

Combining multiple statistical methods to evaluate the performance of process-based vegetation models across three forest stands

Joanna A. Horemans^{1*}, Alexandra Henrot², Christine Delire³, Chris Kollas⁴,
Petra Lasch-Born⁴, Christopher Reyer⁴, Felicitas Suckow⁴, Louis François²,
and Reinhart Ceulemans¹

¹Centre of Excellence PLECO, Department of Biology, University of Antwerp, Universiteitsplein 1, B–2610 Wilrijk, Belgium

²Unité de Modélisation du Climat et des Cycles Biogéochimiques (UMCCB), Université de Liège, Allée du Six Août 17, B–4000 Liège, Belgium

³Centre National de Recherches Météorologiques, Unité Mixte de Recherches UMR3589, CNRS Météo-France, F–31000 Toulouse, France

⁴Potsdam Institute for Climate Impact Research, Telegrafenberg A31, D–14473 Potsdam, Germany

Abstract

Process-based vegetation models are crucial tools to better understand biosphere-atmosphere exchanges and eco-physiological responses to climate change. In this contribution the performance of two global dynamic vegetation models, i.e. CARAIB and ISBA_{CC}, and one stand-scale forest model, i.e. 4C, was compared to long-term observed net ecosystem carbon exchange (NEE) time series from eddy covariance monitoring stations at three old-grown European beech (*Fagus sylvatica* L.) forest stands. Residual analysis, wavelet analysis and singular spectrum analysis were used beside conventional scalar statistical measures to assess model performance with the aim of defining future targets for model improvement. We found that the most important errors for all three models occurred at the edges of the observed NEE distribution and the model errors were correlated with environmental variables on a daily scale. These observations point to possible projection issues under more extreme future climate conditions. Recurrent patterns in the residuals over the course of the year were linked to the approach to simulate phenology and physiological evolution during leaf development and senescence. Substantial model errors occurred on the multi-annual time scale, possibly caused by the lack of inclusion of management actions and disturbances. Other crucial processes defined were the forest structure and the vertical light partitioning through the canopy. Further, model errors were shown not to be transmitted from one time scale to another. We proved that models should be evaluated across multiple sites, preferably using multiple evaluation methods, to identify processes that request reconsideration.

Key words: singular spectrum analysis; residual analysis; wavelets; *Fagus sylvatica*; net ecosystem carbon exchange

Editor: Bohdan Konôpka

1. Introduction

Forest net ecosystem exchange (NEE; all abbreviations also in Table A1), the flux of carbon between the forest and the atmosphere, plays a key role in the global carbon balance (Le Quéré et al. 2016). Therefore, understanding NEE responses to environmental change, to ecosystem management and to site characteristics is essential for predicting future biogeochemical cycles (Law et al. 2002; Pan et al. 2014; Li et al. 2016). To this end, process-based vegetation models (PVMs) of varying complexity are being used, operating at varying scales (Keenan et al. 2012; Fischer et al. 2014; Reyer 2015). Dynamic Global Vegetation Models (DGVMs) are designed in order to

depict the water and carbon fluxes on a global scale (Hickler et al. 2012). On the contrary, smaller scale vegetation models commonly depict one ecosystem only, such as one particular forest stand and can be used to test different management alternatives or understand ecosystem processes at stand scale.

Insufficient knowledge about underlying processes as well as uncertainties about the parameters and the initial conditions of PVMs lead to biased and uncertain model results (Walker et al. 2003). Although more detailed processes are often implemented in the stand scale models, thus increasing confidence in the model realism, the complexity of these models may further increase uncer-

*Corresponding author. Joanna A. Horemans, e-mail: Joanna.Horemans@uantwerpen.be, phone: +32 3265 1724

tainty and/or bias in the results (Larocque et al. 2014). Therefore, PVMs need to be thoroughly evaluated to better understand model uncertainties and the implications for model predictions. Prominent evaluation approaches include model validation and model-model comparisons (e.g. Trudinger et al. 2007).

To achieve valuable model evaluations, the availability and the quality of long-term observational data are crucial (Hollinger & Richardson 2005) as multi-annual responses are less well captured by models than daily and seasonal changes (e.g. Braswell et al. 2005; Stoy et al. 2005; Siqueira et al. 2006). In recent years eddy covariance (EC) measurements and remote sensing data covering long periods have become increasingly available. During the last two decades, NEE has been intensively monitored by use of EC techniques across multiple ecosystems in Europe (Aubinet et al. 2000). Nonetheless, measurement errors in EC data still confine the applicability of model validation with EC data (Aubinet et al. 2012), even though measurement techniques and data processing tools are improving and increasingly standardized.

A range of statistical evaluation methods (SEM) have been used to validate and to compare the performance of process-based models. Scalar statistical measures (SSMs) of error and fit provide an indication of the overall match of model output and data, but they offer only limited insight into the potential of the model to capture the variability in the data and they neglect the temporal dimension. As residual analysis (RA) examines model errors as a function of simulated or observed data and of environmental drivers, it may reveal potential model shortcomings (Medlyn et al. 2005). More complex time series analysis techniques including wavelet analysis (WA; Stoy et al. 2005; Dietze et al. 2011) and singular spectrum analysis (SSA; Mahecha et al. 2007; Mahecha et al. 2010; Wang et al. 2012) effectively provide insight into the model fit at different time scales.

In this contribution, we used two DGVMs and one forest stand-scale model, to simulate NEE. We evaluated model results by comparing model NEE outputs with observed NEE data over a period of at least 16 years for three beech forest sites in Soroe (Denmark), Vielsalm (Belgium) and Collelongo (Italy). The SEMs currently in use, i.e. SSMs, RA, WA and SSA, were combined across multiple sites. The objectives of the study were to define environmental conditions under which the models perform poorly and to identify the related model processes that should be revised for the adequate reproduction of NEE dynamics of forest sites. Further, we aim to highlight the importance of applying multiple SEMs for multiple sites in order to perform a useful evaluation of the performance of PVMs. We focused on the consequences of the model structure on NEE simulations and did not intend a full evaluation of all possible model outputs.

2. Materials and methods

2.1. Model selection, set-up and simulation protocol

We selected three PVMs to simulate NEE for the three forest sites. The first model, CARAIB (CARbon Assimilation In the Biosphere) is a physically based DGVM developed to study the role of the vegetation in the global carbon cycle (Warnant et al. 1994; Gérard et al. 1999) and to study the vegetation distributions in the past (François et al. 1998, 2011; Henrot et al. 2016), in the present and in the future (Laurent et al. 2008; Dury et al. 2011). The second model, ISBA_{cc} (Interaction Soil–Biosphere–Atmosphere – Carbon Cycle) simulates the exchange of water, energy and carbon between the land surface and the atmosphere (Noilhan & Mahfouf 1996; Gibelin et al. 2008; Masson et al. 2013). The third model, 4C (FORESEE – FORESt Ecosystems in a changing Environment) is a stand-scale, process-based model developed to study the effects of environmental change on forest ecosystems (Bugmann et al. 1997; Suckow et al. 2001; Lasch et al. 2005). For the evaluation period, the three models, selected to represent different types of process-based models used for the simulation of the carbon balance of forests, were forced by the meteorological data measured at the sites. At the point of initialisation, models were fed with the site specific soil properties that remained constant over the simulations. For CARAIB and 4C, EC and meteorological data were obtained from the daily aggregated FLUXNET2015 data (FULLSET, <http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/fullset-data-product/>). For ISBA_{cc}, the hourly FLUXNET2015 data were used for meteorological forcing. In order to set the model's carbon pools to steady state and to obtain a mature forest at the beginning of the evaluation period, model spin-ups were run for both DGVMs, CARAIB (using ERA-Interim reanalysis; Uppala et al. 2005) and ISBA_{cc} (by cycling through the available meteorological data). The variables NEE, GPP and R_{eco} were extracted from model outputs and aggregated to daily values if the time steps were smaller, i.e. for CARAIB and ISBA_{cc}. The model parameters were not calibrated for the specific sites, since that could have concealed the actual problems in the model structure. The main features of the models can be found in Table 1 and a more extensive overview of the model processes, model input variables and model specifications for the set-up in Tables A2 (CARAIB), A3 (ISBA_{cc}) and A4 (4C).

2.2. Flux data

2.2.1 Site selection

We used data from three FLUXNET sites: Soroe (Denmark), Vielsalm (Belgium) and Collelongo (Italy). At all sites the dominating vegetation type consisted of mature beech (*Fagus sylvatica* L.) forest (from 76 years to 106

Table 1. Short description of the main features of the models used in this study three.

Model characteristics	CARAIB	ISBAcc	4C
Spatial scale	grid or point scale	grid or point scale	stand-scale
Smallest temporal scale	2-hourly	hourly	daily
Spin up	yes	yes	no
Plant functional type options	multiple	multiple	only tree species
Number of PFT dependent parameters	55	40	99
Number of calibrated tree (PFT) species	15 (26)	8	14
Forest structure	one layer of trees and one layer containing herbs and shrubs	one layer of trees	cohorts with different tree characteristics
Driving variables	CO ₂ , air temperature, amplitude of air temperature ($T_{day_{max}} - T_{day_{min}}$), precipitation, air relative humidity, short-wave incoming radiation, wind velocity	CO ₂ , air temperature, precipitation, air relative humidity, short-wave incoming radiation, long-wave incoming radiation, wind velocity	CO ₂ , air temperature, precipitation, air relative humidity, net radiation, wind velocity
Developed in	University of Liège, Belgium	Météo France/CNRS, Toulouse, France	Potsdam Institute for Climate Impact Research, Germany

years old), with different soil and environmental characteristics. EC data were available for a 16-year period (1997–2012) for Soroe and for an 18-year period (1997–2014) for Vielsalm and Collelongo. A detailed description of the sites including stand and soil characteristics is provided in Table 2.

2.2.2 Eddy covariance measurements

The NEE measurements, i.e. the time series of the carbon exchange between ecosystems and the atmosphere, were monitored at a 10 Hz sampling frequency and aggregated to 30-min averages. NEE measurements were processed using a constant friction velocity threshold across years with the reference selected based on model efficiency. Time series of GPP and R_{eco} were calculated by the partitioning of NEE based on nighttime NEE values (Reichstein et al. 2005). Based on the dataset of observations the random uncertainty and the joint uncertainty in NEE were calculated according to the FLUXNET2015 pro-

ocols (<http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/data-processing/>). An extensive description of the partitioning of the NEE flux and measurement errors, as well as of the EC flux calculation and processing protocols has been previously published (Aubinet et al. 2012)

2.3. Model evaluation methods

The performance of the three models was evaluated by four SEMs, i.e. SSM, RA, WA and SSA.

2.3.1 Scalar statistical measures

Three fit statistics were used. Firstly, the coefficient of determination (R^2), secondly, R^2 multiplied by the slope of the regression line between simulations and observations (bR^2), allowing to account for the systematic discrepancy in the magnitude of two signals as well as for the proportion of variance in the observations pre-

Table 2. Descriptions of the three beech-dominated sites investigated in the current study.

	Soroe	Vielsalm	Collelongo
Location			
Country (region)	Denmark	Belgium (Ardenne)	Italy (Abruzzo)
Lat [deg N]	55.5	50.2	41.9
Long [deg E]	11.7	6	13.6
Elevation [m]	40	450	1550
Climate			
Average daily temperature [°C]	8.4	8.3	7.43
Average yearly sum precipitation [mm]	872	964	1159
Average daily relative humidity [%]	82.6	80.6	72.6
Average daily irradiation [J cm ⁻²]	988	991	1489
Soil			
Soil type	Alfisol or mollisol	Dystric cambisol	Humic alisol -calcarous
Clay in top soil [%]	14.2	15.7	14.2
Sand in top soil [%]	59.2	25	55.3
Average C content of root zone [g m ⁻²]	1963	2457	2605
Average N content of root zone [g m ⁻²]	125.2	113	213.1
Average field capacity [vol%]	19.2	30.8	50.1
Average wilting point [vol%]	9.2	12	26.9
Soil density mineral [g cm ⁻³]	1.6	2.65	0.8
Rooting depth [cm]	75	60	86.5
Stand characteristics			
Year of plantation	1921	1908	1891
Age at the beginning of the study period [yrs]	76	88	106
Age at the end of the study period [yrs]	89	103	122
Initial forest density [trees ha ⁻¹]	354	243	900
Initial average diameter at breast height [cm]	38	31.79	20.2
Initial average height [m]	25	26.79	17.3
Initial average basal area [cm ²]	48.77	19.76	28.86

dicted by the simulation results (Krause et al. 2005), and thirdly, the Nash-Sutcliffe efficiency (NSE), providing the relative magnitude of the residual variance compared to the measured data variance (Nash & Sutcliffe 1970). Two scalar error estimates were calculated, the normalized root mean squared error (RMSE) and the more robust normalized euclidean error (NMEE; Li & Zhao 2006; Mahecha et al. 2010).

2.3.2 Residual analysis

The residuals were calculated as the daily simulated values minus the observed values. The following moments of their distribution were calculated: the mean, here called the mean error (ME), the standard deviation (sdR), expressing the variability in the model errors, the skewness (skR; when skR is between -0.5 and 0.5 the distribution is approximately symmetric) and the excess kurtosis (kurR; positive kurR means fat-tailed and negative kurR means thin-tailed distribution). We analyzed the time course of the residuals and their relation with respect to the observed values. The correlation coefficients between the residuals and the observed values (CorrO) as well as between the residuals of the different models (CorrE) were also calculated.

The relation between the residuals and a number of meteorological variables, i.e. air temperature (T), incoming solar radiation (RAD), vapour pressure deficit (VPD) and the modelled drought index (DRI) were studied. A third order polynomial function was fitted through the daily residuals as a function of each of the meteorological variables to analyze the model performance over the range of those variables. DRI was not an observed variable, but calculated by 4C as the daily ratio of water uptake and demand as follows. The potential canopy transpiration demand (D_{tr}) is calculated from the potential evapotranspiration reduced by the interception evaporation (E_{int}), the unstressed stomatal conductance (g_{tot}) and the maximum stomatal conductance (g_{max}) of the forest canopy [Eq. 1]. The transpiration demand of each cohort is derived by considering its relative conductance. The water uptake per cohort is calculated from the soil water availability, itself depending of the daily precipitation and the relative share of fine roots, and is limited by the transpiration demand of the cohort (D_{tr}^c). The DRI of a cohort is defined as the average of the ratios of uptake and demand over the time period of interest (number of days).

$$D_{tr} = E_{int} * \alpha_m * \left(1 - e^{-\left(\frac{g_{tot}}{g_{max}}\right)} \right) \quad [1]$$

In addition, the residuals of the centered and normalized observed and simulated time series (Res_{cn}) were calculated. The centering and the normalization eliminated the effect of consistent model biases and differences

in variance amplitude, respectively (Dietze et al. 2011). The day of the year (DOY) averages of Res_{cn} over the whole study period were calculated to check for systematic asynchronies between simulations and observations, and were compared with the intra-annual NEE evolution represented in the same way.

2.3.3 Wavelet analysis

A discrete as well as a continuous wavelet transformation (based on the non-orthogonal Morlet mother wavelet function), were performed on the observed meteorological data (T, RAD, wind velocity (u), VPD, DRI) and on the observed and simulated time series of the ecosystem fluxes (package WaveletComp in R version 3.1.2). With this analyses the model's performance to reproduce the spectral properties of the observed fluxes was quantified. The minimum period in the analysis was two days and the maximum was 6475 days (5844 for Soroe). In order to check the significance of the average power on each frequency of the series, a low resolution (1/20) was used. In addition, a WA was performed on Res_{cn} . More details about the wavelet method can be found in Lau & Wang (1995) and Torrence & Compo (1997). The scaling exponent, which is the slope of the regression of the log-log relation between the time period and the wavelet power calculated for the normalized residuals ((simulated-observed)/observed) was calculated to check if errors at one scale were correlated with errors at larger scales (cfr. Richardson et al. 2008).

2.3.4 Singular spectral analysis

We conducted an SSA that quantified the relative amount of variance in the time series explained by specifically defined frequency bins (package Rssa in R version 3.1.2). The following frequency (period) bins were defined: 0 – 0.002 (infinity–500 days, multi-annual trend, bin 1), 0.002 – 0.004 (500 – 250 days, yearly variability, bin 2), 0.004 – 0.006 (250 – 166 days, half-yearly variability, bin 3), 0.006 – 0.009 (166 – 111 days, seasonal variability, bin 4) and 0.009 – 0.015 (111 – 66 days, inter-weekly variability, bin 5) and 0.015 – 0.05 (66 to two days, day to day variability, bin 6). Some of these bins were linked to the evolution of meteorological variables and consequently to physiological processes; others appeared as powerful spectrum peaks for one or more of the environmental variables during the wavelet analysis. A full description with implementation guidance of the SSA method is available (Golyandina & Zhigljavsky 2013). The NMEE quantified the goodness of fit between observed and simulated time series for each of the bins (cfr. Mahecha et al. 2010). To include the uncertainty on the EC data, the SSA was also performed for the observed value plus and minus its joint uncertainty. Additionally, the significance of the extracted subsignals (bins) was tested by a red noise test

(package *dplR* in R version 3.1.2). Based on the Lomb-Scargle Fourier Transform 1000 surrogate time series were generated within the borders of a first order autoregressive (AR(1)) spectrum. Assuming that the background noise could be approximated by an AR(1) model, the hypothesis of a spectrum being purely appointed to noise could be rejected at the chosen confidence levels (95% and 99%). For a detailed description of the method see Schultz & Mudelsee (2002).

3. Results

3.1. Observed NEE values and scalar statistical measures

NEE values diverged largely between the three FLUXNET sites. In Soroe, the annual averages were close to zero, and even positive in the first three years of the study, with an average net carbon storage rate of $-0.42 \mu\text{mol m}^{-2} \text{s}^{-1}$. The other forests were more productive with an average storage rate of $1.26 \mu\text{mol m}^{-2} \text{s}^{-1}$ in Vielsalm and of $1.97 \mu\text{mol m}^{-2} \text{s}^{-1}$ in Collelongo (see Table 3). The evolution of NEE over the year clearly differed between sites, with less extreme values in summer and a longer growing season in Vielsalm compared to the other two sites. In Collelongo and Soroe, the maximum carbon

uptake rate occurred at the same time of the year (absolute maximum on DOY 165 in Soroe and on DOY 173 in Collelongo), but overall, Collelongo was a more active carbon sink than Soroe with a higher maximum and a less rapid activity decline. Although GPP was highest in Soroe (avg. of $234.8 \text{ gC m}^{-2} \text{y}^{-1}$), the extremely high $R_{\text{eco}}/\text{GPP}$ ratio (0.92 for the mean yearly values) undermined the net storage of carbon. For Vielsalm this $R_{\text{eco}}/\text{GPP}$ ratio was 0.74 and for Collelongo 0.48. Under similar meteorological conditions (Table 3) GPP at Vielsalm was approximately 10% lower and R_{eco} 30% lower as compared to Soroe, resulting in a higher carbon storage rate in Vielsalm. The three models underestimated yearly averages of R_{eco} for Soroe consistently over the whole study period. For Collelongo (a forest at high altitude with a high tree density) CARAIB and 4C constantly overestimated R_{eco} . For Vielsalm, 4C and ISBA_{CC} strongly and systematically underestimated R_{eco} , while CARAIB overestimated it for the first 11 years, and turned it into an underestimation later on when the observed R_{eco} values strongly increased (yearly values not shown).

The model-site dependent errors in the simulations for either GPP or R_{eco} or both, resulted in diverged biases and random errors in the NEE simulations (Fig. 1). Overall, ISBA_{CC} and CARAIB underestimated the net storage of carbon for Vielsalm and Collelongo (positive

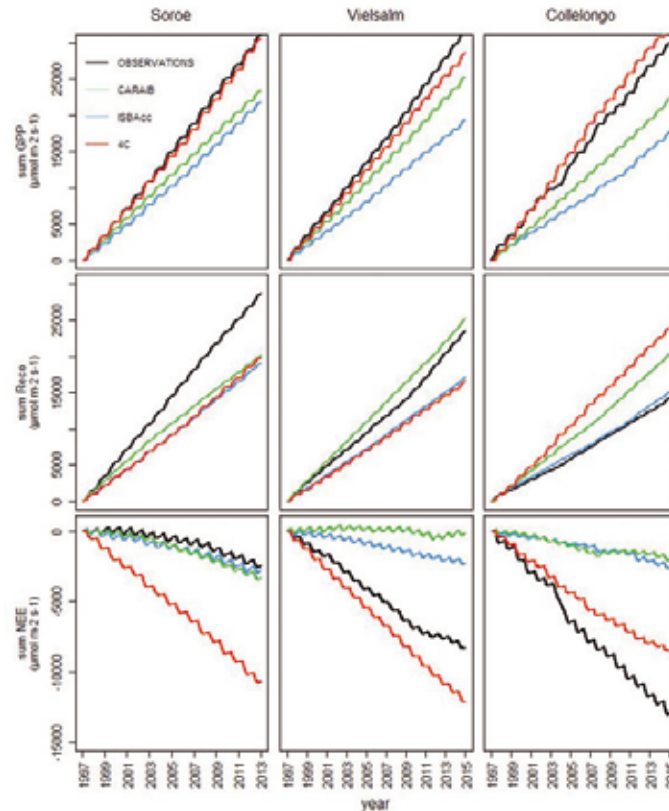


Fig. 1. Cumulative plots of the daily observed and simulated gross primary production (GPP; top), ecosystem respiration (R_{eco} ; middle) and net ecosystem exchange (NEE; bottom) for the study sites Soroe (1997–2012), Vielsalm and Collelongo (both 1997–2014).

ME and skR, Table A5), but not for Soroe. Even though the DGVMs exhibited higher R^2 values for those two sites as compared to 4C, larger biases were observed (lower bR^2 and higher ME). For Soroe all models overestimated the carbon storage over time leading to large values for ME, NRMSE and NMEE and low NSE for the NEE simulations of 4C. For the two other models, however, the combined effect of underestimating both R_{eco} and GPP resulted in an apparently adequate fit with NEE. Although the NEE simulations by ISBA_{cc} and CARAIB were similar, the simulations of its underlying components (GPP and R_{eco}) diverged.

this sink performance of the forests, while, in contrast, at low or negative carbon storage capacity, the models tend to overestimate the carbon uptake. The observed NEE value for which NEE was correctly simulated differed between models. Minimum yearly average residuals were observed for mean yearly NEE values around $-1.5 \mu\text{mol m}^{-2} \text{s}^{-1}$ for 4C and around $-0.5 \mu\text{mol m}^{-2} \text{s}^{-1}$ for ISBA_{cc} and CARAIB. This similar model behaviour indicated some degree of equifinality in the model results on a yearly time scale. Remarkably, for all models, at the onset of the growing season, Res_{cn} started to increase and reached a maximum when NEE values were at half their minimum yearly value, to be low during the small period of minimum NEE (highest carbon uptake). In full sum-

Table 3. Observed and simulated average values of daily net ecosystem exchange (NEE), gross primary production (GPP) and ecosystem respiration (R_{eco}) with their range (of averages over different years; within brackets). The simulations were made with three different models, i.e. 4C, CARAIB and ISBA_{cc} for the three forest sites. All values are in $\mu\text{mol CO}_2 \text{m}^{-2} \text{s}^{-1}$. The dimensionless ratio of R_{eco}/GPP is also indicated.

		NEE		GPP		Reco		R_{eco}/GPP
Soroe	Observations	-0.42	[-0.88, 0.24]	5.32	[4.29, 6.75]	4.9	[4.31, 5.59]	0.92
	4C	-1.84	[-2.43, -1.41]	5.23	[4.08, 6.75]	3.39	[2.65, 4.32]	0.65
	CARAIB	-0.56	[-0.97, -0.13]	4	[3.79, 4.28]	3.43	[4.01, 3.43]	0.86
	ISBA _{cc}	-0.47	[-0.84, -0.13]	3.73	[2.97, 4.49]	3.26	[2.77, 3.75]	0.87
Vielsalm	Observations	-1.26	[-1.89, -0.48]	4.82	[4.18, 5.44]	3.57	[2.81, 4.65]	0.74
	4C	-1.85	[-2.09, -1.49]	4.35	[3.78, 4.93]	3.57	[2.81, 4.65]	0.57
	CARAIB	-0.01	[-0.67, 0.43]	3.84	[3.25, 44.26]	2.5	[2.10, 2.85]	1
	ISBA _{cc}	-0.34	[-0.61, -0.01]	2.94	[2.42, 3.37]	3.83	[3.56, 4.58]	0.89
Collelongo	Observations	-1.97	[-3.69, -0.94]	4.55	[1.44, 7.71]	2.6	[2.33, 2.95]	0.48
	4C	-1.3	[-2.06, -0.66]	4.94	[3.76, 6.62]	2.16	[1.72, 2.67]	0.74
	CARAIB	-0.3	[-0.73, 0.31]	3.39	[2.75, 4.01]	3.09	[2.68, 3.54]	0.91
	ISBA _{cc}	-0.4	[-1.00, 0.07]	2.69	[1.90, 4.00]	2.3	[1.89, 3.00]	0.85

3.2. Residual analysis

For none of the models the daily residuals were normally distributed, with skR and sdR depending on the site. The density distribution of the daily residuals showed high kurR, indicating that a large part of the variance was explained by more extreme values. Overall, the sdR values were highest for Soroe, while skR and kurR were maximum for Collelongo for the three models. 4C differed from the other models in Soroe by producing very high kurR and skR values here, too (Table A5; Fig. 2, top row). The correlations between the daily NEE residuals of the different models were high for some model-site combinations, but not very consistent over sites and models within each site (CorrE in Table A5).

The average yearly residuals changed with time and had a large range including both negative and positive biases (Fig. 2, middle row). The mean yearly bias differed between the models, but the relative size of the residuals fluctuated in parallel. In other words, the models reacted with a comparable magnitude and in the same direction to inter-annual environmental changes. Overall, the average yearly residuals of all models were negatively correlated with the average yearly observed NEE (Table A5; Fig. 3, bottom row). When high yearly carbon storages were observed, the models consistently underestimated

mer (July–September), the carbon storage rate slowed down and Res_{cn} decreased sharply (Fig. 3). However, neither average yearly simulated NEE values nor average yearly NEE residuals were directly correlated with the average yearly or seasonal T, RAD, VPD and DRI (correlations not shown).

On a daily time scale the univariate relation between the NEE residuals and the observed climate variables, i.e. T, RAD, VPD and DRI could not be unambiguously interpreted. These relations were neither consistent between models, nor for each model between sites (Fig. 4). The effect of T on the residuals was low up to temperatures of 10 °C. For higher T, site and model dependent changes in residuals were observed. For Collelongo all models strongly overestimated NEE for high temperatures. The relation between residuals and RAD was less site dependent, although in Collelongo the summer days with high RAD values led again to underestimations of net carbon storage (Fig. 4). Site-to-site inconsistency was also observed for the dependence of the model residuals on VPD observations. However, for each site individually, the models largely reacted in the same way. ISBA_{cc} and CARAIB performed independently of DRI, while 4C residuals varied as a function of DRI.

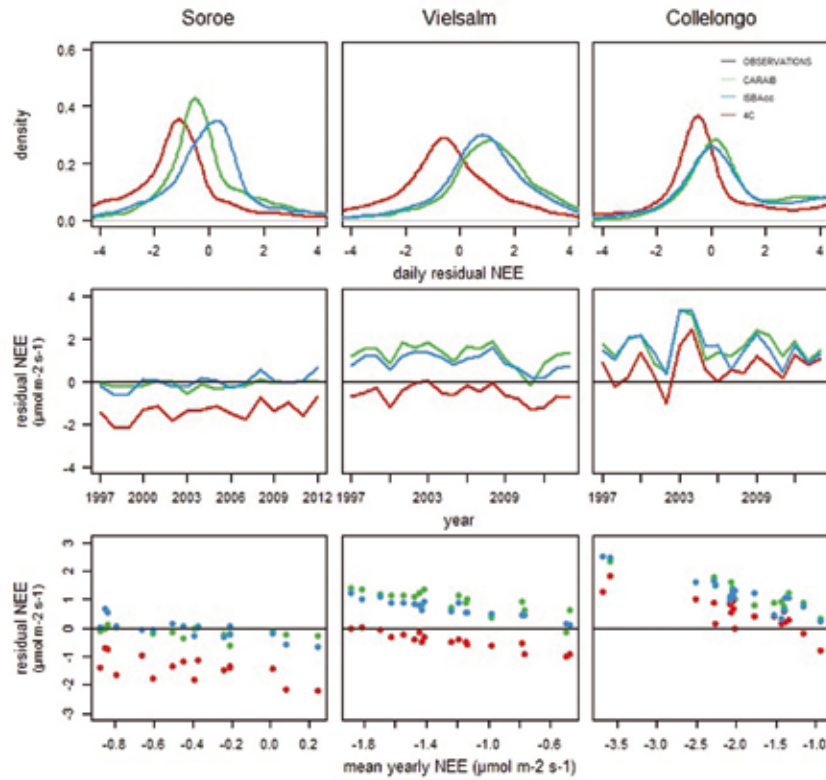


Fig. 2. For the three study sites: (i) the distribution density of the daily net ecosystem exchange (NEE) residuals (top row), (ii) the yearly averaged residuals over the whole study period (middle row), and (iii) yearly average residuals versus average yearly observed values (bottom row).

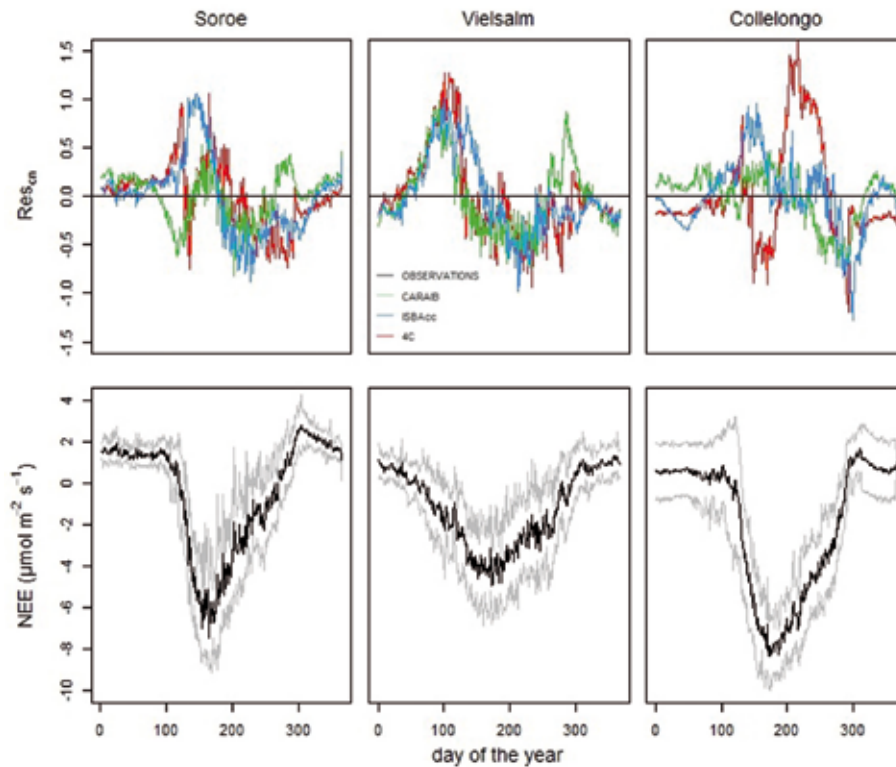


Fig. 3. Day of the year averages of the residuals of the centered and normalized simulations and observations (1997–2012 for Soroe; 1997–2014 for Vielsalm and Collelongo) for the three study sites and for each of the models (top row) and the day of the year averages of the NEE observations (black line) with their standard deviations (grey lines). NEE: net ecosystem exchange.

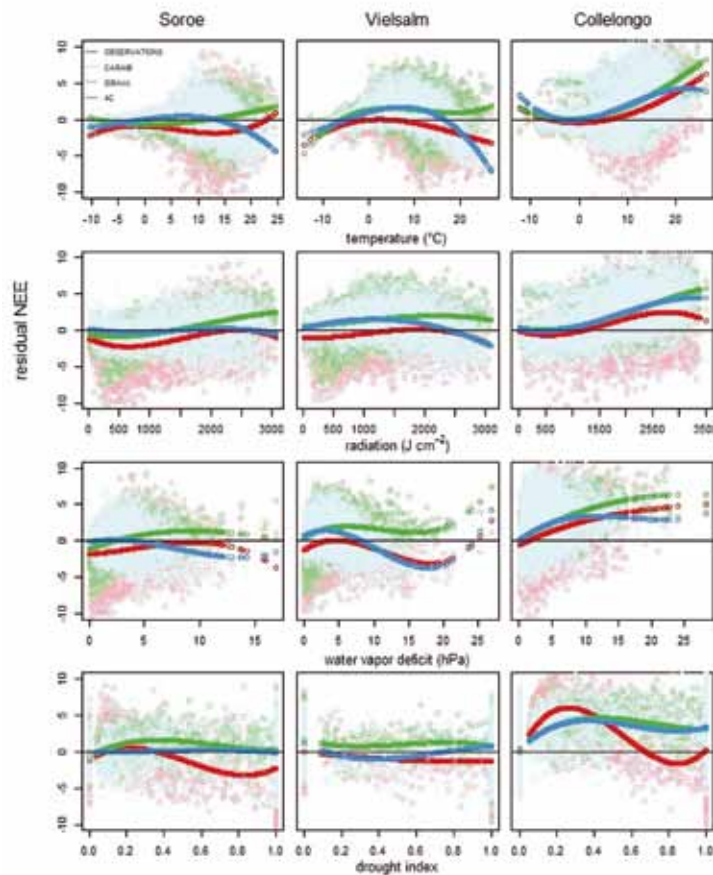


Fig. 4. Daily residuals (light colored circles) of net ecosystem exchange (NEE) as a third order polynomial function (the dots represent the individual daily values) of daily observations of air temperature (top row), incoming radiation (second row from top), vapor pressure deficit (third row from top) and drought index (bottom row) for the three models at the three study sites.

3.3. Wavelet analysis

The meteorological variables T, RAD, u, VPD as well as the DRI all gave, as expected, a significant yearly signal, being least pronounced for u and VPD. For u there was a significant recurrent pattern for higher frequencies (between a few days and two weeks), reflecting the well-known 4-day peak of meteorological processes (Vinichenko 1970). The DRI showed a significant signal around 128 days meaning that there was a pattern in drought peaks approximately three times a year. In Collelongo, a significant half-yearly VPD signal was present (Fig. 5, top row).

The strong power of the yearly signal in NEE observations was depicted by all models (Fig 5, middle row). 4C and ISBA_{cc} underestimated the strength of this yearly signal for Collelongo; CARAIB slightly overestimated its strength at all sites as did 4C for Vielsalm. The strength of the signal simulated by ISBA_{cc} seemed exactly right in Soroe and Vielsalm. A recurrent half-yearly pattern in NEE was furthermore simulated for all sites and was observed in Soroe and Collelongo (Fig. 5, middle row). The inter-site differences in intra-annual observed NEE variability were obvious from the power spectra plotted

in the time-frequency domain (Fig. 6). In Vielsalm there was a higher power at the half-yearly time scale in some of the years, while for the other two sites it was present every year. For Collelongo, the strength of the half-yearly signal simulated by ISBA_{cc} and 4C seemed to be close to the observed one, while CARAIB underestimated this temporal variability (Fig. 5, middle row). For Soroe, 4C underestimated the half-yearly signal and in Vielsalm all models overestimated the intra-annual variability leading to a false half-yearly signal in the simulations. 4C overestimated the strength of the spectral power in the NEE time series at higher frequency (inter-monthly to inter-weekly) at all sites, but especially for Collelongo. There was also a significant NEE signal for Soroe and for Collelongo at four months (period approximately 128 days). For periods below one month there were no significant signals for any of the sites.

The results of the WA for the observed and simulated values showed that the yearly peak, driven by the annual solar cycle, was properly modeled, but the same analysis on the Res_{cn} showed a dominating asynchrony between simulations and observations on this time scale (Fig. 6, bottom row). The annual signal as well as the half-yearly and the quarterly signal observed for NEE and

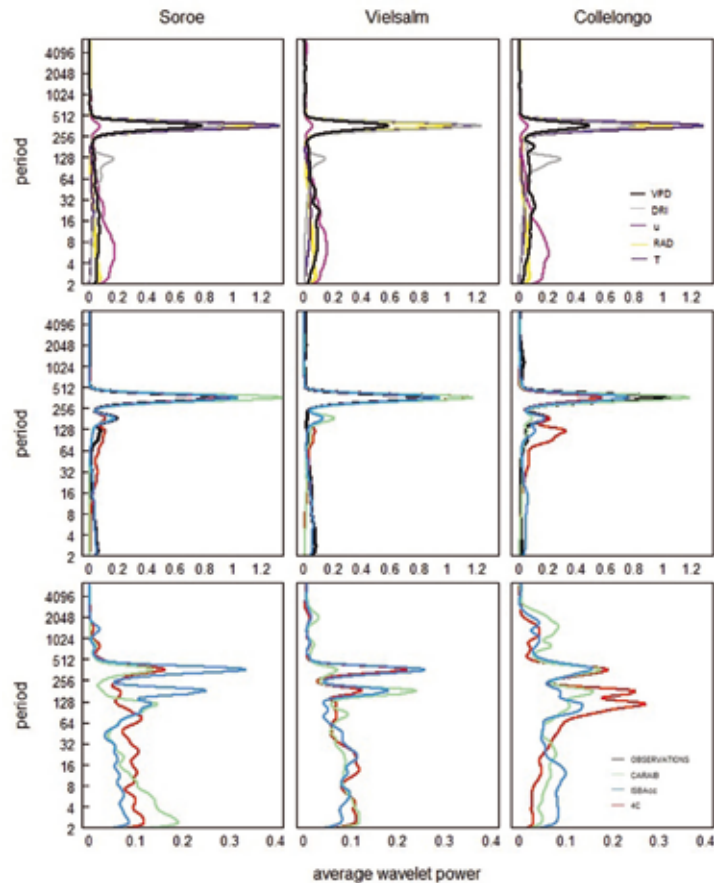


Fig. 5. Average power of the wavelet transform for the three study sites. Top row: as a function of the period for the observed meteorological variables, i.e. temperature (T), incoming radiation (RAD), wind speed (u), vapor pressure deficit (VPD) and drought index (DRI). Middle row: for net ecosystem exchange (NEE) simulated by the three models and from observations. Bottom row: power spectra of the wavelet analysis of the residuals of the centered and normalized simulated and observed values of NEE for the three models.

the meteorological variables were clearly visible in the power spectra of Res_{cn} . The spectral power of the half-yearly asynchronies between simulations and observations were almost as important as that on the yearly time scale. This could not directly be linked to the seasonal pattern of drought neither to the power spectrum of any other environmental variable studied here. Also on higher frequencies (inter-monthly and inter-weekly), where more precipitation related processes and time lagged responses occurred, the model results all significantly deviated from the observed power spectrum. At a frequency of approximately 128 days, where a drought peak was present on all sites, significant shortcomings in model performance were found at two of the three sites, but not in Vielsalm. Since the Res_{cn} for evapotranspiration (data not shown) did not show this 128-days signal but NEE did, the model shortcoming might be linked to an incorrect incorporation of the effects of drought on NEE. For periods between 16 and 64 days, where T, RAD as well as the observed NEE values lacked significant power, unexplained significant periodicity in Res_{cn} still remained. For time periods shorter than 16 days the

power spectra were no longer significant. The calculated scaling exponents were between -1 and -1.16 meaning that there was no memory of errors over time scales, i.e. errors were independent on different time-scales.

3.4. Singular spectrum analysis

The red noise test showed that there was no significant recurrent pattern on the multi-annual NEE time series, but the SSA allocated high variabilities to bin 1. The variability on this time scale was poorly assessed by the models: the variance that was attributed to this bin was incorrect (Fig. 7) and the NMEE values between the bin specific simulations and the observations were large (Table 4). For Soroe less variance was explained by the multi-annual variability than for the other sites. For this site, the variance explained by the multi-annual signal was overestimated by 4C, leading to high NMEE values (Table 4). For Vielsalm and Collelongo, the variance attributed to long-term changes in NEE was generally underestimated, but less by 4C than by the

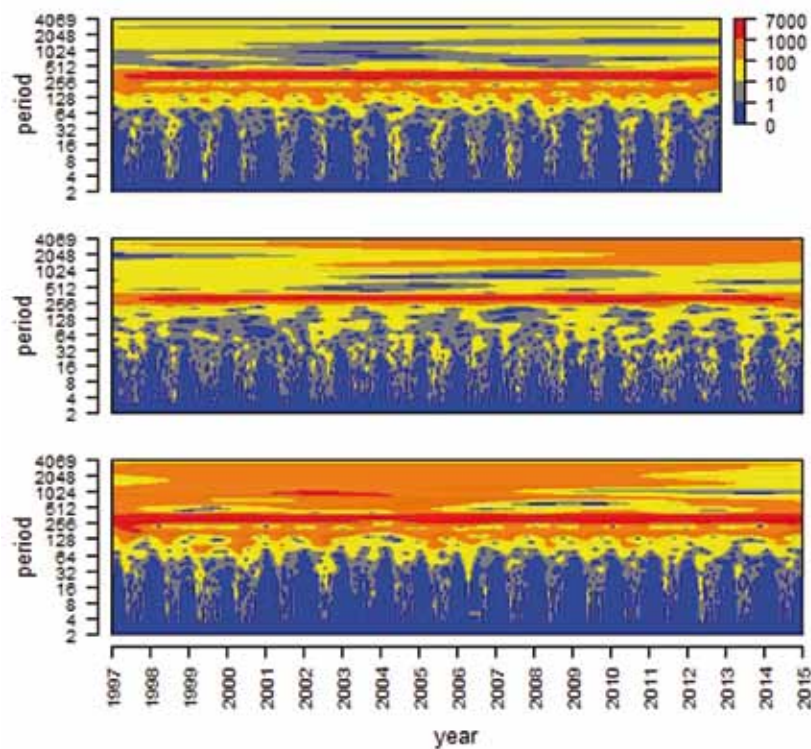


Fig. 6. Plots of the spectral power (colors) as a function of time and period for the three models for the observed net ecosystem exchange at the three study sites i.e. Soroe (top), Vielsalm (middle) and Collelongo (bottom).

DGVMs. For each site, 4C attributed more variance to this bin 1 than the other models. For other frequency bins the NMEE values were site and model dependent without clear patterns. Approximately half of the total variance of the observed NEE time series could be attributed to the annual variability at all sites (bin 2; Fig. 7). CARAIB always overestimated the contribution of the annual variability. However, this did not lead to very high NMEE values (Table 4). ISBA_{cc} overestimated the percentage explained by the yearly signal for Soroe and Vielsalm, while 4C underestimated it for Soroe and Collelongo. Bin 3 (the half-yearly signal) was significant for the observations in Soroe (10.12%) and Collelongo (7.80%). In Vielsalm this bin only explained 0.45% of the observed variance in the NEE time series. Nevertheless, the models did simulate a significant half-yearly signal for Vielsalm. While in the WA the half-yearly signal seemed to be better simulated by ISBA_{cc} and 4C, the variance they attributed to bin 3 was not closer to the observed variance than the variance of that bin estimated by CARAIB. In Soroe bin 4 (inter-monthly or seasonal variability) was significant, even if it explained only 3.90% of the total variance. Higher frequency bins explained small parts of the variance and were not significant for the observations at any of the sites. 4C, giving significant signals for bin 4 and 5 systematically overestimated the variance explained by high frequency bins. CARAIB and ISBA_{cc} had no significant signal in bins 4, 5 and 6.

4. Discussion

4.1. NEE values and scalar statistical measures

This study focused on NEE. This variable is directly measured by EC techniques but the models computed NEE as the net result of R_{eco} and GPP. The NEE values were not always correctly simulated by the models, but the observed model errors for NEE could be caused by either wrongly simulated GPP or R_{eco} values or both. Sometimes NEE values seemed to be adequately simulated, while it was actually the net result of simulated GPP and R_{eco} values, both not completely correctly simulated.

A combination of SSMs showed that the model's predictive ability and their relative error compared to the other models were dependent on the site. None of the models had the best or the worst fit for all three sites. This could partly be due to auto- and heterotrophic respiration processes incorporated in the models. The former is modelled by 4C using the fixed fraction calculation method (Landsberg & Waring 1997), that yields important differences in model results as compared to models (e.g. ISBA_{cc} and CARAIB) incorporating maintenance respiration separate from photosynthesis (Hickler et al. 2015). Generally, there is a lack in our understanding of the soil C–climate interactions, especially over the longer term (Crowther et al. 2016). The models evaluated in this contribution simulated

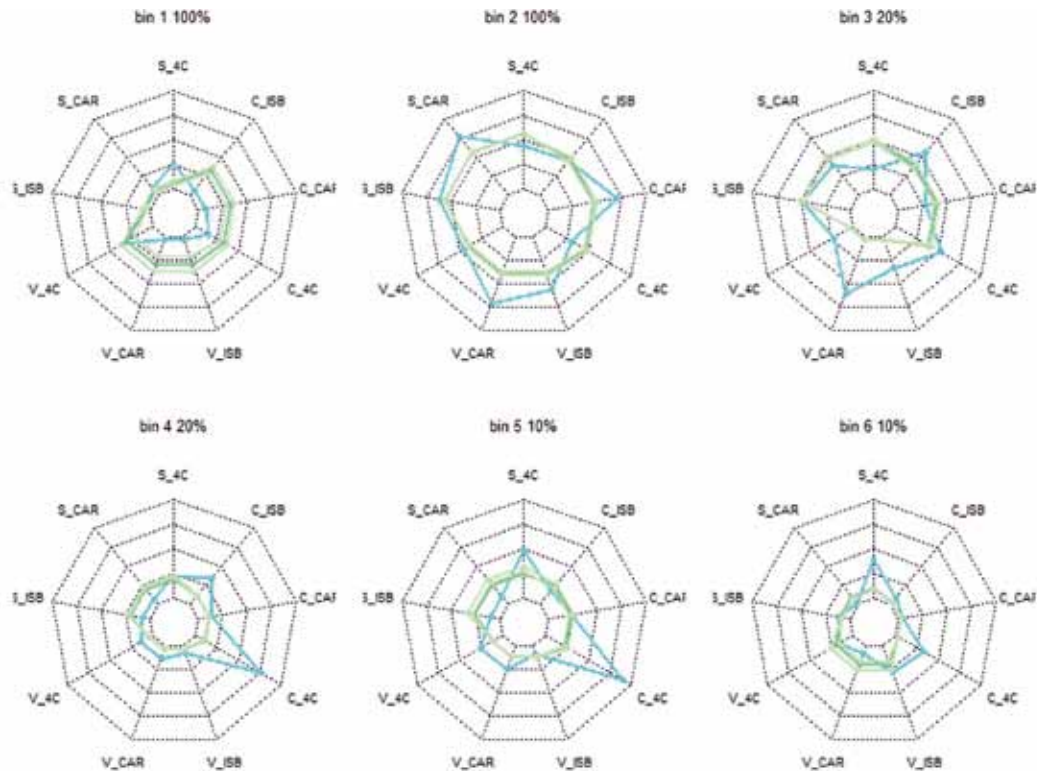


Fig. 7. Radar plot of the singular spectrum analysis for the simulated net ecosystem exchange (NEE) in blue, for the observed NEE (green), and for the observed NEE plus and minus the joint uncertainty (light green). S: Soroe, V: Vielsalm, C: Collelongo, CAR: CARAIB, ISB:ISBA_{CC}.

Table 4. Normalized median Euclidean error (NMEE) for the six bins of the singular spectrum analysis for the three models at the three sites.

	Soroe			Vielsalm			Collelongo		
	4C	CARAIB	ISBA _{CC}	4C	CARAIB	ISBA _{CC}	4C	CARAIB	ISBA _{CC}
bin 1	4.52	0.5	0.52	1.38	3.28	2.37	1.26	1.84	1.78
bin 2	0.27	0.19	0.31	0.42	0.15	0.36	0.44	0.5	0.54
bin 3	0.3	0.35	0.63	3.43	4.18	3.45	0.7	0.66	0.5
bin 4	0.38	0.72	0.59	1.89	1.02	0.98	1.75	0.46	0.73
bin 5	0.73	0.63	0.6	1.14	0.97	0.67	1.03	0.59	0.75
bin 6	1.61	0.68	0.93	1.04	0.76	0.98	2.32	0.46	0.99

soil respiration in different ways. Other processes as the coupling of transpiration and photosynthesis, allocation rules and the sink activity of the trees, as well as the effect of nutrient availability are also of prominent importance for reliable carbon exchange simulations (Hickler et al. 2015). Although some SSMs are more robust than others (Li & Zhao 2006) and give a first impression of how well the simulations fit with the observations, none of them provides information about the specific timing of these model errors and the environmental situations in which they occur. Therefore, more advanced evaluation methods as the ones discussed below, should be used in addition to SSMs for more reliable model evaluations.

4.2. Residual analysis

The different model-data asynchronies between the models partly resulted from the different ways in which the evolution of LAI and of phenology is represented in the different models. Although the included processes were different, all three models showed high NEE residuals at moments of transition phases. Another reason for the asynchronies could be that parameter values are often representing adult leaves and the physiological responses of both young or senescing leaves were not well represented.

Surprisingly, the correlations between residuals and yearly NEE were negative for all models. The models overestimated NEE in periods of extreme large carbon uptake and mostly underestimated in periods of carbon release. This could cause problems for predicting NEE

in more extreme environments and under future climate change scenarios. Since the residuals of the three models were not systematically cross-correlated, we can not conclude that the observed NEE values contained errors caused by the assumptions in data processing.

Since the PVM processes and the errors on their outcome are neither linearly nor univariately related to climate variables, we might get incomplete and possibly misleading information using univariate relations. Responses to environmental input variables are hard to interpret due to multicollinearity between model structure and parameters; a univariate study could reflect the result of the multicollinearity rather than the cause of model errors. The dependence of the model error on the meteorological input variables were highly site dependent, meaning that such relations were strongly affected by the quality of the site characteristic data, the model context and by the inherent differences of the different forest systems. This site dependency calls for caution in the interpretation of such univariate relations and clearly illustrates the need to evaluate models for several sites, even more with the ambition for long-term prediction. Responses to meteorological variables do not only directly influence modelled NEE, but also leaf development and senescence. A thorough improvement of the physiological and phenological process description during these periods and also of winter activity will be helpful. Further, residual analysis techniques including bivariate plots, added variable plots (Medlyn et al. 2005) or principal component analysis, could improve our understanding of the environmental dependency of the model errors on NEE simulations.

4.3. Spectral analyses

Spectral analyses are in use to detect and to quantify temporal patterns in model simulations, in observations (cfr. Stoy et al. 2005), in their dissimilarities (cfr. Dietze et al. 2011) and to test the statistical significance of those patterns (Mahecha et al. 2010). On the multi-annual time scale, we found high bin-specific model errors (NMEE), as did also other authors (e.g. Braswell et al. 2005; Sequera et al. 2006), which were linked to the correlation between residuals and observed NEE values. Apparently, certain modelled negative feedback mechanisms affecting NEE were overestimated with respect to reality, where more extreme values occurred. The effect of nitrogen limitation or of stimulation on photosynthesis is very important on the long-term time-scale and needs further investigation. Models not incorporating this effect might overestimate GPP values as a result of CO₂ fertilization, an effect that is often overestimated by PVMs (Anav et al. 2015). Nitrogen limitation effects were only incorporated by 4C, as was also the effect of forest management during the study period. Eventual forest disturbances (e.g. pest plagues) were included in none of the models. On stand scale, such local high impact events and their

lag-effects have a high impact on the long-term evolution of the carbon balance (Anav et al. 2015). On the multi-annual time scale, the spectral power was shown to be often improperly simulated by PVMs (e.g. Braswell et al. 2005). Furthermore, the extraction reliability of the SSA method for low-frequency modes is low (Mahecha et al. 2010). Finally, the time series were not centered before the SSA analysis; the NMEE values did thus not only reflect asynchronies but the entire model error including the model bias.

Regarding the observed and simulated half-yearly spectral peaks, a link to phenology and LAI development was suggested because it more often appears in deciduous forests than in evergreen forests (Mahecha et al. 2010; Dietze et al. 2011). Indeed, Vielsalm consists for one third of conifers. In this study however, the stand was modelled as a mono-species beech forest. Our study, using the longest available EC time series, supports the hypothesis that the asynchrony between simulations and observations was large in spring and autumn, dropping to lower levels in full summer and full winter (see Fig. 2). Inter-annual phenology variability might explain a large part of yearly NEE fluctuations (Keenan et al. 2012) and the way it is incorporated in simulation models affects model performance (Richardson et al. 2012). Further research on this intra-annual variability of carbon exchange and especially on the effect of both transition phases, remains necessary to improve model performance.

4C often overestimated the importance of high frequency variability (inter-monthly to inter-daily) in NEE. The dependence on T which is used to redistribute weekly simulated NEE values to daily values could be too sensitive. Other reasons for asynchrony on smaller frequency bins, by all models, could possibly be ascribed to the simplifications of the forest structure and the vertical radiation partitioning through the canopy, affecting photosynthesis as well as respiration on small time scales. One possible cause for the significant periodicity in Res_{cn} for periods between 16 and 64 days, could be the influence of VPD (Dietze et al. 2011). Also, NEE observations and model input variables measured at the site are prone to random and/or systematic measurement errors. The former might have large effects on the time scale specific analyses of the highest frequencies (Hollinger & Richardson, 2005). An important part of the uncertainty in NEE observations is ascribed to the assumptions in the NEE calculation procedure (Aubinet et al. 2012). Since observational data are often incomplete and models show context errors, it is important to evaluate the models for several sites to discover the real systematic problems in the model structure. While model residuals were shown to maintain a certain temporal correlation structure varying over sites (Richardson et al. 2008), the scaling exponent from our WA of the normalized residuals did not give evidence for a lingering effect of errors over time scales.

5. Conclusions

This study proved the added value of using a set of different statistical evaluation methods and data from multiple sites (without site-specific calibration) for long-term model evaluations. The evaluation methods not only confirm each other but also lead to new insights. Aspects for which one method only provides speculative evidence can be specified using another method. Our study highlighted the need for reviewing the accuracy of the models at the time of canopy closure in spring and canopy shed in autumn and points to other processes to be reconsidered. The study confirmed the confining effect of model complexity on the model evaluation process. In order to evaluate each proposed model process in depth, we advise to perform additional techniques including parameter sensitivity tests and the evaluation of structural changes in the models on long-term data across different sites.

Acknowledgements

This study was financially supported by the Belgian Science Policy Office [BELSPO BRAIN-be program, contract number BR/121/A2/MASC] as well as by the Methusalem funding to the PLECO Research Centre of Excellence. The first author has been supported by a Short Term Scientific Mission of the European COST Action FP1304 PROFOUND. CPOR acknowledges funding from the German Federal Ministry of Education and Research [BMBF, grant number 01LS1201A1]. We thank Giorgio Matteucci, Tanguy Manise and Andreas Ibrom for having provided additional site specific information.

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

References

- Anav, A., Frielingstein, P., Beer, C., Ciais, P., Harper, A., Jones, C. et al., 2015: Spatio-temporal patterns of terrestrial gross primary production. *Reviews of Geophysics*, 53:785–818.
- Aubinet, M., Grelle, A., Ibrom, A., Rannik, U., Moncrieff, J., Foken, T. et al., 2000 : Estimates of the annual net carbon and water exchange of forests: The EUROFLUX methodology, *Advances in Ecological Research*, 30:113–175.
- Aubinet, M., Heinesch, B., Perrin, D., Moureaux, C., 2005: Discriminating net ecosystem exchange between different vegetation plots in a heterogeneous forest. *Agricultural and Forest Meteorology*, 132: 315–328.
- Aubinet, M., Vesala, T., Papale, D. (eds.), 2012: *Eddy Covariance: A Practical Guide to Measurement and Data Analysis*. Springer Atmospheric Sciences, Dordrecht, Netherlands, 424 p.
- Botkin, D., 1993: *Forest Dynamics: An Ecological Model*. Oxford University Press, Oxford & New York, UK. 309 p.
- Braswell, B. H., Sacks, W. J., Linder, E., Schimel, D. S., 2005: Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. *Global Change Biology*, 11:335–355.
- Bugmann, H., Grote, R., Lasch, P., Lindner, M., Suckow, F., 1997: A new forest gap model to study the effects of environmental change on forest structure and functioning. In: Mohren, G. M. J., Kramer, K., Sabate, S. (eds.): *Impacts of Global Change of Tree Physiology and Forest Ecosystem*. Kluwer Academic Publisher, Dordrecht, The Netherlands, p. 255–261.
- Calvet, J. C., Noilhan, J., Roujean, J. L., Bessemoulin, P., Cabelguenne, M., Olioso, A. et al., 1998: An interactive vegetation SVAT model tested against data from six contrasting sites. *Agricultural and Forest Meteorology*, 92:73–95.
- Carrer, D., Roujean, J. L., Lafont, S., Calvet, J. C., Boone, A., Decharme, B. et al., 2013: A canopy radiative transfer scheme with explicit FAPAR for the interactive vegetation model ISBA-A-gs: impact on carbon fluxes. *Journal of Geophysical Research: Biogeosciences*, 118:888–903.
- Chen, C. W., 1993: The response of plants to interacting stresses: PGSM Version 1.3 Model Documentation. EPRI TR-101880. Electric Power Research Institute, Palo Alto, CA, USA, 128 p.
- Collatz, G. J., Ball, J. T., Grivet, C., Berry, J. A., 1991: Physiological and environmental regulation of stomatal conductance, photosynthesis and transpiration: a model that includes a laminar boundary layer. *Agricultural and Forest Meteorology*, 54:107–136.
- Crowther, T. W., Todd-Brown, K. E. O., Rowe C. W., Wieder, W. R., Carey, J. C., Machmuller, M. B. et al., 2016: Quantifying global soil carbon losses in response to warming. *Nature*, 540:104–108.
- Davidson, R. L., 1969: Effect of root/leaf temperature differentials on root/shoot ratios in some pasture grasses and clover. *Annals of Botany*, 33:561–569.
- Decharme, B., Boone, A., Delire, C., Noilhan, J., 2011: Local evaluation of the interaction between soil biosphere atmosphere soil multilayer diffusion scheme using four pedotransfer functions. *Journal of Geophysical Research: Biogeosciences*, 116:1984–2012.

- De Pury, D. G., Farquhar, G. D., 1997: Simple scaling of photosynthesis from leaves to canopies without the errors of big-leaf models. *Plant, Cell and Environment*, 20:537–557.
- Dietze, M. C., Vargas, R., Richardson, A. D., Stoy, P. C., Barr, A. G., Anderson, R. S. et al., 2011: Characterizing the performance of ecosystem models across time scales: a spectral analysis of the North American Carbon Program site-level synthesis. *Journal of Geophysical Research: Biogeosciences*, 116: G04029.
- Dury, M., Hambuckers, A., Warnant, P., Henrot, A.-J., Favre, E., Ouberdous, M. et al., 2011: Response of the European forests to climate change: a modeling approach for the 21st century. *iForest – Biogeosciences and Forestry*, 4:82–99.
- Dyck, S., Peschke, G., 1995: *Grundlagen der Hydrologie*. Verlag für Bauwesen, Berlin, Germany, 536 p.
- Farquhar, G. D., von Caemmerer, S., Berry, J. A., 1980: A biochemical model of photosynthetic CO₂ assimilation in leaves of C₃ species. *Planta*, 149:78–90.
- Fischer, J. B., Huntzinger, D. N., Schwalm, C. R., Sitch, S., 2014: Modeling the terrestrial biosphere. *Annual Review of Environment and Resources*, 39:91–123.
- François, L. M., Delire, C., Warnant, P., Munhoven, G., 1998: Modelling the glacial–interglacial changes in the continental biosphere. *Global Planetary Change*, 16–17:37–52.
- François, L., Utescher, T., Favre, E., Henrot, A.-J., Warnant, P., Micheels, A. et al., 2011: Modelling late Miocene vegetation in Europe: results of the CARAIB model and comparison with palaeo-vegetation data. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 304:359–378.
- Franke, U., 1990: C- und N-Dynamik beim Umsatz organischer Substanz im Boden. *Akademie der Landwirtschaftswissenschaften der DDR, Dissertation 2*, Berlin, Germany, 140 p.
- Gérard, J. C., Nemry, B., François, L. M., Warnant, P., 1999: The interannual change of atmospheric CO₂: contribution of subtropical ecosystems? *Geophysical Research Letters*, 26:243–246.
- Gibelin, A. L., Calvet, J. C., Viovy, N., 2008: Modelling energy and CO₂ fluxes with an interactive vegetation land surface model, evaluation at high and middle latitudes. *Agricultural and Forest Meteorology*, 148:1611–1628.
- Glugla, G., 1969: Berechnungsverfahren zur Ermittlung des aktuellen Wassergehaltes und Gravitationswasserabflusses im Boden. *Albrecht-Thaer-Archiv*, 13:71–376.
- Golyandina, N., Zhigljavsky, A., 2013 (eds.): *Singular Spectrum Analysis for Time Series*. Springer Briefs in Statistics, Springer, Heidelberg, Germany, 119 p.
- Goudriaan, J., van Laar, H. H., van Keulen, H., Louwerse, W., 1985: Photosynthesis, CO₂ and plant production. In: Day, W., Atkin, R. K. (eds.). *Wheat Growth and Modelling*. Plenum Press, New York, USA. NATO AS/ Series, Series A, 86:107–122.
- Goudriaan, J., 1986: A simple and fast numerical method for the computation of daily totals of crop photosynthesis. *Agricultural and Forest Meteorology*, 38:249–254.
- Grote, R., Suckow, F., Bellmann, K., 1998: Modelling of carbon-, nitrogen-, and water balances in pine stands under changing air pollution and deposition. In: Hüttel, R. F. and Bellmann, K. (eds.): *Changes of Atmospheric Chemistry and Effects on Forest Ecosystems. A Roof Experiment Without Roof*. Kluwer Publishers, Dordrecht, The Netherlands, p. 251–281.
- Haxeltine, A., Prentice I. C., 1996: A general model for the light-use efficiency of primary production. *Functional Ecology*, 10:551–561.
- Henrot, A.-J., Utescher, T., Erdei, B., Dury, M., Hamon, N., Ramstein, G. et al., 2017: Middle Miocene climate and vegetation models and their validation with proxy data. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 467:95–119.
- Hickler, T., Vohland, K., Feehan, J., Miller, P. A., Smith, B., Costa, L. et al., 2012: Projecting the future distribution of European potential natural vegetation zones with a generalized, tree species-based dynamic vegetation model. *Global Ecology and Biogeography*, 21:50–63.
- Hickler, T., Amming, A., Werner, C., 2015: Modelling CO₂ impacts on forest productivity. *Current Forestry Reports*, 1:69–80.
- Hollinger, D. Y., Richardson, A. D., 2005: Uncertainty in eddy covariance measurements and its application to physiological models. *Tree Physiology*, 25:873–885.
- Jacobs, C. M. J., 1994: Direct impact of atmospheric CO₂ enrichment on regional transpiration, PhD thesis, Agricultural University, Wageningen, The Netherlands, 192 p.
- Jansson P. - E., 1991: Simulation model for soil water and heat conditions. Description of the SOIL model (vol. 165). Swedish University of Agricultural Sciences, Department of Soil Sciences, Division of Agricultural Hydrotechnics, Uppsala, Sweden, 72 p.
- Joetzjer, E., Delire, C., Douville, H., Ciais, P., Decharme, B., Carrer, D. et al., 2015: Improving the ISBA_{cc} land surface model simulation of water and carbon fluxes and stocks over the Amazon forest. *Geoscientific Model Development*, 8:1709–1727.
- Jorritsma, I. T. M., van Hees, A. F. M., Mohren G. M. J., 1999: Forest development in relation to ungulate grazing: A modeling approach. *Forest Ecology and Management*, 120:23–34.
- Kartschall, T., Döring, P., Suckow, F., 1989: Simulation of nitrogen, water and temperature dynamics in soil. *Systems Analysis Modelling Simulation Journal*, 6:117–123.
- Keane, R. E., Morgan, P., Running, S. W., 1996: FIRE-BGC – a mechanistic ecological process model for simulating fire succession on coniferous forest landscapes of the northern Rocky Mountains. Research Paper INT-RP-484, United States Department of Agriculture, Forest Service, Intermountain Research Station, Missoula, MT, USA, 122 p.

- Keenan, T. F., Baker, I., Barr, A., Ciais, P., Davis, K., Dietze, M. et al., 2012: Terrestrial biosphere model performance for inter-annual variability of land-atmosphere CO₂ exchange. *Global Change Biology*, 18:1971–1987.
- Koitzsch, R., 1977: Schätzung der Bodenfeuchte aus meteorologischen Daten, Boden- und Pflanzenparametern mit einem Mehrschichtmodell. *Zeitschrift für Meteorologie*, 27:302–306.
- Krause, P., Boyle, D. P., Base, F., 2005: Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, 5:89–97.
- Landsberg, J. J., Waring, R. H., 1997: A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. *Forest Ecology and Management*, 95:209–228.
- Le Quéré, C., Andrew, M., Canadell, J. G., Sitch, S., Korsbakken, J. I., Peters, G. P. et al., 2016: Global carbon budget 2016. *Earth System Science Data*, 8:605–649.
- Larocque, G. R., Bhatti, J., Arsenault, A., 2014: Integrated modelling software platform development for effective use of ecosystem models. *Ecological Modelling*, 288:195–202.
- Lasch, P., Badeck, F. - W., Suckow, F., 2005: Model-based analysis of management alternatives at stand and regional level in Brandenburg (Germany). *Forest Ecology and Management*, 207:59–74.
- Lasch, P., Badeck, F. - W., Suckow, F., Lindner, M., Mohr, P., 2005: Model-based analysis of management alternatives at stand and regional level in Brandenburg (Germany). *Forest Ecology and Management*, 207:59–74.
- Lasch-Born, P., Suckow, F., Gutsch, M., Reyer, C., Hauf, Y., Murawski, A. L. et al., 2015: Forests under climate change: potential risks and opportunities. *Meteorologische Zeitschrift*, 24:157–172.
- Lau, K. M., Wang, H. - Y., 1995: Climate signal detection using wavelet transform: How to make a time series sing? *Bulletin of the American Meteorological Society*, 76:2391–2402.
- Laurent, J. - M., François, L., Bar-Hen, A., Bel, L., Cheddadi, R., 2008: European bioclimatic affinity groups: data-model comparisons. *Global and Planetary Change*, 61:28–40.
- Law, B. E., Falge, E., Gu, L., Baldocchi, D. D., Bakwin, P., Berbigier, P. et al., 2002: Environmental controls over carbon dioxide and water vapor exchange of terrestrial vegetation. *Agricultural and Forest Meteorology*, 113:97–120.
- Li, H., Zhang, F., Li, Y., Wang, J., Zhang, L., Zhao, L. et al., 2016: Seasonal and inter-annual variations in CO₂ fluxes over 10 years in an alpine shrubland on the Qinghai-Tibetan plateau, China. *Agricultural and Forest Meteorology*, 228–229:95–103.
- Li, X. R., Zhao, Z., 2006: Evaluation of estimation algorithms, part I: Incomprehensive measures of performances. *IEEE Transactions on Aerospace and Electronic Systems*, 42:1340–1358.
- Lloyd, J., Taylor, J. A., 1994: On the temperature dependence of soil respiration. *Functional Ecology*, 8:315–323.
- Loehle, C., LeBlanc, D., 1996: Model-based assessments of climate change effects on forests: a critical review. *Ecological Modelling*, 90:1–31.
- Mahecha, M. D., Reichstein, M., Lange, H., Carvalhais, N., Bernhofer, C., Grünwald, T. et al., 2007: Characterizing ecosystem-atmosphere interactions from short to interannual time scales. *Biogeosciences*, 4:743–758.
- Mahecha, M. D., Reichstein, M., Jung, M., Seneviratne, S. I., Zaehle, S., Beer, C. et al., 2010: Comparing observations and process-based simulations of biosphere-atmosphere exchange on multiple time scales. *Journal of Geophysical Research: Atmospheres*, 115:7111–7125.
- Mäkelä, A., 1990: Modeling structural-functional relationships in whole-tree growth: resource allocation. In: Dixon, R. K., Meldahl, R. S., Ruark, G. A., Warren, W. G (eds.): *Process Modeling of Forest Growth Responses to Environmental Stress*. Timber Press, Portland, Oregon, 81–95.
- Masson, V., Le Moigne, P., Martin, E., Faroux, S., Alias, A., Alkama, R. et al., 2013: The SURFEXv7.2 land and ocean surface platform for coupled or offline simulation of Earth surface variables and fluxes. *Geoscientific Model Development*, 6:929–960.
- Medlyn, B. E., Robinson, A. P., Clement, R., McMurtrie, E., 2005: On the validation of models of forest CO₂ exchange using eddy covariance data: some perils and pitfalls. *Tree Physiology*, 25:839–857.
- Mintz, Y., Walker, G., 1993: Global fields of soil moisture and land surface evapotranspiration derived from observed precipitation and surface air temperature. *Journal of Applied Meteorology*, 32:1305–1334.
- Nash, J. E., Sutcliffe, J. V., 1970: River flow forecasting through conceptual models part I – A discussion of principles. *Journal of Hydrology*, 10:282–290.
- Nemry, B., François, L. M., Warnant, P., Robinet, F., Gérard, J. C., 1996 : The seasonality of the CO₂ exchange between the atmosphere and the land biosphere: a study with a global mechanistic vegetation model. *Journal of Geophysical Research-Atmospheres*, 101:7111–7125.
- Noilhan, J., Planton, S., 1989: A simple parameterization of land surface processes for meteorological models. *Monthly Weather Review*, 117:536–549.
- Noilhan, J., Mahfouf, J. F., 1996: The ISBA land surface parameterization scheme. *Global and Planetary Change*, 13:145–159.
- Otto, D., Rasse, D., Kaplan, J., Warnant, P., François, L., 2002: Biospheric carbon stocks reconstructed and the Last Glacial Maximum: comparison between general circulation models using prescribed and computed sea surface temperature. *Global Planetary Change*, 33:117–138.

- Pan S., Tian, H., Dangal, S. R. S., Ouyang, Z., Tao, B., Ren, W. et al., 2014: Modeling and monitoring terrestrial primary production in a changing global environment: toward a multiscale synthesis of observation and simulation. *Advances in Meteorology*, 2014:1–17.
- Parton, W. J., Schimel, D. S., Cole, C. V., Ojima, D. S., 1987: Analysis of factors controlling soil organic matter levels in Great Plains grasslands. *Soil Science Society of America Journal*, 51:1173–1179.
- Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Valentini, R., Aubinet, M. et al., 2005: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global Change Biology*, 11:1–16.
- Reyer, C., 2015: Forest productivity under environmental change – a review of stand-scale modeling studies. *Current Forestry Reports*, 1:53–68.
- Richardson, A. D., Mahecha, M. D., Falge, E., Kattge, J., Moffat, A. M., Papale, D. et al., 2008: Statistical properties of random CO₂ flux measurement uncertainty inferred from model residuals. *Agricultural and Forest Meteorology*, 148:38–50.
- Richardson, A. D., Anderson, R. S., Arian, M. A., Barr, A. G., Bohrer, G., Chen, G. et al., 2012: Terrestrial biosphere models need better representation of vegetation phenology: results from the North American Carbon Program Site synthesis. *Global Change Biology*, 18:566–584.
- Running, S. W., Gower, S. T., 1991: FOREST-BGC, A general model of forest ecosystem processes for regional applications II. Dynamic carbon allocation and nitrogen budgets. *Tree Physiology*, 9:147–160.
- Saxton, K., Rawls, W. J., Romberger, J., Papendick, R., 1986: Estimating generalized soil-water characteristics from texture. *Soil Science Society of America Journal*, 50:1031–1036.
- Schaber, J., Badeck, F. - W., 2003: Physiology based phenology models for forest tree species in Germany. *International Journal of Biometeorology*, 47:193–201.
- Schultz, M., Mudelsee, M., 2002: REDFIT: estimating red-noise spectra directly from unevenly spaced paleoclimatic time series. *Computers & Geosciences*, 28:421–426.
- Shinozaki, K., Yoda, K., Hozumi, K., Kira, T., 1964: A quantitative analysis of plant form – the pipe model theory. I. Basic analysis. *Japanese Journal of Ecology*, 14:97–105.
- Siqueira M. B., Katul, G. G., Sampson, D. A., Stoy, P. C., Juang, J. - Y., McCarthy, H. R. et al., 2006: Multiscale model intercomparisons of CO₂ and H₂O exchange rates in a maturing southeastern US pine forest. *Global Change Biology*, 12:1189–1207.
- Stoy, P. C., Katul, G. G., Siqueira, M. B., Juang, J. - Y., McCarthy, H. R., Kim, H. - S. et al., 2005: Variability in net ecosystem exchange from hourly to inter-annual time scales at adjacent pine and hardwood forests: a wavelet analysis. *Tree Physiology*, 25:887–902.
- Suckow, F., Badeck, F. - W. Lasch, P., Schaber, J., 2001: Nutzung von Level-II-Beobachtungen für Test und Anwendungen des Sukzessionsmodells FORESEE. *Beiträge für Forstwirtschaft und Landschaftsökologie*, 35:84–87.
- Sykes, M. T., Prentice, I. C., 1996: Carbon storage and climate change in Swedish forests: A comparison of static and dynamic modelling approaches. In: Apps, M. J., Price, D. T. (eds.). *Forest Ecosystems, Forest Management and the Global Carbon Cycle*. Springer-Verlag, Berlin, Germany, NATO ASI Series, 40:69–78.
- Torrence, C., Compo, G. P., 1998: A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79:61–78.
- Trudinger, C. M., Raupach, M. R., Rayner, P. J., Kattge, J., Liu, Q., Pak, B. et al., 2007: OptIC project: An inter-comparison of optimization techniques for parameter estimation in terrestrial biogeochemical models. *Journal of Geophysical Research*, 112: G02027.
- Uppala, S. M., Kallberg, P. W., Simmons, A. J., Andrea, U., Da Costa Bechtold, V., Fiorino, M. et al., 2005: The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131:2961–3012.
- Vinnichenko, N., 1970: The kinetic energy spectrum in the free atmosphere – 1 second to 5 years. *Tellus*, 22:158–166.
- Walker, W. E., Harremoes, P. J., Rotmans, J., Van Der Sluis, J. P., Van Asselt, M. B. A., Janssen, P. et al., 2003: Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment*, 4:5–17.
- Wang, T., Brender, P., Ciais, P., Piao, S., Mahecha, M., Chevalier, F. et al., 2012: State-dependent errors in a land surface model across biomes inferred from eddy covariance observations on multiple timescales. *Ecological Modelling*, 246:11–25.
- Warnant, P., François, L. M., Strivay, D., Gérard, J.-C., 1994: CARAIB: a global model of terrestrial biological productivity. *Global Biogeochemical Cycles*, 8:255–270.
- Warnant, P., 1999: Modélisation du cycle du carbone dans la biosphère continentale à l'échelle globale. PhD thesis, Université de Liège, Liège, 276 p.
- Wythers, K. R., Reich, P. B., Tjoelker, M. G., Bolstad, P. B., 2005: Foliar respiration acclimation to temperature variable Q₁₀ alter ecosystem carbon balance. *Global Change Biology*, 11:435–449.

Appendix tables

Table A1. List of abbreviations and acronyms used in the text.

Abbreviation	Explanation
AR(1)	First order autoregressive function
bR ²	Slope of the regression times the coefficient of determination
CorrO	Correlation coefficient between the residuals and the observed NEE values
CorrE	Correlation coefficient between the residuals of the different models
D _{tr} ^c	Potential canopy transpiration demand of a cohort
DOY	Day of the year
DRI	Drought index
Dtr	Potential canopy transpiration demand
EC	Eddy covariance
E _{int}	Interception evaporation
EOF	Empirical orthogonal function
DGVM	Dynamic global vegetation model
GPP	Gross primary production
G _{max}	Maximum stomatal conductance
g _{tot}	Unstressed stomatal conductance
kurR	Kurtosis of the daily residual distribution
LAI	Leaf area index
ME	Mean error
NEE	Net ecosystem exchange
NMEE	Normalized median Euclidean error
NRMSE	Normalized root mean squared error
NSE	Nash-Sutcliffe efficiency
PFT	Plant functional type
PVM	Process-based vegetation model
RA	Residual analysis
RAD	Incoming solar radiation
R ²	Coefficient of determination
R _{eco}	Ecosystem respiration
Rescn	Residuals of the centered and normalized simulated and observed values
sdR	Standard deviation of the daily residual distribution
SEM	Statistical evaluation method
SSM	Scalar statistical measure
skR	Skewness of the daily residual distribution
SSA	Singular spectrum analysis
T	Air temperature
u	Wind velocity
VPD	Vapour pressure deficit
WA	Wavelet analysis

Table A2. Short description of the processes included in the CARAIB global dynamic vegetation model.

Plant functional types	Plant functional type or species level. In this study parameters for beech.
Spatial scale	Grid cell containing different PFTs or point scale.
Soil input initialisation	Sand fraction, clay fraction, rooting depth per PFT/species, soil color.
Forest structure and carbon pools initialisation	Grid cell with one layer of trees and one layer containing herbs and shrubs.
Climate input	CO ₂ , air temperature, amplitude of air temperature ($T_{day_max} - T_{day_min}$), precipitation, air relative humidity, short-wave incoming radiation, wind velocity. [Daily]
Spin-up	Yes, by using ERA-interim re-analysis (Uppala et al. 2005) [Daily] Light interception by big leaf approach with separation of sun and shaded leaves (De Pury & Farquhar 1997). Time step subdivided into sunny and non-sunny portions, depending on the percentage of sunshine hours. Photosynthesis thus calculated 3 times in each time step for each PFT/species: for sun and shaded leaves during the sunny portion and for all leaves during the non-sunny portion. Light use efficiency calculated by the model of Farquhar (1980), as simplified by Collatz (1991). Radiative transfer through the canopy according to Goudriaan et al. (1985) with radiation attenuation by Beer's law. Gross primary production only calculated when air temperature >−10 °C and if LAI >0. [2-hourly]
Photosynthesis	Autotrophic respiration subdivided into growth respiration, as a fixed fraction of the carbon allocated to the growth of carbon pools, and maintenance respiration, as a Q ₁₀ function of temperature and proportional to the biomass and the C:N ratio of that pool (Warant et al. 1999) and decreasing with the average air temperature of the previous 2 months for leaves and the previous 4 years for wood/roots to mimic temperature acclimation process (Wythers et al. 2005). Heterotrophic respiration dependent on soil temperature and soil moisture (Nemry et al. 1996) [2-hourly]
Respiration	Photosynthetic products (GPP) are allocated to the metabolic (leaves and fine roots) and structural (wood and coarse roots) carbon reservoirs. The carbon partitioning between the two pools is species-specific and depends on environmental conditions (temperature and soil water) (Otto et al. 2002).
Allocation	
Carbon nitrogen balance	Constant C:N ratio prescribed at initialisation. Turnover of litter and organic matter vary with temperature and soil water. Three carbon reservoirs are considered: leaf litter, wood litter and humus. [Daily]
Soil water balance	Soil water budget modelled in the root zone. Soil hydraulic conductivity is calculated from soil texture, using the parameterization of Saxton et al. (1986). Soil water can vary from wilting point to saturation. [Daily]
Water interception storage	Parameterization from leaf bucket model run at very high temporal resolution (~1 minute), depending on precipitation, potential evapotranspiration and LAI. [Daily]
Evapotranspiration	Actual evapotranspiration calculated as the sum of snow sublimation, the evaporation of intercepted rain and the evaporation/transpiration from the soil-vegetation system. This sum cannot exceed the potential evapotranspiration calculated over the pixel from Penman's equation (e.g., Mintz & Walker 1993). Transpiration is considered as a supply function for the water transpired by the PFT/species growing on the pixel. [Daily]
Phenology	Regulated purely by evolution of LAI. LAI growth (resp. leaf fall) is initiated when the air temperature is above (resp. below) a prescribed species-dependent threshold. [Daily]
Regeneration/planting	Amount of seeds proportional to NPP. Seeds (here only beech) can colonize gaps in the canopy caused by mortality. [yearly]
Management	A prescribed fraction of biomass (leaf or wood) can be removed. [Daily] Not used in this study.
Mortality	Age dependent natural mortality and mortality caused by thermal and water deficit stress as well as by fire disturbance. [Daily]

Table A3. Short description of the processes included in the ISBA_{cc} global dynamic vegetation model.

Plant functional types	Several plant functional types in one gridcell possible, not interacting and each having their own soil. Here only temperate deciduous forest.
Spatial scale	Grid or point scale.
Soil input initialisation	Per layer: saturation, field capacity, permanent wilting point, sand fraction, clay fraction, carbon content, General: rooting depth, ground water depth.
Forest structure and carbon pools initialisation	Grid cell with one layer of trees.
Climate input	CO ₂ , air temperature, precipitation, air relative humidity, short-wave incoming radiation, long-wave incoming radiation, wind velocity. [Hourly]
Spin-up	Yes by cycling through the available meteorological data. [Hourly]
Photosynthesis	Semi-empirical parametrization of net carbon assimilation and mesophyll conductance following the photosynthesis model of Jacobs (1994) based on Goudriaan et al. (1985) and implemented by Calvet et al. (1998). 10-layer radiative transfer scheme taking into account direct and diffuse radiation and sunlit and shaded leaves to calculate photosynthesis in the canopy (Carré et al. 2013). [Hourly]
Respiration	Maintenance respiration rates of twigs, sapwood and fine root carbon pools depending linearly on biomass of the pool and its temperature, calculated by the Arrhenius temperature function (Lloyd & Taylor, 1994; Joetzer et al. 2015). Growth respiration proportional to the photosynthetic capacity of the leaves (Jacobs, 1994). Heterotrophic respiration based on the CENTURY model (Parton et al. 1987). [Hourly]
Allocation	Assimilated carbon directly allocated to leaves, twigs, aboveground and belowground wood and fine roots following the daily carbon balance of the leaves (Gibelin et al. 2006). [Daily]
Carbon nitrogen balance	Nitrogen not simulated.
Soil water balance	Multilayer (14 layers) solution of the Fourier law and the mixed-form of the Richards equation to calculate the soil energy and water budgets including freezing/thawing (Boone et al. 2000; Decharme et al. 2011). [Hourly]
Water interception storage	Depending on LAI, precipitation and a maximum interception pool. [Hourly]
Evapotranspiration	Sum of snow sublimation, evaporation of intercepted rain, transpiration and soil evaporation (Noilhan & Planton 1989). [Hourly]
Phenology	Directly resulting from the leaf carbon balance. A minimum LAI at all time (0.3 for deciduous trees). Leaves start to grow when the amount of assimilated carbon is larger than the amount of lost carbon through respiration and turnover. This depends on the incoming radiation, the temperature and is only possible when the soil moisture is not limiting. At the end of the growing season the inverse happens. [Daily]
Regeneration/planting	Not explicitly modelled, presence of a minimum LAI allowing plant functional types to grow when climatic conditions are favorable. [Daily]
Management	Not modelled
Mortality	Not explicitly modelled, except for leaves. Biomass decreases through turnover. [Daily]

Table A4. Short description of the processes included in the 4C forest model.

Plant functional types	No plant functional types. Fixed parameters available for 13 tree species, here beech.
Spatial scale	Stand-scale.
Soil input initialisation	Per layer: field capacity, permanent wilting point, soil density, pH, stone fraction, sand fraction, clay fraction, humus fraction, carbon and nitrogen content in the humus fraction, NH_4 and NO_3 content. General: rooting depth, ground water depth, evaporation depth, mineralization constant of humus in litter layer and in mineral soil, nitrification constant.
Forest structure and carbon pools initialisation	Per cohort: species, foliage biomass, fine root biomass, sapwood biomass, heartwood biomass, cross sectional area of heartwood at stem base, tree height, bole height, tree age, number of trees, diameter at crown base, diameter at breast height. Cohorts compete for light and for water and nutrients in the soil.
Climate input	CO_2 , air temperature, precipitation, air relative humidity, net radiation, wind velocity. [Daily]
Spin-up	No.
Photosynthesis	Net photosynthesis as function of environmental drivers and physiological capacity depending on light use efficiency calculated according to Haxeltine & Prentice (1996) based on the mechanistic model of Farquhar (1980) as simplified by Collatz (1991) and limited by water and nitrogen availability and maximum nitrogen uptake per cohort. Net photosynthetic fraction per cohort proportional to its share in the absorbed photosynthetic active radiation, adapted when forest structure changes and with phenology (Lambert-Beer law). [Weekly, redistributed to daily values by a Q_{10} function of air temperature]
Respiration	Autotrophic respiration proportional to photosynthetic capacity (Landsberg & Waring, 1997). Heterotrophic respiration calculated by the carbon dynamics of the soil, dependent on soil temperature and soil moisture. [Weekly, redistributed to daily values by a Q_{10} function of air temperature]
Allocation	Theory of Mäkelä (1990), functional balance hypothesis (Davidson 1969), pipe model theory (Shinozaki, 1964) and mass-conservation law. Allometric relationships dynamically responding to water and nutrient limitations. [Yearly]
Carbon nitrogen balance	Decomposition of primary organic matter to humus described by first order reactions (Grote et al. 1998). Turnover from organic matter depending on water content, soil temperature and pH (Franko 1990; Kartschall 1989). Soil carbon/nitrogen depending on the percentage in the organic matter and their turnover rates (Running & Gower 1991). Outflow of nitrogen from the root zone by plant uptake and its transport by water. [Daily]
Soil water balance	Soil water balance per soil horizon by percolation model, bucket model, water leaching and conductivity parameter depending on soil texture (Glugla 1969; Koitzsch 1997). Link to vegetation is plant available water versus transpiration demand and limited when more than 10 percent difference from field capacity (Chen et al. 1993), divided per cohort depending on its share in fine root biomass. [Daily]
Water interception storage	Depending on precipitation and evapotranspiration (Jansson 1991) and proportional to LAI. [Daily]
Evapotranspiration	Potential evapotranspiration by equation of TURC if air temperature $> 5^\circ\text{C}$ and by an equation of IVANOV if air temperature $< 5^\circ\text{C}$ (Dyck & Peschke 1989). Calculation of potential transpiration takes into account the interception evaporation and partitioned to cohorts considering their relative conductance. [Daily]
Phenology	Interaction of growth-promoting and growth-inhibiting agents driven by temperature and photoperiod (Schaber & Badeck 2003). Leaves appearing and disappearing all together at one time point when the threshold is reached. [Yearly]
Regeneration/planting	Regeneration by seed supply (Rogers & Johnson 1998), seed germination (Jorritsma et al. 1999) Not used in this study. [Yearly]
Management	Thinning (from below or from above), harvest (clear cut, shelterwood) and planting strategies options (method, strength and timing; Lasch et al. 2005). Here used by thinning to target number of trees known during the study period. [Yearly]
Mortality	Intrinsic mortality depending on maximum life span (Botkin 1993) or carbon-based stress mortality, by drought stress or light shortage or by disturbances (Keane et al. 1996; Loehle & LeBlanc 1996; Sykes & Prentice 1996). Not used in this study. [Yearly]

Table A5. Scalar statistical measures (SSM) for simulated (S) versus observed (O) net ecosystem exchange (NEE), moments of the residual distribution (for abbreviations see Table A1) and correlation coefficients between the residuals and the observed values (CorrO) and the residuals of the other models (CorrE) for the three models run for the three sites.

SSM	Interpretation	Soroe			Vielsalm			Collelongo		
		4C	CARAIB	ISBA _{cc}	4C	CARAIB	ISBA _{cc}	4C	CARAIB	ISBA _{cc}
Scalar fit statistics: how well does S reproduce O?										
R ²	Fraction of the variance in O explained by linear relation between S and O; the closer to 1 the better	0.53	0.74	0.7	0.41	0.47	0.52	0.39	0.66	0.44
bR ²	Fraction of the variance in O explained by linear relation between S and O taking into account systematic error; the closer to 1 the better	0.44	0.44	0.59	0.35	0.25	0.34	0.23	0.22	0.15
NSE	S predicts better the O than the mean of O if NSE>0; the closer to 1 the better	0.22	0.7	0.67	0.17	0.11	0.27	0.3	0.37	0.29
Scalar error estimates: how large is the relative error of the models?										
NRMSE	Measure of the relative error between S and O; the closer to 0 the better	13.8	8.6	8.9	13.9	14.4	13	19.1	18.2	19.3
NMEE	Measure of the relative error between S and O; the closer to 0 the better	0.47	0.28	0.26	0.48	0.61	0.48	0.28	0.28	0.31
Moments of the distribution of the daily residuals: how are the model errors distributed?										
ME	Value far from 0 indicates a probable systematic bias or at least many more or more extreme errors in one direction	−1.42	−0.14	−0.05	−0.59	1.25	0.92	0.53	1.48	1.37
sdR	Expresses the variability in the model errors; high variability means high random error	2.69	1.66	1.74	0.41	0.33	0.52	0.55	0.61	0.54
skR	The more negative the heavier the tail of the negative errors; the more positive the more heavy the tail of the positive errors	−26.19	0.58	−0.34	−9.81	13.37	8.12	37.85	48.35	56.79
kurR	The more negative the less peaks in the distribution and the more the variance is dominated by many, but not very extreme errors. The more positive the more peaks; the more the variance is dominated by some rare extreme errors	232.7	36.22	46.46	92.3	81.09	58.45	390.7	300.0	405.7
Are model errors correlated with the observed NEE value and the errors of other models?										
CorrO	Correlation between the residuals and the observed values. The higher the more the error is dependent on the value of the observed flux	−0.65	−0.5	−0.8	−0.93	−0.81	−0.94	−0.84	−0.93	−0.93
CorrE	Correlation between the residuals and the residuals of the other models	4C	0.65	0.43		0.75	0.4		0.42	0.4
	The higher the more the errors between two models are correlated	CARAIB		0.35			0.17			0.75