12-2011

Characterizing the performance of ecosystem models across time scales: A spectral analysis of the North American Carbon Program site-level synthesis

Michael C. Dietze
University of Illinois at Urbana-Champaign

Rodrigo Vargas
Centro de Investigación Científica y de Educación Superior de Ensenada

Andrew D. Richardson
Harvard University, arichardson@oeb.harvard.edu

Paul C. Stoy
Montana State University-Bozeman

Alan G. Barr
Atmospheric Science and Technology Directorate, Saskatoon

Characterizing the performance of ecosystem models across time scales: A spectral analysis of the North American Carbon Program site-level synthesis

Michael C. Dietze,[1] Rodrigo Vargas,[2] Andrew D. Richardson,[3] Paul C. Stoy,[4] Alan G. Barr,[5] Ryan S. Anderson,[6] M. Altaf Arain,[7] Ian T. Baker,[8] T. Andrew Black,[9] Jing M. Chen,[10] Philippe Ciais,[11] Lawrence B. Flanagan,[12] Christopher M. Gough,[13] Robert F. Grant,[14] David Hollinger,[15] R. Cesar Izaurralde,[16,17] Christopher J. Kucharik,[18] Peter Lafleur,[19] Shugang Liu,[20] Erandathie Lokupitiya,[8] Yiqi Luo,[21] J. William Munger,[22] Changhui Peng,[23] Benjamin Poulter,[24,25] David T. Price,[26] Daniel M. Ricciuto,[27] William J. Riley,[28] Alok Kumar Sahoo,[29] Kevin Schaefer,[30] Andrew E. Suyker,[31] Hanqin Tian,[32] Christina Tonitto,[33] Hans Verbeeck,[34] Shashi B. Verma,[31] Weifeng Wang,[23] and Ensheng Weng[21]

Received 20 January 2011; revised 19 September 2011; accepted 22 September 2011; published 20 December 2011.

[1] Ecosystem models are important tools for diagnosing the carbon cycle and projecting its behavior across space and time. Despite the fact that ecosystems respond to drivers at multiple time scales, most assessments of model performance do not discriminate different time scales. Spectral methods, such as wavelet analyses, present an alternative approach that enables the identification of the dominant time scales contributing to model performance in the frequency domain. In this study we used wavelet analyses to synthesize the performance of 21 ecosystem models at 9 eddy

[1]Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.
[2]Departamento de Biología de la Conservación, Centro de Investigación Científica y de Educación Superior de Ensenada, Ensenada, Mexico.
[3]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, USA.
[4]Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, Montana, USA.
[5]Climate Research Division, Atmospheric Science and Technology Directorate, Saskatoon, Saskatchewan, Canada.
[6]Numerical Terradynamic Simulation Group, University of Montana, Missoula, Montana, USA.
[7]School of Geography and Earth Sciences and McMaster Center for Climate Change, McMaster University, Hamilton, Ontario, Canada.
[8]Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, USA.
[9]Faculty of Land and Food Systems, University of British Columbia, Vancouver, British Columbia, Canada.
[10]Department of Geography and Program in Planning, University of Toronto, Toronto, Ontario, Canada.
[11]Centre d'Etudes Orme des Merisiers, Gif-sur-Yvette, France.
[12]Department of Biological Sciences, University of Lethbridge, Lethbridge, Alberta, Canada.
[13]Department of Biology, Virginia Commonwealth University, Richmond, Virginia, USA.
[14]Department of Renewable Resources, University of Alberta, Edmonton, Alberta, Canada.
[15]Northern Research Station, U.S. Department of Agriculture Forest Service, Durham, New Hampshire, USA.
[16]Pacific Northwest National Laboratory, Richland, Washington, USA.
[17]Joint Global Change Research Institute, University of Maryland, College Park, Maryland, USA.
[18]Department of Agronomy and Nelson Institute Center for Sustainability and the Global Environment, University of Wisconsin-Madison, Madison, Wisconsin, USA.
[19]Department of Geography, Trent University, Peterborough, Ontario, Canada.
[20]Earth Resources Observation and Science Center, Sioux Falls, South Dakota, USA.
[21]Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma, USA.
[22]School of Engineering and Applied Sciences and Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts, USA.
[23]Department of Biology Sciences, University of Quebec at Montreal, Montreal, Quebec, Canada.
[24]Swiss Federal Research Institute WSL, Birmensdorf, Switzerland.
[25]Now at Laboratoire des Sciences du Climat et de l'Environnement, Institut Pierre Simon Laplace, Gif-sur-Yvette, France.
[26]Northern Forestry Centre, Canadian Forest Service, Edmonton, Alberta, Canada.
[27]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA.
[28]Climate and Carbon Sciences, Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA.
[29]Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey, USA.
[30]National Snow and Ice Data Center, University of Colorado at Boulder, Boulder, Colorado, USA.
[31]School of Natural Resources, University of Nebraska–Lincoln, Lincoln, Nebraska, USA.
[32]School of Forestry and Wildlife Sciences, Auburn University, Auburn, Alabama, USA.
[33]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York, USA.
[34]Faculty of Bioscience Engineering, Laboratory of Plant Ecology, Department of Applied Ecology and Environmental Biology, Ghent University, Ghent, Belgium.

0148-0227/11/2011JG001661

covariance towers as part of the North American Carbon Program's site-level intercomparison. This study expands upon previous single-site and single-model analyses to determine what patterns of model error are consistent across a diverse range of models and sites. To assess the significance of model error at different time scales, a novel Monte Carlo approach was developed to incorporate flux observation error. Failing to account for observation error leads to a misidentification of the time scales that dominate model error. These analyses show that model error (1) is largest at the annual and 20–120 day scales, (2) has a clear peak at the diurnal scale, and (3) shows large variability among models in the 2–20 day scales. Errors at the annual scale were consistent across time, diurnal errors were predominantly during the growing season, and intermediate-scale errors were largely event driven. Breaking spectra into discrete temporal bands revealed a significant model-by-band effect but also a nonsignificant model-by-site effect, which together suggest that individual models show consistency in their error patterns. Differences among models were related to model time step, soil hydrology, and the representation of photosynthesis and phenology but not the soil carbon or nitrogen cycles. These factors had the greatest impact on diurnal errors, were less important at annual scales, and had the least impact at intermediate time scales.

Citation: Dietze, M. C., et al. (2011), Characterizing the performance of ecosystem models across time scales: A spectral analysis of the North American Carbon Program site-level synthesis, *J. Geophys. Res.*, *116*, G04029, doi:10.1029/2011JG001661.

1. Introduction

[2] Ecosystem models remain our most important tool for diagnosing and forecasting carbon cycle dynamics across space and time. These models also play a critical role in understanding the potential responses of ecosystems to global change [*VEMAP Members*, 1995]. There are a large number of ecosystem models currently in use, but each makes different assumptions and was developed using different study sites and data sets. Observational and experimental data sets that provide detailed information about carbon cycle dynamics are increasingly available for almost every biome [*Baldocchi*, 2008] and represent an important source of data to test for ecosystem models. Despite numerous efforts to test multiple models at single sites [e.g., *Hanson et al.*, 2004] or single models at multiple sites [e.g., *Krinner et al.*, 2005; *Poulter et al.*, 2009], there has not been a major synthesis effort to evaluate the performance of the community of ecosystem models at multiple sites using a standardized protocol.

[3] In order to address these discrepancies, the North American Carbon Program (NACP) has initiated an intercomparison between ecosystem models and eddy covariance flux observations from multiple study sites. Recent NACP analyses have demonstrated that, across many models, the error in monthly NEE was lowest in the summer and for temperate evergreen forests, and was highest in the spring, fall, and during dry periods [*Schwalm et al.*, 2010a]. These analyses also suggest that skill is higher in (1) models where canopy phenology is prescribed by remote sensing rather than prognostic, (2) models where NEE is driven by a calculation of GPP-Re rather than NPP-Rh, and (3) models with a subdaily rather than a daily time step [*Schwalm et al.*, 2010a]. These same analyses have also demonstrated that regardless of model structure there is substantial room for improvement in the performance of all models, which suggests that parameter error is at least as important to model performance as structural errors [*Schwalm et al.*, 2010a].

[4] It is difficult to diagnose the mechanisms responsible for the lack of agreement between model and measurement using conventional model fitting statistics alone. Summary statistics or even residual analyses may not sufficiently explain the model failure that results if one or more carbon cycle processes are not correctly formulated. Model errors are challenging to resolve because ecosystem processes respond to climatic trends at multiple temporal scales. For example, GPP responds not just to the diurnal and annual course of solar radiation, but also to the seasonality in leaf area, short- and long-term drought effects, canopy development and nutrient availability over years to decades, and disturbance and migration at decadal to centennial time scales [*Falge et al.*, 2002; *Arnone et al.*, 2008; *Beer et al.*, 2010; *Schwalm et al.*, 2010b]. In general, models are broadly similar in how they respond to solar radiation, variable in how they represent leaf area index and phenology, and quite diverse in the ways they represent stand development and drought responses. Conventional metrics of model performance, such as root mean squared error, are typically calculated at a single time scale, which for eddy covariance data is often either 30 min or monthly. Such metrics cannot easily separate processes operating at different time scales, and most implicitly emphasize model performance at the fast time scales that dominate model variance at the cost of neglecting performance at longer time scales (e.g., annual-to-decadal variability and trends). An alternative step for model evaluation and improvement is to understand when a model output fails at multiple temporal scales by assessing error in the frequency domain rather than in the time domain [*Braswell et al.*, 2005; *Siqueira et al.*, 2006; *Williams et al.*, 2009; *Vargas et al.*, 2010; *Mahecha et al.*, 2010].

[5] The goal of this study is to evaluate the performance of ecosystem models at multiple time scales at select NACP sites using wavelet decomposition [*Torrence and Compo*, 1998; *Katul et al.*, 2001]. A multimodel analysis using wavelet decomposition has not been performed to date and

introduces new challenges in interpretation, especially given the NACP goal of explicitly considering uncertainty in observations and model output. Rather than attempting to identify "winners" and "losers," we direct our analysis to provide information useful for model improvements by identifying the time scales at which models fail, thus giving insight into the processes responsible for model/measurement mismatch. Previous research has identified diurnal and annual time scales as being disproportionately responsible for the variance in surface atmosphere CO_2 flux observations [*Baldocchi et al.*, 2001; *Katul et al.*, 2001; *Richardson et al.*, 2007; *Stoy et al.*, 2009]. Model-data comparisons conducted using individual models at small numbers of sites have identified both intermediate time scales (weeks to months) and interannual time scales as those in which models tend to fail [*Braswell et al.*, 2005; *Stoy et al.*, 2005; *Siqueira et al.*, 2006; *Vargas et al.*, 2011]. However, it is unclear whether these patterns would be expected to hold over a much wider range of biomes and against the breadth of different model structures considered in the NACP, which range from simple flux-and-pool matrix models to next generation terrestrial biosphere models [*Schwalm et al.*, 2010a]. Based on these previous findings, we test two interrelated hypotheses:

[6] 1. Models will accurately replicate flux variability at the daily and annual time scales, as important biological processes are regulated by diel and seasonal variation (e.g., the photosynthetic response to solar radiation) that is thought to be correctly formulated in the models.

[7] 2. Models will have difficulty in representing variation at intermediate time scales (weeks to months) because synoptic weather events and lagged responses in plant physiology and ecosystem biogeochemistry regulate variation in fluxes at these intermediate temporal scales.

[8] To evaluate these hypotheses we focus on an analysis of NEE model-data residuals to highlight where there is still room for improvement rather than on where models and data agree. We also develop a novel method for assessing the contribution of flux-tower observation error via a Monte Carlo approach and demonstrate that a failure to account for flux observation error leads to a qualitative misidentification of the modes of model failure.

2. Methods

2.1. Models and Data

[9] The NACP site-level model-data intercomparison encompasses 21 ecosystem models and 32 North American eddy covariance flux tower sites. Not all models were run at all sites; in total there are 463 out of 672 possible model-site combinations. What is statistically problematic is that the missing model-site combinations are not randomly distributed, but rather reflect the choices of individual modeling groups, which are undoubtedly influenced by model skill. For example, tundra and wetland model runs are strongly underrepresented suggesting that the models not run at these sites are likely to perform worse than those which were run. The fact that runs are not statistically "missing at random" has the potentially introducing biases in a statistical interpretation. Therefore, since most models were run for nine high-priority sites that had been identified a priori in the original model protocol, we restricted our analyses to these sites (Table 1).

Table 1. Sites and Models Used in Intercomparision[a]

Biome[b]	Site ID	Name	Years	AgroIBIS	BEPS	BIOME_BGC	Can-IBIS	CNCLASS	DLEM	DNDC	ecosys	ED2	EDCM	EPIC	ISOLSM	LoTEC_DA	LPJ_wsl	ORCHIDEE	SiB3	SiBCASA	SiBcrop	SSiB2	TECO	TRIPLEX-Flux
CRO	US-NE3	Mead corn/soy	02–04	X	X	X	X		X	X	X	X	X	X		X				X	X	X	X	X
GRA	CA-Let	Lethbridge	99–07		X	X	X		X		X	X	X		X	X	X	X	X	X		X	X	
DBF	CA-Oas	BERMS Aspen	97–06		X	X	X	X	X		X	X	X		X	X	X	X	X	X		X	X	
DBF	US-Ha1	Harvard Forest	92–05		X	X	X	X	X		X	X	X		X	X	X	X	X	X		X	X	
DBF	US-UMB	U. Michigan	99–03		X		X	X	X		X	X	X		X	X	X	X	X	X		X	X	
ENFB	CA-Obs	BERMS Spruce	00–06		X	X	X	X	X		X	X	X		X	X	X	X	X	X		X	X	
ENFT	US-Ho1	Howland Forest	96–04		X	X	X	X	X		X	X	X		X	X	X	X	X	X		X	X	
ENFT	CA-Ca1	Campbell River	98–06		X	X	X	X	X		X	X	X		X	X	X	X	X	X		X	X	
WET	CA-Mer	Mer Bleue	99–06		X		X		X		X		X		X	X	X	X	X	X		X	X	

[a]See the work of *Schwalm et al.* [2010a] for a more detailed summary of site and model characteristics.
[b]IGBP biome codes: CRO, crop; GRA, grassland; ENFB, evergreen needleleaf forest—boreal; ENFT, evergreen needleleaf forest—temperate; DBF, deciduous broadleaf forest; WET, wetland.

[10] Model runs at each site followed a prescribed protocol to facilitate intercomparison. Each model used a standardized meteorological forcing data based primarily on the observed meteorology at each flux tower. Meteorological data were gap-filled using a combination of nearby met station data and the DAYMET reanalysis as documented by *Ricciuto et al.* [2009]. Ancillary data such as soil texture and management history were available via NACP Biological-Ancillary-Disturbance-Methodology (BADM) templates [*Law et al.*, 2008] to ensure that all models were making the same assumptions about the local environment at each site. In addition a standard subset of GIMMS Normalized Difference Vegetation Index data set [*Tucker et al.*, 2005] data were provided for the subset of models that are diagnostic rather than prognostic. Models were expected to be run to steady state using their standard parameter set; site-specific model tuning was prohibited. The exception to this was the LoTEC model, which was run using a data assimilation scheme. The performance of LoTEC relative to other models thus highlights the contribution of parameter error, rather than models structural errors, to model performance. The full modeling protocol can be found online (http://nacp.ornl.gov/docs/Site_Synthesis_Protocol_v7.pdf).

[11] This analysis focuses on the comparison of observed and simulated net ecosystem exchange (NEE) of CO_2. All analyses were conducted on the finest temporal resolution available, which was 60 min at US-Ha1, US-Ne3, and US-UMB and 30 min for all other sites (Table 1). Models with a daily time step used the daily mean value for all values within a 24 h period and thus we refrain from interpreting the results from these models at time scales less than 2 days. For each model at each site we calculated the normalized residual error ($\varepsilon_{s,m,t}$) in NEE between models and data as

$$\varepsilon_{s,m,t} = \left(\frac{Model_{s,m,t} - \overline{Model_{s,m}}}{\sigma_{s,m}}\right) - \left(\frac{Data_{s,t} - \overline{Data_s}}{\sigma_s}\right) \quad (1)$$

with subscripts s = site, m = model, and t = time and a bar indicating an average over the full length of the time series. This error metric was designed to highlight the synchrony of the model with the data rather than identify persistent model biases, which are generally a reflection of errors in model parameterization not model structure and are reported on in detail elsewhere [*Schwalm et al.*, 2010a]. Data and model output were mean-centered to eliminate biases in the cumulative flux and divided by the standard deviation (σ) across the entire record to normalize the amplitude of variability.

[12] As the continuous wavelet transform does not accommodate missing data, flux data were gap-filled using Marginal Distribution Sampling (MDS) [*Reichstein et al.*, 2005], which is a standard FLUXNET data product [*Moffat et al.*, 2007]. An estimate of NEE observation error at every time point was generated by *Barr et al.* [2009], accounting for uncertainties associated with U* filtering and random measurement error [*Richardson et al.*, 2006; *Richardson and Hollinger*, 2007]. These uncertainties are incorporated in the spectral null model, rather than in the error metric, as described below.

2.2. Spectral Analysis

[13] Spectral analyses are based on the premise that a time series can be decomposed into an additive series of wave functions that have different time scales in a way directly analogous to how a Taylor series decomposes a function into a series of polynomials. These analyses allow one to identify the time scales that dominate a signal because wave functions that match the fluctuations in the data will explain the most variance (i.e., power). In contrast to traditional methods such as Fourier spectra, which are based on using sinusoidal waves, wavelet analyses are based on using wave functions that are finite in length but moved over the time series in a way conceptually similar to a moving window. In this way wavelet analyses are able to identify not only the time scales that dominate a signal but also when in time those time scales are strongest. Wavelet analyses are typically plotted on what is referred to as the wavelet halfplane, where time is along the *x* axis, time scale is along the *y*, and spectral power is indicated by color, for example with hot (red) colors indicating high power and cool (blue) colors indicating low power. Wavelet analysis has been widely applied in the geosciences [*Torrence and Compo*, 1998] for quantifying the spectral characteristics of time series that may be nonstationary and heteroscedastic, thereby offering an improvement over traditional Fourier decomposition (e.g., see the demonstration by *Scanlon and Albertson* [2001]). A continuous wavelet transform was computed in R using the dplR library [*Bunn*, 2008] using the Morlet wavelet basis function and setting the wave number (k0) to six and calculating four suboctaves per octave (four voices per power of two). The Morlet wavelet looks like a sine wave centered on zero that decays rapidly to extinction in both directions. Wavelet power was corrected for biases following *Liu et al.* [2007] to ensure a consistent definition of power in order to enable comparisons across spectral peaks.

[14] The challenge when interpreting the spectral characteristics of model error is to determine when model-data mismatch is statistically significant. For this to be useful it is important that the spectra be compared to the appropriate null model, which for eddy covariance data requires that the null spectra account for the errors in the flux observations. Thus, conventional significance tests using, for example, Monte Carlo analyses on colored noise spectra, are inadequate [*Torrence and Compo*, 1998; *Grinsted et al.*, 2004; *Stoy et al.*, 2009]. Determining appropriate null spectra is further complicated because flux measurement error is non-normally distributed; NEE measurement errors have a double exponential distribution, which is more fat-tailed than the normal, and is highly heteroscedastic, with error increasing linearly with the absolute magnitude of the flux [*Hollinger and Richardson*, 2005; *Richardson et al.*, 2006, 2008; *Lasslop et al.*, 2008]. Because of this error distribution, even a perfect model that exactly predicted the true flux would have a strong diurnal and seasonal error spectrum when compared to data because the magnitude of observation error in the data increases systematically during the day and during the growing season when NEE tends to be larger.

[15] To account for this observation uncertainty we developed a novel Monte Carlo approach to generating the null wavelet spectrum. A stack of 1000 wavelet spectra were calculated using 1000 Monte Carlo replicate "pseudo-data" time series for each site. Replicate data sets were generated using the methods of *Barr et al.* [2009] that account for both uncertainty in the gap-filling algorithm and measurement uncertainty, and sample over the distribution of both

simultaneously. The relative error between the pseudo-data and the original data, and the wavelet spectra of this error, were both calculated in the exact same way as model error. From the 1000 replicate spectra, the mean, median, and a quantile-based confidence interval were calculated and the distribution of these replicate spectra was used as the null model for the model spectra.

[16] Rather than present all wavelet half plane diagrams for all model-site combinations, results are summarized in three ways. First, the global power spectra were calculated for all model-site combinations. To account for the data uncertainty, the peaks in the global spectrum for each model at each site were tested for significance by comparison to the distribution of the global spectra of the 1000 Monte Carlo pseudo-data sets for that site. To simplify presentation, the model spectrum at each site was divided by the one-sided 95% confidence bound generated from the 1000 Monte Carlo replicates. In this approach, any part of the spectrum that is greater than one is interpreted as the model error falling outside the range of data uncertainty. To enable comparisons among models, the global spectrum for each model averaged across sites was calculated as the median of these null-corrected spectra. The spectrum for the multi-model mean was calculated by first calculating the ensemble average time series in the time domain, and then treating this like any other model, rather than averaging spectra across models in the wavelet domain.

[17] Second, to synthesize the proportion of variance in the error metric that was attributable to different time scales, we extracted power from five spectral bands for each site by model combination. These bands and their time scales were the following: subdaily (<0.5 day), daily (0.5–2 days), synoptic (2–180 days), annual (180–700 days), and interannual (>700 days). Bandwidths were determined by examining the wavelet spectra of sinusoidal waves with a "pure" diurnal and annual signal. The analysis of this simple synthetic time series also served to verify that the analytical methods were functioning as expected. The five bands were then summarized on both a by-site and by-model basis in terms of the relative contribution of each band to the overall spectra. As before, all spectra were normalized by the upper confidence interval of the null spectra, which in addition to providing a metric to determine significance, also allows spectra from different sites to be compared despite differences in time series length and total variance. The proportion of spectral energy in each band was compared using three-way ANOVA with site, model, and spectral band as covariates and including all pairwise interactions. Within the ANOVA the interannual time scale was dropped as a response variable because of edge effects within the cone of influence and because its proportional energy is linearly determined by the other four bands. Because this analysis was unbalanced, a second ANOVA was performed on just the 15 models that have completed all runs at the three deciduous forest sites and three conifer forest sites (Table 1). Within this analysis we also included biome as a covariate to test for differences among the conifer and deciduous sites.

[18] The third way model performance was summarized was to examine the across-model composite wavelet spectra for each site. Each model-site spectrum was first normalized so that total power sums to one before calculating the across-model average of the full wavelet spectra for each site (i.e., averaging was performed in the spectral domain). This analysis was done in order to discern the presence or absence of consistent temporal patterns in model performance within a site in order to quantify when models consistently fail when challenged by data from each site. Errors that are common across models are expected to have higher spectral energy than errors that are unique to a single model because random errors will cancel in the resulting power spectra.

[19] In order to identify phenological errors in the wavelet spectrum the beginning and end of the growing season was marked on the wavelet half plane for each site. Phenological boundaries were estimated based on a 10 day moving average of tower-based GPP. We used a threshold of 20% maximum GPP, which gives very similar results as the more common zero NEE threshold at most sites, but was less sensitive to noise for the few sites were the zero NEE threshold gave unrealistic results.

3. Results

[20] Below we demonstrate the wavelet-based uncertainty analysis using the output from one model, the Ecosystem Demography (ED2) model [*Moorcroft et al.*, 2001; *Medvigy et al.*, 2009], at one site, Howland forest (US-Ho1) [*Hollinger et al.*, 2004]. We use this example to explain the Monte Carlo analysis with pseudo-data and discuss one model-site comparison in more detail. We then analyze the spectra from all sites and models and partition the relative error for each site and model among the different time scales (hourly to interannual) that the length of the data records permit us to interpret.

3.1. Wavelet Decomposition of Eddy Covariance and Ecosystem Model Time Series With Explicit Error Accounting

[21] Figure 1 displays the wavelet half plane spectra for one site (US-Ho1, Figure 1a), one model (ED2, Figure 1b) run at this site, and the normalized residual error between the model and the observations (Figure 1c) using data from 1996 to 2003. It is important to note that this analysis focuses on the normalized residual error spectra (Figure 1c), not the spectra of the data (Figure 1a) or the model (Figure 1b) itself, and that this residual (equation (1)) is calculated in the time domain (i.e., it is not the difference between the wavelet coefficients displayed in Figures 1a and 1b). In Figure 1, time is along the x axis, time scale is along the y, and spectral power is indicated the intensity of color on a logarithmic scale with warm colors (dark red) indicating the highest spectral power, which can be interpreted as the strongest match between a Morlet wavelet and the time series. Importantly, the warm colors in Figure 1c indicate regions in the frequency domain of substantial data-model disagreement.

[22] From Figure 1, the dynamics of both the model and the data are dominated by a diurnal signal (1 day) and an annual signal (365 days). We also observe that while the annual signal is present and relatively constant across time, the spectral power for time scales between ~1 h and ~2 weeks is considerably stronger (colors toward red) during the growing season and weaker (blue) during the winter, especially at the 1 day time scale. More subtlety, comparing the data and model spectra suggests that the model may have less variability (visualized as less red) than the observed

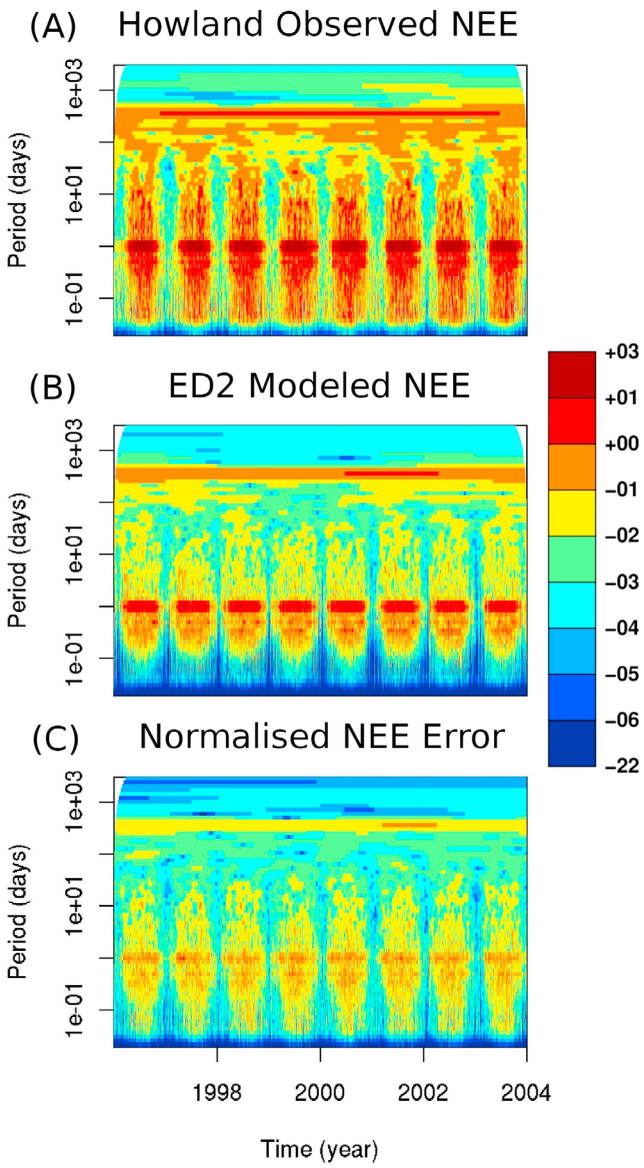

Figure 1. Wavelet coefficients displayed in the wavelet half plane for (a) observed net ecosystem exchange at Howland Forest (US-Ho1), (b) the predictions of the ED2 model, and (c) model-data normalized residual error. Areas of high spectral power are indicated by hot colors, with dark red representing the highest power, while low power is indicated by cool colors, with violet representing the lowest power.

NEE at both the subdaily time scale and at the time scales between daily and annual (henceforth called the intermediate time scales). The lower variability in models at subdaily time scales is a reflection of the fact that the models do not include measurement noise, which arises from instrument error, the stochastic nature of turbulent eddies, and variation in the flux tower footprint. In addition, many of the models can only predict NEP, which is less variable than NEE due to the absence of canopy CO_2 storage, though in all cases we are only considering ecosystem atmosphere CO_2 fluxes not dissolved carbon or organic trace gases. When we look at the residual error spectra (Figure 1c), we should be encouraged by the fact that the wavelet coefficients have a smaller magnitude, suggesting a degree of correspondence between the model and the data. However, clear signals of model-data mismatch at the annual and diurnal time scales remain.

[23] These mismatches between model and measurement in Figure 1c can be further interpreted by accounting for the uncertainties in the observations as in Figure 2. Figure 2a (black line) provides an example of the global power spectrum

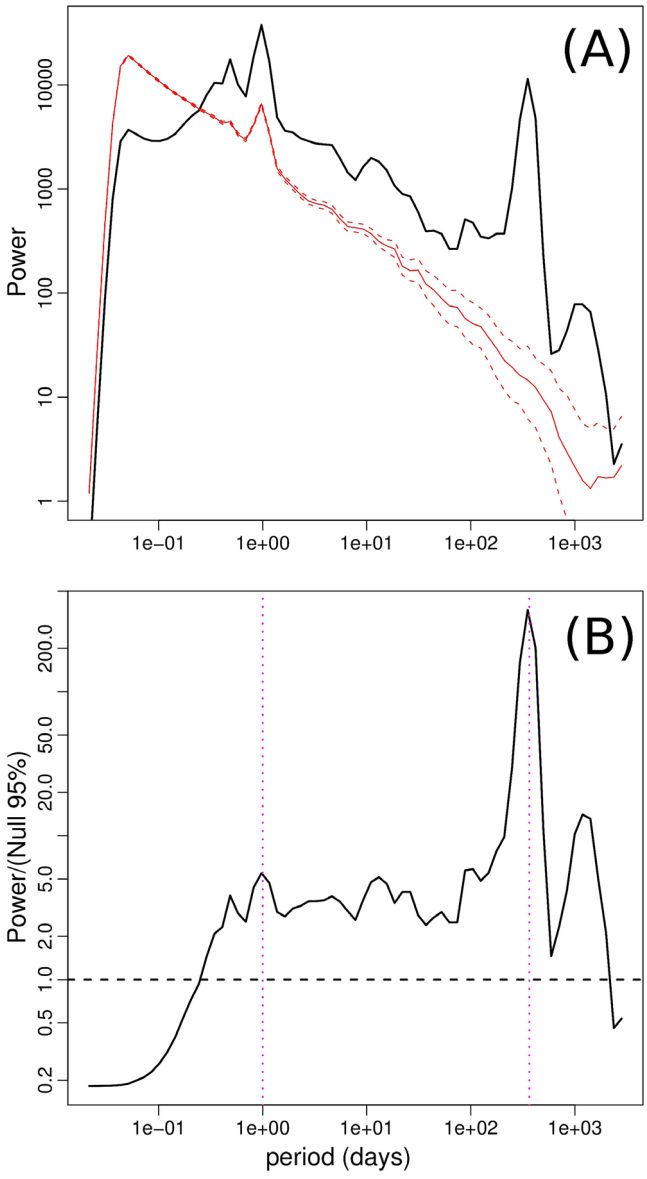

Figure 2. (a) Global power spectra of the normalized residual error for the ED2 model at Howland Forest (US-Ho1) (black line) as compared to the null spectra (mean, solid red line; 95% confidence interval, dashed red line). (b) The ED2/US-Ho1 global power spectra divided by the upper confidence interval of the null spectra. When the model-data error spectrum is greater than one (dashed line), this indicates that the model error has significantly more spectral power at these time scales than would be expected based on observation error.

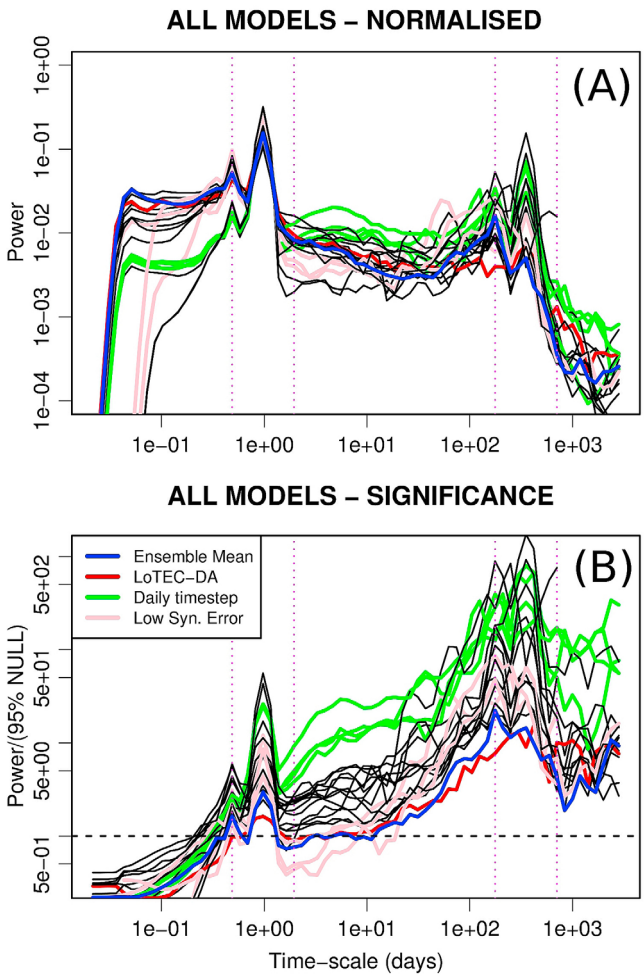

Figure 3. (a) Global power spectra for all models where each line is the median spectra for one model averaged over the nine high-priority sites. (b) Comparison of the global spectra to the null spectra. The blue line indicates the multimodel ensemble mean and the red line indicates the LoTEC data assimilation model, which were the top performing models. The pink lines designate other models that were within the observation error for at least part of the synoptic time scale (AgroIBIS, SiB-crop, SiBCASA0), while models that operate at a daily time step (BEPS, BIOME-BGC, DLEM, LPJ-wsl) are in green.

for the error of the ED2 model at Howland, which is simply the marginal distribution of the full error spectrum in Figure 1c, in comparison to the Monte Carlo estimate of the spectra of the observation error (red line: solid line = mean, dashed lines = 95% CI). In order to facilitate the comparison of these spectra we divided the model-data error spectra by the upper 95% CI of the observation error spectra for each time scale (Figure 2b). In this context any time scale that falls above the horizontal line (>1) indicates a model residual error that is "significantly" higher than the uncertainty in the observed data. This shows, for example, that while the error in the model is greatest at the diurnal time scale, the data uncertainties are also very high at this time scale. By contrast, the absolute error at the annual scale is lower, but the random uncertainty in the net carbon flux at this time scale is also considerably lower, such that the normalized peak at the annual time scale dominates the overall spectra.

3.2. Global Model Spectra

[24] Prior to correcting for the null spectra, the global power spectra, averaging each model across all sites (Figure 3a), suggest that the error in all of the models considered is dominated by errors in the diurnal cycle. The large majority of models also show a second peak at the annual time scale that is almost the same magnitude as the diurnal peak. Based on these overall spectra, there appears to be little consistent structure to model error at the subdaily or intermediate time scales. Power in the error spectra declines at interannual time scales, but this likely reflects the limited length of the time series at these time scales rather than a confirmation of model performance at capturing long-term trends. This interpretation is supported by the large error bounds in the null spectra at this time scale (Figure 2a, dashed red line).

[25] When compared to the null model spectra, which corrects for the observation error in the flux data, there are substantial changes in the overall pattern of model performance (Figure 3b). While models continue to have significant error at the diurnal time scale, this is no longer the dominant peak of the spectra because there is significant structure to the observation error at this time scale. This implies that much, but by no means all, of the dominant diurnal peak in the nonnormalized spectral results from the noise in the data (Figure 3a). Once corrected for observation error, the overall error for most models is dominated by error at the annual time scale and the greatest variability in model performance comes at the intermediate time scales. For a subset of models (AgroIBIS, LoTEC, SIBcrop, SiBCASA), as well as for the ensemble mean, model performance at the daily to monthly time scales falls at or within the uncertainty bounds of the data. Of these, AgroIBIS and SIBcrop are crop models that were only run at the crop site, LoTEC was run using a data assimilation routine, and thus would be expected to have a lower error rate, and SiBCASA is driven in part using remote-sensing products and thus has more information than prognostic models. For another set of models (BEPS, BIOME-BGC, DLEM, LPJ-wsl) model error shows a dip immediately after the diurnal peak but then rises rapidly during the first part of the intermediate scale (daily to monthly) to reach levels that are comparable to error in the daily time scale. The common feature of this group of models is that it consists of all of the noncrop models with a daily time step. The remaining models also show a dip after the diurnal peak but then error stays lower throughout these time scales, with errors slightly larger than would be expected by chance between the daily and monthly scale. All models show substantial increase in error in the second half of the intermediate scale (monthly to seasonal).

[26] In order to further diagnose the drivers of the variability among models within specific time scales, we tested for the effects of model structure on the integrated power within the diurnal, intermediate, and annual bands using ANOVA (Table 2). This analysis was restricted to the group of models that operate at a subdaily time step, as we already identified a distinct pattern for daily models, and repeated both for all models and sites and for just the set of complete

Table 2. The *p* Values for ANOVAs Assessing the Impact of Model Structural Covariates on the Spectral Power Within Each of the Three Time Scales[a]

	All[b]			Forest[c]		
	Diurnal	Intermediate	Annual	Diurnal	Intermediate	Annual
Site	<0.001	<0.001	<0.001	0.006	<0.001	0.021
Soil H_2O layers	0.064	ns	ns	0.049	0.065	0.031
Soil C pools	0.030	ns	ns	ns	ns	ns
Phenology	0.015	ns	0.088	0.018	ns	ns
Photosynthesis	<0.001	ns	0.015	<0.001	ns	ns
Nitrogen	ns	ns	ns	ns	ns	ns

[a]Structural covariates are all categorical and take on the following states: soil water layers (0,1, >1), multiple soil carbon pools (yes/no), phenology (prognostic, prescribed), photosynthesis (enzyme kinetic, stomatal conductance), and soil nitrogen cycle (yes/no).
[b]All study sites and models were used.
[c]Restricted to the six forest study sites and the models that were run at all sites.

forest runs. We also excluded LoTEC because it employed a data assimilation scheme. The inclusion of multiple soil layers in the soil moisture model had a significant effect of increasing error at the diurnal time scale for both all sites and the forest sites. Within the forest sites this effect was also significant at the annual scale and marginally significant at the intermediate time scale. The representation of soil carbon pools was in general not significant across scales and sites with the exception of the diurnal scale when considering all sites, in which case models with multiple pools performed worse than models with one carbon pool or no explicit representation of soil carbon. Canopy phenology (prognostic versus prescribed or semiprognostic) had a significant effect at the diurnal time scale at both all sites and forest sites, with fully prognostic models showing larger error than those which used some amount of external information to control phenology. The choice of photosynthesis scheme (enzyme kinetic versus stomatal conductance) was highly significant at the diurnal time scale, with enzyme kinetic models [e.g., *Farquhar et al.*, 1980] having lower error. For both phenology and photosynthesis these effects were also seen at the annual scale when looking across all models, but not at the forest sites nor at the intermediate scale for either set of sites. Finally, the inclusion of an explicit nitrogen cycle did not have a significance effect on the spectral power at any time scale regardless of whether one considers all sites or just the forest sites. Error spectra on a model-by-model basis (Figure S1) and model structural characteristics (Table S1) are provided in the auxiliary material.[1]

3.3. The Proportion of Model Error at Different Time Scales

[27] Model error was binned into temporal bands representing subdaily, daily, intermediate, annual, and interannual time scales (Figure 4). The full ANOVA suggests that the differences among sites ($p < 0.001$, $F = 5.34$, $df = 8$) and the site by band interactions ($p < 0.001$, $F = 4.82$, $df = 24$) were significant. These results were also significant within the forest-only ANOVA (site: $p < 0.001$, $F = 5.16$, $df = 5$; site-by-band: $p < 0.001$, $F = 4.70$, $df = 15$). Figure 4a, which shows the overall relative error partitioning by site and band, shows that the error at most sites was dominated by the intermediate and annual time scales. There is a comparatively large amount of spectral power in the interannual band at the CA-Oas site but almost none at US-Ne3, the latter of which is a reflection of the fact that the crop site was only a 3 year time series and thus almost all of the interannual band falls outside the cone of influence. Differences among sites do not show any obvious pattern between the three deciduous sites (Figure 4a, left), the three conifer sites (Figure 4a, middle), and the three nonforested sites (Figure 4a, right). This was consistent with both the forest-only ANOVA, which did not find a significant biome effect, and with a subsequent post hoc analysis that did not find interactions between biome and any other term.

[28] The full ANOVA also suggests that there were significant differences among the spectral bands ($p < 0.001$, $F = 943.43$, $df = 3$) and in the band by model interaction ($p < 0.001$, $F = 6.06$, $df = 57$). These results were likewise consistent with the forest-only analysis (band: $p < 0.001$, $F = 531.80$, $df = 3$; band-by-model: $p < 0.001$, $F = 6.02$, $df = 42$). Figure 4b shows the overall relative error partitioning by model and band. The agroecosystem models (AgroIBIS, EPIC, SiBcrop, TRIPLEX) stand out because of their lack of interannual variability, but as noted above this is a characteristic of the crop site not the crop models per se. Among the remaining models, EDCM and SSiB2 were dominated by errors in the annual cycle while BEPS, ED2, and LPJ had the largest fraction of error at the interannual time scale and the smallest fraction at the annual time scale compared to other models. Interestingly LoTEC-DA, the one model to employ data assimilation, was not unusual in terms of the relative contributions among the different bands. Finally, the ANOVA found the model and the model-by-site interactions were significant in neither the full ANOVA, nor the forest-only analysis. Because of a lack of a model effect, no model structural variables were tested.

3.4. The Mean Normalized Spectra of Multiple Models

[29] The across-model error spectra for each site can help ascertain consistent temporal patterns of the model failures on a site-by-site basis (Figure 5). Strong diurnal and annual error spectral signals appear at all sites. Diurnal error is highest during the growing season at all sites (delineated by vertical black lines) and is lowest during the winter, but can also be nontrivial outside the growing season suggesting that these errors cannot be isolated to the GPP calculations within the model. The fact that this seasonal variation appears stronger at the deciduous and nonforested sites (Figures 5 (top) and 5 (bottom), respectively) but is not absent at coniferous sites (Figure 5, middle) suggests that phenology/LAI may contribute to the error. The elevated error during the growing season is not isolated to the diurnal cycle, but is also present but of lower magnitude across the intermediate time scale at all sites as well.

[30] The magnitude of the annual error tends to be large and consistent across seasons, with some variability within and among sites. For example, a large annual and seasonal signal is apparent for the agricultural site (US-Ne3) for 2003. This corresponds to the year maize was planted in a maize-soy rotation and is a reflection of the fact that the majority of the models consistently underpredicted maize

[1]Auxiliary materials are available in the HTML. doi:10.1029/2011JG001661.

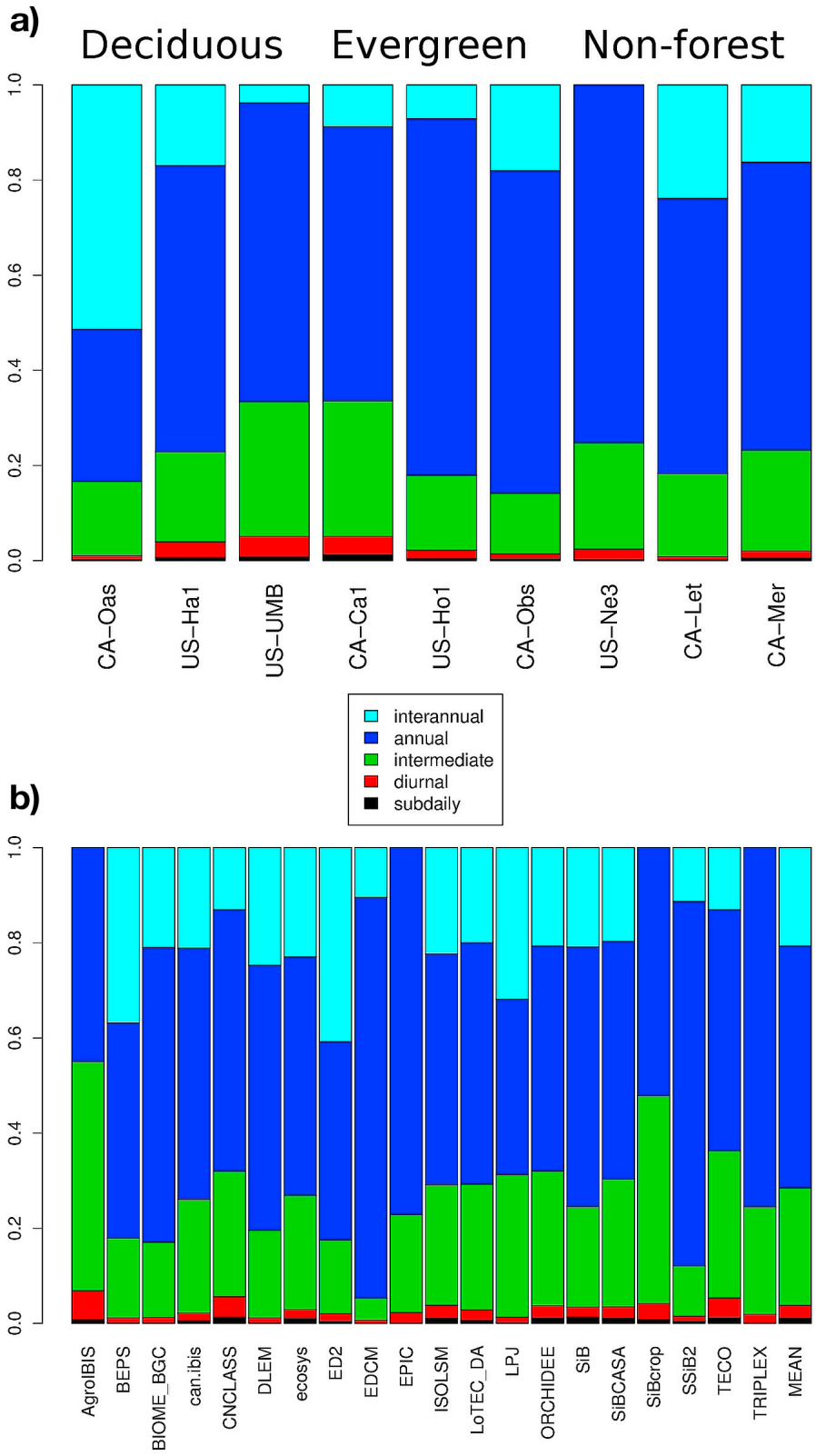

Figure 4. The proportion of model error at different time scales for (a) the nine high-priority NACP study sites and (b) the 20 ecosystem models investigated.

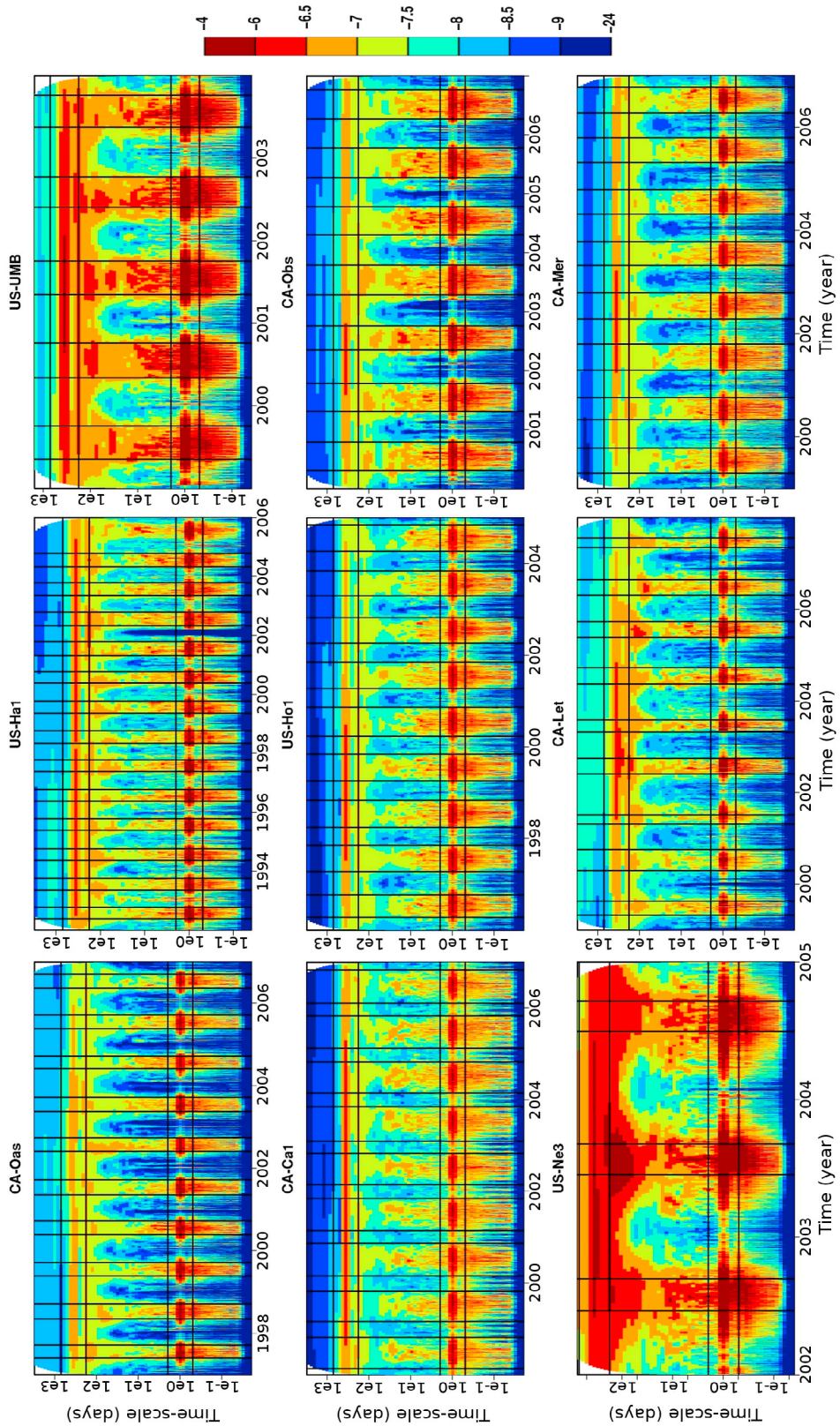

Figure 5. The average normalized wavelet spectra of residual model error for each NACP high-priority site. Spectral power is indicated by color on a base 10 logarithmic scale, which the lowest and highest bins extended to account for the tails of the power distribution. Horizontal lines delineate the bins used to divide interannual, annual, intermediate, daily, and subdaily bins. Vertical lines delineate the bounds of the growing season as determined by when a 10 day smoothed time series reached 20% of maximum GPP. The *y* axes vary slightly from site to site because of the different lengths of each time series.

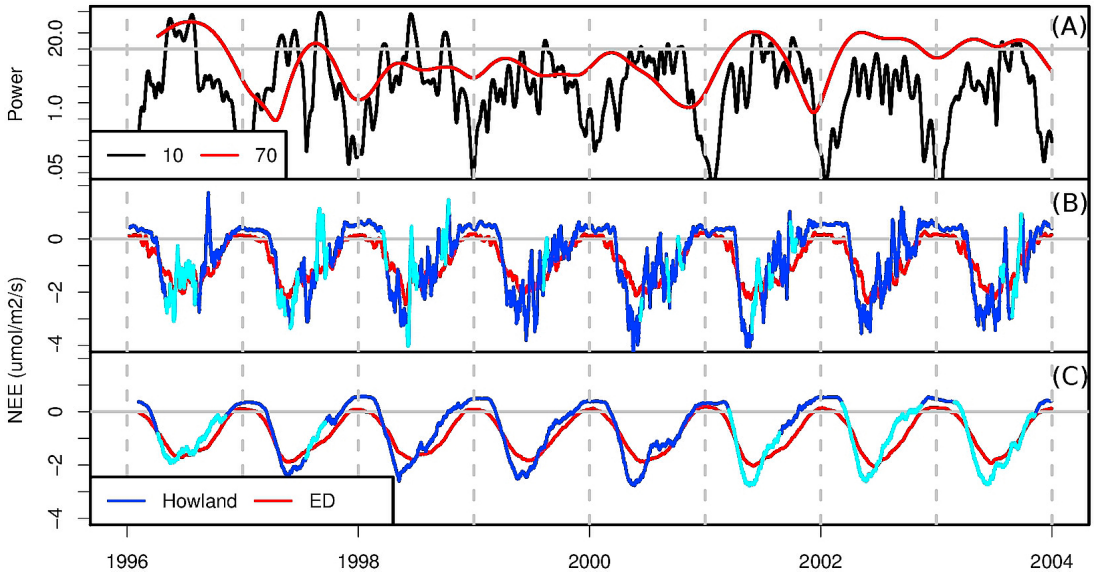

Figure 6. Model errors in ED2 at Howland diagnosed by using wavelet spectra. (a) The spectral power is shown at two time scales, 10 days and 70 days, with a threshold of 10 set to identify peaks in the spectra. The model and flux data are shown with a moving average at a (b) 10 day and (c) 70 day window with times that fall above the threshold highlighted. At both time scales, the model-data discrepancies result from the model being overly smooth and unable to capture the variability present in the data. Full spectra for this site and model are shown in Figures 1 and 2.

GPP and NEE (E. Lokupitiya et al., Evaluation of model-predicted carbon and energy fluxes from cropland ecosystems, submitted to *Global Change Biology*, 2011). Similarly, the strong seasonal to annual signal at the grassland site (CA-Let) for 2002–2006 corresponds to a series of years that had appreciably greater NEE [*Flanagan et al.*, 2002].

[31] Within the intermediate time scale there are also clear indications of brief periods of elevated model error during the growing season across all sites. These error "events" show up as patches or vertical plumes of red and orange in Figure 5. Because these periods tend to be brief and irregular, their contribution to overall error is smaller than dominant annual and diurnal cycle errors, but they do point to systematic errors that are shared across models. These events were investigated further on a model-by-model basis by plotting the wavelet power for individual time scales and comparing this to model and data smoothed to the same time scales. As an example, we return to our previous case of the ED2 model at Howland Forest and investigate the dynamics at two intermediate time scales, 10 days and 70 days (Figure 6). We see that in all cases the intermediate time scale "events" identified by the wavelet analysis correspond to times when there was greater variability in the data than in the model. Examples of other models and other sites generally confirmed this trend (data not shown) that most models were noticeably smoother than the data. As a reminder, these are discrepancies in variability on the order of weeks to months and are thus unlikely to arise from random measurement error in the data, though this does not rule out the possibility of systematic errors in instrumentation. We have not diagnosed the environmental and biotic drivers of these many small events, as this is beyond the scope of this study, but useful examples of this approach can be found in the literature [*Mahecha et al.*, 2010]. Finally, it is worth noting that there does not appear to be any correspondence between error "events" in the intermediate period and the phenological boundaries identified from the tower flux data.

4. Discussion and Conclusions

[32] Our first hypothesis was that models would perform well at the daily and annual time scales because biological processes at both these scales are driven by a solar radiation cycle and corresponding changes in temperature. In contrast to our expectations, model error was overwhelmingly dominated by the annual cycles and also showed a clear diurnal signal. Models captured a significant amount of variability at these time scales (Figure 1b), but these time scales are nonetheless responsible for such a large fraction of the overall variability in NEE (Figure 1a) that errors in their representation dominate the error spectrum and drive overall model performance. Our analysis further reveals that model error on the diurnal cycle predominantly occurs during the growing season regardless of biome, which is not surprising given the larger magnitude of summertime fluxes. These results suggest that further model development focus first and foremost on correctly replicating flux variability and magnitude on the annual and diurnal time scales. This recommendation runs counter to recommendations from site-specific model wavelet analyses where, for example, flux variability was correctly replicated at most time scales but interannual variability was captured for the wrong reasons [*Siqueira et al.*, 2006]. This discrepancy arises because at a single site models can often be calibrated to match observations, but these calibrations may not hold when applied to other sites. It should also be noted that the spectra in this previous analysis [*Siqueira et al.*, 2006] were not corrected

for observation error and on visual inspection appear very similar to the uncorrected spectra in this analysis (Figure 3a).

[33] Probing deeper into the contribution of model structure to diurnal and annual error (Table 2) reveals that model structure is particularly important at the diurnal scale. Within the diurnal scale, the choice of photosynthetic scheme had the greatest impact, with enzyme kinetic models performing best. Counterintuitively, the choice of phenology scheme also had a strong impact at the diurnal scale, though not surprisingly models which predict their own phenology performed worse than those with relied fully or in part on external phenological information. The representation of soil moisture also had a modest, though significant, effect on diurnal errors, though with the somewhat surprising result that the inclusion of multiple soil moisture layers increased error. This may result either from uncertainties in the soil texture causing errors in the predicted depth distribution of moisture itself, or in errors associated with rooting depth distributions and the ability of plants to take up moisture from different layers, neither of which is an issue within a simple single-bucket approach. Finally, the effects of soil carbon representation were modest and inconsistent, while soil nitrogen representation was nonsignificant. At the annual scale both soil C and N representation remain nonsignificant, while the importance of other structural factors were consistent with the diurnal patterns, but the effects were generally weaker and significance varied the between the complete set of forest models and sites and all sites.

[34] Our second hypothesis, that models would have difficulty capturing intermediate time scale processes, was supported by the analysis. The intermediate time scale is difficult for models to capture due to the stochastic nature of weather events and the presence of within season biotic feedbacks. Once data uncertainties were accounted for, error at the intermediate time scale constituted a nontrival contribution to overall error (Figure 4). What was not predicted a priori was that there are actually two different domains within the intermediate time scales, split at a time scale of approximately 20 days (Figure 3b). This 20 day time scale is only slightly longer than the time scale at which the influence of radiation variability was found to decline and vapor pressure deficit variability became more important for modeling carbon flux variability in a coniferous stand in the Duke Forest [*Stoy et al.*, 2005]. Likewise, variability in leaf area index in a deciduous stand became disproportionately important for describing NEE variability at approximately a 20 day time scale [*Stoy et al.*, 2005]. More generally, this split in time scales also corresponds to the approximately 3 week duration of synoptic weather patterns. Unlike the predictable error structure in the annual and diurnal cycles, the error at intermediate time scales was much more variable, with periods of large error appearing within stretches where models performed well (Figure 5). Investigations into model dynamics during these error "events" suggest that in general models are not variable enough and tend to smooth over within-season variability. These intermediate scale failures are more important than their overall contribution to model error would suggest because it is the climatic variability at this scale that gives us insight into a model's capacity to capture stress responses, which are critical for forecasting global change.

[35] The composite full spectra (Figure 5) indicate that these discrete intermediate-scale error "events" appear to be shared among many of the models, suggesting shared structural errors. Such structural errors may arise due to both the sharing of mathematical formulations among models and due to shared false assumptions arising from our incomplete understanding of ecosystem dynamics. That said, there were no significant correlations between model structure and the performance at the intermediate time scale. The hints of structural effects are related to soil processes, specifically the number of soil layers, which is consistent with our expectation that soil moisture plays an important role in synoptic scale responses. However, the fact that models with multiple soil layers performed worse suggests that additional model complexity does not guarantee superior model performance. Further diagnosis of the environmental drivers of these intermediate-scale errors and the structural characteristics of models that avoid them is clearly warranted, as are empirical analyses of these systems at the scales relevant to resolving ubiquitous model uncertainties. Interestingly, the mean of the ensemble of models had lower error in the spectral domain than almost all of the individual models, suggesting that while there may be shared model errors, there are also many errors across models that average out in the ensemble. This result reiterates the common finding in the time domain that a multimodel ensemble frequently has the best predictive skill [*Bates and Granger*, 1969; *Schwalm et al.*, 2010a].

[36] Also noteworthy at the intermediate time scales is an absence of error peaks at the beginning and end of the growing season. The representation of phenological cycles is a known challenge for models [*Richardson et al.*, 2011], but on average this error was not found to dominate on intermediate time scales and instead error is consistently elevated across the growing season. The absence of a clear phenological signal may be due to the lack of synchrony in phenological errors among models or because phenological errors are showing up as part of the larger annual error.

[37] By observing the error contribution across temporal bands, the strong model $\times$ band effect combined with the nonsignificant model $\times$ site effect suggests that individual models are consistent in their error patterns. This is encouraging because it suggests that model failures are not idiosyncratic and site specific. This implies that model improvements are likely to translate to many sites, as opposed to improvements at some sites coming at the expense of reduced performance at others. The significant effects of site and band $\times$ site suggest that, as expected, the models taken as a group are performing differently at different sites. A previous analysis of model error in the time domain showed that absolute error varies with biome [*Schwalm et al.*, 2010a], with the smallest errors in well studied biomes such as deciduous and evergreen forests, and the largest errors in less intensively studies systems, such as tundra and shrublands. The current frequency domain analysis suggests that the relative impacts of different time scales do show consistency among models for a given site. The current pattern of site-to-site differences appears a bit idiosyncratic at this point and the forest-only analysis failed to show a significant difference between deciduous and evergreen sites (Figure 4a). At one deciduous forest site, CA-Oas, models demonstrated an unusually large fraction of error at

the interannual time scale. Further investigation showed that the interannual error was over five times greater than average at this site, while the annual error was less than 18% below average, suggesting that high interannual error rather than low annual error drove the pattern at this site. Future work with a larger number of sites may be able to clarify site-to-site differences but within the NACP analysis this requires addressing the nontrivial statistical problem that missing model/tower combinations are not random. However, the dominance of error in the annual time scale across sites and models is so clear that increasing the sample size would not provide much additional guidance on how to improve models.

[38] One of the most novel and important aspects of this analysis was the inclusion of observation error estimates in the evaluation of model-data mismatch across time scales. Observation error is not randomly distributed (Figure 2a), but has a strong spectral signature that follows flux magnitude [*Richardson et al.*, 2008]. Failing to include the magnitude of observation errors would have resulted in qualitatively different conclusions about the significance and relative importance of the different spectral bands. Specifically it would have resulted in an overestimation of the importance of the diurnal cycle and an underestimation of the importance of both the annual cycle and the longer half of the intermediate time scale.

[39] One time scale that has received little attention in this analysis is the role of interannual to decadal time scales in model error [*Stoy et al.*, 2009]. We are only beginning to have tower data records long enough to assess the ability of models to capture decadal variability and longer term dynamics [*Urbanski et al.*, 2007]. For spectral methods this is particularly problematic at long time scales because the edge effects on the amount of usable data, a region known as the "cone of influence," means that a valid inference about interannual variability can only be made for a fraction of the time series. Given that the applications of most models are focused on longer scales, an intercomparison of model performance at longer time scales is critical but largely beyond the length of most existing eddy covariance data records and the protocol of the NACP site-level intercomparison. This indicates a critical data need for long-term records at single sites. Also of large value for assessing long-term dynamics are multitower chronosequence studies [*Bond-Lamberty et al.*, 2004; *Stoy et al.*, 2008], though such sequences cannot be explicitly combined in a spectral analysis along the temporal domain of the chronosequence as there is a substitution of time for space.

[40] It is difficult to generalize about why certain classes of models fail. Coarse classifications of model structure provided insight into the variation in the diurnal cycle but proved to be largely uninformative about the annual and intermediate-scale errors that dominate the current analysis. Details of model function and parameterization are model specific and a model-by-model diagnosis is beyond the scope of this study. While it is possible that the NACP intercomparison simply failed to identify the model structural characteristics that drive model performance, especially at longer time scales, it is also important that efforts to diagnose model structural errors account for model parameter uncertainties. The current intercomparison protocol makes it difficult to distinguish models that failed due to misparameterization versus inherent structural limitations. The intercomparison did include one model (LoTEC-DA) that made use of data assimilation methods, and not surprisingly this model had the lowest absolute error [*Schwalm et al.*, 2010a]. Likewise, results from previously published spectral analyses of single models optimized to a single site resulted in models with very little diurnal or annual error [*Braswell et al.*, 2005], in contrast with the dominant pattern across models in the current analysis. Both these observations suggest that model parameter error may currently be dominating structural error or that many structural errors can be overcome with sufficient parameter flexibility.

[41] In conclusion, spectral analysis helps clarify when and where models fail, and provides guidelines for prioritizing efforts to improve our collective modeling capacity. Annual errors dominate model error and thus should be the first diagnostic upon which modelers should focus. Afterward, modelers should aim to capture the growing season diurnal cycles. Finally, models should focus on identification and attribution of synoptic error events.

[42] **Acknowledgments.** We would like to thank the North American Carbon Program Site-Level Interim Synthesis team and the Oak Ridge National Laboratory Distributed Active Archive Center for collecting, organizing, and distributing the model output and flux observations required for this analysis. Research by K.S. was partly funded by NOAA award NA07OAR4310115. C.K. was supported by the U.S. Department of Energy's Office of Science through the Midwestern Regional Center for the National Institute for Climatic Change Research at Michigan Technological University under award DE-FC02-06ER64158.

References

Arnone, J. A., et al. (2008), Prolonged suppression of ecosystem carbon dioxide uptake after an anomalously warm year, *Nature*, *455*(7211), 383–386, doi:10.1038/nature07296.

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, doi:10.1071/BT07151.

Baldocchi, D., E. Falge, and K. Wilson (2001), A spectral analysis of biosphere–atmosphere trace gas flux densities and meteorological variables across hour to multi-year time scales, *Agric. For. Meteorol.*, *107*, 1–27, doi:10.1016/S0168-1923(00)00228-8.

Barr A., D. Hollinger, and A. D. Richardson (2009), CO_2 flux measurement uncertainty estimates for NACP, *EoS Trans. AGU*, *90*(52), Fall Meet. Suppl., Abstract B54A-04.

Bates, J., and C. Granger (1969), The combination of forecasts, *Oper. Res.*, *20*, 451–468, doi:10.1057/jors.1969.103.

Beer, C., et al. (2010), Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate, *Science*, *329*(5993), 834–838, doi:10.1126/science.1184984.

Bond-Lamberty, B., C. Wang, and S. T. Gower (2004), Net primary production and net ecosystem production of a boreal black spruce wildfire chronosequence, *Global Change Biol.*, *10*(4), 473–487, doi:10.1111/j.1529-8817.2003.0742.x.

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*(2), 335–355, doi:10.1111/j.1365-2486.2005.00897.x.

Bunn, A. G. (2008), A dendrochronology program library in R (dplR), *Dendrochronologia*, *26*, 115–124, doi:10.1016/j.dendro.2008.01.002.

Falge, E., et al. (2002), Seasonality of ecosystem respiration and gross primary production as derived from FLUXNET measurements, *Agric. For. Meteorol.*, *113*, 53–74, doi:10.1016/S0168-1923(02)00102-8.

Farquhar, G., S. Caemmerer, and J. Berry (1980), A biochemical model of photosynthetic CO_2 assimilation in leaves of C_3 species, *Planta*, *149*, 78–90, doi:10.1007/BF00386231.

Flanagan, L. B., L. A. Wever, and P. J. Carlson (2002), Seasonal and interannual variation in carbon dioxide exchange and carbon balance in a northern temperate grassland, *Global Change Biol.*, *8*(7), 599–615, doi:10.1046/j.1365-2486.2002.00491.x.

Grinsted, A., J. C. Moore, and S. Jevrejeva (2004), Application of the cross wavelet transform and wavelet coherence to geophysical time series,

Nonlinear Processes Geophys., *11*, 561–566, doi:10.5194/npg-11-561-2004.

Hanson, P. J., et al. (2004), Oak forest carbon and water simulations: Model intercomparisons and evaluations against independent data, *Ecol. Monogr.*, *74*, 443–489, doi:10.1890/03-4049.

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

Hollinger, D. Y., et al. (2004), Spatial and temporal variability in forest-atmosphere CO_2 exchange, *Global Change Biol.*, *10*(10), 1689–1706, doi:10.1111/j.1365-2486.2004.00847.x.

Katul, G. G., C. T. Lai, K. V. R. Schäfer, B. Vidakovic, J. D. Albertson, D. S. Ellsworth, and R. Oren (2001), Multiscale analysis of vegetation surface fluxes: From seconds to years, *Adv. Water Resour.*, *24*, 1119–1132, doi:10.1016/S0309-1708(01)00029-X.

Krinner, G., N. Viovy, N. de Noblet-Ducoudré, J. Ogée, 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.

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*(5), 1311–1324, doi:10.5194/bg-5-1311-2008.

Law, B., T. Arkebauer, J. Campbell, J. Chen, O. Sun, M. Schwartz, C. van Ingen, and S. Verma (2008), Terrestrial carbon observations: Protocols for vegetation sampling and data submission, *Rep. GTOS 55*, Terr. Carbon Obs. Panel, Global Terr. Obs. Syst., Rome.

Liu, Y., X. San Liang, and R. H. Weisberg (2007), Rectification of the bias in the wavelet power spectrum, *J. Atmos. Oceanic Technol.*, *24*, 2093–2102, doi:10.1175/2007JTECHO511.1.

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.

Medvigy, D., S. C. Wofsy, J. W. Munger, D. Y. Hollinger, and P. R. Moorcroft (2009), Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem Demography model version 2, *J. Geophys. Res.*, *114*, G01002, doi:10.1029/2008JG000812.

Moffat, A., et al. (2007), Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes, *Agric. For. Meteorol.*, *147*, 209–232, doi:10.1016/j.agrformet.2007.08.011.

Moorcroft, P. R., G. C. Hurtt, and S. W. Pacala (2001), A method for scaling vegetation dynamics: The ecosystem demography model (ED), *Ecol. Monogr.*, *71*, 557–586, doi:10.1890/0012-9615(2001)071[0557:AMFSVD]2.0.CO;2.

Poulter, B., U. Heyder, and W. Cramer (2009), Modeling the sensitivity of the seasonal cycle of GPP to dynamic LAI and soil depths in tropical rainforests, *Ecosystems*, *12*, 517–533, doi:10.1007/s10021-009-9238-4.

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*(9), 1424–1439, doi:10.1111/j.1365-2486.2005.001002.x.

Ricciuto, D. M., P. E. Thornton, K. Schaefer, R. B. Cook, and K. J. Davis (2009), How uncertainty in gap-filled meteorological input forcing at eddy covariance sites impacts modeled carbon and energy flux, *Eos Trans. AGU*, *90*(52), Fall Meet. Suppl., Abstract B54A-03.

Richardson, A. D., and D. Y. Hollinger (2007), A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO_2 flux record, *Agric. For. Meteorol.*, *147*, 199–208, doi:10.1016/j.agrformet.2007.06.004.

Richardson, A. D., et al. (2006), A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes, *Agric. For. Meteorol.*, *136*, 1–18, doi:10.1016/j.agrformet.2006.01.007.

Richardson, A. D., D. Y. Hollinger, J. D. Aber, S. 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*(4), 788–803, doi:10.1111/j.1365-2486.2007.01330.x.

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, doi:10.1016/j.agrformet.2007.09.001.

Richardson, A. D., et al. (2011), Land surface models need better representation of vegetation phenology: Results from the North American Carbon Program Site Synthesis, *Global Change Biol.*, doi:10.1111/j.1365-2486.2011.02562.x, in press.

Scanlon, T. M., and J. D. Albertson (2001), Turbulent transport of carbon dioxide and water vapor within a vegetation canopy during unstable conditions: Identification of episodes using wavelet analysis, *J. Geophys. Res.*, *106*, 7251–7262, doi:10.1029/2000JD900662.

Schwalm, C. R., et al. (2010a), A model-data intercomparison of CO_2 exchange across North America: Results from the North American Carbon Program site synthesis, *J. Geophys. Res.*, *115*, G00H05, doi:10.1029/2009JG001229.

Schwalm, C. R., et al. (2010b), Assimilation exceeds respiration sensitivity to drought: A FLUXNET synthesis, *Global Change Biol.*, *16*(2), 657–670, doi:10.1111/j.1365-2486.2009.01991.x.

Siqueira, M. B. S., G. G. Katul, D. A. Sampson, P. C. Stoy, J.-Y. Juang, H. R. McCarthy, and R. Oren (2006), Multi-scale model intercomparisons of CO_2 and H_2O exchange rates in a maturing southeastern U.S. pine forest, *Global Change Biol.*, *12*(7), 1189–1207, doi:10.1111/j.1365-2486.2006.01158.x.

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 time scales at adjacent pine and hardwood forests: A wavelet analysis, *Tree Physiol.*, *25*, 887–902.

Stoy, P. C., G. G. Katul, M. B. S. Siqueira, J.-Y. Juang, K. A. Novick, H. R. McCarthy, A. C. Oishi, and R. Oren (2008), Role of vegetation in determining carbon sequestration along ecological succession in the southeastern United States, *Global Change Biol.*, *14*(6), 1409–1427, doi:10.1111/j.1365-2486.2008.01587.x.

Stoy, P. C., et al. (2009), Biosphere-atmosphere exchange of CO_2 in relation to climate: A cross-biome analysis across multiple time scales, *Biogeosciences*, *6*, 2297–2312, doi:10.5194/bg-6-2297-2009.

Torrence, C., and G. P. Compo (1998), A practical guide to wavelet analysis, *Bull. Am. Meteorol. Soc.*, *79*, 61–78, doi:10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2.

Tucker, C. J., J. E. Pinzon, M. E. Brown, D. A. Slayback, E. W. Pak, R. Mahoney, E. F. Vermote, and N. El Saleous (2005), An extended AVHRR 8-km NDVI dataset compatible with MODIS and SPOT vegetation NDVI data, *Int. J. Remote Sens.*, *26*, 4485–4498, doi:10.1080/01431160500168686.

Urbanski, S. P., C. Barford, S. Wofsy, C. J. Kucharik, E. H. Pyle, J. Budney, K. McKain, D. Fitzjarrald, M. J. Czikowsky, and J. W. Munger (2007), Factors controlling CO_2 exchange on timescales from hourly to decadal at Harvard Forest, *J. Geophys. Res.*, *112*, G02020, doi:10.1029/2006JG000293.

Vargas, R., M. Detto, D. D. Baldocchi, and M. F. Allen (2010), Multiscale analysis of temporal variability of soil CO_2 production as influenced by weather and vegetation, *Global Change Biol.*, *16*(5), 1589–1605, doi:10.1111/j.1365-2486.2009.02111.x.

Vargas, R., M. S. Carbone, M. Reichstein, and D. D. Baldocchi (2011), Frontiers and challenges in soil respiration research: From measurements to model-data integration, *Biogeochemistry*, *102*, 1–13, doi:10.1007/s10533-010-9462-1.

VEMAP Members (1995), Vegetation/Ecosystem Modeling and Analysis Project: Comparing biogeography and biogeochemistry models in a continental-scale study of terrestrial ecosystem responses to climate change and CO_2 doubling, *Global Biogeochem. Cycles*, *9*(4), 407–437, doi:10.1029/95GB02746.

Williams, M., et al. (2009) Improving land surface models with FLUXNET data, *Biogeosciences*, *6*, 1341–1359, doi:10.5194/bgd-6-2785-2009.

R. S. Anderson, Numerical Terradynamic Simulation Group, University of Montana, 32 Campus Dr., Missoula, MT 59812, USA. (ryan.anderson@ntsg.umt.edu)

M. A. Arain, School of Geography and Earth Sciences, McMaster University, 1280 Main St. W., Hamilton, ON L8S 4K1, Canada. (arainm@mcmaster.ca)

I. T. Baker and E. Lokupitiya, Department of Atmospheric Science, Colorado State University, 1371 Campus Delivery, Fort Collins, CO 80523-1371, USA. (baker@atmos.colostate.edu; erandi@atmos.colostate.edu)

A. G. Barr, Climate Research Division, Atmospheric Science and Technology Directorate, Saskatoon, SK S7N 3H5, Canada. (alan.barr@ec.gc.ca)

T. A. Black, Faculty of Land and Food Systems, University of British Columbia, 2357 Main Mall, Vancouver, BC V6T 1Z4, Canada. (andrew.black@ubc.ca)

J. M. Chen, Department of Geography and Program in Planning, University of Toronto, 100 St. George St., Rm. 5047, Toronto, ON M5S 3G3, Canada. (chenj@geog.utoronto.ca)

P. Ciais, Centre d'Etudes Orme des Merisiers, F-91191 Gif-sur-Yvette, France. (philippe.ciais@cea.fr)

M. C. Dietze, Department of Plant Biology, University of Illinois at Urbana-Champaign, 505 S. Goodwin Ave., Urbana, IL 61801, USA. (mdietze@life.uiuc.edu; 217-265-8020)

L. B. Flanagan, Department of Biological Sciences, University of Lethbridge, 4401 University Dr., Lethbridge, AB T1K 3M4, Canada. (larry.flanagan@uleth.ca)

C. M. Gough, Department of Biology, Virginia Commonwealth University, Box 842012, 1000 W. Cary St., Richmond, VA 23284, USA. (cmgough@vcu.edu)

R. F. Grant, Department of Renewable Resources, University of Alberta, 4-30 Earth Sciences Bldg., Edmonton, AB T6G 2E3, Canada. (robert.grant@afhe.ualberta.ca)

D. Hollinger, Northern Research Station, U.S. Department of Agriculture Forest Service, 271 Mast Rd., Durham, NH 03824, USA. (davidh@hypatia.unh.edu)

R. C. Izaurralde, Joint Global Change Research Institute, University of Maryland, 5825 University Research Ct., Ste. 3500, College Park, MD 20740, USA. (cesar.izaurralde@pnl.gov)

C. J. Kucharik, Department of Agronomy, University of Wisconsin-Madison, 1575 Linden Dr., Madison, WI 53706, USA. (kucharik@wisc.edu)

P. Lafleur, Department of Geography, Trent University, 1600 West Bank Dr., Peterborough, ON K9J 7B8, Canada. (plafleur@trentu.ca)

S. Liu, Earth Resources Observation and Science Center, 47914 252nd St., Sioux Falls, SD 57198, USA. (sliu@usgs.gov)

Y. Luo and E. Weng, Department of Botany and Microbiology, University of Oklahoma, 770 Van Vleet Oval, Norman, OK 73019, USA. (yluo@ou.edu)

J. W. Munger, School of Engineering and Applied Sciences, Harvard University, 20 Oxford St., Cambridge, MA 02138, USA. (jwmunger@seas.harvard.edu)

C. Peng and W. Wang, Department of Biology Sciences, University of Quebec at Montreal, PO Box 8888, Montreal, QC H3C 3P8, Canada. (peng.changhui@uqam.ca)

B. Poulter, Laboratoire des Sciences du Climat et de l'Environnement, Institut Pierre Simon Laplace, Bât. 709, Orme des Merisiers, F-91191 Gif-sur-Yvette, France. (benjamin.poulter@lsce.ipsl.fr)

D. T. Price, Northern Forestry Centre, Canadian Forest Service, 5320 122nd St., Edmonton, AB T6H 3S5, Canada. (dprice@nrcan.gc.ca)

D. M. Ricciuto, Environmental Sciences Division, Oak Ridge National Laboratory, PO Box 2008, MS 6301 Oak Ridge, TN 37831, USA. (ricciutodm@ornl.gov)

A. D. Richardson, Department of Organismic and Evolutionary Biology, Harvard University, 22 Divinity Ave., Cambridge, MA 02138, USA. (arichardson@oeb.harvard.edu)

W. J. Riley, Climate and Carbon Sciences, Earth Sciences Division, Lawrence Berkeley National Laboratory, Bldg. 90-1106, 1 Cyclotron Rd., Berkeley, CA 94720, USA. (wjriley@lbl.gov)

A. K. Sahoo, Department of Civil and Environmental Engineering, Princeton University, E 324, Engineering Quad, Princeton, NJ 08544, USA. (sahoo@princeton.edu)

K. Schaefer, National Snow and Ice Data Center, University of Colorado at Boulder, 449 UCB, Boulder, CO 80309, USA. (kevin.schaefer@nsidc.org)

P. C. Stoy, Department of Land Resources and Environmental Sciences, Montana State University, 616 Leon Johnson Hall, Bozeman, MT 59717, USA. (paul.stoy@montana.edu)

A. E. Suyker and S. B. Verma, School of Natural Resources, University of Nebraska–Lincoln, 806 Hardin Hall, 3310 Holdrege St., Lincoln, NE 68583, USA. (asuyker1@unl.edu; sverma1@unl.edu)

H. Tian, School of Forestry and Wildlife Sciences, Auburn University, 602 Duncan Dr., Auburn, AL 36849, USA. (tianhan@auburn.edu)

C. Tonitto, Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY 14853-2701, USA. (ct244@cornell.edu)

R. Vargas, Departamento de Biología de la Conservación, Centro de Investigación Científica y de Educación Superior de Ensenada, Carretera Ensenada-Tijuana 3918, Ensenada, Baja California 22860, Mexico. (rvargas@cicese.mx)

H. Verbeeck, Faculty of Bioscience Engineering, Laboratory of Plant Ecology, Department of Applied Ecology and Environmental Biology, Ghent University, Coupure Links 653, B-9000 Ghent, Belgium. (hans.verbeeck@ugent.be)