Agricultural and Forest Meteorology 178–179 (2013) 161–172

Contents lists available at SciVerse ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

Random errors in carbon and water vapor fluxes assessed with Gaussian Processes

Olaf Menzer[a,*], Antje Maria Moffat[a], Wendy Meiring[b], Gitta Lasslop[c], Ernst Günter Schukat-Talamazzini[d], Markus Reichstein[a]

[a] Max Planck Institute for Biogeochemistry, P.O. Box 100164, 07701 Jena, Germany
[b] Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA
[c] Max Planck Institute for Meteorology, Bundesstr. 53, 20146 Hamburg, Germany
[d] Department of Mathematics and Computer Science, Friedrich Schiller University, Ernst-Abbe-Platz 1-4, 07743 Jena, Germany

ARTICLE INFO

Article history:
Received 15 January 2013
Received in revised form 18 April 2013
Accepted 25 April 2013

Keywords:
Random uncertainty
Net ecosystem productivity
Evapotranspiration
Eddy covariance
Annual sums
Gaussian Process
Machine Learning

ABSTRACT

The flow of carbon between terrestrial ecosystems and the atmosphere is mainly driven by nonlinear, complex and time-lagged processes. Understanding the associated ecosystem responses is a key challenge regarding climate change questions such as the future development of the terrestrial carbon sink. However, high temporal resolution measurements of ecosystem variables (with the eddy covariance method) are subject to random error, that needs to be accounted for in model-data fusion, multi-site syntheses and up-scaling efforts.

Gaussian Processes (GPs), a nonparametric regression method, have recently been shown to capture relationships in high-dimensional, nonlinear and noisy data. Heteroscedastic Gaussian Processes (HGPs) are a specialized GP method for data with inhomogeneous noise variance, such as eddy covariance measurements.

Here, it is demonstrated that the HGP model captures measurement noise variances well, outperforming the model residual method and providing reasonable flux predictions at the same time. Based on meteorological drivers and temporal information, uncertainties of annual sums of carbon flux and water vapor flux at six different tower sites in Europe and North America are estimated. Similar noise patterns with different magnitudes were found across sites. Random uncertainties in annual sums of carbon fluxes were between 9.80 and 31.57 $\mathrm{g\,C\,m^{-2}\,yr^{-1}}$ (or 4–9% of the annual flux), and were between 2.54 and 8.13 $\mathrm{mm\,yr^{-1}}$ (or 1–2% of the annual flux) for water vapor fluxes. The empirical HGP model offers a general method to estimate random errors at half-hourly resolution based on entire annual records of measurements. It is introduced as a new tool for random uncertainty assessment widely throughout the FLUXNET network.

1. Introduction

Modeling biosphere-atmosphere interactions on the basis of ecosystem measurements is essential for a better understanding of the global carbon cycle. Rising concentrations of greenhouse gases in the atmosphere have been attributed to industrialization, human development and the resulting combustion of fossil fuels and changes in land use. The concentration of carbon dioxide (CO_2) is currently the highest it has been in the last 650,000 years (Siegenthaler et al., 2005). The terrestrial biosphere strongly influences the global carbon cycle by sequestering carbon via photosynthesis while simultaneously releasing carbon via respiration into the atmosphere. Its total annual uptake of carbon is 123 ± 8 Gt (Beer et al., 2010) and some terrestrial ecosystems act as long-term carbon sinks.

Providing high resolution measurements of the net CO_2, H_2O and energy fluxes, the FLUXNET observational network (Baldocchi, 2008) is a fundamental data source toward identifying the contributions of various ecosystems to the global carbon sink. The CO_2 exchange between biosphere and atmosphere is the difference between carbon assimilation by photosynthesis (gross primary production, *GPP*) and release of carbon to the atmosphere (ecosystem respiration, *ER*). In this study, focusing on forest ecosystems, we note the ecosystem net CO_2 flux as Net Ecosystem Production (*NEP*), which is opposite in sign to Net Ecosystem Exchange ($NEP = -NEE$). Positive *NEP* fluxes correspond to a net uptake of CO_2 by the biosphere and negative *NEP* fluxes stand for a net release of CO_2 to the atmosphere. The water vapor flux is hereafter denoted as *LE* (latent heat flux).

* Corresponding author. Present address: University of California, Santa Barbara, United States.
E-mail address: omenzer@bgc-jena.mpg.de (O. Menzer).

0168-1923/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.agrformet.2013.04.024

Data sets obtained by the eddy covariance method have the following properties: multidimensional, fragmented and noisy (Moffat et al., 2010). The noise in the flux data is caused by different sources of error. On one hand, there are *random errors*, which are due to measurement instrument errors, the stochastic nature of turbulence, and the varying footprint of the towers. On the other hand, flux measurements are subject to *systematic errors*, which are caused by inaccurate calibration, and measurements under conditions violating the assumptions of the eddy covariance method, including problems related to advection and non-flat terrain (Aubinet et al., 2012). While systematic errors are generally addressed through data filtering (such as u*-filtering), all data still suffers from random measurement error (Hollinger and Richardson, 2005). It has been noted that in practice it can be hard to disentangle random and systematic errors, since many errors have both a random and systematic component and operate at varying time scales (Richardson et al., 2012). Even under the assumption that all systematic errors have been removed from the data it is known that random errors are considerably large over the short averaging periods of the high-frequency measurements (i.e., minutes or hours) and cannot be ignored even on the annual timescale (Richardson et al., 2012).

Uncertainty estimates are needed to quantify the mismatch between models and data, which is essential for model optimization, ecosystem model validation against flux data, multi-site syntheses and regional-to-continental integration efforts (Raupach et al., 2005; Richardson et al., 2008). Moncrieff et al. (1996) has suggested that eddy covariance studies should always report mean fluxes together with their systematic and random uncertainty components by convention. If uncertainties are not considered in a statistically rigorous and transparent manner, model structures and model predictions are less reliable or might not be reached at all. Recently, a code of best practice for an open assessment of uncertainties in model-data fusion techniques has been proposed (Keenan et al., 2011), advising to evaluate data errors in an open and realistic way and feed forwarding them into the model framework. To start with, random error estimates should be propagated through gap-filling and partitioning fluxes into their components, in particular for carbon and water vapor fluxes. Currently, methods of assessing measurement error differ from site to site and there is no single one method that is widely used throughout the FLUXNET network.

Numerous studies have made efforts to estimate random errors in eddy covariance measurements. The paired tower method (Hollinger et al., 2004) uses simultaneous measurements of two towers separated by only a few hundred meters, in the same ecosystem but with nonoverlapping footprints. The between tower variability was found to be lower than the interannual variability in *NEP*. Uncertainties of the flux data could therefore be estimated by the standard deviations of the differences between the two towers and CO_2 flux random errors were shown to vary seasonally. Since this method cannot be applied widely throughout the network, a 24-h differencing approach (Hollinger and Richardson, 2005; Richardson et al., 2006) was developed on the same idea of paired observations, but with "time traded for space", i.e., flux measurements from a single tower are compared on two successive days at exactly the same time of day. Nearly identical environmental conditions are assured by fixed thresholds of meteorological variables, but are frequently not met, and therefore the number of data points for an analysis is reduced. Key results are that the random error follows rather a double-exponential than a normal distribution and it increases as a linear function of the flux magnitude (i.e., a heteroscedastic noise variance).

The model residuals method (Richardson and Hollinger, 2005; Richardson et al., 2008) infers the first four moments of the error distributions from the model residuals (the difference between predictions from a model and measured fluxes). The result of heteroscedastic errors was confirmed and a spectral analysis of the model predictions suggested autocorrelated model residuals, which exhibit site-specific differences. Lasslop et al. (2008) estimates random errors based on the standard deviations of the moving windows and model residuals produced from the MDS (marginal distribution sampling) gap-filling algorithm of Reichstein et al. (2005) and the autocorrelation of the random errors was found to be usually below 0.6 at a lag of 0.5 h.

While the methods above have in common that they estimate the total flux random error, micrometeorologists have made efforts to estimate the different sources of random error in flux measurements one by one. For example, turbulence sampling errors (Lenschow et al., 1994; Finkelstein and Sims, 2001) are attributed mainly to incomplete sampling of large eddies, which result in 30 minute average fluxes that are not representative of the actual flux. There are more recent methods tackling this problem, such as a spatial filter method (Salesky et al., 2012). The random error component resulting solely from instrument error can be estimated by using data from towers that are in the vicinity of each other (Eugster et al., 1997; Dragoni et al., 2007; Schmidt et al., 2012), similar to the paired tower approach (Hollinger et al., 2004). Often the resulting instrument error estimates are hard to distinguish from other sources of error and also assume that both flux systems are viewing identically functioning footprints. Other studies imply that instrument noise does not have an impact on random error of carbon dioxide fluxes (Rannik et al., 2006). In general, variable source areas of fluxes within individual half-hours due to changes in wind direction and velocity can add to the random error on top of sampling and instrument errors. Billesbach (2011) provides a good review and comparison of several of the aforementioned methods and suggests a new method for estimating instrument noise called random shuffle, which shuffles one of the flux covariates in time and attributes the remaining covariance to instrument noise.

Here, the ability of supervised Machine Learning (ML) methods as a tool to estimate total random errors in flux measurements is explored. ML methods allow to extract ecosystem response mechanisms and associated uncertainties directly from ecosystem data. The term "Machine Learning" refers to a set of methods for the generation of mathematical algorithms by learning characteristics and patterns through generalizing from sample data. Probably the most widely used ML algorithms in the eddy covariance literature are different forms of the Artificial Neural Networks method (Papale and Valentini, 2003; Schmidt et al., 2008; Moffat et al., 2010). Another one of the ML methods, Gaussian Processes (GPs), was shown to be applicable as a nonlinear regression tool on multivariate, noisy data sets with different degrees of nonlinearity (Rasmussen, 1996). More recently, GPs have been applied in biological or financial models (Gao, 2004; Sun et al., 2010; Macke et al., 2011). Heteroscedastic Gaussian Processes (HGPs) are a specialized GP method for data with inhomogeneous noise variance, as often observed in eddy covariance measurements. The objective of this study is to assess HGPs as a tool to explore ecological data sets, focusing at the ability to estimate data uncertainties, and consequently, improve the understanding and estimation of the total random error in eddy covariance measurements.

2. Materials and methods

2.1. Data and site descriptions

This study is based on half-hourly eddy covariance and meteorological data from six different measurement sites taken from the CarboeuropeIP database and from the AmeriFlux database. All sites are forests including different types such as deciduous