Comparing observations and process-based simulations of biosphere-atmosphere exchanges on multiple timescales

M. Mahecha, M. Reichstein, M. Jung, S. Seneviratne, S. Zaehle, C. Beer, M. Braakhekke, N. Carvalhais, H. Lange, G. Le Maire, et al.

JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 115, G02003, doi:10.1029/2009JG001016, 2010

Comparing observations and process-based simulations of biosphere-atmosphere exchanges on multiple timescales

M. D. Mahecha,[1,2] M. Reichstein,[1] M. Jung,[1] S. I. Seneviratne,[2] S. Zaehle,[1] C. Beer,[1] M. C. Braakhekke,[1] N. Carvalhais,[1,3] H. Lange,[4] G. Le Maire,[5,6] and E. Moors[7]

Received 27 March 2009; revised 1 November 2009; accepted 16 November 2009; published 9 April 2010.

[1] Terrestrial biosphere models are indispensable tools for analyzing the biosphere-atmosphere exchange of carbon and water. Evaluation of these models using site level observations scrutinizes our current understanding of biospheric responses to meteorological variables. Here we propose a novel model-data comparison strategy considering that CO_2 and H_2O exchanges fluctuate on a wide range of timescales. Decomposing simulated and observed time series into subsignals allows to quantify model performance as a function of frequency, and to localize model-data disagreement in time. This approach is illustrated using site level predictions from two models of different complexity, Organizing Carbon and Hydrology in Dynamic Ecosystems (ORCHIDEE) and Lund-Potsdam-Jena (LPJ), at four eddy covariance towers in different climates. Frequency-dependent errors reveal substantial model-data disagreement in seasonal-annual and high-frequency net CO_2 fluxes. By localizing these errors in time we can trace these back, for example, to overestimations of seasonal-annual periodicities of ecosystem respiration during spring greenup and autumn in both models. In the same frequencies, systematic misrepresentations of CO_2 uptake severely affect the performance of LPJ, which is a consequence of the parsimonious representation of phenology. ORCHIDEE shows pronounced model-data disagreements in the high-frequency fluctuations of evapotranspiration across the four sites. We highlight the advantages that our novel methodology offers for a rigorous model evaluation compared to classical model evaluation approaches. We propose that ongoing model development will benefit from considering model-data (dis)agreements in the time-frequency domain.

Citation: Mahecha, M. D., et al. (2010), Comparing observations and process-based simulations of biosphere-atmosphere exchanges on multiple timescales, *J. Geophys. Res.*, *115*, G02003, doi:10.1029/2009JG001016.

1. Introduction

[2] Understanding the dynamics of CO_2, H_2O, and energy exchange between the terrestrial biosphere and atmosphere is essential for gaining insight to earth system functioning, and prerequisite for projecting its behavior in the near future [*Barford et al.*, 2001; *Schimel et al.*, 2001; *Ciais et al.*, 2005; *Seneviratne et al.*, 2006a; *Bonan*, 2008; *Heimann and Reichstein*, 2008]. In this context, state of the art diagnostic, empirical or process oriented terrestrial biosphere models are indispensable for analyzing greenhouse gas fluxes in time over geographical space [*Cramer et al.*, 2001; *Friend et al.*, 2007].

[3] One prerequisite for the integrative analysis of modeled spatiotemporal biosphere-atmosphere fluxes is the comparison of different model runs [*Vetter et al.*, 2008; *Jung et al.*, 2008]. Conducting comparative investigations of terrestrial biosphere models ideally reveals the effects of different model structures [*Richardson et al.*, 2006], parameterizations [*Braswell et al.*, 2005; *Zaehle et al.*, 2005], or initial conditions [*Carvalhais et al.*, 2008]. The assessment of the strength and pitfalls of terrestrial biosphere models itself requires accurate qualitative and quantitative site level evaluations [*Baldocchi and Wilson*, 2001; *Moorcroft*, 2006; *Siqueira et al.*, 2006; *Jung et al.*, 2007]. These analyses are essentially pattern oriented model-data comparisons, where the validity of terrestrial biosphere models is indirectly challenged [*Rykiel*, 1996; *Savenije*, 2009]. Model-data comparisons can be very instructive for understanding the behavior of both models and data in a joint perspective [*Betts*, 2004; *Jaeger et al.*, 2009].

[1] Biogeochemical Model-Data Integration Group, Max-Planck-Institut für Biogeochemie, Jena, Germany.
[2] Department of Environmental Sciences, ETH Zurich, Zurich, Switzerland.
[3] Faculdade de Ciência e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal.
[4] Norsk Institutt for Skog og Landskap, Ås, Norway.
[5] CIRAD, UPR Fonctionnement et pilotage des écosystèmes de plantations, UPR 80, Montpellier, France.
[6] LSCE, UMR, CEA-CNRS-UVSQ, Gif-sur-Yvette, France.
[7] Alterra, Wageningen University and Research Centre, Wageningen, Netherlands.

0148-0227/10/2009JG001016

[4] Nowadays, the "Eddy Covariance" technique has become a standard for the in situ monitoring of CO_2 and H_2O fluxes [*Aubinet et al.*, 2000]. Regional to global eddy covariance data compilations (especially FLUXNET), play an important role for our general understanding of ecosystem responses to atmospheric forcing, and serve as an invaluable basis for model performance evaluations [*Baldocchi et al.*, 2001a; *Baldocchi*, 2008]. The increasing availability of eddy covariance data has led to a number of site level model-data comparisons where several models were driven by site meteorology [e.g., *Medlyn et al.*, 2005; *Morales et al.*, 2005; *Kucharik et al.*, 2006; *Siqueira et al.*, 2006].

[5] The simplest way to summarize model-data disagreement is using scalar error estimates. Applying error metrics or quantifying flux biases up to annual flux integrals provides only limited insight into the quality of a model. Depending on the objective, however, these "single misfit numbers" [*Evans*, 2003] might be very useful. More sophisticated studies attempt to localize model-data mismatches in time [*Gulden et al.*, 2008]. Intuitively, this can be achieved by moving window approaches or by varying aggregation levels. The ideas behind the latter strive for the identification of timescales of acceptable model performances [*Abramowitz et al.*, 2008]. Classical model performance evaluation usually also embraces residual analysis [*Medlyn et al.*, 2005]. The rationale behind this is that the coincidence of patterns in the residuals and certain environmental conditions elucidate potential problems of terrestrial biosphere models. Independently of the chosen approach for model-data comparison, the natural limit is set by the influence of both random and systematic errors on the eddy covariance data [*Medlyn et al.*, 2005; *Friend et al.*, 2007]. In the overall view, however, the methodology of model performance evaluation has not progressed substantially over the last decades [*Janssen and Heuberger*, 1995; *Rykiel*, 1996; *Medlyn et al.*, 2005; *Abramowitz et al.*, 2008]. *Gulden et al.* [2008] made the point that "the development of robust metrics for comprehensive model evaluation is in its infancy."

[6] Given this background, the present study revisits the issue of model-performance evaluation taking into consideration that ecosystem-atmosphere fluxes are shaped by a variety of fluctuations on different scales of characteristic variability. The (observable) variability in eddy covariance time series ranges from hourly, diurnal, synoptic, seasonal-annual, to decadal periodicities [*Katul et al.*, 2001; *Stoy et al.*, 2005; *Mahecha et al.*, 2007; *Qin et al.*, 2008; *Stoy et al.*, 2009]. The question of whether state of the art terrestrial biosphere models reproduce these properties has been addressed in the frequency domain by *Braswell et al.* [2005] and *Siqueira et al.* [2006]. These studies show that model-data agreement is a matter of frequency and illustrate that residual time series of CO_2 exchange fluxes systematically contain high relative spectral powers for intermediate frequencies: model-data disagreement affects especially periodicities above diurnal and below annual variability, the "spectral gap" where a minor part of the total variance is allocated [*Baldocchi et al.*, 2001b; *Stoy et al.*, 2005]. The efficiency of a spectral analysis lies in the potential to summarize patterns corresponding to different frequency scales in very few (temporally) global coefficients. On one hand, this leads to a refined analysis compared to the use of scalar error estimates, but on the other hand this does not resolve the shortcoming of being a nonlocal analysis.

[7] This paper extends the available set of tools for model-data comparisons by providing a perspective for time-frequency localized performance evaluations. We separate observed and simulated fluxes into subsignals shaped by characteristic frequencies prior to model-data comparisons. Focusing on fluctuations (amplitude modulations) corresponding to certain scales of variability allows addressing the question of whether pronounced disagreement of state-of-the-art terrestrial biosphere models and eddy covariance data occurs on specific scales of variability within specific time periods.

2. Simulated and Observed Data

[8] Terrestrial biosphere models represent interactions between ecosystems and the lower boundary layer of the atmosphere. Model-data comparisons therefore often focus on net land-atmosphere exchanges of CO_2, H_2O, or energy. Along these lines, special emphasis is put here on the behavior of "Net Ecosystem Exchange, *NEE*." It is the balance of CO_2 uptake, the "Gross Ecosystem Exchange, *GEE*," and the overall release of CO_2, the "Terrestrial Ecosystem Respiration, *TER*." Following micrometeorological convention *GEE* is denoted as a negative flux,

$$NEE = GEE + TER, \tag{1}$$

where

$$TER = R_A + R_H. \tag{2}$$

Here, R_A and R_H are autotrophic and heterotrophic respiration, respectively. Furthermore, water fluxes are considered in terms of "Actual Evapotranspiration, *AET*." Model-based predictions of these fluxes are compared to their counterparts derived from eddy covariance measurements.

[9] Two well established and widely applied terrestrial biosphere models were analyzed here: ORCHIDEE [*Krinner et al.*, 2005] and LPJ [*Sitch et al.*, 2003]. Both are "big leaf models" which means that the canopy is treated as a single compartment. Biomass is further subdivided in functional tissue pools where carbon allocation strategies are parameterized according to the plant functional types under investigation. A comparison of their performance is interesting because LPJ and ORCHIDEE represent each a major group of terrestrial biosphere model with a focus on biogeographical and biogeophysical aspects, respectively. The two approaches differ in model structure and parameterization, and for instance employ different approaches to numerical solution. These differences may lead to divergent simulation results as discussed hereafter, even though the models share a common set of fundamental hypothesis about critical ecosystem processes. The two models vary substantially in runtime behavior and structural complexity, which is partly due to the fact that ORCHIDEE runs on half hourly time steps, while LPJ is based on daily values.

2.1. Modeled and Observed *NEE*

[10] Terrestrial biosphere models provide net carbon fluxes as the sum of the opposing fluxes *TER* and *GEE* (equation (1)), which are explicitly modeled (see below).

From an experimental perspective, *NEE* is the only carbon flux that can be directly observed at ecosystem level: For each site, vertical wind velocity components and CO_2 concentrations are available based on a 20 Hz sampling frequency. Processing these data by the eddy covariance method (as described in the EUROFLUX methodology) [*Aubinet et al.*, 2000] leads to half hourly estimates for the CO_2 flux density, *NEE* [see also *Foken*, 2008a]. These data underwent canopy storage corrections and u^* filtering as described by *Papale et al.* [2006]. An initial gap filling of the half hourly data was realized based on time local empirical flux estimates (MDS [cf. *Reichstein et al.*, 2005]). The time series were then further aggregated to daily flux estimates in order to realize the model-data comparisons on equal sampling frequencies (see section 2.6). However, only high-quality daily data were used in the aggregation step: If a daily aggregate had to be estimated from less than 43 out of 48 original half hourly flux estimates, the datum was treated as missing value and filled by means of "Singular System Analysis" (see Appendix B).

2.2. Modeled and Observed *GEE*

[11] In general, terrestrial biosphere models simulate carbon assimilation according to the original *Farquhar et al.* [1980] model or the simplification by *Collatz et al.* [1992]. From an ecophysiological perspective, canopy conductance g_c is the key linkage between carbon and water fluxes. Understanding and modeling this feedback system is intricate and solved very differently in the models. The principle however, consists in estimating canopy conductance based on an initial g_c scaled by incident radiation, vapor pressure deficit, leaf water potential, and minimum nighttime temperature [*Jarvis*, 1976]. Temperature, atmospheric drying power, and soil moisture, but also mutual dependencies between g_c and carbon assimilation play a fundamental role. ORCHIDEE then uses the empirical "Ball-Berry" relation [*Ball et al.*, 1987; *Collatz et al.*, 1991] (also known as "Ball-Woodrow-Berry" approach) where the assimilation rate determines g_c which in turn affects the assimilation rate by modeling the leaf CO_2 concentration. LPJ employs a similar concept, but assumes a fixed ratio of ambient to leaf CO_2 concentration along with an empirical boundary layer description for the atmospheric coupling [*Haxeltine and Prentice*, 1996]. In both models g_c is additionally linked to soil matrix potential and a soil water stress factor.

[12] "Gross primary productivity" ($GPP = -GEE$) further depends in both models on the absorbed photosynthetically active radiation (*APAR*) and thus on the state of plant phenology. Phenology in terrestrial biosphere models is conventionally formulated as a thresholding problem characteristic to each plant functional type. The parameterization is based on heuristics, for instance regarding the prescription of growing degree days, and ORCHIDEE is additionally calibrated against satellite observations.

[13] Although *GEE* cannot be directly observed, it can be partitioned from the observed (half hourly) *NEE*. Flux partitioning exploits the fact that nighttime *NEE* is purely attributable to respiratory processes. Short-term exponential temperature dependence of the available turbulent nighttime CO_2 exchange is extrapolated from nighttime to daytime *TER* (for details, see *Reichstein et al.* [2005] and equation (1)) leads to a data driven estimate for *GEE*. The aggregation step to daily *GEE* estimates was realized in the same way as for *NEE*.

2.3. Modeled and Observed *TER*

[14] While R_A and R_H can hardly be distinguished experimentally, it is a standard to model these processes separately. R_A is further differentiated into ecosystem carbon losses due to maintenance ($R_{A,\mathrm{m}}$) and growth costs ($R_{A,\mathrm{g}}$). Biomass (and other factors such as the characteristic C:N ratio) determine $R_{A,\mathrm{m}}$ which is furthermore a function of temperature. The latter is represented as an exponential in LPJ, whereas ORCHIDEE assumes a linear dependency. $R_{A,\mathrm{g}}$ is generally a constant fraction (LPJ: 0.25, ORCHIDEE: 0.28) of *GPP* reduced by $R_{A,\mathrm{m}}$.

[15] Heterotrophic respiration (R_H) is also a composite term, integrating the efflux from different soil organic matter (SOM) pools. Model structures differ in the number of soil C and litter pools, the description of SOM biogeochemistry, i.e., in pool specific decay rates. In addition, the linkage and sensitivity of the different SOM pools varies substantially. LPJ simulates the dynamics of SOM pools by different decomposition sensitivities where R_H is temperature driven as described by *Lloyd and Taylor* [1994]. ORCHIDEE follows mainly an arctangent function [*Parton et al.*, 1988] to describe R_H as a function of temperature, and has a more differentiated architecture of SOM pools. In both models the decomposition of soil organic C also depends on the soil water content.

[16] As described for *GEE*, the component fluxes cannot be directly monitored and "observed" *TER* is estimated in tandem with the flux separation. Again, the aggregation step to daily *TER* estimates was realized in the same way as for *NEE*.

2.4. Modeled and Observed *AET*

[17] Also water fluxes are not attributable to unique processes. Rather, actual evapotranspiration (*AET*) is a composite flux integrating transpiration which is partly controlled by canopy conductance, soil evaporation, and interception losses. In ORCHIDEE total evapotranspiration is derived according to the submodel SECHIBA [*Ducoudre et al.*, 1993], which calculates the half-hourly energy and water balance of vegetated and nonvegetated surfaces. Evapotranspiration is here composed from semiempirical descriptions of interception losses, soil evaporation and transpiration. The LPJ model uses the empirical Penman-Monteith combination formulation [*Monteith*, 1995].

[18] The eddy covariance system allows to directly estimate evapotranspiration along with the CO_2 exchange, where the covariate is water vapor density (instead of CO_2). These data are then processed in accordance with the description for *NEE* (see above).

2.5. Site Selection

[19] The site level model-data comparison focuses on four CarboEurope-IP eddy covariance towers that cover a reasonable range of European forest ecosystems: Hainich, Germany (temperate broadleaf), Hyytiälä, Finland (boreal coniferous), Loobos, Netherlands (temperate coniferous), and Puechabon, France (Mediterranean evergreen broadleaf). Precise geographical coordinates, site characteristics, and references are summarized in Table 1.

Table 1. Details on the Four CarboEurope-IP Sites Used in the Present Study[a]

Site-Code	Name	Latitude	Longitude	Elevation	Instrument Height	Orography	Dominant Tree	Last Clear-Cut	Reference
DE-Hai	Hainich	51.0793	10.452	430	43.5	gently sloppy	*Fagus sylvatica* L.	1753	*Knohl et al.* [2003]
FI-Hyy	Hyytiälä	61.8474	24.2948	181	14.0	gently sloppy	*Pinus sylvestris* L., *Picea abies* (L.) Karsten	1967	*Suni et al.* [2003]
FR-Pue	Puechabon	43.7414	3.59583	270	6.5	flat	*Quercus ilex* L.	1942	*Rambal et al.* [2004]
NL-Loo	Loobos	52.1679	5.74396	25	27.0	flat	*Pinus sylvestris* L.	-	*Dolman et al.* [2002]

[a]More overview information is available at http://www.fluxdata.org:8080/SitePages/.

2.6. Simulation Protocol

[20] Models were driven with meteorological data measured on site and using prescribed soil water holding capacity data (based on *Granier et al.* [2007]). Plant functional types were defined according to the prevalent vegetation at the site (Table 1). Long-term daily meteorological time series for the model spin-up were generated by the reanalysis-driven product (NCEP-REMO) from *Feser et al.* [2001], harmonized with the site-level meteorology through a regression approach (see Appendix A). Consistently with previous studies [*Jung et al.*, 2007; *Vetter et al.*, 2008] the carbon pools were brought to steady state in relation to a constant atmospheric CO_2 concentration of 285.2 ppm (year 1850) by recycling over one decade of meteorological data (1958–1967). Subsequently, the terrestrial biosphere models were run from 1850–1957 using the same meteorological data but with measured CO_2 concentrations. The last transient phase (1958–2005) was based on observed meteorological data and CO_2 concentrations. The latter were derived from ice core data by *Etheridge et al.* [1996] and atmospheric observations by *Keeling and Whorf* [2005]. Given the strong sensitivity of ecosystem-atmosphere fluxes to the meteorological forcing, only periods when the terrestrial biosphere models were driven by effectively measured meteorology were considered in all subsequent analyses. Note that corrected NCEP-REMO data were only used for the spin-up runs and to keep the model running in the presence of measurement gaps. Site history in terms of management was not prescribed in the simulations.

[21] For the sake of comparability, the model runs were evaluated on a daily time step. LPJ is originally designed to model monthly flux values. However, since it uses internally daily time steps, only the driver data interface was modified for our purposes: daily meteorology measurements could be directly read in so that reliable daily flux estimates could be obtained. We used the models consistently with their application in the CarboEurope-IP project using the standard parameterizations.

3. Methods

3.1. Separating Subsignals of Characteristic Variability

[22] Observed and modeled time series can be described as sets of additively superimposed subsignals, and the assumption is that these subsignals are shaped by characteristic scales of variability. Any time series $Y = \{y_i\}$, where $i = 1, \ldots N$ is therefore denoted as the sum of its subsignals,

$$Y = \sum_{f=1}^{F} X_f, \tag{3}$$

where f is the index over the contained (and discretely separable) characteristic frequencies. Throughout this study, we use "Singular System Analysis" (SSA [*Broomhead and King*, 1986; *Elsner and Tsonis*, 1996; *Golyandina et al.*, 2001; *Ghil et al.*, 2002]) for extracting the subsignals X_f. SSA already proved to be well suited for exploring daily eddy covariance ecosystem-atmosphere fluxes [*Mahecha et al.*, 2007]. Since recent advances enable SSA applications to fragmented time series [e.g., *Kondrashov and Ghil*, 2006; *Golyandina and Osipov*, 2007], all technical prerequisites for SSA applications to data are fulfilled. Here, we only summarize the two step SSA principle; technical details can be found in Appendix B.

[23] 1. The first step is time series decomposition. Initially, time lagged windows of the time series Y are used to embed the series into its trajectory space, which is an application of Takens' embedding theorem [*Takens*, 1981]. The embedding space can be decomposed into underlying (orthogonal) features in terms of a "Principal Component Analysis in the time domain" [*Ghil et al.*, 2002]. This decomposition identifies a set of empirical orthogonal functions and associated principal components. Each component is usually shaped by one single oscillatory mode, and thus, has a very simple representation in the frequency domain. This allows assigning to each empirical orthogonal function a characteristic frequency as identified by the standard Fourier spectrum.

[24] 2. The second step is time series reconstruction. The time series is partly reconstructed from a set of principal components of the user's choice. Typically, a reconstruction is based on few selected components that are characterized by complementary frequencies. Thus, each time series can be finally described by a set of subsignals X_f each of which belongs to a well defined frequency bin.

[25] Two frequency binning schemes are chosen a priori: a coarse binning to five bands and a finer resolution comprising 10 frequency bins (Table 2). Except from the edge bins, both binning boundaries are approximately equally spaced over the logarithm of the frequencies. The heuristic binning schemes account for two desired properties: First, the bins coincide with scales that are accessible to an ecological discussion, e.g., in the fine binning scheme the day-to-day variability can be distinguished from synoptic variability or the annual cycle is separable from semiannual (= seasonal) components. Also in the coarse binning scheme, frequency ranges are met that are clearly interpretable. Second, the chosen bin widths are sufficiently coarse to avoid misinterpretations due to inaccuracies occurring in the frequency assignments to the SSA modes.

[26] We expect an improved inference on possibly inadequate parts of model structure since the range of thinkable

Table 2. Limits of the Two Applied Frequency Binning Schemes[a]

Bins I	Bins II	Upper Limit p[d] ≤ …	Lower Limit p[d] > …	Denotation
A	a	maximum	5137	low-frequency variability
B	b	5137	2593	annual cycle
	c	2593	1309	semiannual variability (seasonality)
C	d	1309	661	intermonthly variability
	e	661	334	
D	f	334	169	interweekly
	g	169	85	(monthly) variability
E	h	85	43	day-to-day
	i	43	22	(weekly) variability
	j	22	minimum	

[a]The discretization is approximately log-equidistant and provides the basis for all illustrations and analyses. The choice of the binning is a trade-off, taking into account the requirements for an ecological interpretation and the limitations in the frequency definition of the reconstructed components (in the SSA framework).

reasons of model-data disagreements might be considerably reduced if only selected frequency bins are investigated. In this respect, it is important to note that SSA is more than a filtering technique: Subsignal separation does not imply a loss of information, i.e., by choosing the full set of components a time series can be reconstructed entirely.

[27] It has been reported that despite of orthogonal base functions the accuracy of subsignals separability is not guaranteed [*Golyandina et al.*, 2001]. This methodological uncertainty has to be strictly distinguished from the effective model-data disagreement. Here, we quantify the separation inaccuracy by a surrogate technique (the "Iterative Amplitude Adjusted Fourier Transform" [*Schreiber and Schmitz*, 2000]): In brief, a set of surrogates is generated for each residual corresponding to an extracted subsignal of interest. Then the subsignal X_f is reextracted many times (20 times for the fine, and 500 times for the coarse binning scheme). Any subsignal is thus replaced by an array of subsignals, and their deviations quantify the extraction uncertainty. All analyses in this paper rely on this array instead of a single subsignal and form the basis for confidence envelopes for any estimated metric (for details, see Appendix B and Figure S1, available as auxiliary material).[1]

3.2. Error Spectra

[28] Subsignal separation is a prerequisite for the central step of the study: the qualitative and quantitative model-data comparison on different scales. Figure 1 shows conceptually that this leads to a model-data comparison for each defined frequency bin. This first part of the analysis replaces the "single misfit number paradigm" by an "error spectrum." Unlike established model-performance evaluations in the frequency domain [*Braswell et al.*, 2005; *Stoy et al.*, 2005; *Siqueira et al.*, 2006; *Richardson et al.*, 2007], the degree of misfit is still estimated in the time domain but accounting only for fluctuations within a well defined frequency range.

[29] For constructing error spectra we use simple, though robust, estimates: the Median Euclidean Error, *MEE*, its standardized counterpart, and the biweight midcorrelation, *R*. This choice is motivated by previous observations that many conventional measures, for instance the root mean squared error, are highly sensitive to outliers [*Li and Zhao*, 2006]. Insisting on robust properties of misfit estimates bears the risk of overly pessimistic model-data comparisons. However, this property increases the general credibility of the error spectra.

[30] In the analysis, it has to be taken into account that the subsignals are centered (and, if this is not precisely the case, the mean is removed). Thus, this study focuses exclusively on differences in amplitude modulation of the subsignals. The *MEE* quantifies the deviations in amplitude modulation misfit as follows,

$$MEE_f = M\left\{\left|X_{f,\mathrm{mod}} - X_{f,\mathrm{obs}}\right|\right\}, \tag{4}$$

where the index f indicates the frequency bin to which the subsignal X_f corresponds, and $M\{\cdot\}$ denotes the sample median. Standardizing the *MEE* by the standard deviation of the fluctuations in the frequency bin leads to the relative error estimate:

$$\mathrm{rel}.MEE_f = \frac{M\left\{\left|X_{f,\mathrm{mod}} - X_{f,\mathrm{obs}}\right|\right\}}{s\left\{X_{f,\mathrm{obs}}\right\}}, \tag{5}$$

where s is the standard deviation.

[31] Correlation coefficients investigate linear relations between the extracted subsignals. We apply the biweight midcorrelation coefficient [*Wilcox*, 2004] (see Appendix C) which behaves similarly to the classical variants: it is bounded in the range of −1 to 1. In the presence of outliers, however, the coefficient is superior compared to Pearson's product moment correlation and yields lower values, otherwise the estimators are equivalent.

[32] The global model performance measures benefit from the uncertainty assessment of time series decomposition: Each performance measure can be estimated for all combinations of the reextracted modeled and observed subsignals. The estimates can thus be characterized through their distributions. For the sake of interpretation, these distributions are summarized in violin plots [*Hintze and Nelson*, 1998]. A "violin" is formed by the mirrored envelope of a kernel-density estimate of the distribution. It contains furthermore all box plot information: lines that indicate the location of the quartiles.

3.3. Time-Frequency Localized Evaluation

[33] As pointed out above, any global estimate of model-data agreement is unable to reveal temporally local model-data (dis)agreements. The discrete decomposition of the time series has the advantage that model performance can be

[1]Auxiliary materials are available in the HTML. doi:10.1029/2009JG001016.

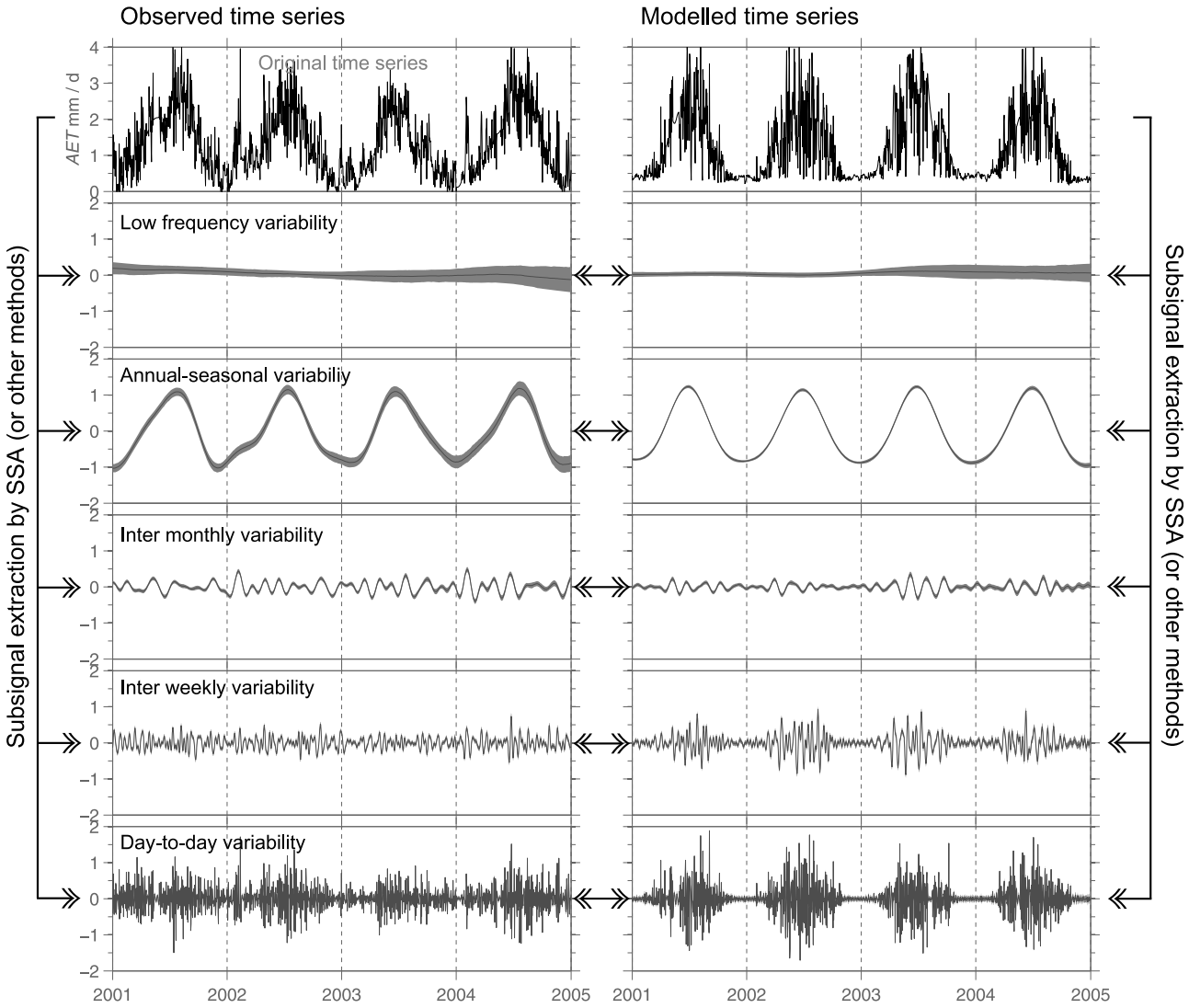

Figure 1. The principle of the model-data comparison on multiple timescales. Both observed and modeled time series are first decomposed into subsignals corresponding to characteristic frequency bins. Qualitative or quantitative model-data comparisons can be carried out on the corresponding pairs of subsignals. Each of the subsignals is represented here by the confidence envelope accounting for the extraction uncertainty. The line within the 95% confidence envelope is the median of the reextracted subsignals. Figure 1 exemplifies the novel model-data comparison strategy with LPJ simulations of *AET* and corresponding observations at the site NL-Loo. The frequency binning corresponds to the coarse discretization as summarized in Table 2.

estimated in the frequency classes at specific times. To the best of our knowledge, the timing of the model-data agreement has not yet been investigated on different scales of characteristic variability, especially not for terrestrial biosphere models. Decidability on the significance of model-data agreement on a specific scale at a time point *i* results from contrasting modeled and observed subsignals. More precisely, we use the 95% confidence envelopes that account for the SSA extraction uncertainty to compare observed and modeled subsignals.

3.4. Time Period of Model-Data Comparison

[34] For both temporally global and local model-data comparisons we have chosen a consistent window for all models at all sites ranging from the 01.01.2001 to 31.12.2004. For all sites, longer observations and model runs were available. Thus, we always use the full length of the time series to derive the subsignals and quantify the extraction uncertainty. However, the effective evaluations only refer to the chosen 4 year long window. This ensures that the effect of the anomalous year 2003 is equally represented at all sites. In this way, also edge effects that appear in SSA and are only partly controlled by the uncertainty assessment, are mostly excluded from the model-data comparison.

4. Results

[35] In the following, we first report on the results from the temporally global model performance evaluation, the

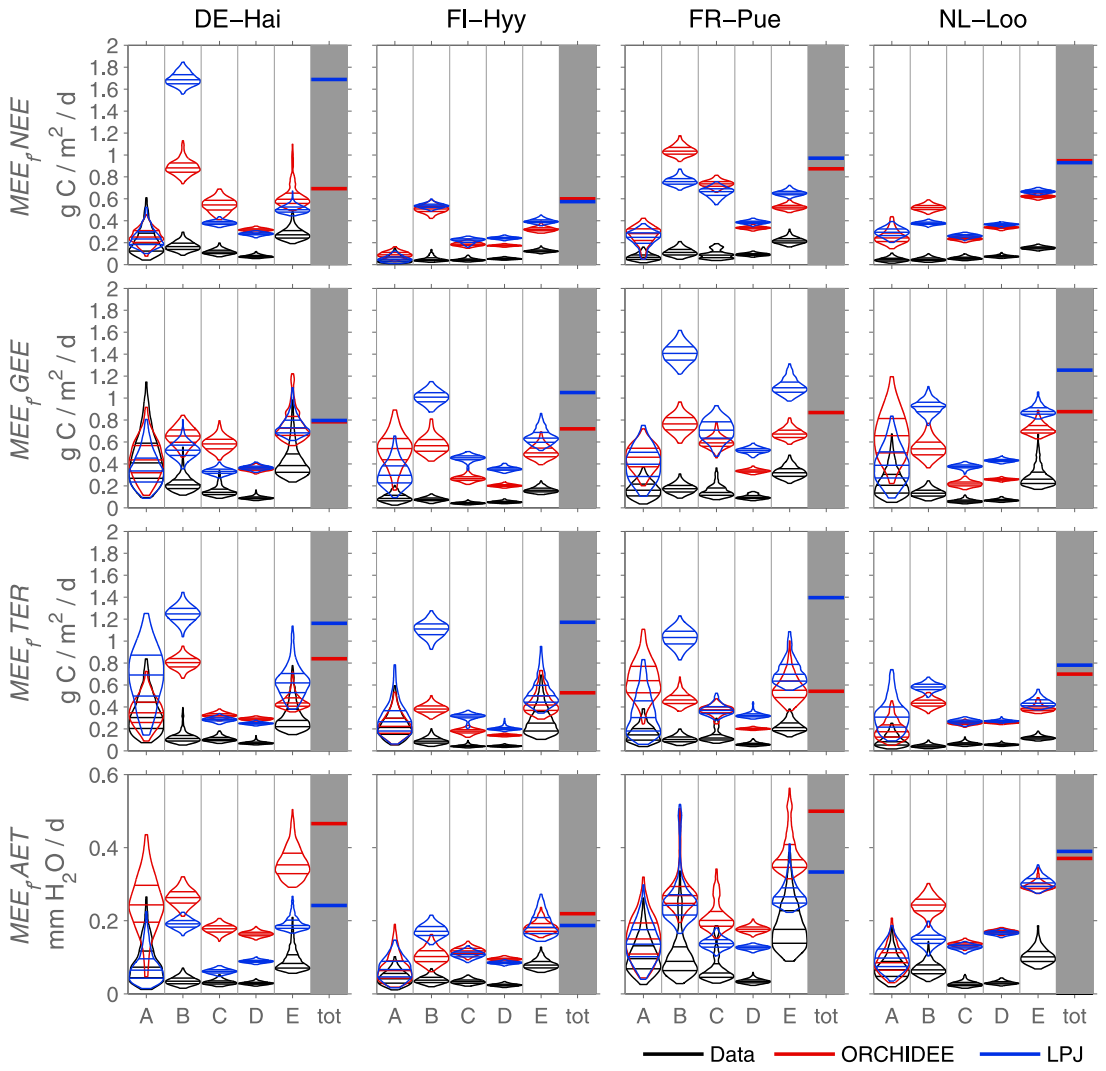

Figure 2. The temporally global median Euclidean errors as a function of frequency, MEE_f, resulting from confronting eddy covariance observations and site level terrestrial biosphere model runs. The error analysis were performed separately in each frequency bin f as described in Table 2 and illustrated by Figure 1. The violins characterize the shape of error distribution. Black violins quantify the methodological uncertainties of subsignal separation attributable to the applied "Singular System Analysis, SSA." The lines in the right gray box (tot) are the reference errors found when directly comparing undecomposed (but centered) time series.

error spectra, and then illustrate the time localization of the frequency-dependent model-data comparisons. It should be recalled that the "errors" disclose model-data disagreements that might originate from problems in models and/or observations.

4.1. Error Spectra: Instantaneous to Weekly Variability

[36] In the joint analysis of the instantaneous (day-to-day) up to weekly variability (coarse binning scheme; bin E in Table 2) the overall model-data disagreement is $0.3 \lessapprox MEE_f \lessapprox 0.7$ g C/m^2/d for *NEE* across sites and models (Figure 2). Generally, the magnitude of these errors is slightly smaller than that of the component fluxes *TER* and *GEE*. This is a generic phenomenon attributable to the relative magnitudes of these fluxes. For the component fluxes, we observe an error range of $0.4 \lessapprox MEE_f \lessapprox 1$ g C/m^2/d with some exceptions. In terms of the C fluxes, LPJ seems to produce consistently larger model-data disagreement in the high frequencies than ORCHIDEE. The most extreme example is the mismatch of the high frequencies in the LPJ simulation of *GEE* at FR-Pue. Regarding the water fluxes, we find a different picture: at DE-Hai and FR-Pue ORCHIDEE is clearly outperformed by LPJ. Nonetheless, the differences among the models are within the range of site-to-site differences. It is noteworthy that model-data mismatches are clearly above the uncertainty attributable to the subsignal extraction which itself produces errors of $\lessapprox$0.2 g C/m^2/d.

[37] A refined analysis of the high frequencies for *NEE* (fine frequency binning h–j in Table 2) reveals that the quantitative errors in the net C flux are dominated by the instantaneous day-to-day variability (frequency bin j,

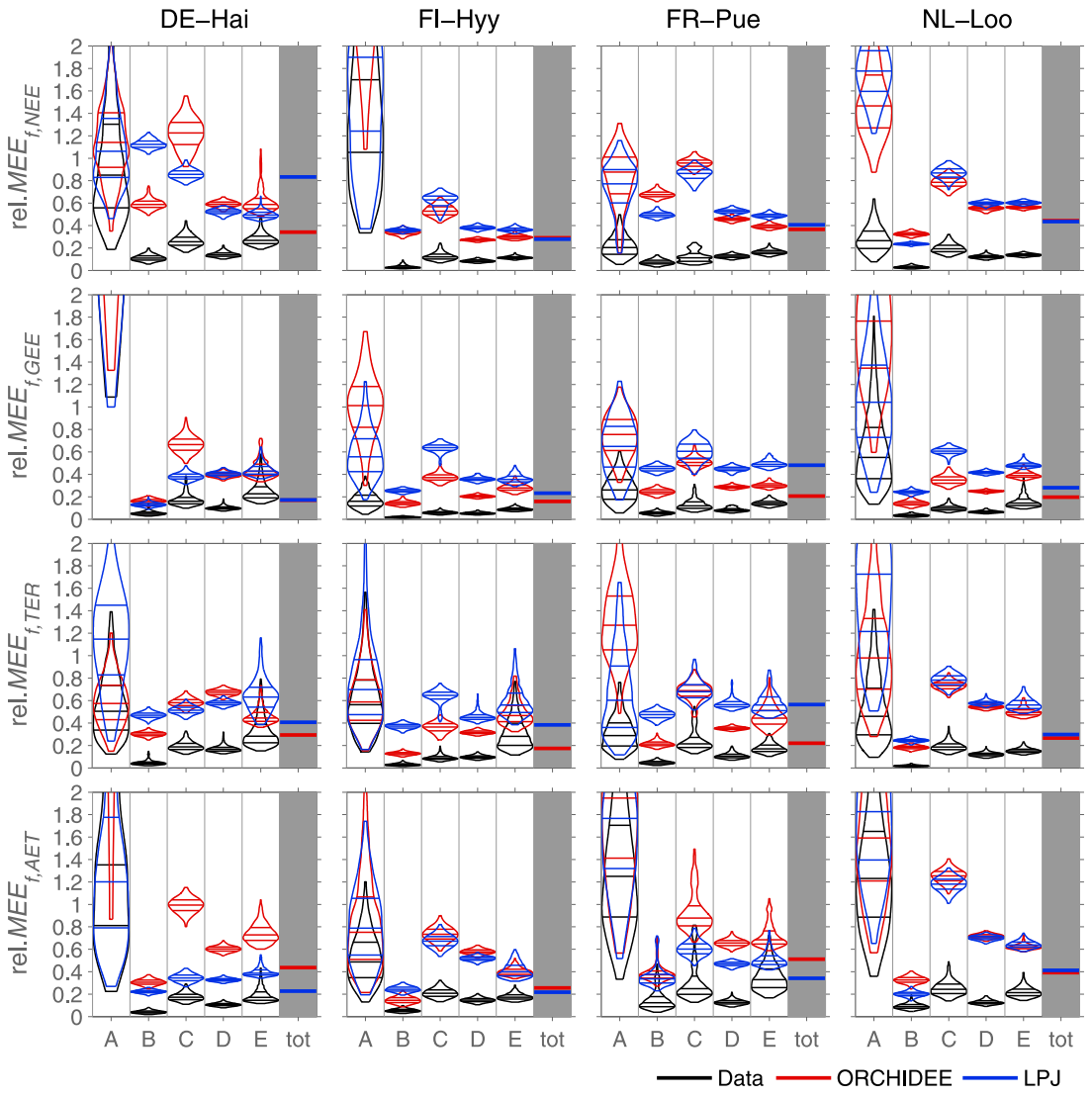

Figure 3. The temporally global relative median Euclidean errors as a function of frequency, rel.MEE_f, resulting from confronting eddy covariance observations and site level terrestrial biosphere model runs. The error analysis were performed separately in each frequency bin f as described in Table 2 and illustrated by Figure 1. Symbols as in Figure 2.

Figure 4). Figure 4 also shows that the highest-frequency bin carries large fractions of time series variance. This explains the quantitative importance of the error. Note also that the values of the correlation analysis decrease in bin j compared to bins h and i.

4.2. Error Spectra: (Intermonthly) Monthly Variability

[38] The temporally global model-data comparison for the two coarse frequency bins that include monthly (bin D in Table 2) to intermonthly (bin C in Table 2) variability is not enlightening in terms of *NEE*: In most cases, no significant difference between both models is found. A notable pattern is, however, that LPJ systematically shows slightly larger disagreements with the observed *GEE*. ORCHIDEE instead tends to produce larger errors regarding *AET*. It is also noticeable that the error magnitudes are very similar across sites for both CO_2 and H_2O fluxes.

[39] While these model-data mismatches are quantitatively small compared to the high-frequency and the seasonal-annual errors (see above, below), Figure 3 further shows that their relative counterparts are not that small compared to the weekly or seasonal-annual frequencies. Poor agreements between simulations and observations are identifiable in the intermonthly variability of all fluxes. The standardization of the errors avoids that these are overlooked because of their low magnitude: Only small fractions of flux variability are induced by these fluctuations; this frequency range is "the spectral gap" [*Baldocchi et al.*, 2001b; *Stoy et al.*, 2005]:

[40] High-resolution *NEE* error spectra corroborate these findings (bins e–h in Table 2 and Figure 4). The analysis reveals on the one hand that less than 20% of the variance in *NEE* is explained by the intermediate frequency bins explaining the quantitatively low error rates. On the other hand, the relative errors are very high within this frequency range, generally above the high-frequency and seasonal-annual

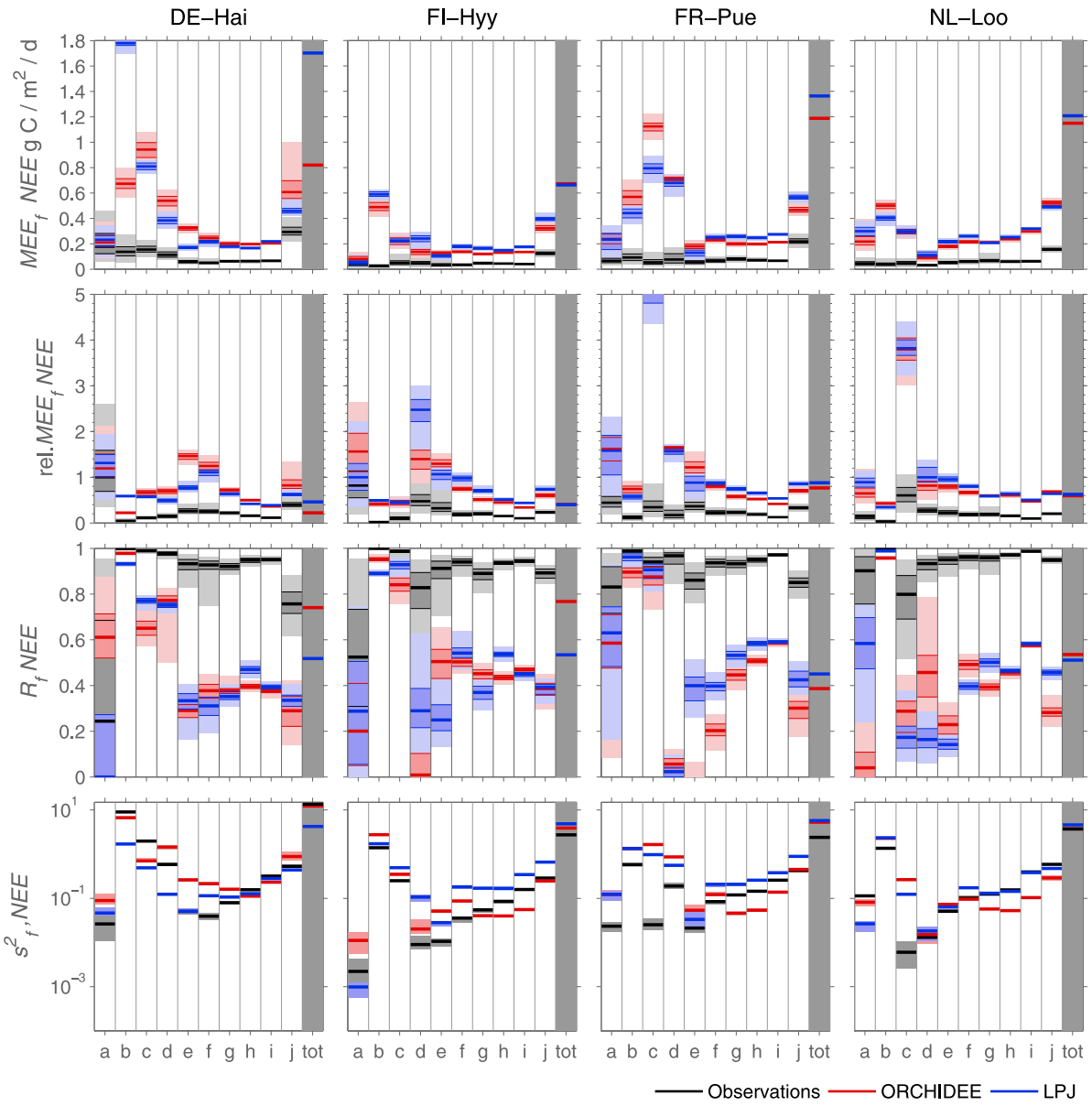

Figure 4. The median Euclidean errors MEE_f, relative MEE_f, robust correlation estimates R_f, variances s_f^2, and ratios of standard deviations corresponding to the frequency bins f (fine discretization, Table 2) for *NEE*. Each estimate is characterized by the median, enveloped by the central quartiles, and the 95% confidence area. The black lines (gray areas) represent the methodical uncertainties of the signal separation by means of "Singular System Analysis, SSA."

errors. An interesting pattern is that within the spectral gap the relative errors tend to increase with decreasing frequency opposed to the correlation coefficients (which however do not exceed values of $R_f \gtrapprox 0.55$) across models and sites. Also the ratios of standard deviations between observed and modeled fluxes are too low/too high in these intermediate bins.

4.3. Error Spectra: Seasonal to Annual Variability

[41] Contrasting the seasonal-annual components (bin B, Table 2) leads to the largest model-data disagreements in the MEE_f spectra. For example, LPJ simulations at DE-Hai diverge from observed *NEE* by a $MEE_f \approx 1.7$ g C/m^2/d (see Figure 2). While ORCHIDEE does a better job at that site it leads to large error rates at FR-Pue and NL-Loo. Tracing the origin of these model-data disagreements back to the quality of *GEE* and *TER* simulations, LPJ is found to depict largest model-data deviations (with the exception of *GEE* at DE-Hai). Obviously large errors in *GEE* and *TER* partly cancel out.

[42] The misrepresentation of CO_2 fluxes appears to be the highest in the seasonal-annual cycles compared to other frequency bins. The corresponding picture for *AET* is, however, less clear. Again, ORCHIDEE is less accurate in terms of water fluxes than LPJ at DE-Hai and NL-Loo, while LPJ produces comparable error rates at FI-Hyy and

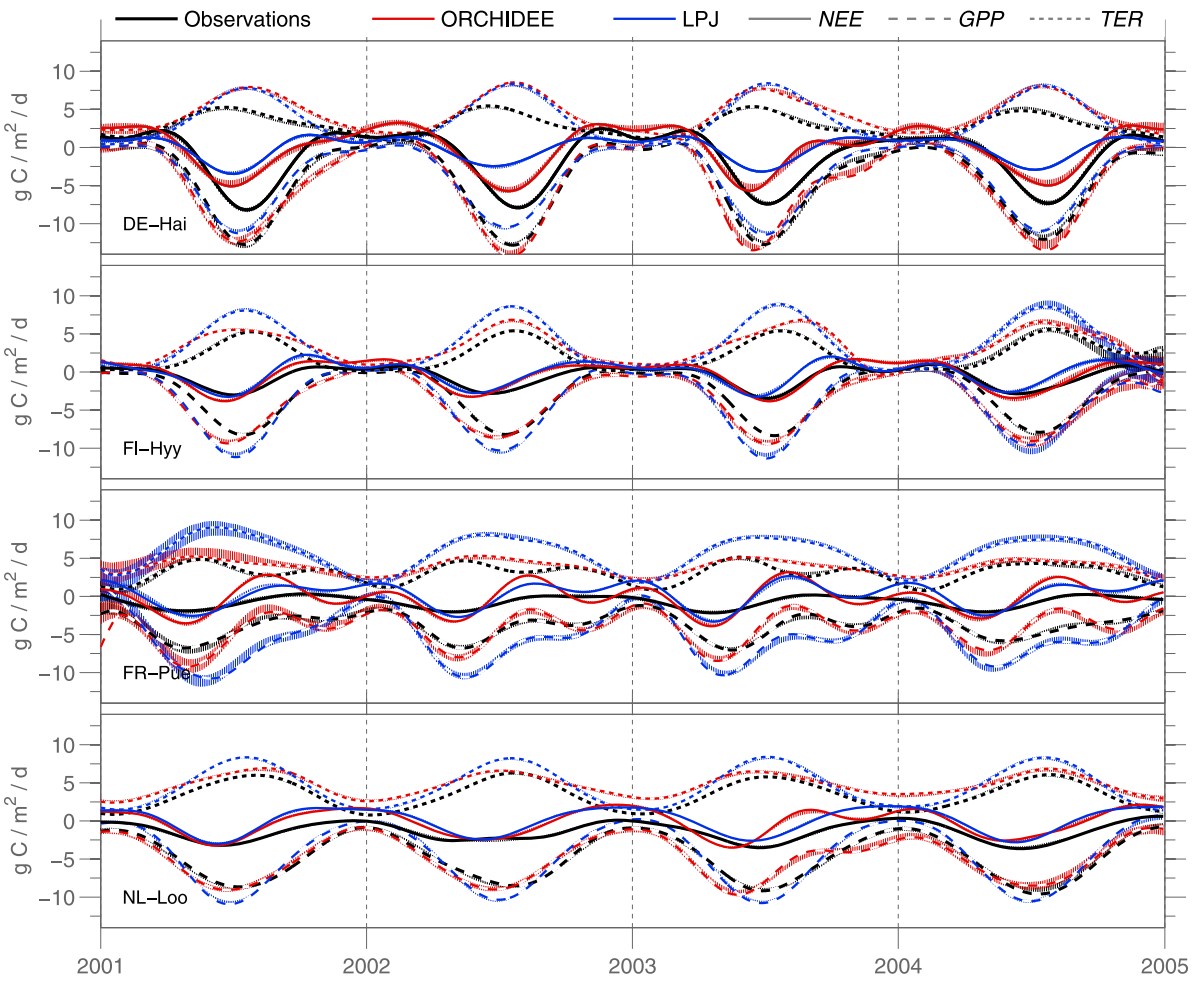

Figure 5. The seasonal-annual subsignals for *NEE* (frequency bin B, Table 2) and the confidence envelopes of the subsignal uncertainty estimation procedure at the four CarboEurope-IP sites extracted from the eddy covariance data and the two terrestrial biosphere models ORCHIDEE and LPJ. Additionally, the corresponding components for *GEE* and *TER*, forming the lower and upper envelopes of the net C flux are displayed. Here, the mean of the original time series was kept, rendering the Singular System Analysis here a filtering technique of the original signal.

FR-Pue. Errors in the simulated seasonal-annual *AET* are smaller compared to the day-to-day variability. This is an interesting observation given that the bulk of the errors in the C fluxes is allocated in the seasonal-annual cycles.

[43] However, these errors appear much less dramatic in the relative perspective. The rel.MEE_f's lie within the range of the errors in the other frequency bins (Figure 3). As anticipated above, the seasonal-annual variability is *relatively* well represented compared to intermediate frequencies.

[44] This finding is supported by the high-resolution *NEE* error spectra. Again, the rel.MEE_f's show a clear depression at the annual cycle. The semiannual cycle, however, which shapes the seasonality, is completely off at the sites FR-Pue and NL-Loo. Nevertheless: in many cases the correlation coefficients are excellent. In the overall picture, the models produce similar high disagreements in the semiannual components.

4.4. Error Spectra: Low-Frequency Variability

[45] In the low-frequency modes (bin A, Table 2), the subsignal separation induces itself errors that are of similar magnitude than the model-data disagreements in terms of MEE_f (Figure 2). Hence, the reliability of SSA for extracting low-frequency modes is not good enough to make any inferences about model performance. At first glance, the overall low values of the performance estimates may be a minor cause of concern: in addition to the small quantitative errors, less than 5% of the total variance is allocated to these components anyhow. However, Figure 3 shows that relative errors in the low frequencies can be as big as those reported for the spectral gap but occurring in tandem with very low (or no) correlation.

4.5. Error Spectra Versus Total Model-Data Mismatches

[46] When comparing the error spectra to the *MEE* of the undecomposed (but centered) time series, the magnitudes of the *MEE* are only sometimes higher compared to the largest MEE_f "peak." This result provides evidence for an error cancelation across different frequency classes. For all variables at most sites, high errors, for instance in the pure

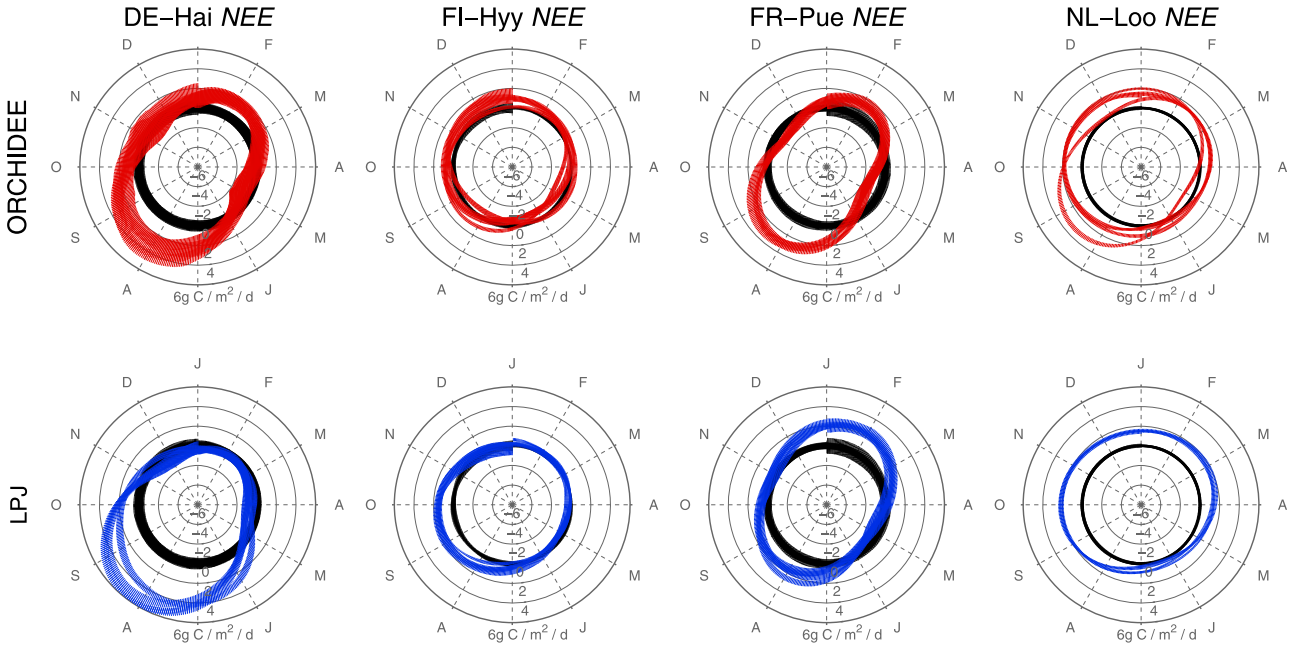

Figure 6. The residuals of the seasonal-annual subsignals (frequency bin B, Table 2) for *NEE* adjusted to the original time series means: $(X_{f_{\mathrm{mod}}} + \overline{Y}_{\mathrm{mod}}) - (X_{f_{\mathrm{obs}}} + \overline{Y}_{\mathrm{obs}})$ for ORCHIDEE and LPJ. Figure 6 shows the residuals in polar plots, where a full circle corresponds to 1 year. The range of the residuals indicates the subsignal separation uncertainty (95% confidence envelopes). The black envelopes indicate the subsignal separation uncertainty when the observations seasonal-annual subsignals are reextracted from the observations only.

seasonal-annual components appear to be partly neutralized by misrepresentations of fluctuations acting on other timescales.

4.6. Time-Frequency Localized Errors

[47] The error spectra reveal model-data disagreement in discrete frequency bins. The obvious question then is if these errors originate from specific periods in time. Time-frequency localized error analyses provide a more sophisticated view on this problem. Figure 5 illustrates the seasonal-annual subsignals of the corresponding simulated and observed time series (bin B in Table 2). The subsignal extraction is used here as a filtering technique where high- and low-frequency variability is removed. Figure 5 discloses the balancing of large deviations in the seasonal-annual variability of *GEE* and *TER*: the resulting *NEE* is met with astonishing accuracy by both models across sites (for example LPJ at FI-Hyy and FR-Pue). This observation is concordant with the question of model equifinality [*Beven*, 2001, 2006]. It becomes also evident that for the selection of sites considered here, error cancelation effects are more important for LPJ because it generally overestimates the magnitudes of both component fluxes.

4.7. Recurrent Model Data Disagreement

[48] The previous time-frequency analysis suggests the existence of recurrent model-data disagreements on different scales of variability. Displaying a time series in a polar plot, where one full circle corresponds to one year, is a good method to address systematically recurrent patterns. Figure 6 shows this for the seasonal-annual residual subsignals of

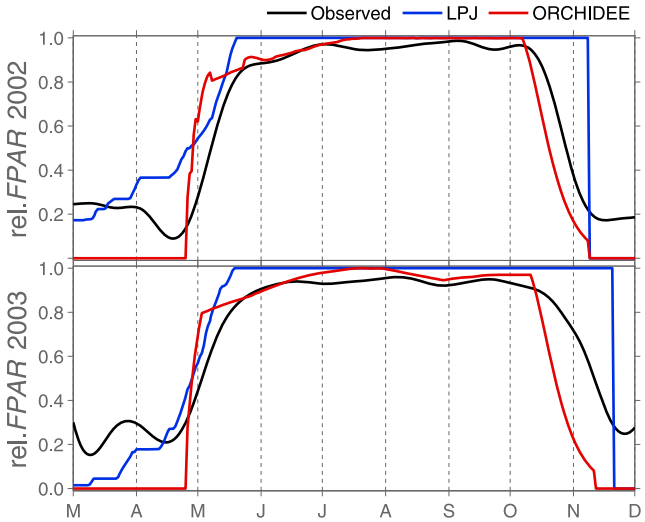

Figure 7. Time series showing the model assumptions on photosynthetically active radiation (*FPAR*) at DE-Hai and corresponding site level estimates. For the sake of comparability, all time series were rescaled to range [0; 1]. The observations were additionally subjected to SSA for removing the high-frequency variability, which (per construction) is not present in the model *FPAR*. Models and data disagree in the timing of spring onset and senescence which is one explanations for divergent simulations of *GEE*.

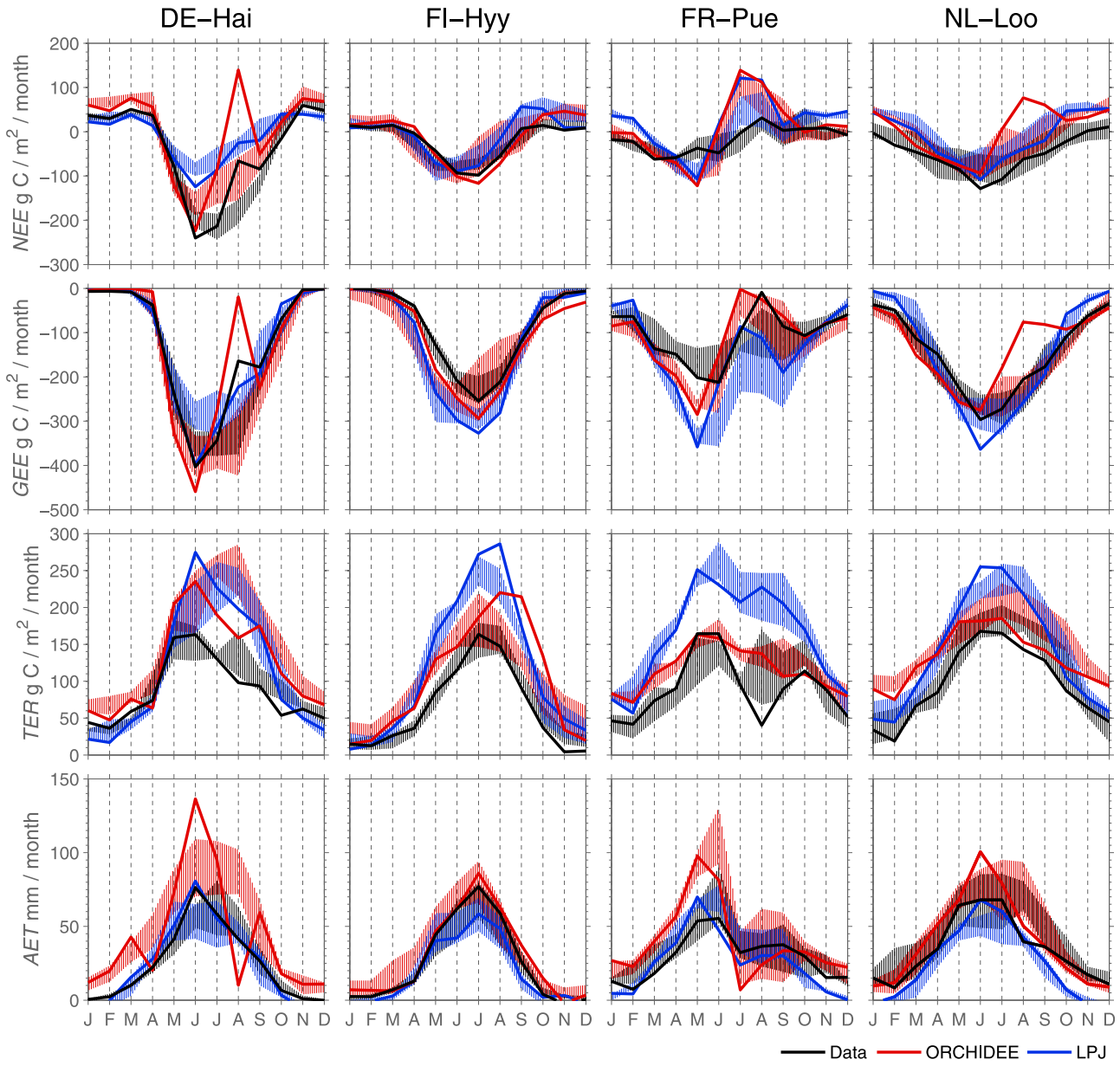

Figure 8. A conventional model-data comparison: Observed and modeled fluxes were aggregated to monthly values and an annual cycle was drawn from the observed monthly flux ranges (hatched areas, DE-Hai (2000–2005), FI-Hyy (1998–2004), FR-Pue (2001–2005), NL-Loo (1998–2005)). However, the anomalous year 2003 where Europe was hit by a severe summer heat wave was excluded from the estimated monthly range estimate. This year is represented by a separate line.

NEE. Both models tend to underestimate the seasonal-annual *NEE* modes during spring and to overestimate it in the later summer months. In the overall view, the seasonal-annual fluctuations in the simulated *NEE* are over pronounced compared to the observations. One reason is that during spring greenup both models systematically underestimate *GEE* (overestimate productivity) except from the site DE-Hai (Figure S2). These patterns hint at problems concerning the internal representation of evergreen phenology in the models: while here ORCHIDEE assumes a dynamic fraction of photosynthetically active radiation (*FPAR*), LPJ does not distinguish any temporal variations. Consequently, LPJ overestimates CO_2 uptake at early stages of vegetation development. A minor, yet relevant problem is that even at DE-Hai modeled *FPAR* disagrees in the timing. Figure 7 shows the model *FPAR* along with site level estimates. Obviously, these observations are also problematic, however, they indicate some differences in the timing of phenology which are propagated into recurrent model-data disagreements regarding *GEE*.

[49] Even more systematic problems than reported for *GEE* affect the simulations of the respiratory processes (Figure S3). While LPJ overestimates *TER* for all sites between June and October, ORCHIDEE shows some site specific problems. In general, *GEE* model-data disagreements are more site than model-specific compared to *TER*

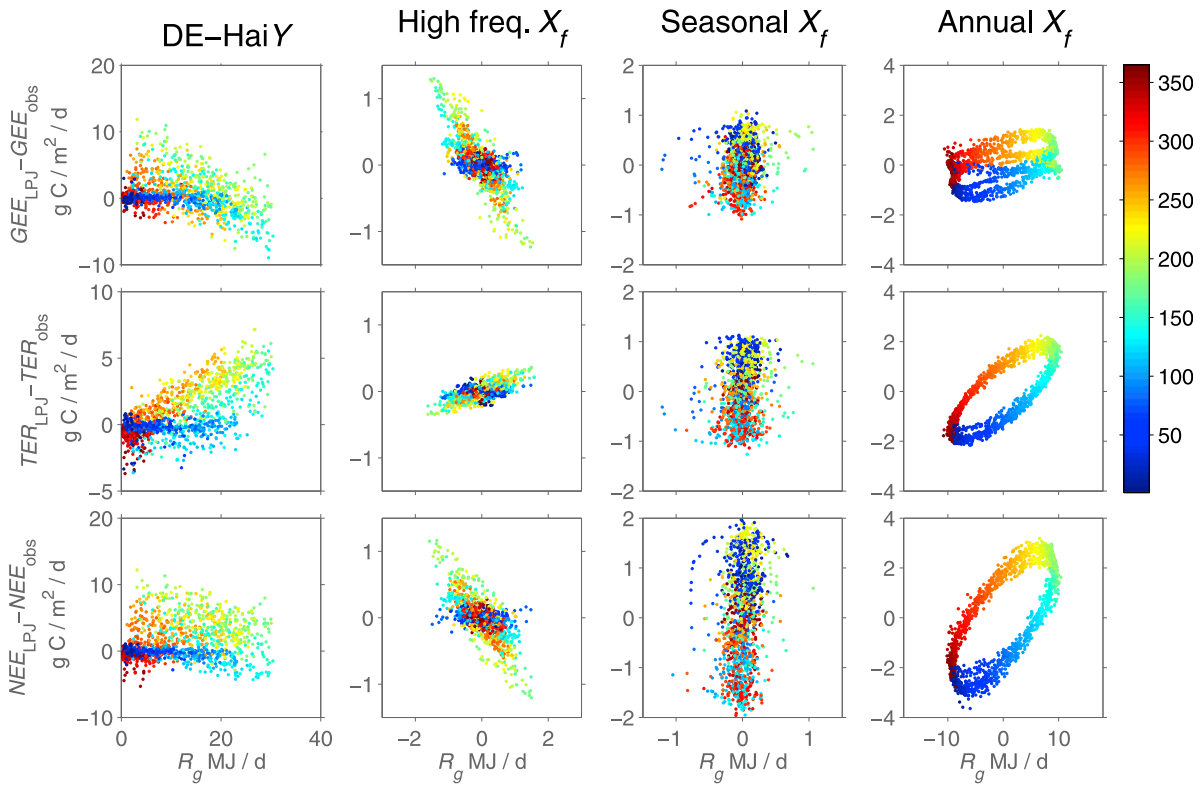

Figure 9. A comparison of a conventional analysis of LPJ residuals and a set of scale-dependent counterparts at the site DE-Hai: The undecomposed time series Y of both model residuals and incident radiation are not clearly related. Using the retrieved subsignals X_f instead, uncovers systematic responses in the residuals to the corresponding fluctuations in the driver. Encoding the scatter by the day of year shows that frequency-dependent model-data disagreements are a function of time. Note that only 3 out of 10 frequencies are exemplarily shown (frequency bins b, c, and j, Table 2). Part of the scatter in the subsignals originates from the uncertainty associated with the subsignal extraction (see section 3.1 and Figure S1 for details).

simulations. This possibly reflects impacts of climate regimes specificities on the model accuracy [*Jaeger et al.*, 2009]. The very systematic seasonal-annual *TER* model-data disagreements instead suggest that the corresponding model concepts are not sufficiently elaborated, which has also been concluded previously by *Kucharik et al.* [2006]. One reason could be the often occurring overestimations of *GEE* influencing the C transfer to SOM over the course of the growing season. However, this cannot explain seasonal-annual *TER* overestimations at all sites, especially not for DE-Hai where *GEE* is well behaved in LPJ and ORCHIDEE simulations. Here, the soil C pools may induce a model bias. One presumption to discuss is if model spin-up and transient runs are suitable for the present sites and account for substantial nonequilibrium conditions [see also *Carvalhais et al.*, 2008]. As a result in the overall *NEE*, none of the modeled seasonal-annual cycle timings is right in the course of the year.

[50] Polar plots for higher-frequency components reveal similar recurrent model-data disagreements. Figure S4 shows the residual intermonthly subsignals (frequency band C) for *NEE* and the component fluxes. Here we find that intermediate frequency variability during the growing season is misrepresented in *NEE*. This may be an indication of issues related to the following three aspects: (1) timing of onset of growing season; (2) possible soil moisture limitation within growing season; (3) timing of end of growing season. Moving on to the highest-frequency class also shows pronounced nonstationary residuals (results not shown). Here, the effect of higher flux variability in the summer is propagated to a higher residual variability in these months. In the high frequencies, the sensitive coupling of C and H_2O fluxes through the canopy conductance, accompanied by a limited understanding of soil moisture-*TER* interactions [*Reichstein et al.*, 2003; *Seneviratne et al.*, 2006a, 2006b] are assumed to limit the performance of the models. At the same time, observational errors become increasingly important in the very high frequency bins as discussed hereafter, limiting the potential to attribute these mismatches to erroneous model parameterization.

5. Discussion

5.1. Time-Frequency Model-Data Comparison

[51] Model-data comparisons traditionally use temporally global performance estimates to quantify (dis)agreements between simulated and observed time series. Drawing a distinction of subsignals according to characteristic frequency classes leads to more differentiated comparisons. In particular, the analyses provided in this paper based on

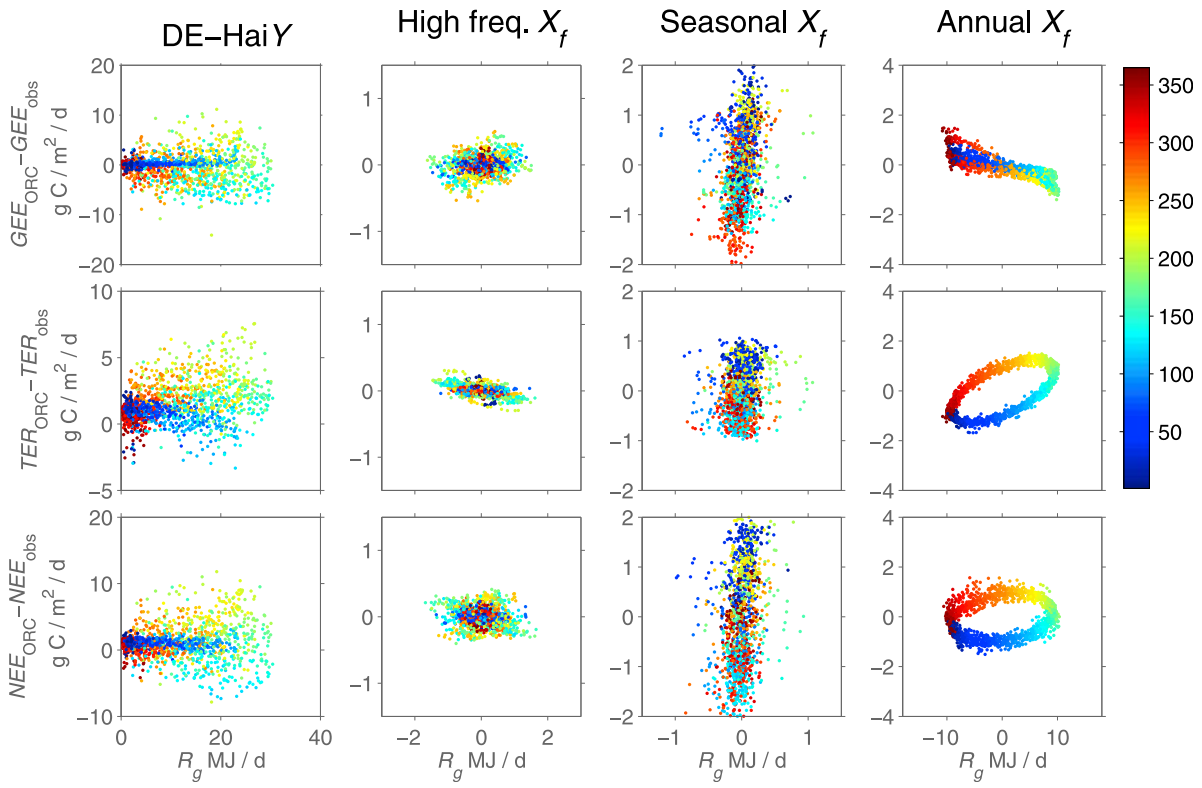

Figure 10. A comparison of a conventional analysis of ORCHIDEE residuals and a set of scale-dependent counterparts at the site DE-Hai: The undecomposed time series Y of both model residuals and incident radiation are not clearly related. Using the retrieved subsignals X_f instead, uncovers systematic responses in the residuals to the corresponding fluctuations in the driver. Encoding the scatter by the day of year shows that frequency-dependent model-data disagreements are a function of time. Note that only 3 out of 10 frequencies are exemplarily shown (frequency bins b, c, and j, Table 2). Part of the scatter in the subsignals originates from the uncertainty associated with the subsignal extraction (see section 3.1 and Figure S1 for details).

"error spectra" reveal that conventional model performance estimates are strongly dominated by the shape of the dominant fluctuations. That is to say that by considering global performance measures of undecomposed time series we overlook model deficiencies at other scales of variability. Localizing frequency dependent model-data disagreement in time further refines the model evaluation.

[52] For further illustrating the advantages of the method, we contrast our approach with a conventional model-data comparison. As exemplified in Figure 8, the conventional approach shows ranges of monthly aggregated fluxes over the available time periods (beyond the four previously analyzed years). The anomalous year 2003 during which Europe was hit by a severe summer heat wave (for details, see *Ciais et al.* [2005] and *Reichstein et al.* [2007]) is excluded from the range estimates and illustrated by separated lines. Clearly the reported effect of recurrent model-data disagreements (Figure 6) due to misinterpretations of the seasonal-annual dynamics cannot be uncovered by the conventional alternative. Figure 8 reveals that the extreme summer deviations in *NEE* (of positive sign, indicating ecosystem C losses) are dramatically overestimated by both models, especially by ORCHIDEE. We could presume then that short-term anomalies are allocated in high-frequency components and deduce that the corresponding frequency classes are not well represented in the simulations. This conclusion, however, remains a vague speculation when relying exclusively on Figure 8. The error spectrum, instead, unambiguously quantifies errors in the high frequencies. The classical model-data comparison can serve as confirmation of our findings, but only provides limited insights as an alone standing analysis.

[53] This effect becomes also evident when investigating the sensitivity of model residuals to the drivers: a classical element of model performance evaluations. In Figures 9 and 10 we visualize a conventional residual analysis along with the corresponding counter parts based on select subsignals of one driver and the residuals. While the standard approach leads to an undefined scatter between the residuals of the carbon fluxes and global radiation, the proposed subsignal analysis allows to identify clear frequency-dependent relationships (further examples are summarized in Figure S5). The residuals respond to the drivers in different intensity and along systematic paths across frequency classes. For instance, the hysteretic residual response of annual subsignals of CO_2 fluxes to radiation supports the previous hypothesis that delayed model responses to meteorological forcing play a central role in model-data mismatch. As in this case, where we highlight systematic differences in time-frequency responses to incident radiation, analogous

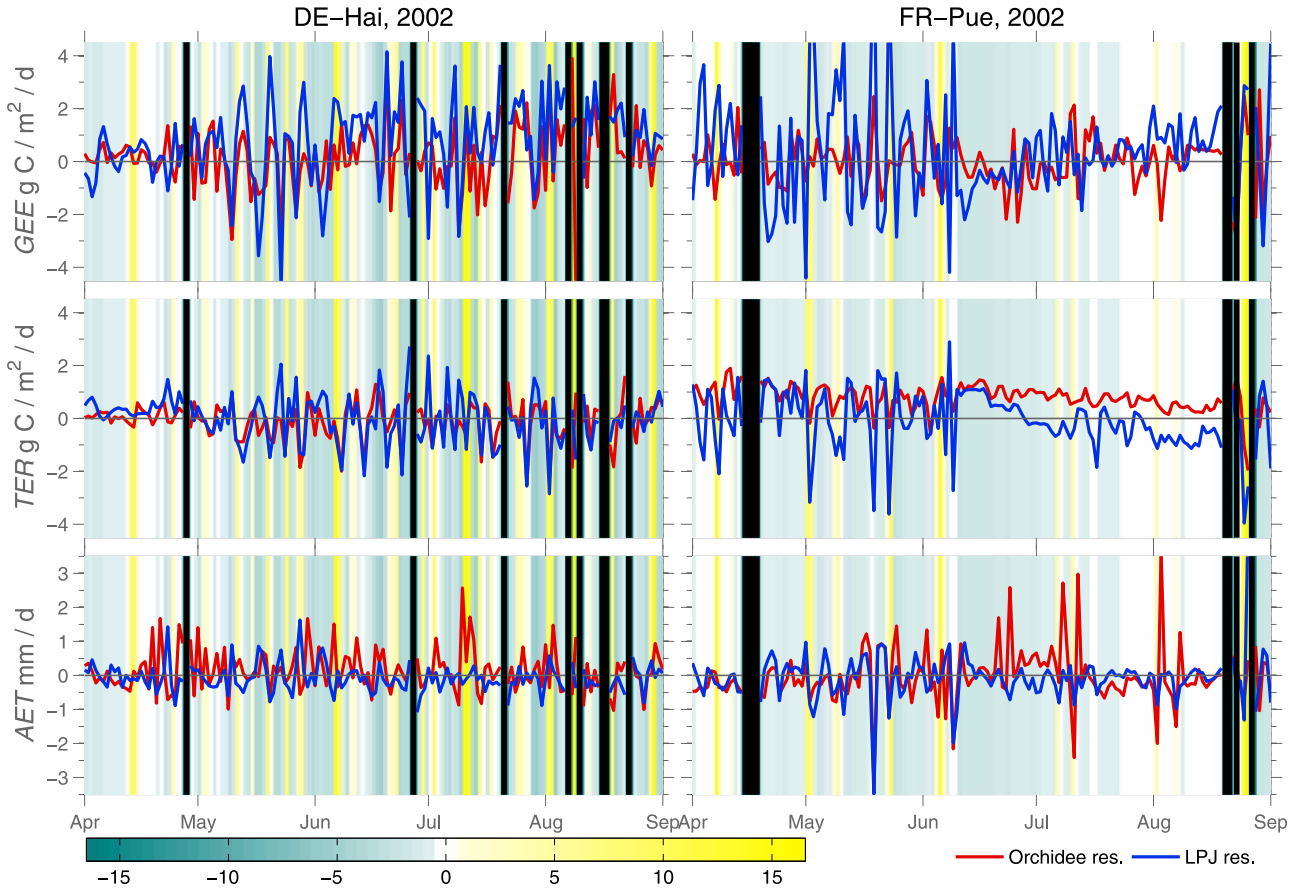

Figure 11. Fluctuation of volumetric soil water content (*SWC*) and high-frequency model-data disagreement. The colored background encodes the difference in *SWC* between 2 days in the time window May–August 2002 at the eddy covariance sites DE-Hai and FR-Pue. The overlaid blue and red lines are the residuals of the high-frequency fluctuations in LPJ and ORCHIDEE simulations (bin E, Table 2), respectively, for gross ecosystem exchange (*GEE*), terrestrial ecosystem respiration (*TER*), and actual evapotranspiration (*AET*). The black lines show where data gaps occurred. Strong fluctuations in *SWC* are clearly accompanied by high-frequency model-data disagreements. For the C fluxes these disagreements are higher in the LPJ simulations and for *AET* ORCHIDEE seems more sensitive.

investigations of model-data disagreement can adumbrate systematically deviating process representations in ongoing model development.

[54] Also Figure 11 exemplifies the potential of such an approach showing how abrupt changes in volumetric soil water content are often, but not always, directly followed by severe high frequency model-data disagreements. This residual analysis uncovers that the instantaneous *TER* and *GEE* responses to alterations in the ecohydrological conditions are more sensitive in LPJ compared to ORCHIDEE at two very different sites. While the C flux residuals of the two models show similar fluctuations, the opposite holds for *AET* simulations: *AET* residuals of LPJ and ORCHIDEE are sometimes even of different sign. ORCHIDEE appears to react more sensitively to sudden hydrological shifts compared to LPJ, which is the consequence of ORCHIDEE's soil surface layer parameterization.

[55] One weakness of the time-frequency model-data comparison is that only the variability of a characteristic scale is preserved and no information on the mean values is given. This intrinsic property of any time series decomposition requires that bias estimations are realized as an additional step. Studies aiming at confining carbon balances of terrestrial ecosystems should thus strive for conventional tools as applied for example by *Jung et al.* [2007]. The two approaches of model-data comparisons can be seen as complementary, where different patterns are addressed. These patterns are, however, only partly independent: a fundamental problem of model biases is their propagation. In terrestrial biosphere models which contain a series of coupled and often nonlinear submodels, biased internal variables might be translated into erroneous amplitude modulations of depending processes via multiplicative terms. In the overall view, however, we find that time-frequency localized model-data comparisons offer a perspective for meticulous analyses of particular interest for model development.

5.2. Observational Limitations

[56] Random and systematic errors in the observations naturally limit the quality of model-data comparisons [*Rykiel*, 1996; *Morales et al.*, 2005; *Friend et al.*, 2007]. The effects of

random errors on eddy covariance data are well investigated [e.g., *Hollinger and Richardson*, 2005; *Richardson et al.*, 2008], and it could be shown that the temporal autocorrelation of eddy covariance measurement errors is very short, usually subdaily [*Lasslop et al.*, 2008]. Being founded on daily aggregates, constructed from up to 48 replicate measurements, we assume the presented model-data comparison not to be strongly affected by error autocorrelation. Thus, observational errors are not expected to have an effect on the low-frequency model-data comparisons. For the highest-frequency bins, however, the quantified errors might result from both model deficiencies and the observational error.

[57] Unlike random errors, systematic biases are poorly constrained. For example the impact of low nighttime turbulence on nocturnal flux estimates [*van Gorsel et al.*, 2007] and the lack of energy balance closure [*Wilson et al.*, 2002; *Foken*, 2008b] might limit the quality of observational data and propagate to subsequent model-data comparisons. No consensus has been achieved so far regarding the treatment of systematic errors is model data synthesis approaches [*Kruijt et al.*, 2004; *Burba et al.*, 2008].

[58] A very different shortcoming of the present model-data comparison is that historical ecosystem alterations remain unconsidered in our simulations. In principle, natural or anthropogenic modifications of stand structure, or changes in the magnitude of nitrogen deposition, to give only two examples, can be incorporated in standard terrestrial biosphere models [*Thornton et al.*, 2002; *Zaehle et al.*, 2006; *Carvalhais et al.*, 2010]. It is data scarcity that largely impedes such modeling exercises [*Morales et al.*, 2005].

[59] An in depth understanding of structural discrepancies of terrestrial biosphere models is favored when singular processes can be encircled in model-data comparisons. However, as introduced in section 2, the experimental setup of eddy covariance observations can resolve only the antipodal gross carbon fluxes of respiration and photosynthesis. Yet, since even these two fluxes cannot be observed independently, the model-data comparison reaches an intrinsic limitation: The applied flux separation turns "data" into simulations and we can only state that here the study is comparing model outputs against diagnostic flux estimates. We hold the view that this is generally a valid reference since we accept these empirical flux estimates as benchmark and best possible approximation of the "true" fluxes (for a general discussion on the role of empirical reference models, see *Abramowitz et al.* [2008]).

[60] An additional limitation to any model-data comparison is set by the limited monitoring period [*Braswell et al.*, 2005; *Kucharik et al.*, 2006]. In the case of time-frequency model-data comparisons this is critical since the accuracy of low-frequency pattern extraction decreases with time series length. Investigating the potential of terrestrial biosphere models to reproduce low-frequency modes is, however, of outstanding relevance under climate change conditions [*Richardson et al.*, 2007]. In this study we proposed a formal method to quantify the uncertainty of subsignal separation and found that the explored observation period does not allow assessing the simulated low-frequency variability. This finding stresses the importance of continued flux monitoring for understanding longer-term carbon and water flux variability in addition to other observations, such as tree ring records [*Briffa et al.*, 2008] or lysimeter data [*Teuling et al.*, 2009], which miss the high-frequency variability.

5.3. Confining Problems in Terrestrial Biosphere Models

[61] The multifactoral effects on *GEE* simulations complicate a precise model diagnostic based on time-frequency localized model-data disagreements, for example how to explain that especially LPJ produced large errors in the seasonal-annual variability? This pattern turns out to be very site specific and especially the *Pinus sylvestris* L. dominated forests FI-Hyy and NL-Loo are clearly misrepresented between April and the beginning of August. In this period *GEE* is underestimated (C uptake is overestimated), which could be due to inaccurate plant functional type specific parameterizations and assumptions which do not account for seasonal changes in the physiological activity of evergreen trees [cf. *Wang et al.*, 2003]. Similar, but less dramatic effects are apparent in the ORCHIDEE simulations of *GEE* which is possibly because of the more differentiated phenology (Figure 7).

[62] Site-to-site differences can help in scrutinizing the origins of identified problems. For example, the fact that at the Mediterranean site FR-Pue the amplitude modulation of the simulated seasonal-annual *GEE* signal (Figure 5) is shaped by overpronounced semiannual fluctuations could hint at an oversensitivity of terrestrial biosphere models to drought situations which shape Mediterranean ecosystems. Indeed, when we correlate the model driver vapor pressure deficit (VPD) and model residuals (Figure S5) we see high correlations ($R_f > 0.6$) between intermonthly and annual periodicities at FR-Pue which are not existent at the other sites.

[63] Another cause of concern is that LPJ simulations of *TER* lead to systematic and recurrent model-data disagreements on a seasonal-annual scale. The highly differentiated architecture of coupled soil C pools and H_2O dynamics, as it is assumed by ORCHIDEE, is apparently less affected by recurrent seasonal-annual *TER* disagreements compared to LPJ. In the latter, the discussed effect of biases, propagated to specific fluctuations, can also play a substantial role. The fact, however, that simulated seasonal-annual *TER* subsignals of both models (especially at DE-Hai) often reach their yearly maximum with a time delay of several days or weeks, indicates an insufficient sensitivity of the SOM decomposition rates. The problem is not explicable by model-data disagreements in *GEE* transferred to incorrect R_A simulations, since sites are affected where the seasonal-annual *GEE* is well matched. Instead, this might be a structural (or parameterization) effect caused by inaccurately defined dependencies of $R_{A,g}$ or $R_{A,m}$ to variations in *GEE*. Also root development at the beginning of the growing season (and the corresponding $R_{A,g}$) is difficult to model and to parameterize.

[64] The observed model-data disagreements for *TER* simulations across sites are, however, not totally unexpected: *Kucharik et al.* [2006] reported overestimations of *TER* during the growing seasons of very different sites. From their comparison (of undecomposed) observed and simulated time series they found disagreements in the range of approximately 0.6 to 3 g C/m^2/d. Figure 5 suggests this number also

Table 3. A Three-Way Analysis of Variance That Traces the Distributions of the Median Euclidean Errors (*MEE*) Back to the Factors[a]

Factor	*NEE*	*GEE*	*TER*	*AET*
Site	1	9	5	4
Model	<1	<1	3	3
Frequency	43	33	38	44
Site × Model	<1	<1	1	5
Site × Frequency	30	37	10	14
Model × Frequency	1	2	<1	3
Unexplained	24	19	43	28

[a]Site-to-site variability, model-to-model differences, and frequency dependencies (summarized in Table 3 as percentages). This three-way ANOVA also considers the interaction terms of the factors. Apparently model performance is first of all a matter of frequency. Regarding *NEE*, strong site-to-site effects are observable which are less dominant in the component fluxes *GEE* and *TER* compared to model-to-model differences.

to hold for our model and site selection. *Kucharik et al.* [2006] trace this mainly back to inappropriate temperature sensitivities of the different respiration subprocesses; we largely agree with this conclusion.

[65] Deficient modulations of simulated seasonal-annual and high-frequency variability in the C fluxes are symptomatic and shape model-data disagreement. Error cancelation in *NEE* generally leads to low error rates for some sites, however, this is not always the case. Especially at the Mediterranean site FR-Pue and the temperate broadleaf site DE-Hai errors in *TER* propagate to the seasonal-annual *NEE* dynamics. These disagreements occur along with imprecise simulations of the variability in H_2O fluxes. However, the general time-frequency analysis of simulated *AET* leads to a very different shape of error spectra (Figure 2). Here, largest errors occur in time-instantaneous responses of *AET* which implicitly shows that the seasonal-annual cycles are quite well represented. Model-to-model differences in *AET* are small compared to some of the specific site-to-site differences. However, the high-frequency responses of *AET* to switches in soil water content (exemplarily shown in Figure 11) uncover very different model responses to driving variables, residing in the parameterization of a skin soil layer in ORCHIDEE, causing the stomata to respond quickly to rain events. These results emphasize the need to further investigate the coupling of water and carbon fluxes on an ecosystem scale.

[66] The model-data differences at synoptic, intermonthly, and low frequencies (period > 1 year) are quantitatively of minor importance compared to errors in seasonal-annual and high-frequency classes. Nonetheless, these disagreements exist and also depict recurrent problems. If we analyze the time-frequency errors for NL-Loo it seems as if the mismatch during anomalous years such as 2003 appears in all models in different frequency bins. The two terrestrial biosphere models seemed to respond on different scales to an unexpected climate anomaly. In view of recent scenarios [e.g., *Yiou et al.*, 2009], alterations of ecosystem variability in response to a changing climate variability are to be expected. Thus, despite of their quantitative minor contributions to the total variability, the intermediate frequencies may gain importance.

[67] As a final point in this study we aim at understanding the relative importance of model choice, site-to-site differences, and target frequency. A qualitative ranking of such factors requires a meta analysis where the variance of the observed (temporally global) error distributions (in terms of *MEE*) can be traced back to its dominant factors. A *n*-way analysis of variance where the impact of analyzed "site," chosen "model," and "frequency band" and their interactions provides an answer (Table 3). This variance partitioning clearly reveals that the major cause of model-data disagreement is a matter of the timescale under investigation. The model-to-model differences play a very minor role in the final net C and H_2O fluxes but are relatively important regarding the component fluxes (this was explained above by error cancelation). Since the applied terrestrial biosphere models encode relatively similar hypotheses of underlying biogeochemical processes, this finding indicates that a common source of model error lies in an incomplete understanding of biospheric responses to climate forcing across timescales. The semiempirical character of dynamical ecosystem theory implies that terrestrial biosphere models differ especially in the detailed formulations and parameterizations of the known principles (section 2), but also that they hardly account for the spatial heterogeneity of real world ecosystems. This is confirmed by the large influence of the site-to-site differences that cannot be captured by these two models of different complexity.

6. Conclusions and Outlook

[68] This paper is an attempt to rethink model-performance evaluations in the context of biosphere-atmosphere exchange fluxes of CO_2 and H_2O. Explicitly locating model-data disagreements in frequency and time helps to identify and explain the essential problems of state-of-the-art terrestrial biosphere models. Since these models encode the current understanding of the dynamics of biosphere-atmosphere CO_2, H_2O, and energy fluxes, this model-data comparison also helps to point out future research and model development needs. The analyzed model-data disagreements are to be read in two ways:

[69] First, they can serve as a guide for assessing the available set of models. As we have demonstrated, we expect improvements of models, for example by considering the reported over sensitivity to short-term fluctuations in the drivers, or through adjustments of unrealistic representations of seasonal-annual variability during the growing season. In this respect, this study provides a technical framework for model assessments to be deployed in test runs accompanying ongoing model development.

[70] Second, the conceptually novel way to analyze ecosystem fluxes on different scales of characteristic variability might serve as a motivation for rethinking the principles of current modeling techniques. We postulate that taking fluctuations on different scales explicitly into account could change the conventional paradigm of model parameterizations. First steps in this direction have been made, where soil or ecosystem respiration models were parameterized considering high-frequency variability only [e.g., *Reichstein et al.*, 2005; *Gu et al.*, 2008]. In this context, but also under conditions where model biases affecting defined timescales are unavoidable, for example due to unknown site history and inappropriate steady state assumptions confounding model parameterizations [*Carvalhais et al.*, 2008],

Table B1. Values of the Normalization Factor M_t and of the Lower L_t and Upper U_t Bounds of Summation

Temporal Locations	M_t	L_t	U_t
For $1 \le t \le P-1$	t^{-1}	1	t
For $P \le t \le K$	P^{-1}	1	P
For $K+1 \le t \le N$	$(N-t+1)^{-1}$	$t-N+P$	P

differentiation according to prevalent timescales offers novel perspectives.

Appendix A: Long-Term Drivers

[71] The product of *Feser et al.* [2001] corresponds to a simulation with the REMO model [*Jacob and Podzun*, 1997] driven by the NCEP reanalysis data set [*Kalnay et al.*, 1996]. The harmonization of NCEP-REMO and site-level meteorology was carried out by site-specific regression models derived from the overlap period in monthly aggregates. Monthly data were chosen to avoid issues related to the exact timing of the passage of frontal systems as simulated in the regional climate model. These would corrupt the regressions on daily time steps. The regression coefficients were then applied to the long NCEP-REMO time series for the respective grid cells at the four EC sites in order to remove possible biases in the reanalysis product. The corrected NCEP-REMO time series were then further used to extend the site meteorology backward and to fill gaps in the meteorological measurements.

Appendix B: Singular System Analysis

[72] In the present study we used "Singular System Analysis" (also known as "Singular Spectrum Analysis," SSA [*Broomhead and King*, 1986; *Elsner and Tsonis*, 1996; *Golyandina et al.*, 2001; *Ghil et al.*, 2002]) as basis for the time-frequency model-data comparison. The application of SSA is justified because of some theoretical advantages compared to other techniques (discussed by *Mahecha et al.* [2007]), and favored by its applicability to fragmented time series. Note however that the question of how to separate superimposed subsignals from a single data stream has been a matter of debate for decades, and is still not resolved [*Ghil et al.*, 2002]. Other methods might be equally applicable in the context of model-data comparisons.

[73] SSA aims at extracting subsignals of a given time series $X(t)$, $t = 1, \ldots, N$ belonging to characteristic scales of variability. Initially, an embedding dimension has to be defined a priori. This is a window of length P which in this study was set equally for all time series to 3 years. Sliding the window along the time series leads to a trajectory matrix consisting of the sequence of $K = N - P + 1$ time-lagged vectors of the original series. The P dimensional vectors of the trajectory matrix $\mathbf{Z}$ are set up as described in equation (B1) [*Golyandina et al.*, 2001].

$$Z_i = (X(i), \ldots, X(i+P-1))^T \qquad 1 \le i \le K \tag{B1}$$

Based on the $K \times P$ trajectory matrix $\mathbf{Z}$ a $P \times P$ covariance matrix $\mathbf{C} = \{c_{i,j}\}$ is built:

$$\mathbf{C} = \frac{1}{K}\mathbf{Z}^T\mathbf{Z} \tag{B2}$$

For constructing the covariance matrix, various approaches have been reported in the literature and we refer the reader to the specialized literature for understanding different variants of SSA [*Vautard and Ghil*, 1989; *Golyandina et al.*, 2001; *Ghil et al.*, 2002]. The entries of the resulting $P \times P$ matrix represent the captured lag covariance. This is used to determine the orthonormal basis by solving equation (B3),

$$\mathbf{E}^T\mathbf{C}\mathbf{E} = \mathbf{\Lambda}, \tag{B3}$$

where, $\mathbf{E}$ is a $P \times P$ matrix of the eigenvectors E_i, also called empirical orthogonal functions (EOFs) of $\mathbf{C}$. The matrix $\mathbf{\Lambda}$ contains the respective eigenvalues in the diagonal, sorted by convention in descending order $diag(\mathbf{\Lambda}) = (\lambda_1, \ldots, \lambda_P)$, where $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_P$. It can be shown that due to the properties of the covariance matrix $\mathbf{C}$, preserving symmetry and being real valued and positive semidefinite, all eigenvectors and eigenvalues are real valued, where the latter are nonnegative scalars. The eigenvalues are proportional to the fraction of explained variance corresponding to each EOF. In analogy to the well known Principal Component Analysis, the decomposition allows the construction of principal components (PCs) as generated time series representing the extracted orthogonal modes (equation (B4)). This is why SSA is often also called a "PCA in the time domain."

$$A^\kappa(t) = \sum_{j=1}^{P} X(t+j-1)E^\kappa(j), \qquad 1 \le \kappa \le P \tag{B4}$$

As it can be seen in equation (B4), the principal components are obtained by simply projecting the time series onto the EOFs. This projection constructs a set of P time series of length K.

[74] The last step in SSA is the reconstruction of the time series through the principal components $A^\kappa(t)$, see equation (B5). The original signal can be fully or partially reconstructed. This is a selective step, and the analyst has to decide which $A^\kappa(t)$ are combined so that one obtains an interpretable combination of principal components. This enables signal-noise separation and the reconstruction of specifically selected frequency components, as illustrated by equation (B5).

$$R^k(t) = \frac{1}{M_t}\sum_{\kappa \in \mathcal{K}}\sum_{j=L_t}^{U_t} A^\kappa(t-j+1)E^\kappa(j) \tag{B5}$$

In this reconstruction procedure, κ is an index set determining the selection of modes used for the reconstruction, M_t is a normalization factor, and the corresponding extension for the series boundaries are given by L_t and U_t (definitions for the boundary terms are given in Table B1; a comprehensive derivation can be found in the work of *Ghil et al.* [2002]).

B1. SSA Gap-Filling Procedure

[75] *Kondrashov and Ghil* [2006] introduced an iterative SSA gap-filling strategy. Their method allows a time series reconstruction for fragmented time series, and thus is a tool for gap filling. Here we outline the fundamental steps of the SSA gap-filling algorithm:

[76] 1. Center the time series to zero mean, the latter being estimated from present data only.

[77] 2. Initiate an inner-loop iteration by applying SSA of the zero-padded time series. The leading reconstructed component (identified through its highest eigenvalue) is used to fill the values in the gaps. This leads to a new estimate of the time series mean, which is used for re-centering the time series. The initially zero-padded values are set to their reconstructions. This procedure is carried out based on the computed and recomputed reconstructed components until a convergence criterion (in terms of the RMSE) is met.

[78] 3. After the first inner-loop iteration meets the convergence criterion, switch to an outer-loop iteration. This is the natural extension of the described procedure above, achieved by simply adding a second (third, etc.) additionally reconstructed component to inner-loop iteration.

B2. Subsignal Separability

[79] *Golyandina et al.* [2001] showed that sometimes a signal carries an overtone that corresponds to a different frequency bin. Thus, we applied SSA first and distribute the subsignals to each frequency bin. The individual subsignals are then resubjected to SSA and if subsignals are detected that do not fit their actual bin these are redistributed to the corresponding frequency class.

B3. Subsignal Confidence Boundaries

[80] It has been reported that despite of orthogonal base functions (the EOFs) the accuracy of subsignals separability is not warranted [*Golyandina et al.*, 2001]. This methodological uncertainty has to be strictly distinguished from the effective model-data disagreement. We quantified the separation inaccuracy by a surrogate technique: For each subsignal in a frequency bin the corresponding residual was retained. For each residual a set of surrogates is generated. Since most of the analysis will focus on the coarse bins, 500 surrogates were created in case of the coarse binning, and 20 in case of the fine binning. A surrogate is a time series which resembles the original counterpart (here the residuals) in two fundamental aspects: the distribution and spectral properties. The latter warrants as identical autocorrelation structure. We followed the technique proposed by *Schreiber and Schmitz* [1996] known as "Iterative Amplitude Adjusted Fourier Transform, IAAFT" and we refer the reader to the original paper for more details.

[81] One problem inherent to IAAFT is that when the difference of start and end points are large, the corresponding spectral power of the "jump" is spread over all frequencies. This affects especially smooth time series with low powers in the high frequencies. The corresponding surrogate time series appear more noisy in all frequency ranges compared to the reference [*Schreiber and Schmitz*, 2000]. Fortunately, in our set up a precise definition of the spectral content is provided by the frequency binning (Table 2). Thus in a final step, surrogates undergo itself SSA, warranting that the surrogates accurately match the frequency structure of the residuals.

[82] After a set of surrogates has been generated and added to the subsignal of interest, the latter are reextracted. Any subsignal is thus replaced by an array of subsignals and their deviations quantify the extraction uncertainty. All analyses in this paper rely on this array instead of a single subsignal and form the basis for confidence envelopes for any estimated metric. Figure S1 conceptually summarizes the procedure in a flow chart.

Appendix C: Biweight Midcorrelation

[83] The correlation between an observed (X_{obs}) and a modeled signal (X_{mod}) can be estimated by means of the biweight midcorrelation coefficient [*Wilcox*, 2004]. First, we need to find the median euclidean deviation from the sample median, *MED*,

$$MED = M\{|X - M\{X\}|\}. \tag{C1}$$

Based upon these estimates, the data are rescaled as follows,

$$p_i = \frac{x_{i,\mathrm{mod}} - M\{X_{\mathrm{mod}}\}}{9MED_{\mathrm{mod}}}, \quad \text{and} \qquad q_i = \frac{x_{i,\mathrm{obs}} - M\{X_{\mathrm{obs}}\}}{9MED_{\mathrm{obs}}}. \tag{C2}$$

These two quantities are used to encode whether the rescaled data do meet the following constraints:

$$a_i = \begin{cases} 1 & \text{if} \quad |p_i| \le 1 \\ 0 & \text{if} \quad |p_i| > 1 \end{cases}, \quad \text{and} \qquad b_i = \begin{cases} 1 & \text{if} \quad |q_i| \le 1 \\ 0 & \text{if} \quad |q_i| > 1 \end{cases}. \tag{C3}$$

Finally, the terms

$$c_i = 1 - p_i^2, \quad \text{and} \quad d_i = 1 - q_i^2, \tag{C4}$$

are to be defined, based upon which the biweight midcovariance is found,

$$s_{\mathrm{mod,obs}} = \frac{N \sum_{i=1}^{N} a_i b_i c_i^2 d_i^2 \left(x_{i,\mathrm{mod}} - M\{X_{\mathrm{mod}}\}\right)\left(x_{i,\mathrm{obs}} - M\{X_{\mathrm{obs}}\}\right)}{\left(\sum_{i=1}^{N} a_i c_i (1 - 5p_i^2)\right)\left(\sum_{i=1}^{N} b_i d_i (1 - 5q_i^2)\right)}. \tag{C5}$$

[84] In analogy to the other correlation estimates, the variances and covariances are used to obtain the coefficient:

$$R_{\mathrm{mod,obs}} = \frac{s_{\mathrm{mod,obs}}}{\sqrt{s_{\mathrm{mod,mod}} s_{\mathrm{obs,obs}}}}. \tag{C6}$$

The biweight midcorrelation is bounded in the range of −1 to 1. Its use was recently advocated by *Cannon and Hsieh* [2008], who showed that in the presence of outliers the coefficient is superior compared to Pearsons' product moment correlation, otherwise it retrieves comparable values.

[85] **Acknowledgments.** The authors thank W. Kutsch, S. Rambal, and T. Vesala for supplying eddy covariance data. We gratefully acknowledge G. Lasslop, M. Migliacca, two anonymous reviewers, and the editors for very constructive comments. This study was supported by the CarboEurope-Integrated Project GOCE-CT-2003-505572. C.B., N.C. M.C.B., M.J., M.D.M., and M.R. thank the Max Planck Society for supporting the independent junior research unit "Biogeochemical Model-Data Integration."

References

Abramowitz, G., R. Leuning, M. Clark, and A. Pitman (2008), Evaluating the performance of land surface models, *J. Clim.*, *21*, 5468–5481.

Aubinet, M., et al. (2000), Estimates of the annual net carbon and water exchange of forests: The EUROFLUX methodology, *Adv. Ecol. Res.*, *30*, 113–175.

Baldocchi, D. (2008), Turner Review No. 15. 'Breathing' of the terrestrial biosphere: Lessons learned from a global network of carbon dioxide flux measurement systems, *Aust. J. Bot.*, *56*, 1–26.

Baldocchi, D., and K. Wilson (2001), Modeling CO_2 and water vapor exchange of a temperate broadleaved forest across hourly to decadal timescales, *Ecol. Modell.*, *142*, 155–184.

Baldocchi, D., et al. (2001a), FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities, *Bull. Am. Meteorol. Soc.*, *82*, 2415–2434.

Baldocchi, D., E. Falge, and K. Wilson (2001b), A spectral analysis of biosphere-atmosphere trace gas flux densities and meteorological variables across hour to multi-year timescales, *Agric. For. Meteorol.*, *107*, 1–27.

Ball, J. T., I. E. Woodrow, and J. A. Berry (1987), A model predicting stomatal conductance and ist contribution to the control of photosynthesis under different environmental conditions, in *Progress in Photosynthesis Research*, vol. 4, *Proceeding of the 7th International Photosynthesis Congress*, edited by J. Bingins, pp. 221–224, Martinus Nijhoff, Dordrecht, Netherlands.

Barford, C. C., S. C. Wofsy, M. L. Goulden, J. W. Munger, E. H. Pyle, S. P. Urbanski, L. Hutyra, S. R. Saleska, D. Fitzjarrald, and K. Moore (2001), Factors controlling long- and short-term sequestration of atmospheric CO_2 in a mid-latitude forest, *Science*, *294*, 1688–1691.

Betts, A. K. (2004), Understanding hydrometeorology using global models, *Bull. Am. Meteorol. Soc.*, *85*, 1673–1688.

Beven, K. (2001), On modelling as collective intelligence, *Hydrol. Processes*, *15*, 2205–2207.

Beven, K. J. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*, 18–36.

Bonan, G. B. (2008), Forests and climate change: Forcings, feedbacks, and the climate benefits of forests, *Science*, *320*, 1444–1449.

Braswell, B. H., W. J. Sacks, E. Linder, and D. S. Schimel (2005), Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations, *Global Change Biol.*, *11*, 335–355.

Briffa, K. R., V. V. Shishov, T. M. Melvin, E. A. Vaganov, H. Grudd, R. M. Hantemirov, M. Eronen, and M. M. Naurzbaev (2008), Trends in recent temperature and radial tree growth spanning 2000 years across northwest Eurasia, *Philos. Trans. R. Soc. London, Ser. B*, *363*, 2269–2282.

Broomhead, D. S., and G. P. King (1986), Extracting qualitative dynamics from experimental data, *Physica D*, *20*, 217–236.

Burba, G. G., D. K. McDermitt, A. Grelle, D. J. Anderson, and L. K. Xu (2008), Addressing the influence of instrument surface heat exchange on the measurements of CO_2 flux from open–path gas analyzers, *Global Change Biol.*, *14*, 1854–1876.

Cannon, A. J., and W. W. Hsieh (2008), Robust nonlinear canonical correlation analysis: Application to seasonal climate forecasting, *Nonlinear Processes Geophys.*, *15*, 221–232.

Carvalhais, N., et al. (2008), Implications of the carbon cycle steady state assumption for biogeochemical modeling performance and inverse parameter retrieval, *Global Biogeochem. Cycles*, *22*, GB2007, doi:10.1029/2007GB003033.

Carvalhais, N., M. Reichstein, P. Ciais, G. J. Collatz, M. D. Mahecha, L. Montagnani, D. Papale, S. Rambal, and J. Seixas (2010), Identification of vegetation and soil carbon pools out of equilibrium in a process model via eddy covariance and biometric constraints, *Global Change Biol.*, *9999*, doi:10.1111/j.1365-2486.2010.02173.x.

Ciais, P., et al. (2005), Europe-wide reduction in primary productivity caused by the heat and drought in 2003, *Nature*, *437*, 529–533.

Collatz, G. J., J. T. Ball, C. Grivet, and J. A. Berry (1991), Physiological and environmental regulation of stomatal conductance, photosynthesis and transpiration: A model that includes a laminar boundary layer, *Agric. For. Meteorol.*, *54*, 107–136.

Collatz, G. J., M. Ribas-Carbo, and J. A. Berry (1992), A coupled photosynthesis-stomatal conductance model for leaves of C_4 plants, *Aust. J. Plant Physiol.*, *19*, 519–538.

Cramer, W., et al. (2001), Global response of terrestrial ecosystem structure and function to CO_2 and climate change: Results from six dynamic global vegetation models, *Global Change Biol.*, *7*, 357–373.

Dolman, A. J., E. J. Moors, and J. A. Elbers (2002), The carbon uptake of a midlatitude pine forest growing on sandy soil, *Agric. For. Meteorol.*, *111*, 157–170.

Ducoudre, N., K. Laval, and D. Perrier (1993), SECHIBA, a new set of parameterizations of the hydrologic exchanges at the land-atmosphere interface within the LMD atmospheric general circulation model, *J. Clim.*, *6*, 248–273.

Elsner, J. B., and A. A. Tsonis (1996), *Singular Spectrum Analysis: A New Tool in Time Series Analysis*, 164 pp., Plenum, New York.

Etheridge, D. M., L. P. Steele, R. L. Langenfelds, R. J. Francey, J.-M. Barnola, and V. I. Morgan (1996), Natural and anthropogenic changes in atmospheric CO_2 over the last 1000 years from air in Antarctic ice and firn, *J. Geophys. Res.*, *101*(D2), 4115–4128.

Evans, G. T. (2003), Defining misfit between biogeochemical models and data sets, *J. Mar. Syst.*, *40–41*, 49–54.

Farquhar, G., S. von Caemmerer, and J. Berry (1980), A biochemical model of photosynthetic CO_2 assimilation in leaves of C_3 species, *Planta*, *149*, 78–90.

Feser, F., R. Weisse, and H. von Storch (2001), Multi-decadal atmospheric modeling for Europe yields multi-purpose data, *Eos Trans. AGU*, *82*, 305–310.

Foken, T. (2008a), *Micrometeorology*, 308 pp., Springer, Heidelberg, Germany.

Foken, T. (2008b), The energy balance closure problem—An overview, *Ecol. Appl.*, *18*, 1351–1367.

Friend, A. D., et al. (2007), FLUXNET and modelling the global carbon cycle, *Global Change Biol.*, *13*, 610–633.

Ghil, M., et al. (2002), Advanced spectral methods for climatic time series, *Rev. Geophys.*, *40*(1), 1003, doi:10.1029/2000RG000092.

Golyandina, N., and E. Osipov (2007), The "Caterpillar"-SSA method for analysis of time series with missing values, *J. Stat. Plann. Inference*, *137*, 2642–2653.

Golyandina, N., V. Nekrutkin, and A. Zhigljavsky (2001), *Analysis of Time Series Structure: SSA and Related Techniques*, *Monogr. Stat. Appl. Probab.*, vol. 90, 305 pp., CRC Press, Boca Raton, Fla.

Granier, A., et al. (2007), Evidence for soil water control on carbon and water dynamics in European forests during the extremely dry year: 2003, *Agric. For. Meteorol.*, *143*, 123–145.

Gu, L., P. J. Hanson, W. Mac Post, and Q. Liu (2008), A novel approach for identifying the true temperature sensitivity from soil respiration measurements, *Global Biogeochem. Cycles*, *22*, GB4009, doi:10.1029/2007GB003164.

Gulden, L. E., E. Rosero, Z.-L. Yang, T. Wagener, and G.-Y. Niu (2008), Model performance, model robustness, and model fitness scores: A new method for identifying good land-surface models, *Geophys. Res. Lett.*, *35*, L11404, doi:10.1029/2008GL033721.

Haxeltine, A., and I. C. Prentice (1996), A general model for the light-use efficiency of primary production, *Funct. Ecol.*, *10*, 551–561.

Heimann, M., and M. Reichstein (2008), Terrestrial ecosystem carbon dynamics and climate feedbacks, *Nature*, *451*, 289–292.

Hintze, J. L., and R. D. Nelson (1998), Violin plots: A box plot-density trace synergism, *Am. Stat.*, *52*, 181–184.

Hollinger, D. Y., and A. D. Richardson (2005), Uncertainty in eddy covariance measurements and its application to physiological models, *Tree Physiol.*, *25*, 873–885.

Jacob, D., and R. Podzun (1997), Sensitivity studies with the Regional Climate Model REMO, *Meteorol. Atmos. Phys.*, *63*, 119–129.

Jaeger, E. B., R. Stöckli, and S. I. Seneviratne (2009), Analysis of planetary boundary layer fluxes and land-atmosphere coupling in the regional climate model CLM, *J. Geophys. Res.*, *114*, D17106, doi:10.1029/2008JD011658.

Janssen, P. H. M., and P. S. C. Heuberger (1995), Calibration of process-oriented models, *Ecol. Modell.*, *83*, 55–66.

Jarvis, P. G. (1976), The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field, *Philos. Trans. R. Soc. London. Ser. B*, *273*, 593–610.

Jung, M., et al. (2007), Uncertainties of modelling gross primary productivity over Europe: A systematic study on the effects of using different drivers and terrestrial biosphere models, *Global Biogeochem. Cycles*, *21*, GB4021, doi:10.1029/2006GB002915.

Jung, M., M. Verstraete, N. Gobron, M. Reichstein, D. Papale, A. Bondeau, M. Robustelli, and B. Pinty (2008), Diagnostic assessment of European gross primary production, *Global Change Biol.*, *14*, 2349–2364.

Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*, 437–471.

Katul, G., C.-T. Lai, K. Schäfer, B. Vidakovic, J. Albertson, D. Ellsworth, and R. Oren (2001), Multiscale analysis of vegetation surface fluxes: From seconds to years, *Adv. Water Resour.*, *24*, 1119–1132.

Keeling, C. D., and T. P. Whorf (2005), Atmospheric CO_2 records from sites in the SIO air sampling network, in *Trends: A Compendium of Data on Global Change*, Carbon Dioxide Inf. Anal. Cent., Oak Ridge Natl. Lab., U.S. Dep. of Energy, Oak Ridge, Tenn.

Knohl, A., E.-D. Schulze, O. Kolle, and N. Buchmann (2003), Large carbon uptake by an unmanaged 250-year-old deciduous forest in Central Germany, *Agric. For. Meteorol.*, *118*, 151–167.

Kondrashov, D., and M. Ghil (2006), Spatio-temporal filling of missing data points in geophysical data sets, *Nonlinear Processes Geophys.*, *13*, 151–159.

Krinner, G., N. Viovy, J. de Noblet-Ducoudre, N. Ogeée, J. Polcher, P. Friedlingstein, P. Ciais, S. Sitch, and I. C. Prentice (2005), A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, *Global Biogeochem. Cycles*, *19*, GB1015, doi:10.1029/2003GB002199.

Kruijt, B., J. A. Elbers, C. von Randow, A. C. Araujo, P. J. Oliveira, A. Culf, A. O. Manzi, A. D. Nobre, P. Kabat, and E. J. Moors (2004), The robustness of eddy correlation fluxes for Amazon rain forest conditions, *Ecol. Appl.*, *14*, 101–113.

Kucharik, C. J., C. C. Barford, M. El Maayar, S. C. Wofsy, R. K. Monson, and D. D. Baldocchi (2006), A multiyear evaluation of a Global Vegetation Model at three AmeriFlux forest sites: Vegetation structure, phenology, and seasonal and interannual CO_2 and H_2O vapor exchange, *Ecol. Modell.*, *196*, 1–31.

Lasslop, G., M. Reichstein, J. Kattge, and D. Papale (2008), Influences of observation errors in eddy flux data on inverse model parameter estimation, *Biogeosciences*, *5*, 1311–1324.

Li, X. R., and Z. Zhao (2006), Evaluation of estimation algorithms, part I: Incomprehensive measures of performance, *IEEE Trans. Aerosp. Electron. Syst.*, *42*, 1340–1358.

Lloyd, J., and J. A. Taylor (1994), On the temperature dependence of soil respiration, *Funct. Ecol.*, *8*, 315–323.

Mahecha, M. D., M. Reichstein, H. Lange, N. Carvalhais, C. Bernhofer, T. Grünwald, D. Papale, and G. Seufert (2007), Characterizing ecosystem-atmosphere interactions from short to interannual timescales, *Biogeosciences*, *4*, 743–758.

Medlyn, B. E., A. P. Robinson, R. Clement, and R. E. McMurtrie (2005), On the validation of models of forest CO_2 exchange using eddy covariance data: Some perils and pitfalls, *Tree Physiol.*, *25*, 839–857.

Monteith, J. L. (1995), Accomodation between transpiring vegetation and the convective boundary layer, *J. Hydrol.*, *166*, 251–263.

Moorcroft, P. R. (2006), How close are we to a predictive science of the biosphere?, *Trends Ecol. Evol.*, *21*, 400–407.

Morales, P., et al. (2005), Comparing and evaluating process-based ecosystem model predictions of carbon and water fluxes in major European forest biomes, *Global Change Biol.*, *11*, 2211–2233.

Papale, D., et al. (2006), Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: Algorithms and uncertainty estimation, *Biogeosciences*, *3*, 571–583.

Parton, W. J., J. W. B. Stewart, and C. V. Cole (1988), Dynamics of C, N, P, and S in grassland soils: A model, *Biogeochemistry*, *5*, 109–131.

Qin, Z., Y. Ouyang, G. Su, Q. Yu, J. Li, J. Zhang, and Z. Wu (2008), Characterization of CO_2 and water vapor fluxes in a summer maize field with wavelet analysis, *Ecol. Informatics*, *3*, 397–409.

Rambal, S., R. Joffre, J. M. Ourcival, J. Cavender-Bares, and A. Rocheteau (2004), The growth respiration component in eddy CO_2 flux from a Quercus ilex mediterranean forest, *Global Change Biol.*, *10*, 1460–1469.

Reichstein, M., et al. (2003), Modelling temporal and large-scale spatial variability of soil respiration from soil water availability, temperature and vegetation productivity indices, *Global Biogeochem. Cycles*, *17*(4), 1104, doi:10.1029/2003GB002035.

Reichstein, M., et al. (2005), On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm, *Global Change Biol.*, *11*, 1–16.

Reichstein, M., et al. (2007), Reduction of ecosystem productivity and respiration during the European summer 2003 climate anomaly: A joined eddy covariance, remote sensing and modeling analysis, *Global Change Biol.*, *13*, 634–651.

Richardson, A. D., et al. (2006), Comparing simple respiration models for eddy flux and dynamic chamber data, *Agric. For. Meteorol.*, *141*, 219–234.

Richardson, A. D., D. Y. Hollinger, J. D. Aber, S. V. Ollinger, and B. H. Braswell (2007), Environmental variation is directly responsible for short- but not long-term variation in forest-atmosphere carbon exchange, *Global Change Biol.*, *13*, 788–803.

Richardson, A. D., et al. (2008), Statistical properties of random CO_2 flux measurement uncertainty inferred from model residuals, *Agric. For. Meteorol.*, *148*, 38–50.

Rykiel, E. J., Jr. (1996), Testing ecological models: The meaning of validation, *Ecol. Modell.*, *90*, 229–244.

Savenije, H. H. G. (2009), The art of hydrology, *Hydrol. Earth Syst. Sci.*, *13*, 157–161.

Schimel, D. S., et al. (2001), Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems, *Nature*, *414*, 169–172.

Schreiber, T., and A. Schmitz (1996), Improved surrogate data for nonlinearity tests, *Phys. Rev. Lett.*, *77*, 635–638.

Schreiber, T., and A. Schmitz (2000), Surrogate time series, *Physica D*, *142*, 346–382.

Seneviratne, S. I., D. Lüthi, M. Litschi, and C. Schär (2006a), Land-atmosphere coupling and climate change in Europe, *Nature*, *443*, 205–209.

Seneviratne, S. I., et al. (2006b), Soil moisture memory in AGCM simulations: Analysis of Global Land-Atmosphere Coupling Experiment (GLACE) data, *J. Hydrometeorol.*, *7*, 1090–1112.

Siqueira, M. B., G. G. Katul, D. A. Sampson, P. C. Stoy, J.-Y. Juang, H. R. McCarthy, and R. Oren (2006), Multiscale model inter-comparisons of CO_2 and H_2O exchange rates in a maturing southeastern US pine forest, *Global Change Biol.*, *12*, 1189–1207.

Sitch, S., et al. (2003), Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ Dynamic Global Vegetation Model, *Global Change Biol.*, *9*, 161–185.

Stoy, P. C., G. G. Katul, M. B. S. Siqueira, J.-Y. Juang, H. R. McCarthy, H.-S. Kim, A. C. Oishi, and R. Oren (2005), Variability in net ecosystem exchange from hourly to inter-annual timescales at adjacent pine and hardwood forests: A wavelet analysis, *Tree Physiol.*, *25*, 887–902.

Stoy, P. C., et al. (2009), Biosphere-atmosphere exchange of CO_2 in relation to climate: A cross-biome analysis across multiple timescales, *Biogeosciences*, *6*(10), 2297–2312.

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

Takens, F. (1981), Detecting strange attractors in turbulence, in *Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, vol. 898, pp. 366–381, Springer, New York.

Teuling, A. J., et al. (2009), A regional perspective on trends in continental evaporation, *Geophys. Res. Lett.*, *36*, L02404, doi:10.1029/2008GL036584.

Thornton, P. E., et al. (2002), Modeling and measuring the effects of disturbance history and climate on carbon and water budgets in evergreen needleleaf forests, *Agric. For. Meteorol.*, *113*, 185–222.

van Gorsel, E., R. Leuning, H. A. Cleugh, H. Keith, and T. Suni (2007), Nocturnal carbon efflux: Reconciliation of eddy covariance and chamber measurements using an alternative to the u*-threshold filtering technique, *Tellus, Ser. B*, *59*, 397–403.

Vautard, R., and M. Ghil (1989), Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series, *Physica D*, *35*, 395–424.

Vetter, M., et al. (2008), Analyzing the causes and spatial pattern of the European 2003 carbon flux anomaly using seven models, *Biogeosciences*, *5*, 561–583.

Wang, Q., J. Tenhunen, E. Falge, C. Bernhofer, A. Granier, and T. Vesala (2003), Simulation and scaling of temporal variation in gross primary production for coniferous and deciduous temperate forests, *Global Change Biol.*, *10*, 37–51.

Wilcox, R. R. (2004), *Introduction to Robust Estimation and Hypothesis Testing*, 2nd ed., Academic, San Diego, Calif.

Wilson, K., et al. (2002), Energy balance closure at fluxnet sites, *Agric. For. Meteorol.*, *113*, 223–243.

Yiou, P., D. Dacunha-Castelle, S. Parey, and T. T. Huong Hoang (2009), Statistical representation of temperature mean and variability in Europe, *Geophys. Res. Lett.*, *36*, L04710, doi:10.1029/2008GL036836.

Zaehle, S., S. Sitch, B. Smith, and F. Hattermann (2005), Effects of parameter uncertainties on the modeling of terrestrial biosphere dynamics, *Global Biogeochem. Cycles*, *19*, GB3020, doi:10.1029/2004GB002395.

Zaehle, S., S. Sitch, C. Prentice, J. Liski, W. Cramer, M. Erhard, T. Hickler, and B. Smith (2006), The importance of age-related decline in forest NPP for modeling regional carbon balances, *Ecol. Appl.*, *16*, 1555–1574.

C. Beer, M. C. Braakhekke, M. Jung, M. D. Mahecha, M. Reichstein, and S. Zaehle, Biogeochemical Model-Data Integration Group, Max-Planck-Institut für Biogeochemie, PO Box 10 01 64, D-07701 Jena, Germany. (miguel.mahecha@bgc-jena.mpg.de)

N. Carvalhais, Faculdade de Ciência e Tecnologia, Universidade Nova de Lisboa, P-2829-516 Caparica, Portugal.

H. Lange, Norsk Institutt for Skog og Landskap, N-1431 Ås, Norway.

G. Le Maire, CIRAD, UPR Fonctionnement et pilotage des écosystèmes de plantations, UPR 80, F-34398 Montpellier, France.

E. Moors, Alterra, Wageningen University and Research Centre, NL-6700 AA Wageningen, Netherlands.

S. I. Seneviratne, Department of Environmental Sciences, ETH Zurich, CH-8092 Zurich, Switzerland.