Geoderma 181–182 (2012) 91–101

Contents lists available at SciVerse ScienceDirect

Geoderma

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

Analyzing spatiotemporal variability of heterotrophic soil respiration at the field scale using orthogonal functions

Alexander Graf [a,*], Michael Herbst [a], Lutz Weihermüller [a], Johan A. Huisman [a], Nils Prolingheuer [a], Ludger Bornemann [b], Harry Vereecken [a]

[a] Agrosphere (IBG-3), Jülich Research Centre, 52425 Jülich, Germany
[b] Institute of Crop Science and Resource Conservation – Soil Science, Bonn, Germany

ARTICLE INFO

Article history:
Received 11 November 2010
Received in revised form 2 November 2011
Accepted 12 February 2012
Available online 6 April 2012

Keywords:
Closed chamber
Empirical orthogonal functions
Principal component analysis
Semivariogram
Soil CO_2 efflux

ABSTRACT

Soil CO_2 efflux was measured with a closed chamber system along a 180 m transect on a bare soil field characterized by a gentle slope and a gradient in soil properties at 28 days within a year. Principal component analysis (PCA) was used to extract the most important patterns (empirical orthogonal functions, EOFs) of the underlying spatiotemporal variability in CO_2 efflux. These patterns were analyzed with respect to their geostatistical properties, their relation to soil parameters obtained from laboratory analysis, and the relation of their loading time series to temporal variability of soil temperature and moisture. A particular focus was set on the analysis of the overfitting behaviour of two statistical models describing the spatiotemporal efflux variability: i) a multiple regression model using the k first EOFs of soil properties to predict the n first EOFs of efflux, which were then used to obtain a prediction of efflux on all days and points; and ii) a modified multiple regression model based on re-sorting of the EOFs based on their expected predictive power. It was demonstrated that PCA helped to separate meaningful spatial correlation patterns and unexplained variability in datasets of soil CO_2 efflux measurements. The two PCA analyses suggested that only about half of the total variance of efflux could be related to field-scale spatial variability of soil properties, while the other half was "noise" attributed to temporal fluctuations on the minute time scale and short-range spatial heterogeneity on the decimetre scale. The most important spatial pattern in CO_2 efflux was clearly related to soil moisture and the driving soil physical properties. Temperature, on the other hand, was the most important factor controlling the temporal variability of the spatial average of soil respiration.

1. Introduction

CO_2 efflux from the soil is one of the largest fluxes in the atmospheric greenhouse gas balance and of particular interest due to its potential positive feedback to global warming (IPCC, 2007). However, the environmental factors controlling the magnitude of CO_2 efflux remain difficult to disentangle, even though an increasing number of case studies have been published during past decades (for an overview, see e.g. Bond-Lamberty and Thomson, 2010). The reasons for this are rooted in the numerous interactions between environmental factors and CO_2 efflux, in combination with the different scales on which they vary in space and time (Briones, 2009; Davidson and Janssens, 2006; Mahecha et al., 2010; Wixon and Balser, 2009). Point measurements of soil CO_2 efflux have repeatedly been reported to exhibit a poor spatial dependence and strong variability at short distances (i.e. a high nugget effect, see Herbst et al., 2009; La Scala et al., 2000; Rochette et al., 1991; Rodeghiero and Cescatti, 2008). Consequently, correlations with expected driving variables in space appear to be low (Herbst et al., in press).

We hypothesize that the difficulty of understanding the driving factors of spatial variability of soil respiration is partly caused by short-term temporal fluctuations that inevitably occur during the acquisition of a spatial data set of soil CO_2 efflux. Recently, we showed that by repeating a survey with a sufficiently high frequency, the raw measurements can be decomposed by simple averaging procedures into estimates of the time-stable part of the spatial pattern of efflux, and fast fluctuations of area-averaged efflux (Graf et al., 2011). However, this study also reported that the spatial patterns were only stable for a few days. Often, measurements are only repeated at larger time intervals and the decomposition approach reported in Graf et al. (2011) cannot meaningfully be applied. Alternatively, underlying spatial patterns present in the entire data set can be investigated using empirical orthogonal functions (EOFs) derived by principal component analysis (PCA, cf. Korres et al., 2010; Perry and Niemann, 2008, for soil water content) or canonical correlation analysis. These EOFs can be related to explanatory variables such as the spatio-temporal variability of soil properties, including soil temperature and moisture amongst others. Unlike classical regression, which would link the spatial pattern of efflux to explanatory variables independently for each snapshot in time, PCA provides insight into the combined spatiotemporal dependencies of soil CO_2 efflux. PCA has also been shown to efficiently

* Corresponding author. Tel.: +49 2461 61 8676; fax: +49 2461 61 2518.
E-mail address: a.graf@fz-juelich.de (A. Graf).

0016-7061 © 2012 Elsevier B.V. Open access under CC BY-NC-ND license.
doi:10.1016/j.geoderma.2012.02.016

separate noise from signal (e.g. Perry and Niemann, 2008), a property that would be of particular interest for soil CO_2 efflux datasets.

The aim of this study is to test whether PCA can be used to identify spatio-temporal patterns of soil CO_2 efflux with statistically significant relations to explanatory variables. We used measured CO_2 efflux from a bare soil with a gentle slope and a gradient in various soil properties, which serve as explanatory variables. To identify and describe returning spatial patterns in the efflux time series and the explanatory variables without redundancy, both datasets are turned into EOFs independently using PCA. Two types of regression models are considered: one for the two sets of EOFs in their original order, and one where EOFs are re-ordered according to their expected predictive power. For each regression model, performance on unknown samples as a function of the number of EOFs was determined by cross-validation.

2. Theoretical background

Consider the dependent variable $Y_{m,n}$ known for M sampling points and N measurement times. In addition, there are K explanatory variables contained in $X_{m,k}$. These explanatory variables vary in space, but are assumed to be persistent in time. Therefore, they are available at the same M sampling points, but without repetition in time. To determine to what extent Y can be explained by X, canonical correlation analysis (CCA, Hotelling, 1935) is frequently used. However, standard methods for solving CCA require that $min\{N, K\}<M$. If this is not the case, a common approach is to perform a principal component analysis (PCA, Hotelling, 1933) independently on both X and Y before further analysis (Muller, 1982). PCA transforms a set of variables into a set of new variables, called principal components (PCs) or empirical orthogonal functions (EOFs), that are linearly independent of each other. They are ordered by the portion of total variance in the original data that they explain (see appendix for more details). If N or K is larger than M, PCA reduces the number of non-zero new variables to M. Prior application of PCA on both X and Y reduces a subsequent CCA to a rotation (Muller, 1982). Often, the CCA step is omitted altogether and the prediction of Y from X is done by regression. Because the EOFs determined from X and Y are orthogonal, the regression coefficients can be independently determined by bivariate regression between each possible pair of EOFs. This intermediate approach between multiple regression and CCA (Jolliffe, 1982) is here referred to as PCA-based regression.

If the number of explanatory variables K is large compared to the number of sampling points M, there is a danger of overfitting. Adding an additional explanatory variable will always improve the ability of the model to fit the data (in-sample performance). However, overfitting has occurred when at the same time the ability of the model to predict independent data decreases (out-of-sample performance). In multiple regression, adjusted goodness-of-fit indices such as R^2_{adj} or Aikake´s information criterion are often used to estimate the optimum number of explanatory variables (e.g. Herbst et al., in press), or a significance test is performed for each candidate explanatory variable. For EOFs, a number of significance tests have been suggested. However, their results are often inconsistent (Peres-Neto et al., 2005; Perry and Niemann, 2008), may require prior knowledge of the correlation length in order not to overestimate the number of independent samples (Korres et al., 2010), and are not necessarily related to predictive power. Jolliffe (1982) summarized four examples demonstrating that predictive success, rather than explained variance, should be used to determine the EOFs to be included in PCA-based regression problems. Nevertheless, and in particular to ensure the relevance of the predicted EOFs of Y, we will report results of two significance tests for comparison. According to Peres-Neto et al. (2005), both are recommendable, but differently conservative.

The most direct, assumption-free, and intercomparable method to estimate out-of-sample performance, is cross-validation. A subset of the available data is excluded before parameter determination, and the goodness-of-fit indices are calculated between the predictions and measurements of Y in this unused subset only. A prerequisite for cross-validation is that the independent data set must be large enough to reliably determine the goodness-of-fit indices, but at the same time the data set remaining for model parameterisation must also be large enough. In case of a small number of sampling sites, this problem can be circumvented by the leave-one-out version of cross validation. One at a time, each of the M rows of X and Y are removed from the dataset, and the remaining M-1 rows are used to estimate the unknown model parameters. Then, each of the M alternative model versions is used to predict the row of Y values that was left out. Leave-one-out cross-validation enables us to quantify the effect of including each EOF of both the X and the Y set in the regression model, starting with the first EOF. As an EOF of X may describe a large portion of the variance of X, but not predict well any of the EOFs of Y (Jolliffe, 1982), we also test an approach where the EOFs of both X and Y are re-sorted according to the amount of variance in Y that they help to explain. This approach adds the strength of CCA to PCA-based regression, while avoiding its predictive weakness. CCA tends to assign strong weights to few or even one X and Y pair(s), if they are correlated considerably stronger to each other than the majority, independent of the portion of variance in Y they explain (Mishra, 2009). An intermediate solution between PCA and CCA was proposed by Mishra (2009) to solve this problem, but the application of this method is beyond the scope of this study because of the lack of a closed-form solution for this method. We performed CCA on our dataset and found that it did not improve out-of-sample performance as compared to PCA-based regression. For reasons of conciseness, CCA is not discussed further here.

3. Methods

3.1. Study site

Measurements were taken at the FLOWATCH test site (50°52'09''N, 06°27'01''E, 104.5 m a.s.l.), a 60 m by 190 m bare soil field (Graf et al., 2008; Weihermüller et al., 2007). In its longitudinal direction, the field is subject to a gentle slope and a strong gradient in coarse material content (Fig. 1). At the centre of the field, the fine texture (<2 mm) is classified as a silt loam. The climate is warm temperate, with an average air temperature of 9.9 °C and an annual precipitation of 698 mm (1961–2009, data taken from the climate station of the Forschungszentrum Jülich at a distance of 5.3 km from the test site). The two years of the experimental study were slightly warmer and wetter (2006: 11.0 °C, 723 mm; 2007: 11.1 °C, 878 mm). Historically, the field was typically ploughed annually up to a depth of 30 cm. Directly before and once during the study period, a grubber to a depth of 15 cm and a harrow were applied. With this treatment and several applications of glyphosate, weeds were controlled on the field site during our measurements.

3.2. Field measurements

Soil CO_2 efflux measurements were performed using a manual closed chamber system (LI-8100, Li-Cor, Lincoln, NE, USA; Xu et al., 2006) in intervals of one to two weeks between summer 2006 and autumn 2007. At each measurement point, a polypropylene collar of 10 cm depth and 20 cm inner diameter was permanently installed such that the upper edge protruded 2 cm above the average soil surface. Collars were kept free of plants as much as possible and were removed only for soil grubbering and harrowing. The location of each measurement point was determined using a differential GPS system (GPS-702-GG/Propak V3, NovAtel, Calgary, Alberta, Canada).

In this analysis, we use efflux data from 18 points spaced 10 m apart in a transect following the main height and stone content gradient of the field site (Fig. 1). For this transect, complete efflux records