with a bootstrap approach, in order to take draws from a suitable approximation of posterior distribution of model parameters. Armed with these instruments, we are able to introduce three new multiple imputation models, characterized by an increasing level of complexity, and built on top of multivariate normality assumption: 1) MLR, which imputes EC missing values using a static multiple linear regression of observed values of suitable input variables; 2) ADL, which enriches with dynamic properties the static specification of MLR, by considering an autoregressive distributed lag specification; 3) PADL, which adds further complexity by embedding the ADL model in a panel-data perspective. Under several artificial gap scenarios, we show that PADL has a better ability in modeling the complex dynamics of ecosystem fluxes and reconstructing missing data points, thus providing unbiased imputations and preserving the original sampling distribution. The added flexibility arising from the time series cross section structure of PADL warrants improved performances, outperforming those of other imputation methods, as well as of the marginal distribution sampling algorithm (MDS), a widely used gap-filling approach introduced by Reichstein et al. (2005), especially in the case of nighttime flux data. It is expected that the strategy proposed in this paper will become useful in creating multiple imputations for a variety of EC datasets, providing valid inferences for a broad range of scientific estimands (such as annual budgets).

Keywords: eddy covariance, net ecosystem exchange, carbon budget, missing data, multiple imputations, Expectation-Maximization (EM) algorithm, panel autoregressive distributed lag model (PADL).

1. Introduction

The major sinks of atmospheric carbon dioxide (CO_2) are represented by terrestrial ecosystems. Among these systems, forests sequester about one-third of the total anthropogenic emissions and play a major role in global carbon dynamics by exchanging trace gases between the atmosphere and the biosphere. A better understanding of the potentials of ecosystems to reduce the rise of atmospheric CO_2 levels, as well as a better ability to properly quantify the terrestrial carbon stocks and model the temporal and spatial variation in carbon uptake, are crucial in order to develop mitigation strategies in response to climatic changes.

For these reasons, an important research frontier in ecology is directed toward measuring exchange rates of trace gases over natural ecosystems and agricultural fields. The eddy covariance (EC) technique is nowadays the most reliable and direct method for this purpose (Aubinet et al., 2012), because it allows scientists to readily calculate the main fluxes of Net Ecosystem Exchange (NEE) of CO_2, Latent Heat (LE) and Sensible Heat (H) at ecosystem scale. In particular, NEE is expressed as the difference between the CO_2 assimilated by photosynthetic activities, and the CO_2 released to the atmosphere through ecosystem respiration processes.

Despite improved accuracy of measurement devices, EC datasets are characterized by a large amount of missing data. Breakdowns and damage of measurement instruments, wrong system calibrations and ordinary maintenance interventions are unavoidable events, resulting in the presence of gaps and missing data in the sequence of measurements over time. In addition, missing data are also caused by quality control (QC) procedures (Aubinet et al., 2012; See chapters 3-5 for examples of data-filtering procedures), which aim to discard bad data acquired under non-ideal conditions with respect to the characteristics of instruments and physical assumptions behind the EC technique (in particular, those assumptions related to well-developed and stationary turbulence regimes). To give an idea

* Corresponding author. Tel.: +39-(0)761-357044; fax: +39-(0)761-357389.
E-mail address: domvit@unitus.it (D. Vitale).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.201800391

of the scale and the importance of missing data issues, Falge et al. (2001) reported that the average NEE data coverage during a year, across 18 sites from the EUROFLUX project and the AmeriFlux network, was only 65% due to system failures or data rejection. Similarly, Papale et al. (2006) estimated that the percentage of half-hourly data rejected under different conditions varied from 20% to 60%, depending on the quality of the raw data as well as on the severity of QC procedures.

Several ad hoc gap-filling methods in EC measurements have been developed until today (Aubinet et al., 2012; see Chapter 6 and references therein). One of the earliest examples is reported in Hui et al. (2004), which propose a multiple imputation (MI) algorithm based on a multivariate normal (MVN) model. In synthesis, MI is a Monte Carlo simulation technique imputing missing values M times to obtain M multiple copies of a complete data set, and then combining parameter estimates for all M complete data analyses to have a single point estimate, with associated uncertainty properly reflecting the presence of missing data. On the contrary, single imputation (SI) yields a single value per missing datum. For example, mean substitution is a standard SI imputation method with which the missing values are imputed with the mean value based on the observed values. Also in this case, many SI gap-filling algorithms have been proposed to reconstruct the missing data in EC datasets. In a classic review paper, Moffat et al. (2007) provided an extensive comparison of both SI and MI selected techniques, by evaluating their performance for different artificial gap scenarios on a set of 10 benchmark datasets. According to the results of the simulation experiments, Moffat et al. (2007) concluded that SI methods, such as artificial neural network based techniques and marginal distribution sampling (MDS; Reichstein et al., 2005), generally showed a good overall performance, whereas the MI algorithm proposed by Hui et al. (2004) showed high biases and markedly underperformed in terms of NEE annual sum estimates.

This apparently awkward behavior can be explained by introducing the notion of *proper* MI algorithm (Rubin, 1987; van Buuren, 2012). Any complete data analysis procedure $\hat{Q}$ for estimating a scientific estimand Q is said to be *valid* if: 1) the average of the MI estimate $\hat{Q}$ over all possible complete samples Y is unbiased; 2) the actual coverage of the associated confidence intervals, based on estimated variances, equals (at least approximately) the nominal coverage (Rubin, 1996). Any MI procedure is said to be *proper* if we can convert an incomplete sample M times into a complete sample and compute M different point estimates for Q under the complete data analysis procedure, combining them according to rules introduced by Rubin, without introducing any further bias. When a MI procedure is proper, the imputation model preserves those aspects of the distribution that are relevant to the analysis model, and imputed values act like the observed values when used in the analysis stage, yielding valid inferences in the sense defined above. It is not always easy to check analytically whether a certain procedure is proper (sufficient conditions are given, for example, in van Buuren, 2012) and numerical experiments are often the only resort. Although 'crude' MVN modelling has proven to be useful even in some cases of violation of normality assumption, it was often unable to correctly reproduce the data generation process (DGP) of EC data, and produced low-quality imputations characterized by high out-of-sample performance. Proper imputation methods are, therefore, needed to guarantee that the estimates of interest will be unbiased in the presence of missing data.

In other words, we can say that when the goals of analysis are limited, a crude normal model can often be useful. To be more precise, a gap-filling algorithm is crude if no use is made of special time series characteristics, such as the presence of temporal autocorrelation, as well as of heavy-tailed and time varying random errors. However, EC data are a special challenge because of their complex stochastic structure (Richardson et al., 2012), so that broad imputation models are more likely to be ineffective and to produce biased inferences. It is thus sensible to design imputation algorithms especially tailored to the above-mentioned characteristics. With this goal in mind, in this paper we propose a data analysis strategy based on the algorithm recently proposed by Honaker and King (2010). Similar to Hui et al. (2004), the underlying imputation model assumes that the complete data likelihood is MVN. However, the unique computational strategy, henceforth labelled EMB, combines the classic Expectation-Maximization (EM) algorithm with a bootstrap approach, in order to draw simulated values from the approximate posterior distribution of parameters. The increased computational efficiency makes possible the implementation of suitable conditional MI models, where imputations of missing data in the flux time series of interest are typically based on the relationship between the incomplete variable and the observed part of some suitable input variables (predictors), possibly including deterministic polynomial functions of time, as well as lagged endogenous and exogenous variables, in order to enrich the generative process of complete data with dynamical characteristics. Armed with these ideas, we propose three new conditional MI models, based on the multivariate normality assumption: M1. MLR, a 'baseline' model, which imputes EC missing values using a static multiple linear regression of observed values of input variables; M2. ADL, which enriches with dynamic properties the static specification of MLR by considering an autoregressive distributed lag (ADL) specification; M3. PADL, which adds further complexity by embedding the ADL model in a panel-data perspective. Reproducibility of our results is greatly facilitated by the availability of the *Amelia* R package (Honaker et al., 2011), which provides an interface to the Amelia II program for MI imputation of incomplete datasets under the EMB approach outlined above.

The paper is organized as follows. Subsections 2.1 and 2.2 briefly review some missing data theory and terminology. Subsections 2.3 and 2.4 describe both EM- and data-augmentation algorithms behind the joint MVN imputation model. Section 2.5 introduces three new conditional imputation models based on the MVN complete data likelihood and, finally, Subsection 2.6 shows how imputations are combined to obtain a final estimate of scientific estimands of interest. Study sites, data collection and basic pre-processing tools are introduced in Section 3, where we also describe suitable in-sample indicators to evaluate the quality of imputed values from MI, as well as the design of

a simulation experiment to assess the out-of-sample performance. Section 4 presents a detailed multi-site comparison (including the analysis of some issues that can arise from computational difficulties). Finally, Section 5 refocuses on the purpose of the research, draws conclusions and proposes future developments.

2. Methods

In this section, we review some missing data terminology and the theory behind our algorithms; we refer the reader to existing literature on the topic for more details (Schafer, 1997; Schafer and Graham, 2002; Little and Rubin, 2002; van Buuren, 2012).

2.1. Definitions and Basic Notations

Let z denote the complete $T \times K$ data matrix, including K variables (NEE, LE, H and other micrometeorological and soil variables described in Section 3) at T equally spaced half-hourly timestamps. Let z be not fully observed and partitioned as $z = (z^{obs}, z^{mis})$, where z^{obs} denotes those entries actually observed and z^{mis} those missing. For any complete data set z we define a fully observed set of indicator variables R, referred to as the *missingness*, which is a matrix data structure of the same dimension of z, indicating whether the corresponding measurement is observed $(R = 1)$ or missing $(R = 0)$. Missingness is conveniently described as a probabilistic phenomenon, as it is not realistic to describe accurately all potential causes for missing data.

The probability distribution of missingness (or *missing data model*), say $p(R \mid z)$, can depend on either the observed or the missing data, and can be classified according to the nature of relationship between the missingness itself and the observed data. If $p(R \mid z) = p(R \mid z^{obs})$ the missing data are defined to be *missing at random* (MAR). In other words, MAR allows probabilities of missingness to depend on observed data but not on missing data. This terminology is unfortunate and particularly confusing, as MAR does not actually indicates that missing data are distributed at random. If the data are MAR, the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data (e.g. NEE is missing when the friction velocity is below some threshold value; see Papale et al., 2006). If $p(R \mid z) = p(R)$ the missing data are said to be *missing completely at random* (MCAR). MCAR is a special case of MAR, occurring when the distribution of missingness does not depend on observed data either. The underlying idea is more appealing in this case, as the missing data are just a random subset of the complete data. Finally, when the distribution of missingness also depends on unobserved information, that is on z^{mis}, then the missing data are said to be missing not at random (MNAR). For example, MNAR occurs when missingness depends on the missing value itself (e.g. NEE is missing if NEE is greater than a given threshold, expressed in μmol CO_2 $m^{-2}s^{-1}$) or when missingness depends on an unobserved variable. Among these possibilities, at least MAR assumption is required for missingness to be ignorable, in the sense that we can draw valid inferences without knowing the process that generates the missingness (Rubin, 1987). Unfortunately, MAR assumption is not testable, being a condition depending on unobservable data. In Appendix B, we discuss an experimental arrangement under which MAR can reasonably hold (or, at least, that a non-ingnorable missingness pattern is less likely).

2.2. Multiple Imputation Under a Joint MVN Model

In what follows, we assume that the complete dataset z is a random sample of size T from a K-dimensional MVN distribution $N_K(.\mid \mu, \Sigma)$, with K-dimensional mean vector μ and positive definite covariance matrix Σ, where off-diagonal elements of Σ allow marginal components to depend on each other. For $t = 1, \ldots, T$, let z_t denotes a generic row of the complete data matrix z. Under the assumption of multivariate normality, the complete data likelihood has the following form:

$$L(\theta \mid z^{obs}, z^{mis}) = \prod_{t=1}^{T} N_K(z_t^{obs}, z_t^{mis} \mid \theta), \quad \text{with } \theta = (\mu, \Sigma) \tag{1}$$

This exchangeable assumption would seem a crude approximation to the true distribution of the data. However, as will be explained later, with a careful choice of the K variables entering the complete-data likelihood, we can design relatively narrow-scope imputation methods such that the imputed values act like the observed values when used in the analysis stage.

As we said before, following arguments reported in Schafer (1997), Schafer and Graham (2002), it can be shown that under MAR assumption posterior Bayesian inference about θ can be performed without regard for the missing-data mechanism (provided that θ and nuisance parameters pertaining to the probability distribution of R are independent a-priori), and all information about θ is summarized in the observed-data posterior:

$$p(\theta \mid z^{obs}) \propto L(z^{obs} \mid \theta)\pi(\theta) \tag{2}$$

where $\pi(\theta)$ is the prior distribution over model parameters θ, and $L(z^{obs} \mid \theta)$ is the observed data likelihood:

$$L(\theta \mid z^{obs}) \propto \int \left[\prod_{t=1}^{T} N_K(z_t^{obs}, z_t^{mis} \mid \theta) \right] dz_t^{mis} \tag{3}$$

Under a flat non-informative prior distribution, $\pi(\theta) \propto 1$, the observed data posterior reduces to the observed data likelihood, and posterior and likelihood inference turn out to be equivalent. As in Hui et al. (2004), assumption (1) is the basis for drawing from the complete data posterior, and making imputations by drawing value of z^{mis} from its distribution conditional on z^{obs} and the draws of θ (to account for estimation uncertainty). The naïve algorithm runs as follows:

- Use the MVN distribution as an approximation to the joint posterior distribution of model parameters $\theta = (\mu, \Sigma)$ under a flat prior, and find approximate posterior modes by maximum likelihood estimates of mean vector μ and covariance matrix Σ using the EM algorithm for incomplete data.

- Draw imputations for the missing values from the normal model using a Gibbs sampling over the joint posterior distribution of model parameters and missing values, using the EM modal estimates from the previous step as a starting point for the sampler.

In what follows we briefly review the algorithms, in order to better appreciate the characteristics of the approach proposed in Hui et al. (2004). Those readers uninterested in the statistical details may safely skip ahead to Subsection 2.5.

2.3. The EM Algorithm for Incomplete Data

The EM algorithm for maximum likelihood estimation from incomplete data has been originally proposed by Dempster et al. (1977). Here, we present a modern 'variational' exposition, closely following Tzikas et al. (2008). A convenient starting point is the following decomposition of the observed data log-likelihood (Blei et al., 2017):

$$\ln p(z^{\text{obs}} \mid \theta) = F(q,\theta) + \text{KL}(p \,\|\, q) \tag{4}$$

where

$$F(q,\theta) = \int q(z^{\text{mis}}) \ln\left(\frac{p(z^{\text{obs}}, z^{\text{mis}} \mid \theta)}{q(z^{\text{mis}})} \right) dz^{\text{mis}} \tag{5}$$

where $q(z^{\text{mis}})$ denotes any probability density function over the missing data, while $\text{KL}(p \,\|\, q)$ is the Kullback-Leibler divergence between the posterior distribution of missing data and $q(z^{\text{mis}})$:

$$\text{KL}(p \,\|\, q) = -\int q(z^{\text{mis}}) \ln\left(\frac{p(z^{\text{mis}} \mid z^{\text{obs}}, \theta)}{q(z^{\text{mis}})} \right) dz^{\text{obs}} \tag{6}$$

It is important to note that $\text{KL}(p \,\|\, q) \geq 0$, with $\text{KL}(p\|q) = 0$ only for the special choice $q(z^{\text{mis}}) \equiv q_0(z^{\text{mis}}) = p(z^{\text{mis}} \mid z^{\text{obs}}, \theta)$. Moreover, from $\text{KL}(p \,\|\, q) \geq 0$ it follows that $\ln p(z^{\text{obs}} \mid \theta) \geq F(q,\theta)$. In other words, $F(q,\theta)$ is a lower bound of the observed data log-likelihood. Based on this result, the EM algorithm can be presented as a two-step iterative algorithm that maximizes the observed data log-likelihood by maximizing the lower bound $F(q,\theta)$. To verify this claim, we assume that the current state of the parameter vector is $\theta^{(s)}$. In the E-step, the lower bound $F(q,\theta^{(s)})$ is maximized with respect to $q(z^{\text{mis}})$, and it is straightforward to verify that this occurs when $\text{KL}(p \,\|\, q) = 0$ or, equivalently, when $q(z^{\text{mis}})$ is set equal to $q_0(z^{\text{mis}}) = p(z^{\text{mis}} \mid z^{\text{obs}}, \theta^{(s)})$. In the M-step, $q_0(z^{\text{mis}})$ is held fixed and $F(q_0,\theta)$ is maximized with respect to θ to give some updated value $\theta^{(s+1)}$. We can summarize these two steps by the following chain of inequalities, showing that the observed data log-likelihood increases monotonically to a local maximum:

$$\begin{aligned} &\ln p(z^{\text{obs}} \mid \theta^{(s+1)}) \geq^{\text{M-step}} F(q_o, \theta^{(s+1)}) \geq \\ &\geq^{\text{E-step}} F(q_0, \theta^{(s)}) = \ln p(z^{obs} \mid \theta^{(s)}) \end{aligned} \tag{7}$$

Sufficient conditions to ensure that the algorithm converges to a global maximum rather than to a stationary point are given by Wu (1983), and these conditions are known to hold under our MVN likelihood. It is worth noting that if we substitute $q_0(z^{\text{mis}}) = p(z^{\text{mis}} \mid z^{\text{obs}}, \theta^{(s)})$ into $F(q,\theta)$ we obtain:

$$\begin{aligned} F(q_0,\theta) &= \int p(z^{\text{mis}} \mid z^{\text{obs}}, \theta^{(s)}) \ln p(z^{\text{obs}}, z^{\text{mis}} \mid \theta) dz^{\text{mis}} - \\ &- \int p(z^{\text{mis}} \mid z^{\text{obs}}, \theta^{(s)}) \ln p(z^{\text{mis}} \mid z^{\text{obs}}, \theta^{(s)}) dz^{\text{mis}} = \\ &= Q\left(\theta \mid \theta^{(s)}\right) - \text{constant with respect to } \theta \end{aligned} \tag{8}$$

Therefore, the EM can be summarized as an iterative algorithm involving the following two steps:

- E-step: Compute the posterior predictive distribution of the missing data $p(z^{\text{mis}} \mid z^{\text{obs}}, \theta^{(s)})$ and $Q(\theta \mid \theta^{(s)})$.
- M-step: Update the current guess of model parameters by $\theta^{(s+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(s)})$.

The E-step and the M-step are repeated alternately until $\ln p(z^{\text{obs}} \mid \theta^{(s+1)}) - \ln p(z^{\text{obs}} \mid \theta^{(s)}) < \delta$, where δ is a pre-assigned tolerance. Calculations under the complete data MVN likelyhood (1) are straightforward. Explicit expressions of $p(z^{\text{mis}} \mid z^{\text{obs}}, \theta^{(s)})$, $Q(\theta \mid \theta^{(s)})$ and $\theta^{(s+1)}$ are provided, for example, by Hui et al. (2004), Section 2.1, and Gelman et al. (2013, pp. 454).

2.4. Filling in Missing Data with Gibbs Sampling

To fill in missing values z^{mis} we can exploit the posterior predictive distribution:

$$p(z^{\text{mis}} \mid z^{\text{obs}}) = \int p(z^{\text{mis}} \mid z^{\text{obs}}, \theta) p(\theta \mid z^{\text{obs}}) d\theta \tag{9}$$

Sampling from the posterior predictive distribution (9) is accomplished using a data augmentation (DA) algorithm (Tanner and Wong, 1987) over the augmented parameter space (z^{mis}, θ), treating missing values as latent variables. This algorithm consists of alternately drawing z^{mis} and θ from their conditional posterior distributions, which both have closed form under multivariate normality (Schafer, 1997):

- I-step: draw $z^{\text{mis},(g+1)}$ from $p(z^{\text{mis}} \mid z^{\text{obs}}, \theta^{(g)})$.
- P-step: draw $\theta^{(g+1)}$ from $p(\theta \mid z^{\text{obs}}, z^{\text{mis},(g+1)})$.

This iterative algorithm is a Gibbs sampler that generates a Markov chain for $g = 1,\ldots,G$, converging in distribution to the joint posterior of parameters and missing values, $p(z^{\text{mis}}, \theta \mid z^{\text{obs}})$, after a transient burn-in period (Gelman et al., 2013). In particular, the MI method proposed by Hui et al. (2004) uses precisely the sampling-based algorithm described above. As we said before, modal estimates of θ outputted by the EM algorithm are used as a starting point for the first iteration of the I-step. Missing values are imputed by storing, after the burn-in period, M draws from $p(z^{\text{mis}} \mid z^{\text{obs}}, \theta)$, thinning the sample using only every n^{th} step to reduce the strong autocorrelation usually present in the Gibbs sampler output. Alternatively, at the computational cost of running $M-1$ additional independent

chains, the DA algorithm can be run M times in parallel to draw a single imputation from every chain.

2.5. The Proposed Imputation Models

A first concern of the above described approach is the efficiency of the algorithm. The Gibbs sampler can converge very slowly to its limiting stationary distribution in the event that the true posterior of model parameters describes a highly correlated multidimensional random variable. In addition, building M complete data sets involve running M simulations, thus generating a large amount of random draws, most of which are discarded and not needed anymore in any subsequent analyses.

To overcome the computational overloading of Gibbs sampling, Honaker and King (2010) propose a computational strategy combining the Expectation-Maximization algorithm with Bootstrap, henceforth labelled EMB (see also Schomaker and Heumann, 2016). The combined EMB algorithm runs as follows:

- Bootstrap-step: use non-parametric bootstrap to draw M sample with replacement (including missing data) of dimension $T \times K$ from z.
- EM-step: for each bootstrap sample, find approximate posterior modes by maximum likelihood (ML) estimates of μ and Σ, using the EM algorithm for incomplete data under a flat prior of model parameters, as described before.

Bootstrapping the complete data matrix is aimed at simulating estimation uncertainty, and it is carried out by considering each row as one multivariate observation. When incomplete data are resampled, each bootstrap sample has high probability of being incomplete, and thus posterior estimates of θ can be approximated using the EM algorithm under incomplete data, as seen before. Interestingly, Efron (2012) points out that ML estimates from the bootstrap samples are asymptotically equivalent to a sample from the posterior distribution of θ (also in the case when $\pi(\theta)$ is not flat), thus propagating correctly uncertainty in estimating θ.

Once estimates of complete data parameters are available, imputations of missing values are drawn, as shown below, conditional on observed values of explanatory variables and each of the M estimates of θ. The idea of substituting ML estimates $\tilde{\theta}^{(m)} = (\tilde{\mu}^{(m)}, \tilde{\Sigma}^{(m)})$ from M bootstrapped samples (for $m = 1, \ldots, M$) to draw missing values from the approximate posterior predictive distribution $p(z^{\text{mis}} \mid z^{\text{obs}}, \tilde{\theta}^{(m)})$ dates back to Efron (1994). To put this idea to work with EC flux data, we partitioned complete data vectors as $z_t = (y_t, x_t)$ for $t = 1, \ldots, T$ where y_t denotes the dependent variable, while x_t includes explanatory variables to be used in the gap-filling phase. In most case studies, the dependent variable y_t is one among NEE, LE or H. Inputs x_t of the imputation phase can include the remainning fluxes and soil and micro-meteorological variables, as well as suitably defined additional synthetic inputs aiming at improving basic imputation models, in order to reflect the special nature of time series data. For example, we can include the information that some variables have smooth trends by supplementing z with new input variables (columns) contructed prior to running the algorithm, based on q-order polynomials or function bases, such as splines or wavelets (which have good approximation capabilities for any functional form of t; Hastie et al., 2009). Another way of handling time series information is to include lagged variables (Honaker and King, 2010; Honaker et al., 2011).

Armed with this machinery, we consider three novel imputation models which are able (with different degrees of effectiveness) to accommodate for some dynamic characteristics of EC flux data. The final objective is to provide proper MI gap-filling procedures, thus reducing biases occurring with the crude MVN-based algorithm proposed by Hui et al. (2004).

MLR

The first imputation model is a static multiple linear regression (labelled as MLR) of input variables, conditional on the observed part, with parameters that can be calculated directly from θ. It is a commonly accepted practice to estimate separate imputation models from the qualitatively different daytime and nighttime data subsets (Moffat et al., 2007; See also Appendix B), a difference due to variation in fluxes in response to changes in meteorological conditions, often leading to very different performance of gap-filling techniques. Moreover, as will be better explained in Appendix B, conducting separate analysis for daytime and nighttime has a regularizing effect over the MI procedure, in the sense that MAR hypothesis is more likely to hold. We have, therefore, the following switching regression model (daytime ≡ diurnal regime, where assimilation processes are prevalent; nighttime ≡ nocturnal regime, where respiration processes are prevalent):

$$\tilde{y}_t^{(m)} = \begin{cases} x_t^{\text{obs}} \tilde{\beta}_1^{(m)} + \tilde{\varepsilon}_{1t}^{(m)} & t \in \text{daytime} \\ x_t^{\text{obs}} \tilde{\beta}_2^{(m)} + \tilde{\varepsilon}_{2t}^{(m)} & t \in \text{nighttime} \end{cases} \tag{10}$$

where $\tilde{y}_t^{(m)}$ indicates the mth imputed value, t indicates half-hourly timestamps, $\tilde{\beta}_1^{(m)}$ and $\tilde{\beta}_2^{(m)}$ are vectors with $K-1$ elements. Under the MVN likelihood (1), the distribution of y_t conditional on x_t^{obs} is Gaussian, with conditional expectation that can be expressed as a linear function of parameters $\theta = (\mu, \Sigma)$. This fact justifies the linear specification of the conditional imputation model (10). Random draws $\tilde{\beta}_i^{(m)}$ of the regression coefficient vector can therefore be calculated directly by boots-trapped ML estimates $\tilde{\mu}_i^{(m)}$ and $\tilde{\Sigma}_i^{(m)}$ (with $i = 1, 2$ corresponding to daytime and nighttime estimates, respectively; see, e.g., Honaker and King, 2010, pp. 576), based on standard expressions of conditional distributions of a multivariate normal distribution. Similarly, random error $\tilde{\varepsilon}_{it}^{(m)}$ is a normal random variable with zero mean and variance equal to the corresponding diagonal element of $\tilde{\Sigma}_i^{(m)}$.

It is worth noting that randomness in $\tilde{y}_t^{(m)}$ is generated by estimation uncertainty due to not knowing parameters, as well as by irreducible uncertainty in the DGP, since the diagonal elements of $\tilde{\Sigma}_i^{(m)}$ are not null. This means that even if we had an infinite sample, thus replacing $\tilde{\beta}_i^{(m)}$ with 'true' value, there would still be a source of uncertainty taken into account by drawing from the distribution of $\tilde{\varepsilon}_{it}^{(m)}$. However, MLR model

defined by equations (10) is *static*, in the sense that input variables have an instantaneous impact on the imputed flux. In agreement with Hui et al. (2004), it must be stressed that MLR model (10) is not estimated by a full data set spanning a single calendar year, but separate models are fitted for each hydro-ecological regime identified on the basis of a data-driven procedure described in Appendix A.

It is important to note that MLR can be considered as a 'baseline' conditional MI model, forming a basis for comparesons with other more structured models. Apart of its computational efficiency, no special time series characteristics are taken into account by generative equations (10). In other words, MLR is essentially equivalent to crude joint MVN model, and quality of MIs is not expected to improve dramatically.

ADL

In order to add dynamic features to the static specification outlined above, we propose a narrower conditionally Gaussian linear imputation model, based on a first-order trend-stationary switching autoregressive distributed lag specification (labelled as ADL):

$$\tilde{y}_t^{(m)} = \begin{cases} \tilde{c}_{10}^{(m)} + \tilde{c}_{11}^{(m)}\mathrm{DoR}(t) + \tilde{c}_{12}^{(m)}\mathrm{DoR}(t)^2 + \\ \quad + \tilde{c}_{13}^{(m)}\mathrm{DoR}(t)^3 + x_t^{\mathrm{obs}}\tilde{\beta}_1^{(m)} + \\ \quad + \tilde{y}_{t-1}^{(m)}\tilde{\xi}_1^{(m)} + x_{t-1}^{\mathrm{obs}}\tilde{\gamma}_1^{(m)} + \tilde{\varepsilon}_{1t}^{(m)} & t \in \text{daytime} \\ \tilde{c}_{20}^{(m)} + x_t^{\mathrm{obs}}\tilde{\beta}_2^{(m)} + \tilde{y}_{t-1}^{(m)}\tilde{\xi}_2^{(m)} + \\ \quad + x_{t-1}^{\mathrm{obs}}\tilde{\gamma}_2^{(m)} + \tilde{\varepsilon}_{2t}^{(m)} & t \in \text{nighttime} \end{cases} \quad (11)$$

In dynamic equations (11), t indicates half-hourly timestamps ($t = 1,\ldots,T$, where T is the total sample size) and $\mathrm{DoR}(t)$ is the corresponding day of the regime which is being reconstructed (thus $\mathrm{DoR}(t)$ ranges from 1 to T_R, where T_R is the total number of days in the regime, and remains constant throughout each calendar day), $\tilde{\xi}_i^{(m)}$ is a scalar, $\tilde{\beta}_i^{(m)}$ and $\tilde{\gamma}_i^{(m)}$ are vectors with $K-1$ elements. Also in this case, the ADL specification (11) was separately estimated for each of regime detected following the procedure described in Appendix A. For daytime subsets, the deterministic term included in the ADL model corresponds to a cubic function of time. A cubic trend is flexible enough to capture, at a modest computational cost, the medium-long term component of diurnal flux time series. During nighttime, the high percentage of missing data, the low signal-to-noise ratio and the presence of spikes (which is attributable, in most cases, to the presence of extreme values rather than measurement errors) limited the use of deterministic terms to a simple intercept term. Note that this choice does not prevent the possibility of modelling the trend component during nighttime periods, but means that any temporal dynamic in EC flux variables is fully driven by the relationships with other meteorological factors, if any.

It is interesting to study the characteristics of the stochastic part of model (11) for fixed m (that is, for a single imputation). Therefore, for notation simplicity, from now on until the end of this subsection, we can suppress both indices m and i (daytime/nighttime indicator). Stable dynamic behaviour requires that input variables x_t are second-order stationary and $\tilde{\varepsilon}_t$ is an uncorrelated White Noise, and finally that $|\tilde{\xi}|<1$ (Zivot and Wang, 2006; Hassler and Wolters, 2006). It is also useful to observe that the ADL model (11) is quite general, as it encompasses as a special case several other basic models. By momentarily neglecting deterministic terms, a static regression with independent and identically distributed (IID) errors is obtained when $\tilde{\xi} = \tilde{\gamma} = 0$, corresponding to (10); A static regression with stable AR(1) disturbances is obtained when $\tilde{\gamma} = -\tilde{\xi}\tilde{\beta}$; The AR(1) model corresponds to $\tilde{\beta} = \tilde{\gamma} = 0$; A first-difference model is obtained when $\tilde{\xi} = 1$ and $\tilde{\beta} = -\tilde{\gamma}$. Finally, it can be shown that, after some algebraic manipulation, the ADL model can be re-parameterized as an error-correction model (Hassler and Wolters, 2006).

The lacking of dynamic properties of MLR model (10) can be conveniently described in terms of the *immediate impact multiplier*, $\partial\tilde{y}_t / \partial x_t^j$, consisting of a spike of length $\tilde{\beta}^j$ associated with the instantaneous change of $\tilde{y}_t$ in response to a unit change in one component of x_t, say x_t^j (we use superscripts to denote components of vectors, such as $\tilde{\beta}$). The remaining lagged multipliers $\partial\tilde{y}_t / \partial x_{t-k}^j$ are null for $k > 1$, and the immediate impact thus coincides with the long-run effect. Similarly, the immediate impact multiplier of ADL shows a spike of length $\tilde{\beta}^j$. However, lagged multipliers are not null, as the impact of x_{t-1}^j on $\tilde{y}_t$ is $\tilde{\beta}^j\tilde{\xi} + \tilde{\gamma}^j$, and thereafter the equilibrium is progressively restored toward a new long-run level, as the effect of x_{t-k}^j on $\tilde{y}_t$ dies out geometrically at rate $\tilde{\xi}$ as $k \to +\infty$. This nice transient dynamical feature may be not enough to capture special characteristics of EC time series data, but it does represent a marked improvement over static imputation models such as MLR (10).

PADL

With the aim of allowing more flexibility in modelling the complex diurnal cycle, we consider a panel data perspective sharing many similarities with the approach proposed in Huisman et al. (2007), which introduces a panel model for hourly electricity prices in day-ahead markets and examines their characteristics. By taking this approach, we have S consecutive cross-sectional units ($h = 1,\ldots,S$), entering the following panel autoregressive distributed lag model (hereafter PADL; Beck and Katz, 1995; Hsiao, 2007):

$$\tilde{y}_{(h)t}^{(m)} = \begin{cases} \tilde{c}_{(h)10}^{(m)} + \tilde{c}_{(h)11}^{(m)}\mathrm{DoR}(t) + \tilde{c}_{(h)12}^{(m)}\mathrm{DoR}(t)^2 \\ \quad + \tilde{c}_{(h)13}^{(m)}\mathrm{DoR}(t)^3 + x_{(h)t}^{\mathrm{obs}}\tilde{\beta}_1^{(m)} + \\ \quad + \tilde{y}_{(h)t-1}^{(m)}\tilde{\xi}_1^{(m)} + x_{(h)t-1}^{\mathrm{obs}}\tilde{\gamma}_1^{(m)} + \tilde{\varepsilon}_{(h)1t}^{(m)} & t \in \text{daytime} \\ \tilde{c}_{(h)20}^{(m)} + x_{(h)t}^{\mathrm{obs}}\tilde{\beta}_2^{(m)} + \tilde{y}_{(h)t-1}^{(m)}\tilde{\xi}_2^{(m)} + \\ \quad + x_{(h)t-1}^{\mathrm{obs}}\tilde{\gamma}_2^{(m)} + \tilde{\varepsilon}_{(h)2t}^{(m)} & t \in \text{nighttime} \end{cases} \quad (12)$$

It is unusual, in standard panel data econometric theory,

that $T >> S$. However, under this condition, it is possible to estimate a separate linear imputation model for each cross-sectional unit, which is indeed not possible in the small T case, and it becomes natural to consider heterogeneous panel models where the parameters can differ over units (Smith and Fuertes, 2016). The data structure implied by the model (12) can be better appreciated by representing, for the first day of the year, a generic flux time series y_t in a panel data form $y_{(h)t}$ (we consider $S = 12$ in this example, thus each row corresponds to a two-hour interval):

		00 : 30	01 : 00	01 : 30
00 : 00	$y_{(1)1}$	$y_{(1)2}$	$y_{(1)3}$	$y_{(1)4}$
		02:30	03:00	03:30
02 : 00	$y_{(2)5}$	$y_{(2)6}$	$y_{(2)7}$	$y_{(2)8}$
	⋮	⋮	⋮	⋮
	$y_{(h)t}$	$y_{(h)t+1}$	$y_{(h)t+2}$	$y_{(h)t+3}$
	⋮	⋮	⋮	⋮
		22 : 30	23 : 00	23 : 30
22 : 00	$y_{(12)45}$	$y_{(12)46}$	$y_{(12)47}$	$y_{(12)48}$

Implicitly, when timestamp t indicates the first column, $y_{(h)t}$ is regressed over $y_{(h-1)t-1}$ and $x_{(h-1)t-1}$, so that each cross-sectional unit is augmented with the last observation of the preceding unit. Similar to the ADL model (11), the effects of explanatory variables are assumed to be stable inside each regime, and for each regime they do not vary across cross-section units and over time. On the contrary, deterministic terms are stable inside each regime, but are allowed to change across cross-section units (adding more flexibility). In particular, for nighttime data we have a *fixed effect* model where every cross-sectional unit has its own estimated constant term (also in this case we considered only a fixed constant for the nighttime imputation model, because numerical stability issues arising with the MLR model can be even more severe in this case), while for daytime data a cubic function of time was included. By construction, error terms are IID over *h* and within *t*, and are independent of input variables. Panel data contain many degrees of freedom and more sample variability than time series data, which is a panel with $S = 1$. These unique characteristics are expected to further improve the quality of imputations.

2.6. Annual Budget Estimation

One of the most used scientific estimand of interest is the annual sum (or *annual budget*) of NEE, LE and H fluxes, each one of the three being denoted by *Q*. With *M* complete data sets, we can compute *M* different point estimates for *Q* and combine them according to Rubin's rules, to obtain valid inferences when the MI procedure is proper. Specifically, let $\tilde{Q}^{(m)}$ the cumulative annual sum, i.e. the sum over all half-hourly measured and gap-filled values in a given year from the *m*th imputed data set, $m = 1, \ldots, M$. The final combined estimate is defined as (Rubin, 1987):

$$\tilde{Q} = \frac{1}{M}\sum_{m=1}^{M}\tilde{Q}^{(m)} \tag{13}$$

The estimate $\tilde{V}$ of the variance of $\tilde{Q}$ can be obtained by combining a within component term $\bar{U}$ and a between component term *B*. The within term accounts for sample variability, and it is the average of the variance estimates $\hat{U}^{(m)}$ for complete data, for $m = 1, \ldots, M$, that is $\bar{U} = M^{-1}\sum_{m=1}^{M}\hat{U}^{(m)}$. The between term $B = (M-1)^{-1}\sum_{m=1}^{M}(\hat{Q}^{(m)} - \tilde{Q})^2$ measures the uncertainty due to imputations. The total variance of $\tilde{Q}$ is thus estimated as $\tilde{V} = \bar{U} + (1 + M^{-1})B$.

Assuming that, under repeated sampling, parameter estimates $\tilde{Q}$ are normally distributed around the population value (that is, $\tilde{Q}$ is unbiased for *Q*), it follows that $(Q - \tilde{Q})/\tilde{V}^{1/2} \sim t_\nu$, where $\nu = (M-1)(1 + r^{-2})$ indicates the degrees of freedom (DOF) and $r = (1 + M^{-1})B/\bar{U}$ represents the relative increase in variance due to missing values (van Buuren, 2012; see also Barnard and Rubin, 1999, for an adjusted version of ν, valid when the complete data DOF is small and the percentage of missing data is not too high). Thus $\tilde{Q} \pm t_{\nu,1-\alpha/2}\tilde{V}^{1/2}$ provides the $100(1-\alpha)\%$ confidence interval of the annual budget *Q*. Finally, it is worth mentioning the fraction ρ of information about *Q* missing due to nonresponse (sometimes simply referred to as 'fraction of missing information'), defined as $\rho = \left((r+2)/(\nu+3)\right)/(1+r)$. If $\rho > 0.5$, statistical inferences are highly dependent on the way in which the missing data were handled, and the influence of the imputation model is much larger than that of the complete data model (van Buuren, 2012, pp. 41-42).

3. Data, Simulation Design and Evaluation Criteria

3.1. Eddy-Covariance Study Sites

Data used in this work are part of the FLUXNET2015 dataset, and subject to a highly standardized data pre-processing and QC pipeline, that generates uniform and high quality derived data products suitable for studies requiring inter-comparability of data from multiple sites (the interested reader can consult the documentation reported at:

http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/data-processing/, and references therein).

Ten benchmark sites were selected, with a brief description of site characteristics being shown in Table 1 (AT-Neu: Wohlfahrt et al., 2008. AU-Cpr: Meyer et al., 2015. AU-How: Beringer et al., 2007. DK-Sor: Pilegaard et al., 2011. FI-Hyy: Suni et al., 2003. FR-Pue: Rambal et al., 2004. GF-Guy: Bonal et al., 2008. IT-CA1: Sabbatini et al., 2016. US-Los: Sulman et al., 2009. US-Ne2: Verma et al., 2005). The site selection was done to ensure the representativeness of different climates and ecosystem types. The years under investigation were selected to guarantee the highest coverage of NEE time series in order to facilitate the simulation of the macro-scenarios described in Section 3.2.

On average, the missing data percentages of NEE, LE and H fluxes were about 50, 19 and 18%, respectively. Regarding NEE, the percentage of missing values was higher during nighttime (~60%) than daytime (~30%) because low turbu-

lence conditions occur most often during these periods, and QC procedures aimed to verify the validity of EC assumptions (e.g. u* filtering) flag and discard a large amount of data.

Several micrometeorological variables and soil parameters (abiotic factors) were taken into consideration: shortwave (SW_IN, Wm^{-2}) and longwave (LW_IN, Wm^{-2}) incoming radiation, net radiation (NETRAD, Wm^{-2}), air (TA, °C) and soil temperature (TS, °C), relative humidity, (RH, %), soil water content (SWC, %), friction velocity (USTAR, ms^{-1}), wind speed (WS, ms^{-1}) and precipitation (P, mm). Vapour pressure deficit (VPD, hPa) was calculated from TA and RH. As a consequence of system maintenance and data rejection after QC post-processing, most of abiotic parameters showed a percentage of missing data less than 5%. Larger percentages were commonly due to instrument breakdowns resulting in long gaps.

As an example, Figure 1 depicts half-hourly time series for the IT-CA1 use case. The IT-CA1 site was a 2-year rotation-cycle-managed poplar plantation of 11 ha (Poplar cultivar was *Populus x Canadensis*, for more detail see Sabbatini et al., 2016) cultivated in the *Gisella ed Elena Ascenzi S.A.S.* private farm, located in Castel d'Asso (Viterbo, Italy, lat: 42°38' N, lon: 12°03' E). The climate is Mediterranean, with mild winters and hot-dry summers, which is responsible for water stress conditions, clearly reflected in both NEE and LE flux dynamics.

3.2. Simulation Design and Performance Measures

Evaluating the quality of an imputation strategy is not an easy task. An informal approach to evaluating whether MI's can provide valid statistical inferences (in the sense discussed in the introductory section) can be based on the overimputation procedure discussed in Honaker et al. (2011), which involves treating observed values as if they had been actually missing. For each observed value several hundred imputed values are generated, a large number that allows us to calculate a mean imputation and construct a confidence interval of imputed values, given the imputation model. In particular: (i) for each imputation model, the averaged in-sample bias error (BE) can be computed, defined as the average difference between observed and mean imputed values. Inference is considered valid when BE is close to zero, and thus bias is negligible; (ii) in the same way, the mean absolute error (MAE) defined as the average of the absolute differences between observed and mean imputed values can be used to evaluate the in-sample model performance. Another valid measure useful for judging the quality of an imputation model is (iii) the *coverage rate* (CR), defined as the percentage of cases where the observed value falls within the 95% confidence limits. Honaker et al. (2011) recommend that CR should be around 90%. Finally, (iv) the average confidence interval width (W) is defined as the average length of the

Table 1. Information about the ten selected FLUXNET benchmark sites (Latitude, Longitude, IGBP designation, climate classification), percentages of missing data in Net Ecosystem Exchange (NEE) of CO_2 time series during the whole year (Y), and separately for daytime (D) and nighttime (N) subsets, and estimates of the friction velocity threshold value (u*th).

						Missing (%)				
Site ID	Country	Lat	Alt	IGBP	Clim	Y	D	N	u*th	Ref
Year	Location	Long	m asl	(a)	(b)	(c)	(c)	(c)	ms^{-1}	
AT-Neu 2010	Austria, Neustift	47.12°E, 11.32°N	970	GRA	Dfb	69	29	49	0.092	Wohlfart et al. 2008
AU-Cpr 2012	Australia, Calperum	34.00°W, 140.59°N	53	SAV	BSk	39	20	57	0.216	Meyer et al. 2015
AU-How 2011	Australia, Howard Springs	12.49°W, 131.15°N	na	WSA	Aw	58	34	83	0.222	Beringer et al. 2007
DK-Sor 2009	Denmark, Soroe	55.49°E, 11.64°N	40	DBF	Cfb	23	14	31	0.255	Pilegaard et al. 2009
FI-Hyy 2007	Finland, Hyytiala	61.85°E, 24.30°N	181	ENF	Dfc	56	49	62	0.406	Suni et al. 2003
FR-Pue 2008	France, Puechabon	43.74°E, 3.60°N	270	EBF	Csb	56	44	68	0.296	Rambal et al. 2004
GF-Guy 2008	French Guayana Guayaflux	5.28°E, 52.92°S	48	EBF	Af	53	35	70	0.160	Bonal et al. 2008
IT-CA1 2012	Italy, Castel d'Asso	42.38°E, 12.03°N	200	DBF	Csa	61	45	78	0.180	Sabbatini et al. 2016
US-Los 2006	USA, Lost Creek	46.08°E, 89.98°S	480	WET	Dfb	40	21	60	0.134	Sulman et al. 2009
US-Ne2 2012	USA, Lincoln (NE)	41.16°E, 96.47°S	362	CRO	Dfa	38	21	56	0.114	Verma et al. 2005

(a) International Geosphere-Biosphere Programme (IGBP) designations. CRO: Croplands; DBF: Deciduous Broadleaf Forests; EBF: Evergreen Broadleaf Forests; ENF: Evergreen Needleleaf Forests; GRA: Grasslands; SAV: Savannas; WET: Permanent Wetlands; WSA: Woody Savannas.
(b) Köppen climate classification (Clim). Af: Tropical, Rainforest; Aw: Tropical, Savanna; BSk: Arid, Steppe, Cold; Cfb: Temperate without dry season and warm summer; Csb: Temperate with dry and warm summer; Csa: Temperate with dry and hot summer; Dfa: Cold (continental) with hot summer; Dfb: Cold (continental) without dry season and warm summer; Dfc: Cold (continental) without dry season and cold summer.
(c) Missing data percentage refers to NEE time series across the whole year (Y), for daytime (D) and nighttime subsets (N). Daytime and nighttime subset are defined using a global radiation threshold set equal to 10 Wm^{-2}.

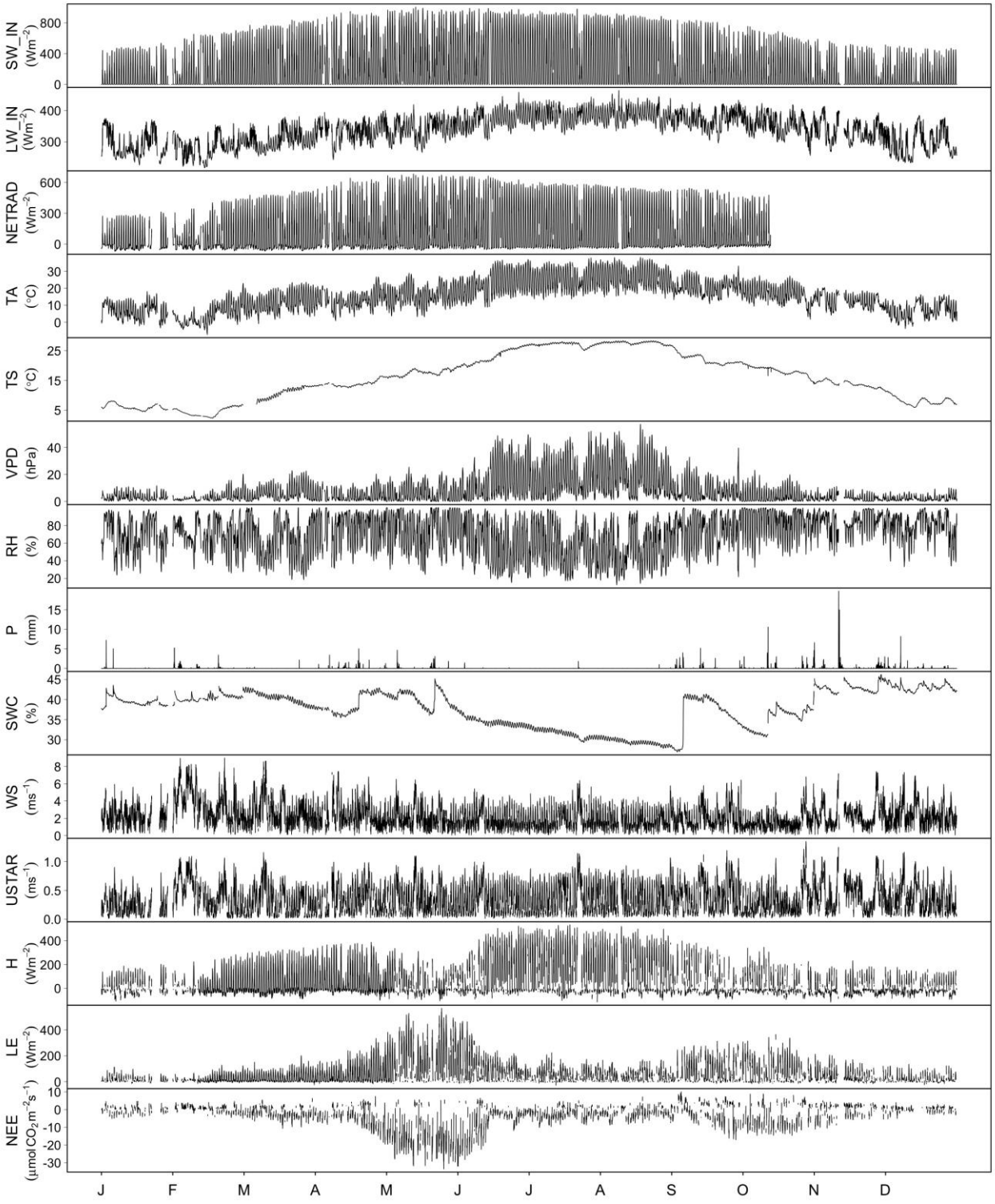

Figure 1. The eddy-covariance dataset collected at IT-CA1 use case during 2012.

95% confidence interval. Parameter W should be as small as possible and, all other things being equal, imputation methods providing narrower confidence intervals should be preferred. However, W should not be too small as to affect the CR.

We have also compared the out-of-sample performances of the three proposed imputation models (MLR, ADL and PA-DL). The comparison was achieved by superimposing several artificial gaps on the already incomplete NEE, LE and H time

series. In particular, similarly to Moffat et al. (2007) but adding a longer gap, we simulated the following 5 macro-scenarios:

- S1 – additional 5% of randomly distributed single hour or half-hour missing data points.
- S2 – additional 5% of randomly distributed short gaps of 4 consecutive hours missing data points.
- M5 – medium gaps of 5 consecutive days randomly distributed across the year.
- L10 – long gaps of 10 consecutive days randomly distributed across the year.
- L20 – long gaps of 20 consecutive days randomly distributed across the year.

On the basis of a large collection of EC time series data, Falge et al. (2001) concluded that over 50% of the gaps present in the analysed data sets were less than 2 hours, and less than 4% were longer than 21 days. S1 and S2 scenarios aim to simulate missing data points due to QC filtering procedures and system maintenance. On the contrary, M5 scenario simulates gaps that are often present, and due to system failure. L10 and L20 scenarios test the stability of MI procedures under extreme conditions, that are however also present in the real data, generally due to important damages to the sensors. All the above scenarios were permuted 10 times for each benchmark site (giving a total 100 simulations for each scenario). It must also be noted that gap locations were always identical in each flux variable. This setting allowed for considering the possibility that missing values caused by real system failure or maintenance simultaneously affect each flux variable (with the only exception being data discarding in consequence of QC procedures). Gap co-occurrence has also the obvious consequence that any of the three flux variables cannot be used as an input variable in a MI model (for example, LE and H, both or individually considered, cannot be used to reconstruct NEE).

For each imputation model and each of the 500 simulated gaps, we computed the out-of-sample bias error (BE) and mean absolute error (MAE):

$$\mathrm{BE} = \frac{1}{T_{gap}} \sum \left(y_{t,gap} - i_{t,gap} \right) \tag{14}$$

$$\mathrm{MAE} = \frac{1}{T_{gap}} \sum \left| y_{t,gap} - i_{t,gap} \right| \tag{15}$$

where $y_{t,gap}$ indicates any observed value that has been flagged as missing, while $i_{t,gap}$ is the corresponding imputed value. The statistical metrics were then grouped and averaged along each of the 5 artificial macro-scenarios defined above, to aid in comparisons.

For an overall evaluation of the proposed MI models, we applied the Friedman test (Friedman, 1940) using a significance level $\alpha = 0.05$, followed by a post-hoc test based on the procedure introduced in Nemenyi (1963). The Friedman test is a non-parametric statistical test, equivalent to repeated-measures ANOVA, which can be used to compare the performances of several models on multiple data sets (Demšar, 2006). In order to do that, ranks are assigned to models. For each data set, the model with the best performance gets the lowest (best) average rank. The null hypothesis of the Friedman test is that there are no significant differences between the mean out- and in-sample performances of all the considered models.

Provided that significant differences were detected by the Friedman test (that is the null hypothesis is rejected) the Nemenyi test can be used for pairwise multiple comparisons of the considered algorithms (Demšar, 2006). Nemenyi test is similar to the post-hoc Tukey test for ANOVA, and its output consists of a critical difference (CD) threshold. The mean performance of two imputation models is judged to be signifycantly different if the corresponding average ranks differ by at least the critical difference (the graphical output of Nemenyi test was implemented using tools provided in the *TStools* R package; Kourentzes and Svetunkov, 2017).

3.3. Benchmark Gap-Filling Algorithm

In order to better evaluate the in-sample and out-of sample performances of the three MI models, results obtained using the marginal distribution sampling (MDS) method proposed by Reichstein et al. (2005) were also added to the comparison. The choice of MDS algorithm as a benchmark is motivated by its good performances in the simulation study by Moffat et al. (2007), despite a simple logic and implementation that made the MDS one of the most used tools in EC data gap-filling. In synthesis, MDS replaces any missing values by the average value under similar meteorological conditions within a time-window constructed around each missing value and with the minimum length possible (starting from 7 days). It is assumed that similar meteorological conditions are present inside that window if SW_IN, TA and VPD do not deviate by more than 50 Wm^{-2} (when SW_IN > 50 Wm^{-2}, otherwise 20 Wm^{-2} are used), 2.5 °C, and 5.0 hPa, respectively. If no sufficient data points under similar conditions are found, less restrictive conditions are imposed in a hierarchical way, increasing the temporal window size, defining the similar meteorological conditions only on the basis of SW_IN or applying the mean diurnal variation method (Falge et al., 2001), i.e. by the arithmetic mean of valid values measured on adjacent days at the same time of the day. More details on how the different conditions are combined can be found in Reichstein et al. (2005). As proposed by Lasslop et al. (2008), we used the standard deviation of observation measured under similar meteorological conditions as a measure of uncertainty of the imputed values. In this work we used the implementation of the MDS algorithm implemented in the *REddyProc* R package.

4. Results and Discussion

In this Section, we report the performance of the three proposed imputation models, where each model has been compared with respect to each other and with the MDS baseline algorithm. Finally, we show annual budget estimates with the associated uncertainty, as an example of the application of Rubin's rules during the complete data analysis stage.

4.1. Algorithmic Details and Additional Data Pre-Processing

All the proposed imputation models were fitted and checked by the *Amelia* R package (Honaker et al., 2011; Yucel, 2011; R Core Team, 2017) which provides an interface to the Amelia II program for MI of incomplete datasets under the EMB approach depicted in Section 2.5.

We set $M = 30$ (the number of imputed datasets), despite the fact that the classic advice is to use a low number of imputations, between 3 and 5, for moderate amounts of missing information (van Buuren, 2012). However, as clarified by a large Monte Carlo study reported in Graham et al. (2007), the number of imputations required is substantially greater (between $M = 20$ and $M = 100$) than previously thought, if we are not willing to tolerate power falloff and abnormally wide confidence intervals of scientific estimands. To speed up the computational time we ran the EMB algorithm in parallel mode. With $M = 30$, the execution times for a complete data analysis per site, including the overimputation procedure, were respectively of about 4, 7 and 10 minutes for MLR, ADL and PADL models, using a 2.2 GHz Intel Core i7 CPU.

We checked the convergence of the EM algorithm at each iteration by monitoring the number of parameters that had significantly changed since the last iteration. Convergence problems can arise when data contains a high degree of missingness, very strong correlation among the variables and/or too many parameters to estimate in respect to the sample size. These problems happened more frequently in the case of nighttime subsets, where the percentage of missing data can go beyond 80%, and became worse with PADL specification. To circumvent this drawback, we added a ridge prior over the covariance matrix Σ, shrinking toward zero the covariances among variables, thus preventing quasi-singular posterior estimates and helping with numerical stability (see Honaker et al., 2011, for details). The level of the empirical ridge prior was dynamically set equal to 0.5% the number of dataset rows when the percentage of miss-ing data in NEE time series was $\geq 70\%$, otherwise it was set to 0.25%, in order to prevent instability of the algorithm in case of multicollinearity among the variables. For the PADL model, the number of cross-sectional units was allowed to vary from $S = 3$ to 4 during daytime, whereas two cross-sectional ob-servations, the first extending until midnight, the second after midnight, were used during nighttime, to prevent the possibility of empty cross-sectional observations.

Within each regime, abiotic variables affected by long consecutive gaps (with a fraction of missing data $\geq 40\%$) were discarded, because of their heavy impact over the computational burden of MI algorithms, an impact not associated with any appreciable improvement in the quality of imputations.

The EMB algorithm is not limited to flat priors over model parameters. With a few modifications the EM algorithm can incorporate prior information over the parameter space, in order to obtain maximum-a-posteriori (MAP) estimates of θ (Gelman et al., 2013). In particular, Amelia has a number of methods of setting priors over the mean and the standard deviation of one or more missing input data cells, and to derive the implied priors over μ and Σ (Girosi and King, 2008;

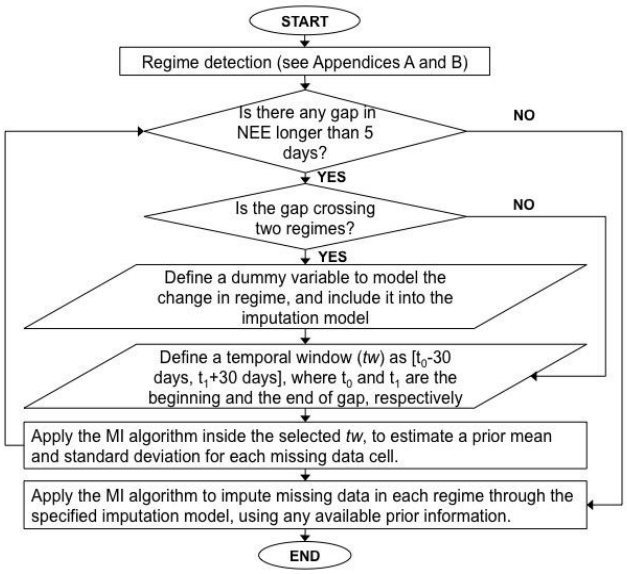

Figure 2. Work Flow of the MI strategy developed for EC datasets.

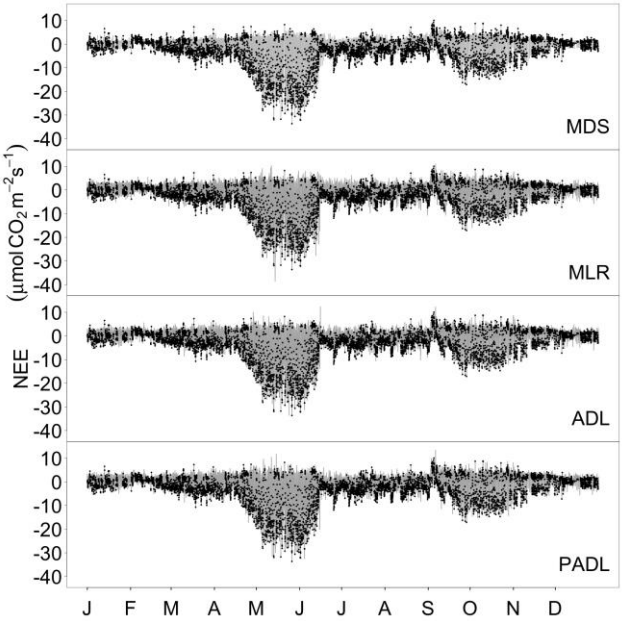

Figure 3. Observed and gap-filledNEEfor the IT-CA1 use case (for MI models one of the M=30 complete time series).

Honaker and King, 2010; Honaker et al., 2011). The use of such observation-level priors can vastly increase the computational efficiency of the algorithm and improve the quality of imputations (if unbiased and accurate information is provided). We used this functionality for those input variables which are difficult to impute from other variables, showing high persistence, low variability and smooth behavior. In particular, TS and SWC were preliminarily gap-filled by linear inter-polation, and each missing data point was subsequently endowed with a Gaussian prior having mean equal to the interpolated value and standard

deviation equal to 10% of the same value. By exploiting this machinery, imputed values become a compromise between model imputations and prior means, with imputations increasingly reflecting only the observed data and ignoring the prior values when the model predicts missing values with high accuracy.

As both ADL and PADL include autoregressive terms, missing data falling around the boundary between nighttime and daytime regimes would be imputed twice, from two very different generative models. Suppose, for example, that the beginning of the daytime regime is determined at 6:00AM. In this case a missing flux at 5:30AM would be first imputed as the output variable of the nighttime model, and then it would also be imputed by the daytime generative model, because it enters such model as a lagged dependent variable. To prevent this double estimate, missing data falling at the beginning of the daytime regime were endowed by informative priors having mean and standard deviation based on data imputed during the nighttime regime. In this way we were able to link nighttime and daytime missing data imputations, and the algorithm became more stable, given the high percentage of missing data falling near the transition from nighttime to daytime periods.

To further improve the quality of missing data imputations, we augmented the covariate set with additional input variables, the downscaled time series from ERA-interim reanalysis (Vuichard and Papale, 2015), consisting of SW_IN, LW_IN, TA and VPD variables. Finally, the contribution of precipitation was taken into account by adding a new predictor to the imputation model given by the logarithm of the cumulative sum of measured rainfall (or gap-filled with ERA-interim product when some measurements were missing) in the past 24 hours, i.e. $\log(\sum_{k=0}^{47} P_{t-k} + 1)$ in the case of half-hourly time series.

Further computational issues occurred with some of the macro-scenarios described in Section 3.2. In particular, with M5 and L10 we found that deterministic term parameter estimation might become unstable when gaps occurred near the boundaries of the time intervals defining the regimes. On the contrary, L20 simulations convergence issues were predominantly driven by the high degree of missingness. Several strategies are useful to overcome this issue. A first possibility consists in modifying the detected regime break dates (see Appendix A), by shortening or extending the temporal window of the regime. Another possibility consists in a preliminary imputation of any gap longer than 5 days, by considering a buffer temporal window beginning at 30 days before the beginning time of the gap, and ending at 30 days after the end of gap. After some empirical testing, we adopted this last choice. When the

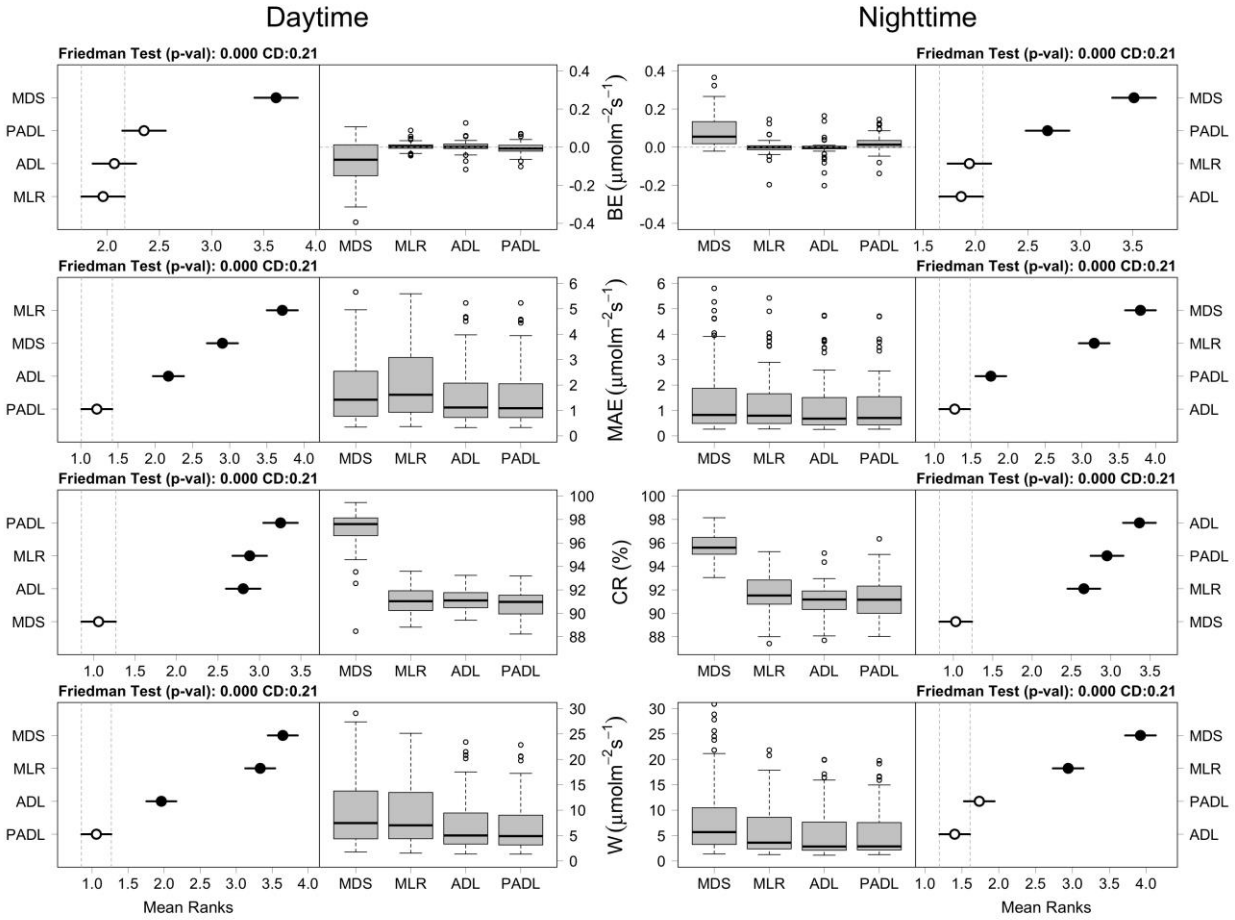

Figure 4. Graphical visualization of the Nemenyi's test for the evaluation of the in-sampleaccuracy measures for NEE.

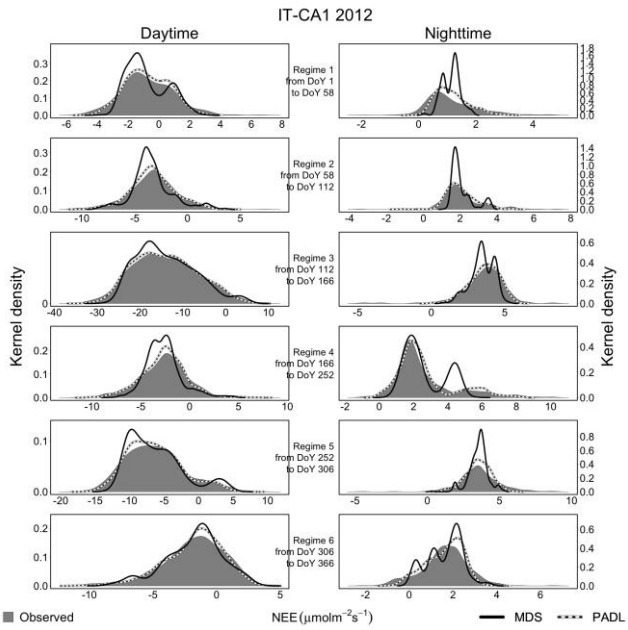

Figure 5. Kernel density estimates comparing observed and overimputed data by MDS algorithm and PADL model for the IT-CA1 use case.

gap crossed two regimes, a dummy variable was introduced in the set of input variables to keep into account the change in regime. This first preliminary run of the MI algorithm was used to estimate the average and the standard deviation of multiply imputed values in each missing data cell inside the buffer time window. Subsequently, a new global run of the MI algorithm was carried out, using informative priors having mean and standard deviation over the missing data cells as determined before.

The proposed MI strategy developed for EC datasets can then be summarized in three main steps (see also Figure 2):

1. Identification of homogeneous ecological regimes, as well as of daytime and nighttime subsets (Appendices A and B).
2. Preliminary imputation of any gap present in NEE, LE and H flux series, longer than 5 consecutive days, in order to estimate informative priors in missing data cells using a sort of empirical Bayes procedure.
3. Multiple imputation of EC dataset, separately for each regime, and daytime/nighttime subsets (possibly using prior information determined in step 2, or any available prior information on missing data cells).

An example of gap-filled NEE time series for the IT-CA1 use case is shown in Figure 3. Visual inspection cannot highlight any significant difference between the algorithms used. Therefore, in the next section we carefully inspect both in-sample and out-of-sample accuracies.

4.2. In-Sample and Out-of-Sample Accuracy

In this section, we report both the in-sample and out-of sample performance of the three proposed MI models, as well as of the baseline MDS algorithm. For each of the 10 selected FLUXNET sites and the four gap-filling algorithms under consideration, out-of-sample metrics (14) and (15) were calculated, then the four algorithms were compared as a whole, on the basis of a new 100-dimensional vector of out-of-sample accuracies (one for each of the 10 sites times 10 simulations) for each synthetic gap macro-scenario, using Friedman non-parametric ANOVA (see Section 3.2). In-sample indicators were

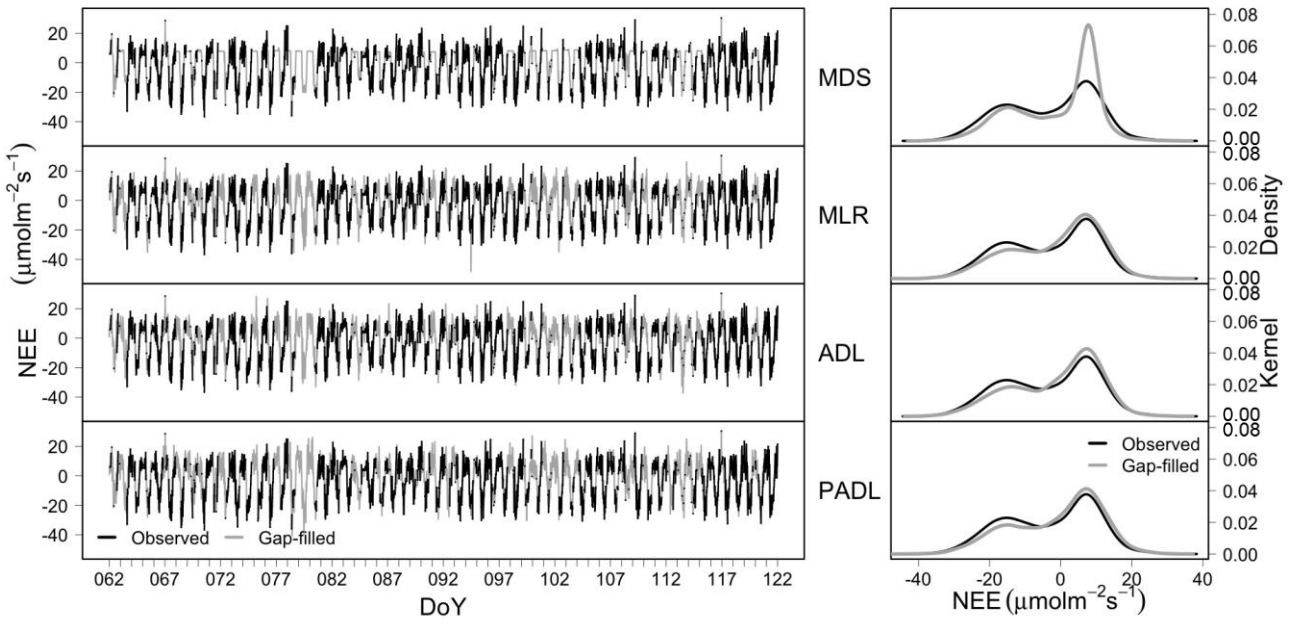

Figure 6. Kernel density estimates of observed and gap-filled NEEfor the GF-Guy use case, over a short period of 60 days (for MI models one of the M = 30 complete time series).

calculated inside each hydro-ecological regime (see Appendix A). The results were summarized through standard box-plots and Nemenyi Critical Difference (CD) plots (Demšar, 2006), separately for daytime and nighttime (daytime and nighttime regimes are defined by means of a global radiation threshold, set equal to 10 Wm^{-2}).

The in-sample bias error (BE), mean absolute error (MAE), coverage rate (CR) and interval width (W) for NEE flux are summarized in Figure 4 (plots related to LE and H fluxes are reported in the supplementary material, due to space limitations). On average, the three proposed multiple imputation models provided less unbiased estimates than the MDS algorithm, this last showing much more variability than MI models. In any case the maximum absolute value of bias was <1/100 than the range of observed data, and can be considered negligible. For NEE, LE and H flux data, the lowest in-sample MAE was always achieved by PADL model during daytime and by ADL during nighttime. As far as the coverage rate (CR), the approach proposed by Lasslop et al. (2008) for the uncertainty estimation of imputed value through the MDS algorithm leads to significantly higher values than those obtained through the three MI models. However, this difference can be explained by the largest interval width (W) of the MDS algorithm. Conversely, both ADL and PADL provided slightly narrower confidence intervals (W) at about the same actual coverage rate (CR).

In order to further investigate the performance of the imputation models from a purely data-analytic point of view, we plotted kernel density estimates comparing observed and over-imputed data. Figure 5 shows the comparison results for the IT-CA1 use case (similar results were found for all the benchmark sites considered in this paper). It was evident how the PADL method preserves the sampling variability, both during daytime and nighttime, under almost all regimes. In comparison, the MDS-based imputations, albeit at least approximately unbiased

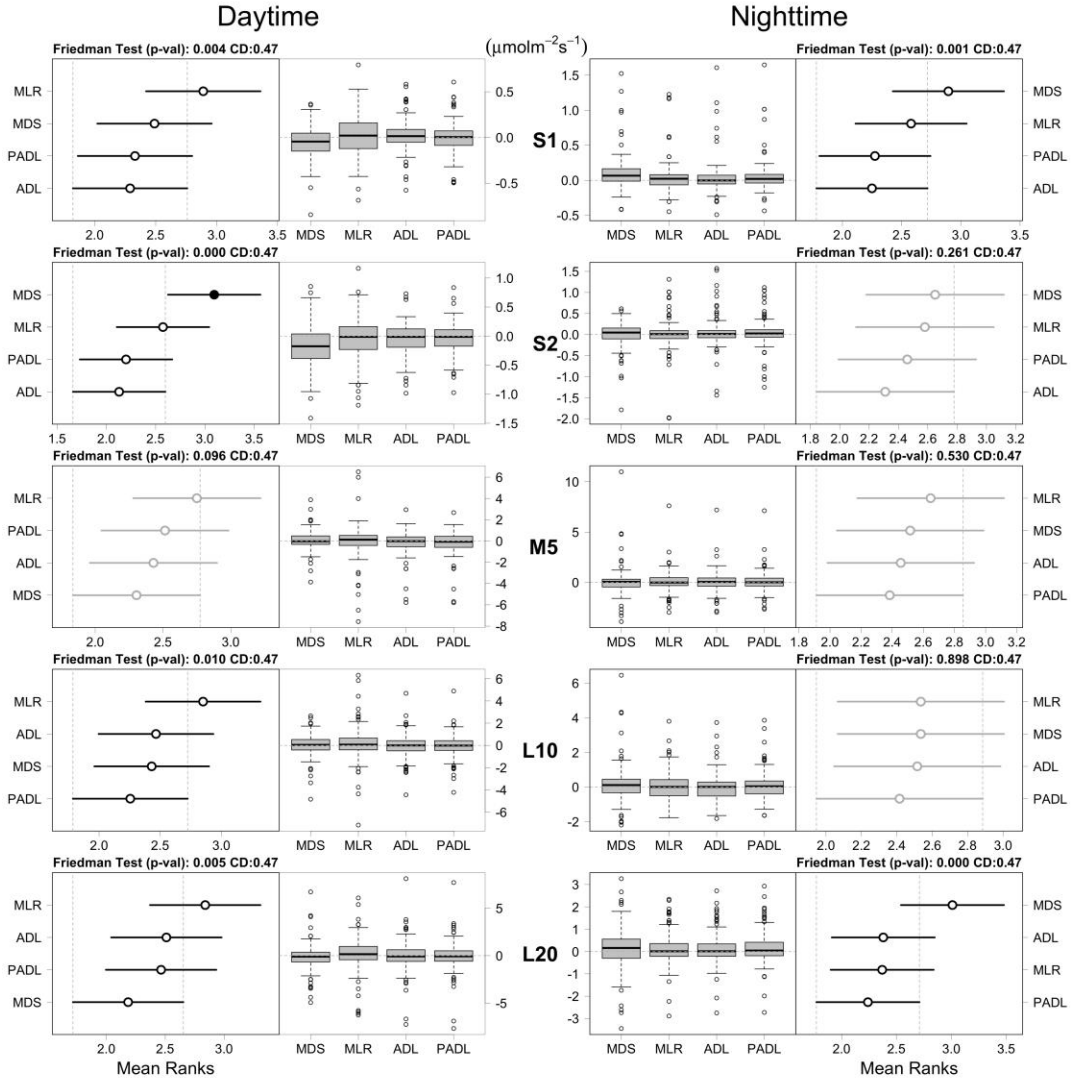

Figure 7. Graphical visualization of the Nemenyi's test for the comparison of the out-of-sample bias error for NEE in different scenarios.

(the estimated densities were not markedly asymmetric), did not preserve sampling variability, in particular at nighttime.

To better appreciate one of the key advantages of MI based procedures with respect to methods based on averaging procedures (here represented by the MDS), in Figure 6 we plotted the complete NEE time series and kernel density estimates of observed and imputed time series by each method (GF-Guy use case). We zoomed into a section of the plot, showing a short temporal window of 60 days (we randomly selected one of the complete multiply-imputed time series). As can be seen, also in this case the distributions of observed and imputed values are almost identical in the three MI methods. The MDS instead shows a peak in the distribution in correspondence of data values. This effect increases progressively as the signal-to-noise ratio is decreased. An explanation of this behavior relies upon the fact that MI procedures are designed not only reproduce the 'true' signal, but also to properly manage both the uncertainty arising from both estimation of model parameters and random error affecting observed data, thus preserving the original variability of the DGP (Kunwor et al., 2017). For these reasons, MI algorithms can be considered a valid and more appropriate alternative to SI methods (van Buuren, 2012, Chapter 1).

Out-of-sample bias error (BE) and mean absolute error (MAE) for NEE flux, under the 5 synthetic macro-scenarios and for daytime and nighttime separately, are respectively presented in Figures 7 and 8 (plots showing results for H and LE flux variables are reported in the supplementary material). By looking at the reconstructed NEE flux, both ADL and PADL models showed less unbiased estimates and lower MAE than MLR and MDS imputation methods, although no marked bias differences are present. The BE showed higher variability in M5, L10 and L20 scenarios reaching the highest values, in absolute terms, in cropland (US-Ne2) and tropical (GF-Guy) sites. The lowest daytime MAE was achieved by PADL model

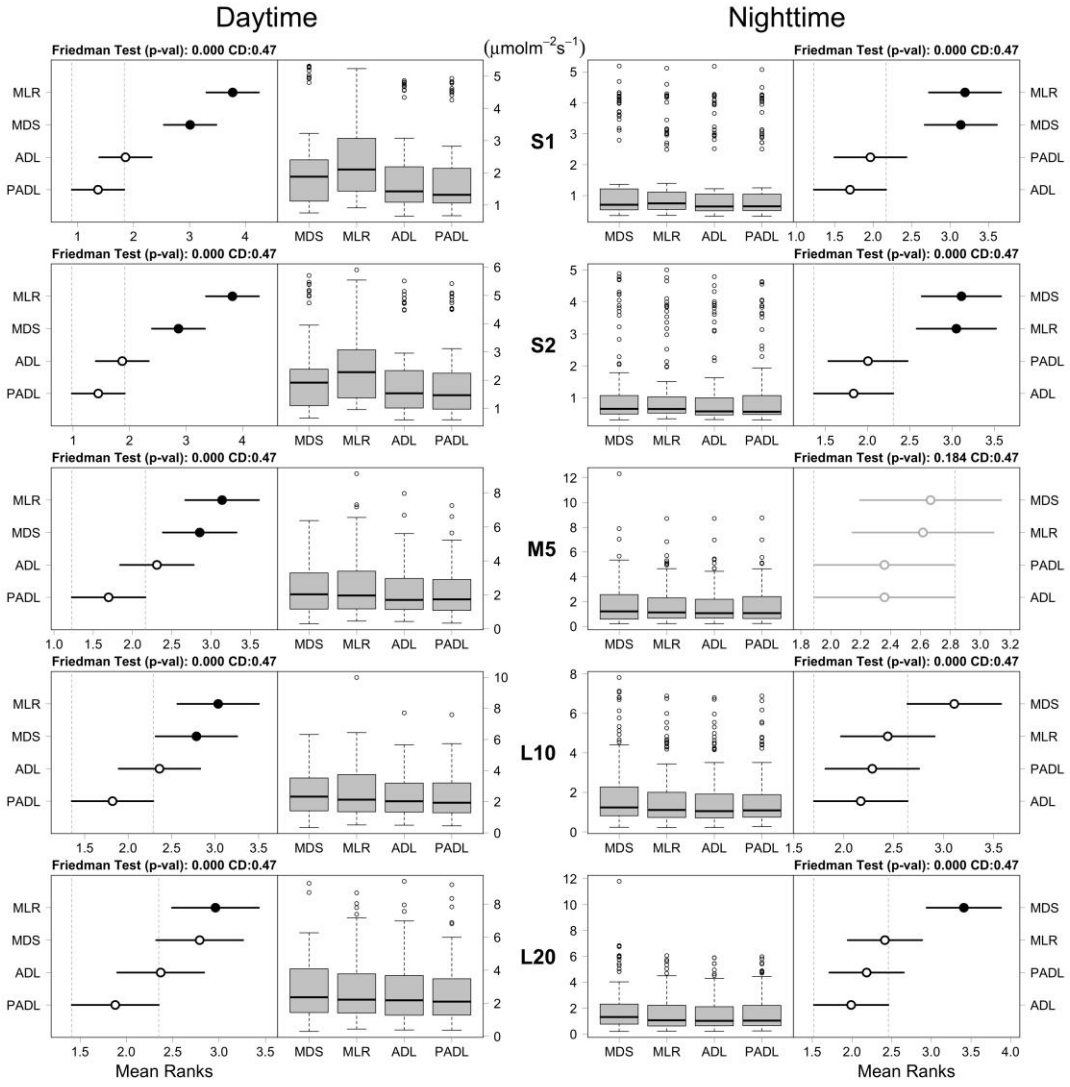

Figure 8. Graphical visualization of the Nemenyi's test for the comparison of the out-of-sample mean absolute error for NEE in different scenar.

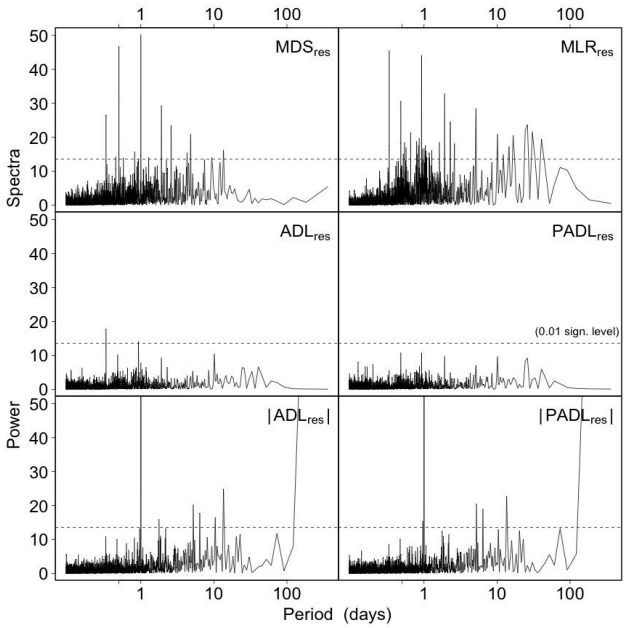

Figure 9. Lomb-Scargleperiodograms of residual time series for the FI-Hyy use case.

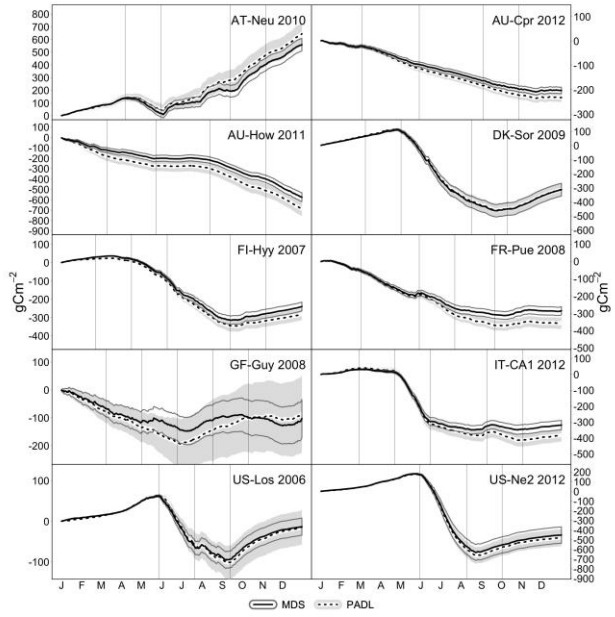

Figure 10. Cumulative NEE for each of the 10 benchmark use cases.

under all the 5 macro-scenarios, while the better performances during nighttime were achieved by both ADL and PADL model. Similar consideration holds for LE and H flux variables.

These results demonstrate that the introduction of lagged endogenous variables, entering both ADL and PADL specifications, substantially improves flux reconstruction. This reinforces the hypothesis that flux variability (or, at least, a part of it) is most likely to be not only explained by exogenous variables (e.g. meteorological factors), but also by the ecosystem state itself, which is well suited to be represented by lagged endogenous variables as those introduced in both ADL and PADL models.

Yet another confirmation that both PADL and ADL methods better reproduce the data distribution and implement a correct DGP can be obtained by looking at the lack of temporal autocorrelation in the residual component (observed minus over-imputed values). This has been observed in most of the use cases under investigation. As an example, Figure 9 shows power spectra (estimated by Lomb-Scargle periodogram) of the NEE flux residual time series at FI-Hyy site. While MDS and MLR residuals showed significant peaks at a period of 1 day, both ADL and PADL spectral estimates closely resembled the typical flat pattern of a white noise process. On the contrary, power spectra of absolute residuals invariably showed a significant daily peak, a fact indicating the presence of a correlation structure in the second moment (heteroscedastic variance) of the data, that would need further development and refinement of the imputation models.

The low performance of the MDS algorithm at nighttime can be attributed to several concurrent factors, such as: (i) the criteria used to define similar meteorological conditions could often be too simplistic, as only two variables are involved in the algorithm (SW_IN is always 0 at nighttime); (ii) the lack of robustness occurring when the temporal search window is getting wider, or when the flowchart of the algorithm has reached the step consisting in the application of mean diurnal variation method; (iii) the leverage effect of anomalous values on the calculation of the mean. In addition, the presence of sites where management activities are present (e.g. crop harvesting) is particularly challenging for methods where there are no indication of changes in the regime (like the one introduced in the PADL and ADL methods).

4.3. Annual Budget Estimation

As an example of a complete data analysis, in this Section we report on annual budget estimates and associated uncertainties obtained through the PADL model and the MDS baseline algorithm. As far as LE and H were concerned, annual estimates from both algorithms showed fairly good agreement in all the use cases under investigation (numerical details are provided in supplementary material). On the contrary, a slight to moderate discrepancy was found between annual carbon budget estimates, even though in most cases the respective 95% CIs were overlapping (see Figure 10).

The most interesting differences can be appreciated by analyzing the uncertainty associated to the annual budget values. Table 2 shows the total uncertainty at 95% confidence level obtained by combining the within uncertainty $\bar{U}^{1/2}$, with the between uncertainty due to missing data imputation $B^{1/2}$, as described in Subsection 2.6. For MDS algorithm the 95% CI for the annual budgets was estimated by doubling the within uncertainty (the B component is obviously undefined in this case). The $\bar{U}^{1/2}$ component was estimated through the square-root-of-time rule as $T^{1/2}s_{\mathrm{NEE}}$, where T is the sample size (e.g.

17520 half-hours in a year) and s_{NEE} is the sample standard deviation of half-hourly NEE time series (comparable estimates were obtained by ordinary bootstrap resampling, but our simple rule of thumb has an apparent advantage from a computational point of view).

As a result, the net effect of including the between uncertainties was an increase in the total uncertainties, which were on average higher by about 50% than those estimated by doubling the within uncertainty. Highest differences were observed for GF-Guy (+100%), US-Los (+98%) and AT-Neu (+78%), while negligible discrepancies were obtained for DK-Sor (+5%) and US-Ne2 (+2%).

These differences are also reflected in the estimates of the fraction of missing information. In fact, when the between uncertainty is higher than the within, ρ increases proportionally to their ratio. As we said before, significant values ($\rho > 0.5$) indicate that statistical inferences are highly dependent on the way in which the missing data were handled. This means that the influence of the imputation model is much larger than that of the complete data model (van Buuren, 2012), and further investigations (with special attention to data quality issues, in particular during the nighttime period) are necessary to reach a conclusion.

Anyway, all these quantities provide a useful tool for the evaluation of the reliability of the annual budget estimates. It is worth noting that as annual budgets obtained through the MDS algorithm cannot be endowed with a between-imputation variability estimate (like any other SI algorithm), this often results in underestimating the total uncertainty.

5. Concluding Remarks

In this paper we have presented three new imputation models, built on top of multivariate normality assumption and characterized by an increasing level of complexity. All these procedures are based on the hybrid EM-Bootstrap algorithm, in short EMB, introduced by Honaker and King (2010), which has several advantages over its direct competitors: (i) its high computational efficiency allows it to cope with large datasets, (ii) effective imputations making use of special time series characteristics becomes possible, (iii) a large number of diagnostic checks, based on overimputing the observed data, are natively available in the *Amelia* R package, which provides an interface to the Amelia II program for MI of incomplete datasets under the EMB algorithm.

Under several synthetic gap-scenarios, the PADL model showed the best out-of-sample performance in imputing missing values, producing unbiased imputations and preserving the original variability in the data. To a large extent, the ADL model also shares these merits. In fact, both models have an improved ability to capture temporal dynamics of EC fluxes, since temperature and water supply, which are considered the primary controlling factors of photosynthetic response, not only have an instantaneous impact, but also play a role by way of cumulated and/or lagged effects. At the same time, the role of the deterministic trend and of its added flexibility must be stressed, because of its ability to model irregular seasonal and diurnal cycles. In view of all these considerations, we can conclude that natural variability of half-hourly EC flux time series cannot be

Table 2. Annual Budget Estimates, Associated Uncertainties, and Fraction of Missing Information ρ of Net Ecosystem Exchange (NEE, gC $m^{-2}y^{-1}$) Gap-filled Flux, Reconstructed through MDS and PADL Algorithms

Site ID / Year	Model	$\widetilde{Q}$ (gC $m^{-2}y^{-1}$)	Uncertainty (gC $m^{-2}y^{-1}$)					ρ
			Within $\overline{U}^{1/2}$	Between $B^{1/2}$	Total $2\widetilde{V}^{1/2}$	95% CI Lower	Upper	
AT-Neu 2010	MDS	558	24.7		49.4	509	608	
	PADL	645	27.2	34.2	88.4	556	733	0.63
AU-Cpr 2012	MDS	-203	6.6		13.2	-216	-190	
	PADL	-232	6.7	6.1	18.2	-250	-214	0.46
AU-How 2011	MDS	-576	21.2		42.4	-618	-533	
	PADL	-689	21.5	25.7	67.6	-756	-621	0.60
DK-Sor 2009	MDS	-314	22.7		45.4	-359	-268	
	PADL	-312	22.8	7.2	47.9	-360	-264	0.09
FI-Hyy 2007	MDS	-240	12.1		24.3	-264	-216	
	PADL	-282	11.9	10.2	31.5	-314	-251	0.44
FR-Pue 2008	MDS	-285	11.9		23.9	-309	-261	
	PADL	-355	12.1	10.8	32.6	-387	-322	0.46
GF-Guy 2008	MDS	-103	34.4		68.8	-172	-34	
	PADL	-85	37.7	56.6	137.6	-223	52	0.71
IT-CA1 2012	MDS	-319	14.8		29.6	-349	-290	
	PADL	-383	14.9	12.8	39.5	-423	-344	0.44
US-Los 2006	MDS	-13	10.6		21.2	-34	8	
	PADL	-15	10.7	17.7	41.8	-57	27	0.75
US-Ne2 2012	MDS	-452	42.2		84.3	-537	-368	
	PADL	-480	42.4	6.5	85.9	-566	-395	0.02

adequately described using a linear relationship with abiotic exogenous factors. In contrast, the effect of lagged variables, in particular of biotic endogenous factors, plays a key role in explaining the complex dynamics of EC fluxes.

Improvements in correctly reproducing the DGP of EC fluxes are indeed possible if we take into account that the absolute value of overimputed residuals, under both the ADL and PADL models, show significant intra-day correlations. This empirical evidence suggests that variability in EC flux time series varies with time. For this purpose, on the basis of the results found in Richardson et al. (2008), which reported that random flux errors more closely followed a fat-tailed non-Gaussian distribution, we considered a stochastic volatility model for high-frequency data proposed by Beltratti and Morana (2001). This model showed an improved ability in modelling both persistence and intra-day cyclical components in flux variability. These results will be published elsewhere.

Notwithstanding that there still remains a vast room for exploration of more flexible models, we expect that the strategy proposed in this paper will become useful in creating multiple imputations for a variety of EC dataset, and providing valid inferences for a broad range of scientific estimands (such as annual budgets).

Acknowledgements. Abbreviations: Domenico Vitale (DV), Massimo Bilancia (MB), Dario Papale (DP).

This paper has been started in the context of the ICOS-INWIRE research project funded by the European Community's 7th Framework Program (FP7/2007-2013) under the agreement no 313169 and continued and finalized under the ENVRIPLUS H2020 European project (Grant Agreement 654182), that the authors thank for the support.

DV conceived the study; DV and DP contributed to the study design; DV, MB and DP wrote the first draft of the manuscript. All authors equally contributed to the writing of Section 1 and Appendices (Appendices A and B are reported in the Supplementary Material). DV wrote Subsections 2.1, 2.5, 4.2, 4.3, and cared about the overall paper structure; MB wrote Subsections 2.2, 2.3, 2.4, 2.6, 3.2 and 4.1; DP wrote Subsections 3.1, 3.3 and Section 5. All authors reviewed and revised the manuscript, approved the final version, and agreed to submit the manuscript for publication.

This work used eddy covariance data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE. The FLUXNET eddy covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices. The authors thank in particular the PIs that shared the data used in the study.

The R code implementing the proposed MI strategy is available at http://www.icos-etc.eu.

References

Aubinet, M., Vesala, T., and Papale, D. (2012). Eddy covariance: A practical guide to measurement and data analysis. *Springer Atmos. Sci.,* https://doi.org/10.1007/978-94-007-2351-1

Barnard, J., and Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika,* 86(4), 948-955. https://doi.org/10.1093/biomet/86.4.948

Beck, N. and Katz, J.N. (1995). What to do (and not to do) with time-series cross-section data. *Am. Polit. Sci. Rev.,* 89(3), 634-647. https://doi.org/10.2307/2082979

Beltratti, A. and Morana, C. (2001). Deterministic and stochastic methods for estimation of intra-day seasonal components with high frequency data. *Economic Notes,* 30(2), 205-234. https://doi.org/10.1111/j.0391-5026.2001.00054.x

Beringer, J., Hutley, L.B., Tapper, N.J., and Cernusak, L.A. (2007). Savanna fires and their impact on net ecosystem productivity in North Australia. *Global Change Biol.,* 13(5), 990-1004. https://doi.org/10.1111/j.1365-2486.2007.01334.x

Blei, D.M, Kucukelbir, A., and McAuliffe, J.D. (2017). Variational inference: A review for statisticians. *J. Am. Stat. Assoc.,* 112(518), 859-877. https://doi.org/10.1080/01621459.2017.1285773

Bonal, D., Bosc, A., Ponton, S., Goret, J.Y., Burban, B., Gross, P., Bonnefond, J.M., Elbers, J., Longdoz, B., Epron, D., Guehl, J.M., and Granier, A. (2008). Impact of severe dry season on net ecosystem exchange in the Neotropical rainforest of French Guiana. *Global Change Biol.,* 14(8), 1917-1933. https://doi.org/10.1111/j.1365-2486.2008.01610.x

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Ser. B.,* 39(1), 1-38. http://www.jstor.org/stable/2984875.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learning Res.,* 7(1), 1-30.

Doove, L.L., Van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Comput. Stat. Data Anal.,* 72, 92-104. https://doi.org/10.1016/j.csda.2013.10.025

Efron, B. (1994). Missing data, imputation, and the bootstrap. *J. Am. Stat. Assoc.,* 89(426), 463-475. https://doi.org/10.1080/01621459.1994.10476768

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *Ann. Appl. Stat.,* 6(4), 1971-1997. https://doi.org/10.1214/12-AOAS571

Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., and Burba, G. et al. (2001). Gap filling strategies for defensible annual sums of net ecosystem exchange. *Agric. For. Meteorol.,* 107(1), 43-69. https://doi.org/10.1016/S01681923(00) 00225-2

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Stat.,* 11(1), 86-92. https://doi.org/10.1214/aoms/1177731944

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013). *Bayesian Data Analysis*, Third Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Girosi, F. and King, G. (2008). *Demographic Forecasting*, Princeton University Press.

Graham, J.W., Olchowski, A.E., and Gilreath, T.D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.,* 8(3), 206-213. https://doi.org/10.1007/s11121-007-0070-9

Kourentzes, N. and Svetunkov, I. (2017). TStools: Time series analysis tools and functions. http://kourentzes.com/forecasting/2014/04/19/tstools-for-r/.

Kunwor, S., Starr, G., Loescher, H.W., and Staudhammer, C.L. (2017). Preserving the variance in imputed eddy-covariance measurements: Alternative methods for defensible gap filling. *Agric. For. Meteorol.,* 232(15), 635-649. https://doi.org/10.1016/j.agrformet.2016.10.018

Hassler, U. and Wolters, J. (2006). Autoregressive dstributed lag models and cointegration. *Allgemeines Stat. Arc.,* 90(1), 59-74. https://doi.org/10.1007/s10182-006-0221-5

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. 2nd ed, Springer Series in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Honaker, J. and King, G. (2010). What to do about missing values in time-series cross-section data. *Am. J. Polit. Sci.,* 54(2), 561-581. https://doi.org/10.1111/j.1540-5907.2010.00447.x

Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A program for missing data. *J. Stat. Software,* 45(7), 1-47. https://doi.org/10.18637/jss.v045.i07

Hsiao, C. (2007). Panel data analysis-advantages and challenges. *Test,* 16(1), 1-22. https://doi.org/10.1007/s11749-007-0046-x

Hui, D., Wan, S., Su, B., Katul, G., Monson, R., and Luo, Y. (2004). Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. *Agric. For. Meteorol.,* 121(1-2), 93-111. https://doi.org/10.1016/S0168-1923(03)00158-8

Huisman, R., Huurman, C., and Mahieu, R. (2007). Hourly electricity prices in day-ahead markets. *Energy Econ.,* 29(2), 240-248. https://doi.org/10.1016/j.eneco.2006.08.005

Lasslop, G., Reichstein, M., Kattge, J., and Papale, D. (2008). Influences of observation errors in eddy flux data on inverse model parameter estimation. *Biogeosci. Discuss.,* 5(1), 751-785. https://doi.org/10.5194/bg-5-1311-2008

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Second Edition. John Wiley & Sons.

Meyer, W.S., Kondrlovà, E., and Koerber, G.R. (2015). Evaporation of perennial semi-arid woodland in southeastern Australia is adapted for irregular but common dry periods. *Hydrol. Process.,* 29(17), 3714-3726. https://doi.org/10.1002/hyp.10467

Moffat, A.M., Papale, D., Reichstein, M., Hollinger, D.Y., Richardson, A.D., and Barr, A.G. et al. (2007). Comprehensive Comparison of Gap-Filling Techniques for Eddy Covariance Net Carbon Fluxes. *Agric. For. Meteorol.,* 147(3-4), 209-232. https://doi.org/10.1016/j.agrformet.2007.08.011

Nemenyi, P.B. (1963). *Distribution-free Multiple Comparisons*, PhD thesis, Princeton University.

Papale, D., Reichstein, M., Aubinet, M. Canfora, E., Bernhofer, C., Kutsch, W., and Longdoz, B. et al. (2006). Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: algorithms and uncertainty estimation. *Biogeosciences.,* 3(4), 571-583. https://doi.org/10.5194/bg-3-571-2006

Pilegaard, K., Ibrom, A., Courtney, M.S., Hummelshøj, P., and Jensen, N.O. (2011). Increasing net CO_2 uptake by a Danish beech forest during the period from 1996 to 2009. *Agric. For. Meteorol.,* 151(7), 934-946. https://doi.org/10.1016/j.agrformet.2011.02.013

R Core Team (2017). R: A language and environment for statistical computing. vienna, Austria. https://www.r-project.org/.

Rambal, S., Joffre, R., Ourcival, J.M., Cavender-Bares, J., and Rocheteau, A. (2004). The growth respiration component in eddy CO_2 flux from a Quercus ilex mediterranean forest. *Global Change Biol.,* 10(9), 1460-1469. https://doi.org/10.1111/j.1365-2486.2004. 00819.x

Reichstein, M., Falge, E., Baldocchi, D., and Papale, D. (2005). On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global Change Biol.,* 11(9), 1-16. https://doi.org/10.1111/j.1365-2486.2005.001002.x

Richardson, A.D., Mahecha, M.D., Falge, E., Kattge, J., Moffat, A.T., Papale, D., Reichstein, M., Stauch, V.J., Braswell, B.H., Churkina, G., and Kruijt, B. (2008). Statistical properties of random CO_2 flux measurement uncertainty inferred from model residuals. *Agric. For. Meteorol.,* 148(1), 38-50. https://doi.org/10.1016/j.agrformet.2007.09.001

Richardson, A.D., Aubinet, M. Barr, A.G., Hollinger, D.Y., Ibrom, A., Lasslop, G., and Reichstein, M. (2012). Uncertainty quantification. *Eddy Covariance,* 173-209. Springer Netherlands. https://doi.org/10.1007/978-94-007-2351-1_7

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley Series in Probability and Statistics. John Wiley & Sons. https://doi.org/10.1002/9780470316696

Rubin, D.B. (1996). Multiple imputation after 18+ years. *J. Am. Stat. Assoc.,* 91(434), 473-489. https://doi.org/10.2307/2291635

Sabbatini, S., Arriga, N., Bertolini, T., Castaldi, S., Chiti, T., Consalvo, S., Njakou Djomo, S., Gioli, B., Matteucci, G., and Papale, D. (2016). Greenhouse gas balance of cropland conversion to bioenergy poplar short-rotation coppice. *Biogeosciences,* 13(1), 95-113. https://doi.org/10.5194/bg-13-95-2016

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. C&H/CRC Monographs on Statistics & Applied Probability, Chapman & Hall. https://doi.org/10.1201/9781439821862

Schafer, J.L. and Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychol. Methods,* 7(2), 147-177. https://doi.org/10.1037/1082-989X.7.2.147

Schomaker, M. and Heumann, C. (2016). Bootstrap inference when using multiple imputation. http://arxiv.org/abs/1602.07933.

Smith, R.P. and Fuertes, A. (2016). *Panel Time-Series*. Technical Report, Birbeck University of London.

Sulman, B.N., Desai, A.R., Cook, B.D., Saliendra, N., and Mackay, D.S. (2009). Contrasting carbon dioxide fluxes between a drying shrub wetland in Northern Wisconsin, USA, and nearby forests. *Biogeosciences,* 6(6), 1115-1126. https://doi.org/10.5194/bg-6-111 5- 2009

Suni, T., Rinne, J., Reissell, A., Altimir, N., Keronen, P., Rannik, Ü., Maso, M.D., Kulmala, M., and Vesala, T. (2003). Long-term measurements of surface fluxes above a Scots pine forest in Hyytiälä, southern Finland, 1996-2001. *Boreal Environ. Res.,* 8(4), 287-301.

Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.,* 82(398), 528-540. https://doi.org/10.1080/01621459.1987.10478458

Thomas, C.K, Law, B.E., Irvine, J., Martin, J.G., Pettijohn, J.C., and Davis, K.J. (2009). Seasonal hydrology explains interannual and seasonal variation in carbon and water exchange in a semiarid mature ponderosa pine forest in central oregon. *J. Geophys. Res.,* 114(G4), G04006. https://doi.org/10.1029/2009JG001010

Tzikas, D., Likas, A., and Galatsanos, N. (2008). The variational approximation for bayesian inference. *IEEE Signal Proc. Mag.,* 25(6), 131-146. https://doi.org/10.1109/MSP.2008.929620

van Buuren, S. (2012). *Flexible Imputation of Missing Data,* Chapman & Hall/CRC Interdisciplinary Statistics Series. Chapman and Hall/CRC. https://doi.org/10.1201/b11826

Verma, S.B., Dobermann, A., Cassman, K.G., Walters, D.T., Knops, J.M., Arkebauer, T.J., Suyker, A.E., Burba, G.G., Amos, B., Yang, H., Ginting, D., Hubbard, K.G., Gitelson, A.A., and Walter-Shea, E.A. (2005). Annual carbon dioxide exchange in irrigated and rainfed maize-based agroecosystems. *Agric. For. Meteorol.,* 131(1-2), 77-96. https://doi.org/10.1016/j.agrformet.2005.05.003

Vuichard, N. and Papale, D. (2015). Filling the gaps in meteorological continuous data measured at FLUXNET sites with ERA-Interim reanalysis. *Earth Syst. Sci. Data,* 7(2), 157-171. https://doi.org/10.5194/essd-7-157-2015

Wohlfahrt, G., Hammerle, A., Haslwanter, A., Bahn, M., Tappeiner, U., and Cernusca, A. (2008). Seasonal and inter-annual variability of the net ecosystem CO2 exchange of a temperate mountain grassland: Effects of weather and management. *J. Geophys. Res.,* 113(D8), D08110. https://doi.org/10.1029/2007JD009286

Wu, J.C.F. (1983). On the convergence properties of the EM algorithm. *Ann. Stat,* 11(1), 95-103. https://doi.org/10.1214/aos/1176346060

Yucel, R.M. (2011). State of the multiple imputation software. *J. Stat. Software,* 45(1). https://doi.org/10.18637/jss.v045.i01

Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). Strucchange : An R package for testing for structural change in linear regression models. *J. Stat. Software,* 7(2). https://doi.org/10.18637/jss.v007.i02

Zivot, E. and Wang, J. (2006). *Modeling Financial Time Series with S-PLUS®*, Springer New York. https://doi.org/10.1007/978-0-387-32348-0