CONCEPTS & SYNTHESIS
EMPHASIZING NEW IDEAS TO STIMULATE RESEARCH IN ECOLOGY

Ecological Monographs, 83(2), 2013, pp. 133–154
© 2013 by the Ecological Society of America

Facilitating feedbacks between field measurements and ecosystem models

David S. LeBauer,[1,3] Dan Wang,[1] Katherine T. Richter,[2] Carl C. Davidson,[2] and Michael C. Dietze[1,2,4]

[1] *Energy Biosciences Institute, University of Illinois, Urbana, Illinois 61801 USA*
[2] *Department of Plant Biology, University of Illinois, Urbana, Illinois 61801 USA*

Abstract. Ecological models help us understand how ecosystems function, predict responses to global change, and identify future research needs. However, widespread use of models is limited by the technical challenges of model–data synthesis and information management.

To address these challenges, we present an ecoinformatic workflow, the Predictive Ecosystem Analyzer (PEcAn), which facilitates model analysis. Herein we describe the PEcAn modules that synthesize plant trait data to estimate model parameters, propagate parameter uncertainties through to model output, and evaluate the contribution of each parameter to model uncertainty. We illustrate a comprehensive approach to the estimation of parameter values, starting with a statement of prior knowledge that is refined by species-level data using Bayesian meta-analysis; this is the first use of a rigorous meta-analysis to inform the parameters of a mechanistic ecosystem model.

Parameter uncertainty is propagated using ensemble methods to estimate model uncertainty. Variance decomposition allows us to quantify the contribution of each parameter to model uncertainty; this information can be used to prioritize subsequent data collection. By streamlining the use of models and focusing efforts to identify and constrain the dominant sources of uncertainty in model output, the approach used by PEcAn can speed scientific progress.

We demonstrate PEcAn's ability to incorporate data to reduce uncertainty in productivity of a perennial grass monoculture (*Panicum virgatum* L.) modeled by the Ecosystem Demography model. Prior estimates were specified for 15 model parameters, and species-level data were available for seven of these. Meta-analysis of species-level data substantially reduced the contribution of three parameters (specific leaf area, maximum carboxylation rate, and stomatal slope) to overall model uncertainty. By contrast, root turnover rate, root respiration rate, and leaf width had little effect on model output; therefore trait data had little impact on model uncertainty.

For fine-root allocation, the decrease in parameter uncertainty was offset by an increase in model sensitivity. Remaining model uncertainty is driven by growth respiration, fine-root allocation, leaf turnover rater, and specific leaf area. By establishing robust channels of feedback between data collection and ecosystem modeling, PEcAn provides a framework for more efficient and integrative science.

Key words: Bayesian approach; ecoinformatics; ecological forecast; ecophysiology; Ecosystem Demography model; ecosystem model; meta-analysis; plant traits; sensitivity analysis; variance decomposition.

Introduction

In the face of unprecedented global change there is growing demand for predictions of ecosystem responses that provide actionable information for policy and management (Clark et al. 2001). Currently, the response of the terrestrial biosphere remains one of the largest

Manuscript received 26 January 2012; revised 1 August 2012; accepted 2 August 2012; final version received 12 October 2012. Corresponding Editor: A. M. Ellison.

[3] E-mail: dlebauer@illinois.edu

[4] Present address: Department of Earth and Environment, Boston University, Boston, MA 02215 USA.

sources of uncertainty in projections of climate change (Denman et al. 2007).

This uncertainty comes from a combination of the uncertainties about our conceptual understanding of ecological systems as captured by the structure and assumptions of the models used to make ecological forecasts, the uncertainties in the parameters of these models, and the uncertainties associated with the underlying data (McMahon et al. 2009). Reducing these uncertainties requires that we be able to synthesize existing information, efficiently identify the dominant sources of model uncertainty and target them with further field research.

Despite the acknowledged importance of these activities, there is often a disconnection between model simulation and data collection. Both model–data synthesis and the investigation of uncertainty remain challenging, while the use of models to quantitatively inform data collection is extremely rare. Most modeling uses a single point estimate for each parameter, effectively treating each parameter value as completely certain. However, such point estimates do not account for the degree to which we understand a parameter based on observations. Furthermore, the rationale for a particular estimate is often unclear, as is the degree to which the estimate represents the process being observed or its representation in a model. In many cases, parameter values are chosen iteratively to "tune" or "calibrate" the model output to observations. A first step toward constraining model uncertainty is to account for uncertainty in model parameters instead of relying on point estimates.

More rigorous approaches to estimating parameter values include model optimization and data assimilation (Reichstein et al. 2003, Medvigy et al. 2009), as well as Bayesian model–data fusion (Luo et al. 2011). However, these approaches have generally started with uninformative or vague prior estimates of model parameters. These vague priors ignore available data that could directly inform parameter values; the most commonly used vague prior distributions are uniform. A uniform prior assigns equal probability to parameter values over its entire range, in many cases over many orders of magnitude.

The use of such vague priors often exacerbates problems with equi-finality (Richardson and Hollinger 2005, Luo et al. 2009, Williams et al. 2009) which can produce unidentifiable parameters, as well as biologically unrealistic parameter sets that generate the right model output for the wrong reasons (Beven and Freer 2001, Beven 2006, Williams et al. 2009).

Another reason to use informed priors is to take advantage of one of the key strengths of the Bayesian paradigm: the ability to synthesize multiple sources of information in a rigorous and consistent framework. For example, plant traits related to leaf stoichiometry and photosynthetic capacity are often well constrained by previous research (Wullschleger 1993, Reich and Oleksyn 2004, Wright et al. 2004, Skillman 2008), whereas other traits, such as root respiration rate, are more difficult to measure and data are sparse. Informed priors allow existing information to be formally integrated into model parameterization, even if there are no data for the particular species or plant functional type (PFT) being measured; the level of confidence in a parameter value is reflected in the parameter's variance.

Models have rarely been used to quantify the value of data with respect to reducing uncertainty. Instead, data collection is often focused on answering specific questions in specific spatial, temporal, and taxonomic contexts. In these contexts, the value of a particular data set is based on the ability to answer a particular question. However, the same data set may have a very different value in the context of reducing model uncertainty. For example, a single data point used to inform a poorly understood but influential model parameter can reduce model uncertainty more than a large collection of data on a trait that is relatively well studied. In a modeling context, the value of an additional data point depends both on how much it constrains parameter uncertainty and on the sensitivity of model output to the parameter. Thus, the ability to comprehensively utilize available data in model parameterization can help to identify gaps in existing knowledge, improve the ability of models to account for current understanding, and inform data collection efforts by identifying the knowledge gaps most responsible for uncertainty.

Although the increasing sophistication of model–data fusion and uncertainty accounting is a critical step in the right direction, the complexity of such analyses can make models even less accessible. One of the reasons to make models more accessible, and to make them better at synthesizing existing data, is that models are fundamentally a formal, quantitative distillation of our current understanding of how a system works. As such, models can be used to identify gaps in our understanding and target further research.

This feedback between models and data could be improved if models were routinely evaluated in a way that quantifies the value of data with respect to reducing uncertainty.

We fundamentally believe that streamlining the informatics of modeling, the need to track, process, and synthesize data and model output, will make the development and application of ecological data and models more accessible, transparent, and relevant.

In this paper we present the Predictive Ecosystem Analyzer (PEcAn) as a step toward meeting these objectives. PEcAn is a scientific workflow that manages the flows of data used and produced by ecological models, and that assists with model parameterization, error propagation, and error analysis. PEcAn accomplishes two goals: first, it synthesizes data and propagates uncertainty through an ecosystem model; second, it places an information value on subsequent data collection to efficiently reduce uncertainty. In addition to quantifying the information content of any prediction or assessment, these techniques also help to identify the gaps in our knowledge of ecological and biogeochemical processes (Saltelli et al. 2008).

PEcAn addresses the challenge of synthesizing plant trait data from the literature in a way that accounts for the different scales and sources of uncertainty. Available data are synthesized using a Bayesian meta-analysis, and the meta-analysis posterior estimates of plant traits are used as parameters in an ecosystem model.

A model ensemble is a set of model runs with parameter values drawn from the meta-analysis posterior estimates of plant traits. Output from a model ensemble represents the posterior predictive distributions of ecosystem responses that account for trait parameter uncertainty. Hereafter "model posterior" refers to the "model ensemble output." Sensitivity analysis and variance decomposition help to determine which traits (model parameters) drive uncertainty in ecosystem response (model posterior) (Larocque et al. 2008, Saltelli et al. 2008). These analyses help to target parameters for further constraint with trait data, forming a critical feedback loop that drives further field research and provides an informative starting point for data assimilation. Here we illustrate an application of PEcAn to the assessment of aboveground yield in a perennial grass monoculture.

In the sections that follow, we provide an overview of the components of PEcAn's integrated framework for data synthesis and ecological prediction. We start with a description of the methods implemented in the workflow (*Implementation*). This includes descriptions of the database, Bayesian meta-analysis, ensemble analysis, sensitivity analysis, and variance decomposition. Finally we present an example of the application of the system (*Application*) to analyze the aboveground biomass of switchgrass (*Panicum virgatum* L.), by the Ecosystem Demography model version 2.1 (ED2) (Moorcroft et al. 2001, Medvigy et al. 2009).

Implementation

PEcAn workflow.—The Predictive Ecosystem Analyzer (PEcAn) manages the flow of information into and out of ecosystem models. PEcAn is not a model itself, but a scientific workflow consisting of discrete steps, or modules. Individual modules are building blocks of the workflow, represented by the rectangles in Fig. 1, whereas flows of information are represented by arrows. This makes PEcAn an encapsulated, semi-automated system for model parameterization, error propagation, and analysis.

A central objective of the PEcAn workflow is to make the entire modeling process transparent, reproducible, and adaptable to new questions (sensu Ellison 2010, Stodden et al. 2010). To achieve this objective, PEcAn's adheres to "best practice" guidelines for ecological data management and provenance tracking (Jones et al. 2006, Michener and Jones 2012).

PEcAn uses a database to track data provenance and a settings file to control workflow analyses and model runs. The database records the site, date, management, species, and treatment information for each trait observation used in the meta-analysis. Settings related to the experimental design and computation are set and recorded in a separate file for each analysis.

The code, inputs, outputs, and a virtual machine required to reproduce the analysis described in the section *Application* are provided in the PEcAn version 1.1 archive (*available online*).[5]

The PEcAn "virtual machine" minimizes the effort required to run PEcAn and begin using an ecosystem model. The virtual machine contains the PEcAn software, ecosystem models, and other software dependencies in a pre-configured desktop environment that can be run on any standard operating system using a freely available virtualization software such as VirtualBox (*available online*)[6] or VmWare Player (*available online*).[7] We use the virtual machine to support investigation, development, and education.

The PEcAn software is primarily written in R and developed in a Linux environment. It also relies on a MySQL database, JAGS, and contributed R packages. PEcAn has a family of model-specific functions that manage the details of launching of model runs and reading model output.

Although PEcAn does not depend on any specific model, it was developed to support ecosystem models that run in high-performance computing environments; for this reason, it is capable of running models locally, remotely, or through queuing systems regardless of whether PEcAn is compiled locally or run as a virtual machine. The PEcAn 1.1 release described herein runs with ED2 (see footnote 5). More recent versions include web interfaces to the database, model execution, and the R programming environment (Dietze et al. 2013) plus support for two additional models: Sipnet (Moore et al. 2008) and BioCro (Miguez et al. 2009). Current work includes multi-model inference and intercomparison.

Trait database.—Model parameters are associated with corresponding prior distributions, and in many cases, with species-level data. Both prior distributions and data are stored in a relational database (Supplement 1). PEcAn directly accesses the database, which contains additional meta-data for each data set, including site descriptions, measurement conditions, experimental details, and citations.

Trait priors.—A fundamental component of the Bayesian approach to parameter estimation is the use of priors. Priors formally incorporate knowledge of a parameter based on previous studies into a new analysis.

[5] https://www.ideals.illinois.edu/handle/2142/34655

[6] www.virtualbox.org/

[7] www.vmware.com

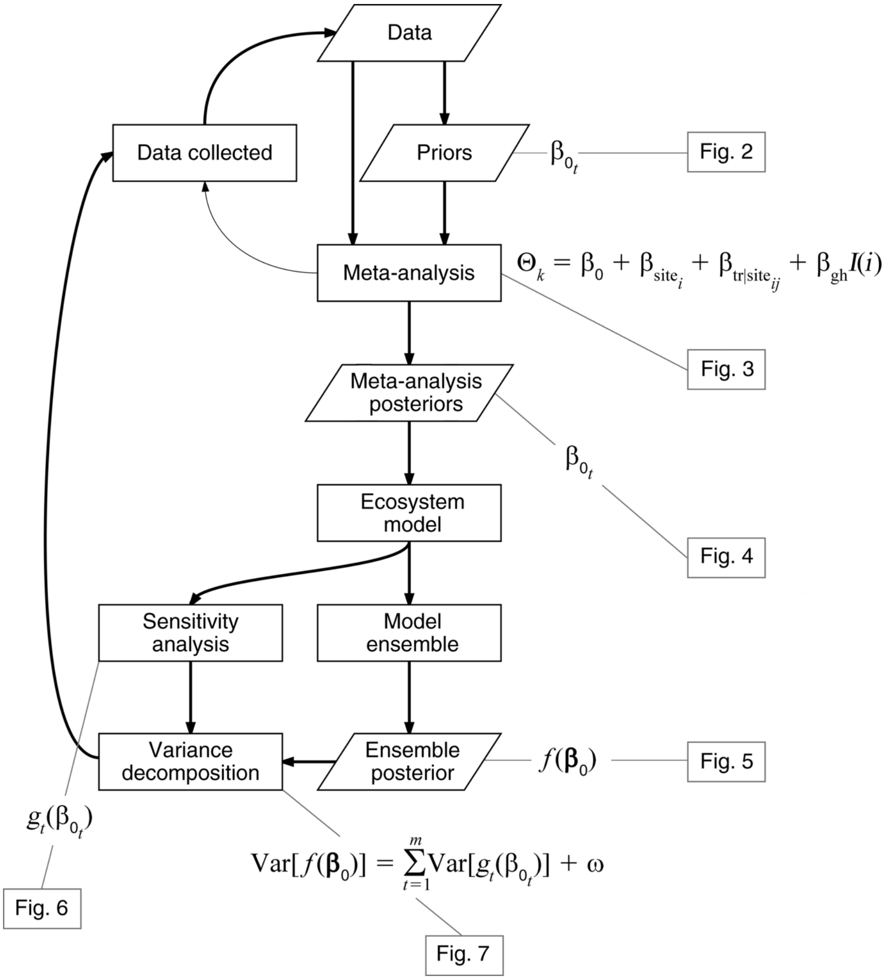

FIG. 1. Overview of the PEcAn (Predictive Ecosystem Analyzer) workflow. The synthesis of plant trait data begins by querying a database of plant trait data for data on a single species or a plant functional type and then mapping these data to the model parameters that they inform. The database also provides probability distributions that describe our prior information about the range of values that a model parameter can take. Next, this information is synthesized in a Bayesian meta-analysis, resulting in a posterior trait distribution that summarizes the uncertainty in each parameter. The ensemble of model runs produces the posterior distribution of model outputs, representing a probabilistic assessment or forecast that accounts for input parameter uncertainty. The final steps in the workflow are the sensitivity analysis and variance decomposition; these steps gives insight into the relative contribution of each parameter to the uncertainty in the model output and can be used to guide additional data collection that will most efficiently reduce model uncertainty.

In the current study, we leverage previously collected data from nontarget species to place biologically informed constraints on the distributions of plant trait parameters. When additional data for a specific species or plant functional type are available, priors are further constrained before being used as model parameters. When no additional data are available, these priors are used directly to parameterize the model.

For the *P. virgatum* example, priors were set using data from all plant species, from only grass species, or from just C_4 grass species depending on available data. Sources of this prior information included data from previous and ad hoc syntheses, expert knowledge, and biophysical constraints (Table 1).

Prior distributions used in the meta-analysis were fit to one of four types of information: (1) data from multiple species, (2) the posterior predictive distribution for a new species from a meta-analysis of data (when error estimates were available), (3) a central tendency informed by data with expert constraint on the confidence interval, or (4) expert constraints on both the central tendency and confidence intervals. In case (2), the across-species meta-analysis "posterior" informs the prior for the species-level meta-analysis. In all cases, maximum likelihood estimation was used to fit a prior distribution. When more than one candidate distribution was considered, Aikake's information criterion (AIC) was used to select the best-fit distribution. The

TABLE 1. Prior distributions used in meta-analysis and model parameterization.

Parameter (units)	Clade	Distribution	*a*	*b*	*N*	Mean	LCL	UCL	Source
Specific leaf area (m^2/kg)	grass	Gamma	2.06	19.00	125	17	3.2	36	Wright et al. (2004)
Leaf turnover rate (yr^{-1})	grass	Weibull	2.90	0.63	40	4.6	0.91	11	Wright et al. (2004)
Root turnover rate (yr^{-1})	grass	Gamma	1.67	0.66	66	0.59	0.073	1.4	Gill and Jackson (2000)
Quantum efficiency (proportion)	C_4 grass	Weibull	90.90	1580.00	56	0.058	0.046	0.07	Skillman (2008)
Stomatal slope (ratio)	C_4 grass	Gamma	3.63	3.81	4	3.4	1.4	5.5	data from this study; Supplement 1
$V_{c,max}$ (μmol $CO_2 \cdot m^{-2} \cdot s^{-1}$)	graminoid	Gamma	3.49	24.70	97	22	8.6	36	Wullschleger (1993), Kubien and Sage (2004), Massad et al. (2007), Wang et al. (2011)
Leaf width (mm)	C_4 grass	Weibull	26.10	5.94	18	4.4	2.9	6.2	Oyarzabal et al. (2008)
Root respiration rate (μmol $CO_2 \cdot m^{-2} \cdot s^{-1}$)	C_4 grass	*F*	5.61	2.33	35	5.6	1	10	Tjoelker et al. (2005)
Fine-root allocation (ratio)	grass	log-normal	0.80	0.81	0	3.1	0.46	11	Saugier et al. (2001)
Seed dispersal (proportion)	grass	Beta	20.10	74.90	30	0.21	0.14	0.3	Jongejans and Schippers (1999)
Photosynthesis min. temp. (°C)	C_4 grass	*F*	10.00	1.02	0	10	8	12	D. S. LeBauer and D. Ort (*personal communication*)
Growth respiration (proportion)	grass	Beta	2.63	6.52	0	0.29	0.062	0.6	†, D. S. LeBauer and M. C. Dietze
Seedling mortality (proportion)	monocots	Beta	3.61	0.43	0	0.89	0.5	1.0	†, D. S. LeBauer and M. C. Dietze
Mortality coefficient (yr^{-1})	plants	Weibull	1.47	0.06	0	25	1.8	80	†, D. S. LeBauer and M. C. Dietze
Reproductive allocation (proportion)	plants	Beta	2.00	4.00	0	0.33	0.053	0.72	†, D. S. LeBauer and M. C. Dietze

Notes: Terms *a* and *b* are the first and second parameters of the probability distribution; *N* is the number of distinct species represented in data used to estimate the prior; LCL and UCL are the upper and lower 95% credible limits, respectively. Sources are listed in the right-most column: citations, authors of the present study, or default ED2 parameterizations. The "Clade" column indicates the group of plants for which the priors were derived.

† Default ED2 parameterizations, as described in *Methods: Trait priors*.

choice of prior was confirmed by visually inspecting the prior density functions overlain by data or expert constraints (Fig. 2).

Meta-analysis.—A hierarchical Bayes meta-analytical model (Fig. 3) formally synthesizes available trait data from multiple studies while accounting for various sources of uncertainty. This hierarchical Bayes approach integrates prior information and provides a flexible approach to variance partitioning and parameter estimation.

The meta-analytical framework is useful for summarizing data sets that include summary statistics. The trait data queried by PEcAn consist of a trait name, sample mean, sample size, and a sample error statistic. PEcAn transforms error statistics to exact or conservative (i.e., erring toward inflating the variance) estimates of precision ($\tau = 1/\mathrm{SE}^2$) (Appendix).

The sample mean is drawn from a normal distribution:

$$Y_k \sim \mathcal{N}(\Theta_k, \tau_k) \qquad (1)$$

where Y_k is the sample mean of the kth unique site × treatment combination (sample unit) and Θ_k is the unobserved "true" value of the trait for the kth sample unit.

The meta-analysis partitions trait variability into among-site, among-treatment, and within-unit variance. The unobserved "true" trait mean Θ_k is a linear function of the global trait mean, β_0, plus random effects for study site (β_{site_j}) and treatment ($\beta_{\mathrm{tr}\,|\,\mathrm{site}_{ij}}$), and a fixed effect for greenhouse (β_{gh}):

$$\Theta_k = \beta_0 + \beta_{\mathrm{site}_i} + \beta_{\mathrm{tr}\,|\,\mathrm{site}_{ij}} + \beta_{\mathrm{gh}} I(i) \qquad (2)$$

where i indexes study site, j indexes each treatment within a study, and $I(i)$ is an indicator variable set to 0 for field studies and 1 for studies conducted in a greenhouse, growth chamber, or pot experiment. The parameter used in the ecosystem model is the posterior estimate of the global mean trait value, β_0, which has an informed prior functional form and parameter specification that varies by trait and species or PFT. Methods used to elicit priors for the present study are provided in the *Application* section under *Priors*.

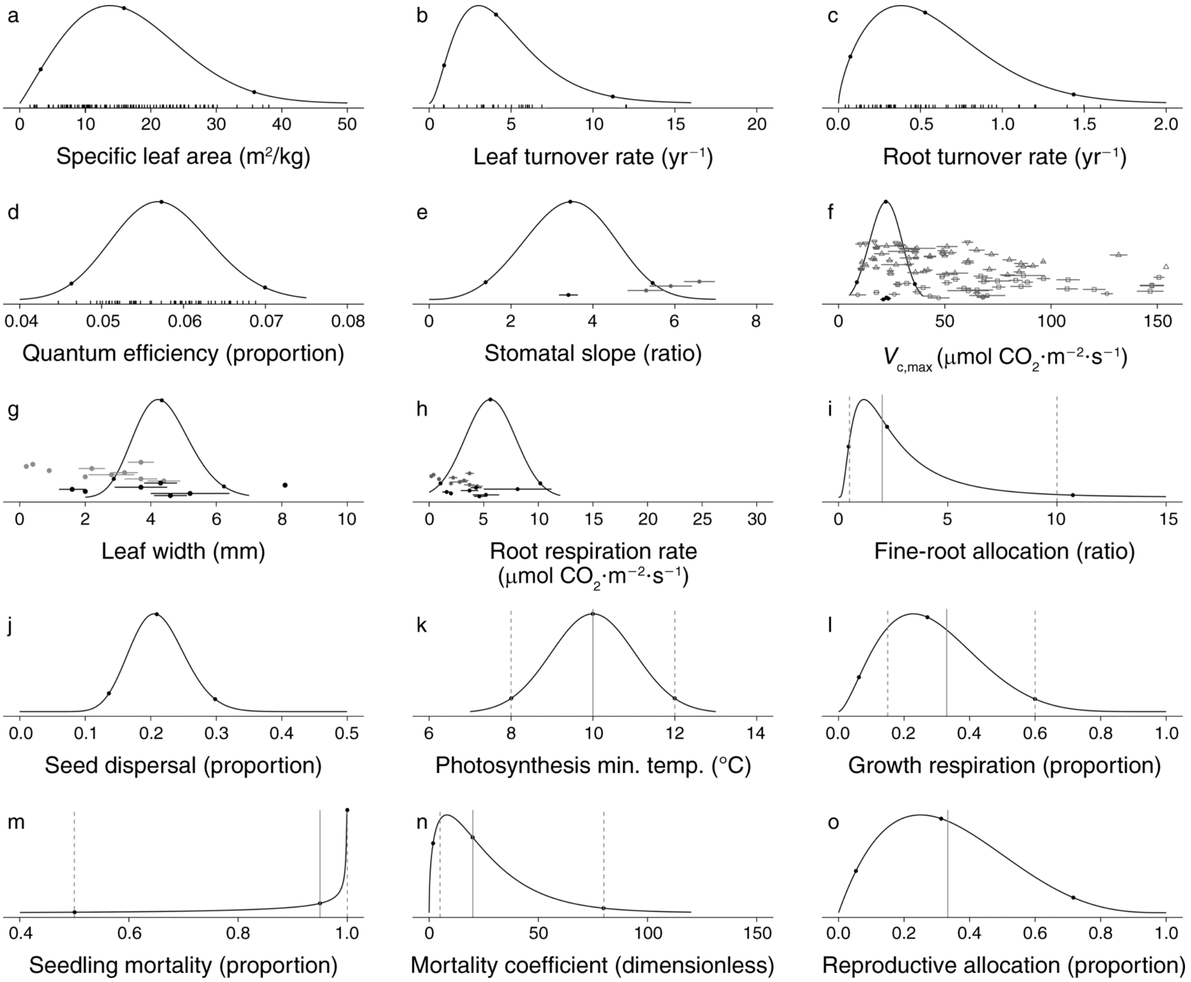

FIG. 2. Prior distributions (probability density functions, PDF) of priors with data constraints, based on the traits of plants within broad taxonomic or functional type groupings (e.g., all grasses, gray density lines). The parameter value is on the *x*-axis, probability density is on the *y*-axis, and the area under each curve equals 1. Three points on each line, from left to right, indicate the 2.5th, 50th, and 97.5th quantiles. (a–d) Four priors fit to data (data points shown as rug plot) using maximum likelihood: specific leaf area and leaf turnover rate (Wright et al. 2004), root turnover rate (Gill and Jackson 2000), and quantum yield (Skillman 2008). (e–h) Four priors fit to the posterior predictive distribution of an unobserved C_4 grass species using Bayesian meta-analysis of data from multiple plant functional types (C_4 data shown in black, other functional types in gray): stomatal slope (*Miscanthus*, black circles; three woody species, gray circles; data collected for the present study and provided in Supplement 1), $V_{c,max}$ of C_3 plants (gray symbols; Wullschleger 1993) and C_4 grasses (black circles; Kubien and Sage 2004, Massad et al. 2007, Wang et al. 2011), leaf width from C_3 and C_4 grass species (Oyarzabal et al. 2008), and root respiration (black circles, C_4 grass; gray symbols, C_3 plants; Tjoelker et al. 2005). (i–o) Priors fit to 95% CI (dashed gray line) and median (solid gray line) based on ED2 (Ecosystem Demography model, version 2) defaults and expert opinion as described in the text: fine-root to leaf ratio (Saugier et al. 2001), seed dispersal (Ernst et al. [1992] model parameterized with site-level data), minimum temperature of photosynthesis (D. Ort, *personal communication*), growth respiration, seedling mortality factor, mortality factor, and reproductive allocation.

The "site" random effects (β_{site}), account for the spatial (among-site) heterogeneity of a parameter. The "treatment" random effect ($\beta_{\text{tr}\,|\,\text{site}}$) accommodates differences among experimental treatments. These random effects of treatment and site are assumed to have normal distribution with zero mean and they have diffuse Gamma priors on precision τ_{site} and τ_{tr}. Control treatments and observational studies have $\beta_{\text{tr}\,|\,\text{site}} = 0$. PEcAn dynamically adjusts the meta-analysis model specification to include terms for each level of site and treatment, or greenhouse studies as required by available data. To ensure that the prior on precision remains sufficiently diffuse when the magnitude of a parameter is small, the scale parameters in the Gamma priors on random effect precision terms (τ_{site} and $\tau_{\text{tr}\,|\,\text{site}}$) are scaled to $(\bar{\beta}_0^2/1000)$ when the prior on β_0 has a mean $\bar{\beta}_0 < \sqrt{10}$.

A "greenhouse" fixed effect accounts for potential biases associated with plants grown in a greenhouse, growth chamber, pot, or other controlled environment. This "greenhouse" effect, β_{gh}, has a diffuse normal prior with a mean of zero and a precision of 0.01.

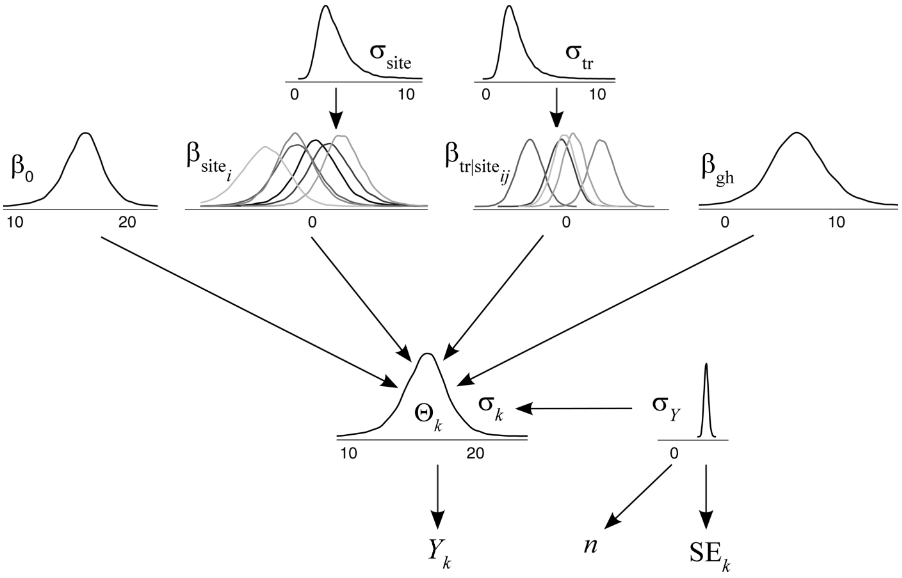

FIG. 3. Overview with a modified Kruschke (2010) diagram of the hierarchical Bayesian meta-analysis model used to estimate species-level trait parameters. For each trait, the posterior estimate of the global trait mean (β_0) is used as an input parameter in the sensitivity analysis and model ensemble (Figs. 5 and 6). Results from the meta-analysis of specific leaf area are used as an illustrative example; *x*-axes have units of m^2/kg, and all plots are on the same scale; the *y*-axis is the probability density function (area under the curve = 1). Each of the *k* sample means (Y_k) were taken from published articles and unpublished field measurements, and may be associated with a sample standard error and sample size. When sufficient data were available, site, treatment, and greenhouse effects were estimated. The within-unit standard deviation, σ_Y, is estimated from SE and *n*. Site and treatment random effects, represented by β_{site} and $\beta_{\text{tr|site}}$, are estimated for each site and treatment within site from normal distributions with mean zero and standard deviations σ_{site} and $\sigma_{\text{tr|site}}$, respectively. Greenhouse is a fixed effect. Table 2 summarizes the global mean, variance terms, and greenhouse effect for the seven model parameters informed by species-level data.

The observation precision (precision = 1/variance) for the *k*th sample mean, τ_k, is determined based on the within-unit precision, τ_Y, and the sample size, *n*, as $\tau_k = n \times \tau_Y$ (since $\text{SE} = \text{SD}/\sqrt{n}$). A common within-sample-unit precision, τ_Y, is assumed in order to accommodate literature values with missing sample sizes or variance estimates. The sample standard error, SE_k, is drawn from a Gamma distribution with parameters informed by the sample size, *n*, and within-site precision, τ_Y:

$$\frac{1}{n \times \text{SE}_k^2} \sim \text{Gamma}\left(\frac{n}{2}, \frac{n}{2\tau_Y}\right) \tag{3}$$

where τ_Y has a diffuse Gamma prior. Unlike the mean and variance parameters, missing values of *n* cannot be estimated and are conservatively set either to 2 (when existence of a variance estimate indicates $n \geq 2$) or to 1 (if no variance estimate is given).

The random and fixed effects and the among-study, among-treatment, and within-unit precisions are used to evaluate the importance of different sources of uncertainty.

The meta-analysis module in PEcAn is fit using JAGS software version 2.2.0 (Plummer 2010) called from within R code that transforms data and specifies the meta-analysis model in JAGS. JAGS uses standard Markov chain Monte Carlo (MCMC) methods (Gelman and Rubin 1992) to approximate the posterior distribution of the terms in the meta-analysis. To overdisperse the *n* MCMC chains, initial values of β_0 are set to the $1/(n+1), \ldots, n/(n+1)$ quantiles of the prior on β_0; for the default $n = 4$ chains, this would be the {0.2, 0.4, 0.6, 0.8} quantiles. Following Gelman and Shirley (2011), PEcAn discards the first half of each chain, thins each chain to 5000 samples, and then combines the chains into a single vector of samples for each term in the meta-analysis model. Trace plots and the Gelman-Rubin (1992) convergence diagnostic are used to assess chain convergence. Density plots (Fig. 4) are used to visually compare the β_0 chain to data and priors. The significance of the greenhouse effect is evaluated by calculating a two-sided probability that $\beta_{\text{gh}} \neq 0$.

When species-level data are unavailable, the posterior distributions are equivalent to the priors.

Each term in the meta-analysis is represented as a vector of MCMC samples from the posterior distribution. Statistical summaries of the parameters can be calculated from these chains, and chains can also be directly sampled for use in ecosystem model parameterization. When the β_0 chains are sampled for the ecosystem model ensemble, the meta-analysis posteriors become the model ensemble priors.

Model analysis

Ensemble analysis.—Typically, ecosystem models are run for a single model parameterization. For example,

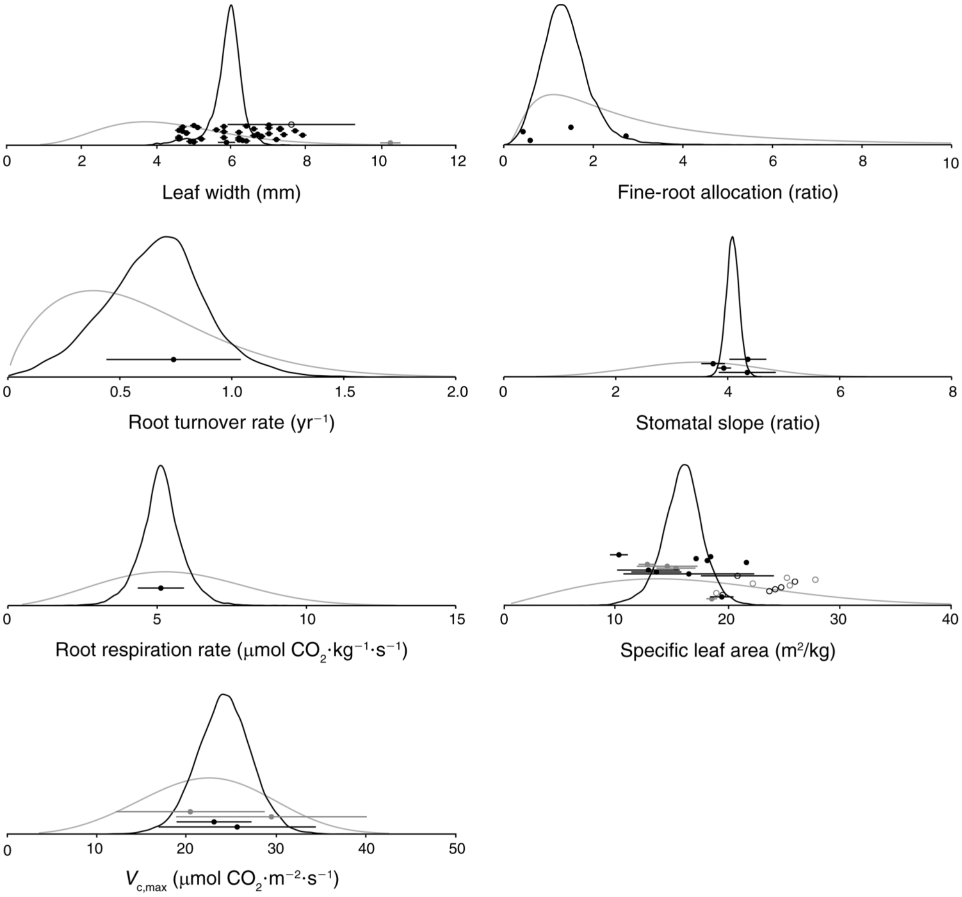

FIG. 4. Curves show prior (gray) and posterior (black) densities of trait parameters used in the analysis. Prior distributions are based on the traits of plants within broad taxonomic or functional type groupings (e.g., all grasses; Fig. 2, Table 1); the *y*-axis is the probability density function (area under the curve = 1). When species-level data were available, they were used in a hierarchical Bayesian meta-analysis, and the posterior estimate of the mean parameter value is shown. Data (mean ± SE) used in the meta-analysis come from both published and our own measurements of the trait on the perennial C_4 grass switchgrass (*Panicum virgatum*). The plots illustrate mismatch between data and the posterior estimate of the global trait mean results from site, treatment, and greenhouse effects. Data from plants grown under an experimental treatment are presented in gray; data from plants grown in a controlled environment (e.g., in a pot or greenhouse) are presented as open circles; data from field-grown plants under control treatments are in black. In the specific-leaf-area and leaf-width plots, site-level effects account for disparity between the black circles and the meta-analysis posterior (black density line).

the model could be evaluated at the median value of each parameter. However, this approach only provides a point estimate, with no accounting for parameter uncertainty.

To propagate parameter uncertainty through the ecosystem model, PEcAn uses standard ensemble-based Monte Carlo approaches. An ensemble of model runs is a set (e.g., 500 or 1000) of model runs that are parameterized by sampling from the trait parameter distributions. For each ensemble member, parameter sets are sampled from the full joint parameter distribution of $\boldsymbol{\beta}_0$, the vector of all model parameters. As a result, the model ensemble approximates the posterior distribution of the ecosystem model output. The model ensemble produces a posterior distribution of the ecosystem model output that can be summarized with standard statistics (e.g., mean, standard error, and credible interval).

Sensitivity analysis.—Sensitivity analyses are used to understand how much a change in a model parameter affects model output; sensitivity is the derivative, $df/d\beta_{0_t}$, of the model (f) with respect to the estimate of β_0; here we use the subscript t for trait and vector $\boldsymbol{\beta}$ as the vector of parameters to distinguish univariate from multivariate calculations. PEcAn approximates the sensitivities based on univariate perturbations of model parameters. These approximations are necessary because analytical solutions for sensitivity are not tractable for most ecosystem models, and PEcAn is designed to be flexible and applicable to any such model. One disadvantage of traditional perturbation-based sensitivity analyses is that the perturbations are usually arbitrary, for example varying each parameter by a fixed percentage of its value (Larocque et al. 2008) rather than over a meaningful range of the parameter. These traditional approaches make interpretation of sensitivities difficult because they fail to acknowledge the distribution or uncertainty of each parameter. In this regard, PEcAn offers a distinct advantage over traditional sensitivity analyses because parameters are varied based on the meta-analysis posterior parameter distributions.

Based on initial exploratory analyses, we found a local perturbation to be inadequate for capturing the responses in most parameters, so we instead estimate sensitivities using a global univariate sensitivity analysis. By default, PEcAn evaluates each parameter at the posterior median and at the six posterior quantiles equivalent to ±[1, 2, 3]σ in the standard normal, while holding all other parameters constant at their posterior median. The relationship between model output and each model parameter β_{0_t} is then approximated by a natural cubic spline $g_t(\beta_{0_t})$ that interpolates through the evaluation points. The model sensitivity to each parameter is approximated by the derivative of the spline ($dg_t/d\beta_{0_t}$) at the parameter median. In addition to the sensitivity analysis, this set of spline approximations is used in the variance decomposition, in partitioning residual variance, and in evaluating the effect of ensemble size on the estimate of model variance.

To facilitate comparisons among the trait sensitivities despite differences in the units on different traits, we tabulate the coefficient of variation (normalized parameter variance) and the elasticity (sensitivity with terms df and $d\beta_{0_t}$ standardized by the mean model output and parameter mean, respectively).

Variance decomposition.—Variance decomposition aims to explain how much each input parameter contributes to uncertainty in model output (Cariboni et al. 2007). Although the present analysis focuses on model parameters, these methods can be extended to address uncertainty in initial conditions or model drivers.

The Delta Method uses Taylor series expansion to approximate the probability distribution of a continuous function of random variables (Oehlert 1992, Casella and Berger 2001:240–245). In this study, the model output $f(\boldsymbol{\beta}_0)$ is a function of a vector of the full set of parameters. After approximating the distribution of $f(\boldsymbol{\beta}_0)$, it is possible to estimate the variance of the model output. The first step is to derive the Taylor series approximation of the variance of a function (Casella and Berger 2001: Eq. 5.5.9):

$$\mathrm{Var}[f(\boldsymbol{\beta}_0)] \approx \sum_{t=1}^{m} \mathrm{Var}\left[f(\overline{\beta_{0_t}}) + \frac{df}{d\beta_{0_t}}(\beta_{0_t} - \overline{\beta_{0_t}}) + \dots\right] \tag{4}$$

$$= \sum_{t=1}^{m} \left(\frac{df}{d\beta_{0_t}}\right)^2 \mathrm{Var}[\beta_{0_t}] + \omega \tag{5}$$

where m is the number of parameters in the model, the error term ω accounts for higher order terms in the Taylor series, and β_{0_t} is the estimate of β_0 from the meta-analysis (Eq. 2) for each trait, t.

With this approximation, it is straightforward to estimate the variance contributed by each parameter. The terms in this form of the variance decomposition can be estimated directly from the preceding analyses: $\mathrm{Var}[f(\boldsymbol{\beta}_0)]$ is the variance of the model ensemble; $\mathrm{Var}[\beta_{0_t}]$ is the posterior variance of trait β_{0_t} from the meta-analysis (Eq. 2); and $df/d\beta_{0_t}$ is the model sensitivity at the parameter mean $\overline{\beta_{0_t}}$. The resulting assertion is that the variance of model output is equal to the sum over the variance of each trait times its sensitivity squared plus a closure term, ω.

We found that the traditional Taylor polynomial approach to variance decomposition produced a poor closure of the total variance of the model output: for more sensitive parameters, a linear approximation of $f(\boldsymbol{\beta}_0)$ provided unrealistic estimates of the sensitivity function that overestimated variance. Increasing the order of the Taylor series expansion actually exacerbated this problem (results not shown). One problem with the polynomial approximation is that, unlike polynomials, most response variables in ecosystems and ecosystem models tend to be asymptotic at both high and low values of a trait. For example, when assessing aboveground biomass, there is a lower bound of zero biomass and most parameters become progressively less sensitive, if not genuinely asymptotic, at their upper bound. This asymptotic behavior is poorly approximated by a polynomial because polynomials are unbounded at extreme parameter values. Therefore, we sought a better approximation for the variance decomposition.

First, we formulated a more generalized form of the variance decomposition (Eq. 4):

$$\mathrm{Var}[f(\boldsymbol{\beta}_0)] = \sum_{t=1}^{m} \mathrm{Var}[g_t(\beta_{0_t})] + \omega. \tag{6}$$

The spline $g_t(\beta_{0_t})$ is a statistical emulator of the model response to trait t that transforms β_{0_t} from the parameter domain to the model domain. The univariate contribution of each parameter to variance of the model output is thus $\mathrm{Var}[g_t(\beta_{0_t})]$.

Eq. 6 only requires β_{0_t} from the preceding meta-analysis, $g_t(\beta_{0_t})$ from the sensitivity analysis, and $\mathrm{Var}[f(\boldsymbol{\beta}_0)]$ from the ensemble analysis.

The final term, ω, is the closure between the right-hand side and the left-hand side of the variance decomposition; ω represents the effects of the higher order terms in the Taylor approximation and the covariance terms between parameters. This closure term is intended to represent parameter interactions that are excluded from the univariate variance decomposition (Eq. 6). Negative trade-offs among physiological traits would result in ω less than zero. However, our estimate of ω also includes errors associated with using finite sample sizes, the spline approximation in each $g_t(\beta_{0_t})$, and biological range restrictions on model output that are not reflected in the variance decomposition (Eq. 6).

One approach to partition the error in the closure term is to use the univariate spline functions from the sensitivity analysis to estimate what the model output would be for each of the parameter sets used in the model ensemble; we call this estimate the "spline ensemble":

$$\mathbf{g}_\ell(\boldsymbol{\beta}_0) = \mathbf{g}(\hat{\boldsymbol{\beta}}_0) + \sum_{t=1}^{m}[g_t(\beta_{0_{t\ell}}) - g_t(\hat{\beta}_{0_t})]. \tag{7}$$

In this equation, $\mathbf{g}_\ell(\boldsymbol{\beta}_0)$ is the spline estimate of the model output for the ℓth ensemble member, where $\mathbf{g}$ is a vector of functions and $\hat{\beta}_{0_t}$ is the posterior median parameter value of trait t.

Although the individual splines may respect range restrictions on output variables (e.g., biomass values cannot fall below zero), combinations of the splines evaluated for a set of unfavorable traits can fall outside these ranges. For parameter sets that give a biologically implausible estimate of negative biomass ($\mathbf{g}_\ell(\boldsymbol{\beta}_0) < 0$), the estimate is set to zero. The only difference between the variance of the spline ensemble (Eq. 7) and the variance decomposition (Eq. 6) is that range restrictions are not corrected for in the variance decomposition. Therefore, the spline ensemble allows us to estimate the effect of using combinations of spline estimates that do not respect the zero bound on biomass in the variance decomposition. The difference between the model ensemble and the spline ensemble provides an estimate of parameter interactions in the model because the spline ensemble does not include the parameter interactions that exist in the model.

The precision of the estimate of model ensemble variance is affected by the number of runs in the ensemble. When the computational expense of the model itself limits the ensemble size, there can be significant uncertainty in the estimate of ensemble variance.

The uncertainty in a sample variance is estimated as

$$\mathrm{Var}(s^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right) \tag{8}$$

(Mood et al. 1974:239), where μ_4 is the fourth central moment. $\mathrm{Var}(s^2)$ scales inversely with sample size. The effect of the limited model ensemble size on uncertainty in the estimate of ensemble variance is measured in two ways. The first way is to calculate $\mathrm{Var}(s^2)$ for the model ensemble ($n = 500$). The second way is to compare $\mathrm{Var}(s^2)$ of the spline ensemble with 500 and 10 000 runs. The 95% credible interval for s^2 is calculated as

$$s^2 \pm 1.96 s_{s^2}, \text{where } s_{s^2} = \sqrt{\mathrm{Var}(s^2)}.$$

The error introduced from using a spline approximation of the model response cannot be estimated based on the existing output, but it is small in comparison to the other effects, given the range restrictions imposed by the spline interpolation.

The results of a model ensemble are posterior estimates of aboveground biomass.

However, we also distinguish between ensembles depending on the nature of model parameters. First, we ran a "prior model ensemble" using an ensemble of parameter sets drawn from prior distributions, and then a "posterior model ensemble" drawn from meta-analysis posteriors.

APPLICATION: SWITCHGRASS MONOCULTURE

We demonstrate the application of PEcAn to estimate the aboveground yield of an experimental switchgrass (*Panicum virgatum*) monoculture. The first step to applying PEcAn was to construct an appropriate set of priors based on data syntheses and expert knowledge. These priors were conservative estimates of the plant trait parameters based on information other than species-level data. Next, switchgrass trait data from both previous studies and field measurements were summarized using meta-analysis to constrain the prior parameter estimates. The Ecosystem Demography model version 2.1 (Moorcroft et al. 2001, Medvigy et al. 2009) was used to simulate plant growth.

The model ensemble and sensitivity analysis were performed using both the prior and posterior parameter estimates. By comparing the prior model ensemble to the posterior model ensemble, we are able to evaluate the ability of species-level data to reduce model uncertainty.

To evaluate model performance, we compare the ensemble estimates of aboveground biomass with observed yields (Heaton et al. 2008, Wang et al. 2010) in Fig. 5.

Site

Switchgrass (*Panicum virgatum*) is a perennial grass native to North America that has received attention as a potential cellulosic biofuel crop (McLaughlin and Kszos 2005, Wang et al. 2010). We modeled the aboveground biomass production of a switchgrass monoculture and compared model estimates to a monoculture planted in 2002 at the University of Illinois Agricultural Research and Education Center in Urbana, Illinois, USA (40.09 N, 88.2 W; see Plate 1). The climate at this site is characterized by hot, humid summers and cold winters, with a 50-year (1959–2009) mean annual temperature of 11°C and mean annual precipitation of 1000 mm/yr (Angel 2010). Meteorological data used to drive the model were downloaded from the North American Regional Reanalysis (Mesinger et al. 2006). Soil is a silt loam from the Drummer-Flanagan soil series; texture data were obtained through the USDA NRCS web soil survey website (*available online*).[8] The yield and other

[8] websoilsurvey.nrcs.usda.gov

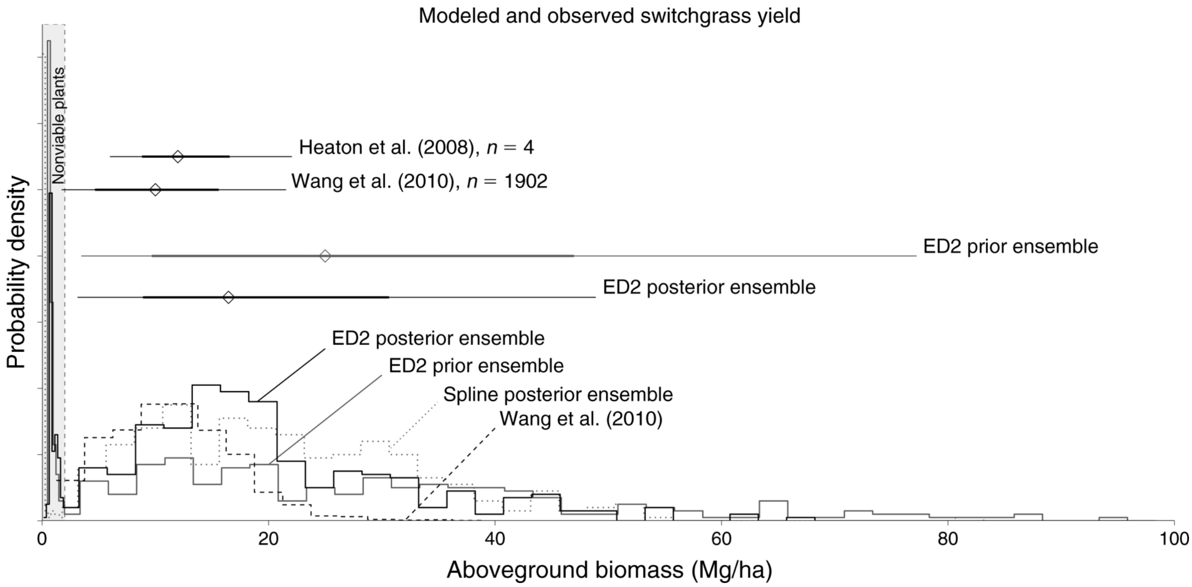

FIG. 5. Ensemble average 2004–2006 post-senescence yield of switchgrass. The histogram shows results from prior ensemble runs (solid gray line), posterior ensemble runs (solid black), the spline posterior ensemble (dotted gray), and observations (dashed black line) from Wang et al. (2010); the *y*-axis is the probability density (area under the curve $=$ 1). The vertical gray box on the left represents nonviable ensemble members (yield $\leq$ 2Mg/ha; see the *Ensemble* section in both *Results* and *Discussion*). Horizontal bars provide a summary of yields: a three-year trial (4 observations) at the modeled site (Heaton et al. 2008), all 1902 observations included in a recent meta-analysis (Wang et al. 2010), and viable runs from the ED2 ensemble based on prior and posterior parameterizations. Open diamonds indicate the median; thick and thin lines indicate the 68% and the 95% credible intervals, respectively. Histogram-style plots provide comparison of the distributions of observations and model runs. For clarity, nonviable and viable runs are plotted with different width bins.

aspects of this ecosystem have previously been reported (Heaton et al. 2008).

Ecosystem Demography model

We used the Ecosystem Demography model, version 2 (hereafter ED2) to model the productivity and soil carbon pools in this switchgrass agroecosystem. ED2 is a terrestrial biosphere model that couples age- and stage-structured plant community dynamics with ecophysiological and biogeochemical models. The biophysical land surface model in ED2 allows plant uptake and growth to respond dynamically to changes in weather and soil hydrology (Medvigy et al. 2009). ED2 has the ability to link short-term, physiological responses to environmental conditions with realistic, long-term successional changes in ecosystem structure and composition (Moorcroft et al. 2001). Although other models have both succession and physiology, ED2 also has explicit spatial scaling, a sub-daily time step, and the ability to couple with a land surface model (Dietze and Latimer 2011).

ED2 incorporates a mechanistic description of plant growth that accounts for the fast temporal responses of plants to changes in environmental conditions. In this study, we vary 15 model parameters based on observable plant traits that control carbon uptake, carbon allocation, turnover, and reproduction (Table 1, Figs. 2 and 4).

ED2 calculates photosynthetic rates using the enzyme kinetic model developed for C_3 plants (Farquhar and Sharkey 1982, Ball et al. 1987) and the modifications for C_4 plants (Collatz et al. 1992). $V_{\mathrm{c,max}}$ sets the upper bound on the rate of Rubisco-limited photosynthesis in C_3 plants and PEP-limited photosynthesis in C_4 plants, whereas light-limited photosynthesis is constrained by the quantum efficiency parameter, and a threshold parameter controls the minimum temperature at which photosynthesis will occur. Stomatal conductance is calculated using the Leuning variant of the Ball-Berry model (Leuning 1995) and is controlled by the stomatal slope parameter. Leaf boundary layer conductance depends on the leaf width parameter.

Together, stomatal conductance and leaf boundary layer conductance affect carbon and moisture fluxes and the leaf energy balance. Specific leaf area (SLA) determines the amount of leaf area produced per unit leaf biomass investment.

In addition to photosynthesis, ED2 also accounts for carbon allocation to growth, respiration, and for the turnover rate of carbon pools. These parameters include: one to partition between leaf and fine-root growth; one for allocation to reproduction; two respiration parameters associated with growth respiration and root maintenance respiration; and two parameters to control the rates of leaf and root turnover.

Finally, three demographic parameters control seed dispersal, seedling mortality, and adult mortality due to carbon limitation (Table 1).

Priors

Priors from data.—Priors were estimated by finding the best-fit distribution to raw data sets, including SLA ($n = 125$) and leaf turnover rate ($n = 40$) from the GLOPNET database (Wright et al. 2004); root turnover rate ($n = 66$; Gill and Jackson 2000); and quantum yield ($n = 56$; Skillman 2008). Candidate distributions for these priors were Gamma, Weibull, log-normal, and F because each of these traits is bound at zero. In all cases, we are interested in using the full distribution of across-species data as our prior constraint on what one individual species is capable of doing, as opposed to using the estimate of the mean of this distribution as our prior.

Quantum yield data represent a survey of published values of quantum yield in C_4 monocots (Skillman 2008); original data were provided by the author and restricted to measurements made under photorespiratory conditions (ambient CO_2 and O_2) (J. Skillman, *personal communication*). Given the narrow range of data (CV = 11%), the normal distribution was also considered but was not the best fit.

Priors from meta-analysis.—We used meta-analysis to calculate a prior from data when summary statistics and sample sizes were available. The meta-analysis model used to calculate prior distributions is similar to the one used by PEcAn to summarize species-level data (Eq. 2), with three differences. First, there were no site, treatment, or greenhouse effects. Second, data from multiple species were used. Third, we generated a posterior predictive distribution to predict the distribution of trait values for an unobserved C_4 plant species, unlike the species-level meta-analysis, which estimated the global mean parameter value. Thus, the model included plant functional type (PFT) as a random effect:

$$\Theta_{\text{species}} = \beta_0 + \beta_{\text{PFT}}. \tag{9}$$

Stomatal slope is the empirical slope coefficient in the Leuning (1995) model of stomatal conductance. Estimates of this parameter are sparse, so we collected new data for this study. Stomatal slope was estimated using measurements of four leaves from each of five field-grown energy crop species during the 2010 growing season (Supplement 1). The five species included two C_4 grasses: *Miscanthus* (*Miscanthus* × *giganteus*) and switchgrass (*Panicum virgatum*) planted in 2008, and three deciduous tree species: red maple (*Acer rubrum*), eastern cottonwood (*Populus deltoides*), and Sherburne willow (*Salix* × *Sherburne*) planted in 2010 as 2-year-old saplings. All plants were grown at the Energy Biosciences Institute Energy Farm (40°10′ N, 88°03′ W). We used the data from the three tree species and *Miscanthus* to calculate the posterior predictive distribution of an unobserved C_4 grass species, and used this distribution as the prior estimate for switchgrass stomatal slope.

Maximal carboxylation rate ($V_{c,\max}$) data consist of 94 C_3 species (Wullschleger 1993) plus three C_4 species (Kubien and Sage 2004, Massad et al. 2007, Wang et al. 2011). To express $V_{c,\max}$ at a common temperature of 25°C for all species, we applied an Arrhenius temperature correction (Appendix). The Wullschleger (1993) data set included a 95% credible interval (CI) and an asymptotic error calculated by the SAS nlin procedure (S. Wullschleger, *personal communication*). We used the asymptotic error as an estimate of SE, the 95% CI to estimate SD ($\text{SD} = (0.5 \times \text{CI})/1.96$), and then estimated n as $\hat{n} = (\text{SE}/\text{SD})^2$. Plant species were classified into five functional types (C_3 grass, C_4 grass, forb, woody non-gymnosperm, and gymnosperm) based on species records in the USDA PLANTS Database (USDA NRCS 2011). Ambiguous species (those with both forb and woody growth forms, $n = 15$) were excluded.

Leaf width data represent 18 grass species grown in a common garden greenhouse experiment (Oyarzabal et al. 2008). *Panicum virgatum* was among the 18 species, and was excluded from the prior estimation but used as raw data in the meta-analysis. The remaining 17 species were divided into C_3 and C_4 functional types. Relative to the small sample of C_4 grasses, switchgrass leaf width was an outlier; inflating the variance fourfold reduced the prior information content to account for this discrepancy.

Root respiration rate values were measured on 36 plants representing five functional types, including six C_4 grass species (Tjoelker et al. 2005). As before, *P. virgatum* data were excluded from the prior estimation and used as raw data in the species-level meta-analysis.

Priors from limited data and expert knowledge.—When available data were limited to a few observations, these were used to identify a central tendency such as the mean, median, or mode, while expert knowledge was used to estimate the range of a confidence interval. An optimization approach was used to fit a probability distribution to this combination of data and expert constraint.

In order to estimate the fine-root to leaf ratio for grasses, we assume that fine roots account for all belowground biomass (Jackson et al. 1997) and that leaves account for all above ground biomass. Roots account for approximately two-thirds of total biomass across temperate grassland biomes (Saugier et al. 2001: Table 23.1), so we constrained a beta prior on the root fraction to have a mean of 2/3 by setting $\alpha = \beta/2$ because the mean of a beta distribution is defined as $\alpha/(\alpha + \beta)$. To convert from proportion to ratio, we used the following identity:

$$\text{if} \quad X \sim \text{Beta}\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$$

$$\text{then} \quad \frac{X}{1-X} \sim F(\alpha, \beta) \times \frac{\alpha}{\beta}.$$

We constrained the 95% CI = [1/3, 10/11], equivalent to a fine-root to leaf ratio with a mean fixed at 2 and a 95% CI = [1/2, 10]. We simulated the distribution of the fine-root to leaf ratio by drawing 10 000 samples from a $F(2\alpha, \alpha)$ distribution and multiplying these samples by 2.

Seed dispersal in ED2 represents the proportion of seed dispersed outside of a 7.5 m radius plot, which we approximate as a beta distribution. Although no direct measurements of seed dispersal were available, it was straightforward to parameterize a ballistic model of seed dispersal (Ernst et al. 1992): $D = V_w H/V_t$. This model relates dispersal distance D to terminal velocity V_t, wind speed V_w, and seed height H. Although more sophisticated treatments of dispersal exist and are important for estimating low-probability long-distance dispersal events (Clark et al. 1999, Thompson and Katul 2008), the Ernst et al. (1992) model is sufficient for the relatively short dispersal distances required in the present context.

Values of terminal velocity, V_t, of grass seeds were taken from two references (Ernst et al. 1992, Jongejans and Schippers 1999) and these data were best described as $V_t \sim$ Gamma(2.93, 1.61). Next the heights from which the seeds are dropped were estimated from calibrated photographs of reproductively mature switchgrass at a field site in Urbana, Illinois, USA: $H \sim \mathcal{N}(2, 0.5)$. Finally, wind speed data observed at the site were fit to a Weibull distribution (Justus et al. 1978). $V_w \sim$ Weibull (2.4, 0.712) (M. Zeri, *unpublished data*). Given these distributions of V_w, H, and V_t, sets of 100 dispersal distances were simulated 10 000 times to calculate the fraction of seeds in each simulation dispersed beyond 7.5 m.

Priors informed by expert knowledge.—When no data were available, expert knowledge was used to estimate the central tendency and confidence interval for a trait parameter. Again, optimization was used to fit a probability distribution to these constraints.

The minimum temperature of photosynthesis for C_4 grasses was given a prior based on expert knowledge with a mean of 10°C and a 95% CI = [8, 12]°C that fits a normal ($\mu = 10$, $\sigma = 1.02$) distribution (D. Ort, *personal communication*).

The growth respiration factor is the proportion of daily carbon gain lost to growth respiration. Because it is a proportion, the beta distribution was fit with a mean set equal to the ED2 default parameter value, 0.33, and a 95% CI = [0.05, 0.60] was conservatively based on the range of construction costs reviewed by Amthor (2000).

The seedling mortality factor represents the proportion of carbon allocated to reproduction that goes directly to the litter pool. Given that the default ED2 parameter is 0.95, we stated a beta prior with a median at 0.95 and a 95% CI = [2/3, 1].

The mortality factor in ED2 is the rate parameter in the negative exponential relationship between carbon balance and mortality (Medvigy et al. 2009). The default parameter for all plant functional types (PFTs) in ED2 is 20, and our weakly informed Gamma prior sets this as the median and gives a 95% CI = [5, 80].

Reproductive allocation represents the proportion of carbon in the storage pool allocated to reproduction. This parameter is a proportion and has a default value of 0.33 in ED. The Beta(2, 4) distribution has a mean of 1/3 and a 95% CI = [0.05, 0.72] representing relatively high uncertainty. This distribution implies that the plant may allocate any fraction of the carbon pool to reproduction between, but not equal to, 0 and 1, and has an 80% probability that less than half of the carbon pool will be allocated to reproduction.

Switchgrass trait meta-analysis

Switchgrass trait data used to constrain model parameters are stored in the Biofuel Ecophysiological Trait and Yield database, a database designed to support research on biofuel crops (BETYdb, *available online*; see Supplement 2).[9] BETYdb includes both previously published and primary data. Prior to entry in the database, data were converted to standard units chosen for each variable (Table 1).

Trait data available for *Panicum virgatum* include $V_{c,max}$, SLA, leaf width, fine-root to leaf ratio, root respiration, stomatal slope, and root turnover rate (Fig. 4, Table 2). Methods used to collect these data and site descriptions are available in the source references (Supplement 1). Each meta-analysis was run with four 50 000-step MCMC chains.

Model analysis

We executed a 10-year, 500-run ensemble of ED2 using parameter values drawn from the prior or posterior parameter distributions. The model was run for the years 1995–2006 to simulate the field trials conducted by Heaton et al. (2008). The model output of interest was the December mean aboveground biomass (AGB) during the years 2004–2006, simulating the yields of the mature stand (Heaton et al. 2008). The ensemble estimate was also compared to the larger set of all reported switchgrass yield data (Wang et al. 2010).

Runs resulting in yields <2 Mg/ha were considered nonviable parameter combinations. To test if prior and posterior parameter sets resulted in different fractions of nonviable runs, we estimated the posterior probability of a nonviable run as a binomial posterior from a beta-binomial model with a flat (Beta(1, 1)) prior. Then we calculated the two-tailed probability that the difference between these binomial posteriors $\neq 0$.

Results

Trait meta-analysis

Switchgrass data were collected from the literature and field for seven of the model parameters: specific leaf area, SLA ($n = 24$ independent estimates), leaf width ($n = 39$), $V_{c,max}$ ($n = 4$), fine-root to leaf allocation ratio ($n = 4$), stomatal slope ($n = 4$), root respiration rate ($n = 1$), and root turnover rate ($n = 1$). Here, n corresponds to the k site × treatment combinations in the meta-analysis. Table 2 summarizes the meta-analysis for each of these

[9] www.betydb.org

TABLE 2. Results of meta-analysis of switchgrass data for six physiological traits.

Parameter (units)	k	β_0	σ_Y	σ_{site}	$\sigma_{\text{tr}\mid\text{site}}$	β_{gh}
Specific leaf area (m^2/kg)	24	16 (12, 20)	2.8 (2.5, 3.2)	3.2 (1.6, 7.3)	2.4 (1.1, 6)	6.5 (1, 12)
Leaf width (mm)	39	6 (4.7, 6.6)	0.46 (0.44, 0.48)	0.47 (0.2, 2.1)	6.4 (1.9, 130)	1.6 (−0.033, 3.5)
$V_{\text{c,max}}$ (μmol $CO_2 \cdot m^{-2} \cdot s^{-1}$)	4	24 (18, 30)	12 (8.1, 17)		1.2 (0.098, 47)	
Fine-root allocation (ratio)	4	1.3 (0.5, 2.6)	2.2 (1.2, 6.2)			
Root respiration rate (μmol $CO_2 \cdot m^{-2} \cdot s^{-1}$)	1	5.1 (3.7, 6.6)	1.2 (0.39, 2.3)			
Root turnover rate (yr^{-1})	1	0.67 (0.2, 1.1)	0.45 (0.14, 0.88)			
Stomatal slope (ratio)	4	4.1 (3.9, 4.3)	0.33 (0.23, 0.45)			

Notes: The number of sample units (number of site × treatment combinations) in the meta-analysis is given by k. The global mean parameter, β_0, is used to parameterize the Ecosystem Demography model and is described in more detail by Fig. 4; σ_Y, σ_{site}, and $\sigma_{\text{tr}\mid\text{site}}$ represent among-site, within-site, and treatment–site variability in random effects, respectively; subscript "gh" refers to greenhouse, a fixed effect. The variance components are transformed from precision to the standard deviation scale for ease of interpretation. Values are reported as the parameter median with the 95% credible interval in parentheses.

parameters, including the posterior mean and 95% CI of the global mean, the fixed greenhouse effect, and each of the variance components (reported as standard deviations).

SLA and leaf width data were from from multiple sites, but the meta-analysis provided no evidence for among-site variability in excess of within-site variability (σ_Y and σ_{site}, respectively; Table 2). For the remaining traits, there was insufficient spatial sampling to assess site-to-site variability. Greenhouse growing conditions had a positive effect on both SLA ($P = 0.027$), and leaf width ($P = 0.052$).

Fig. 4 compares parameters before and after incorporating data in the meta-analysis. A reduction in parameter uncertainty is seen as the reduction in the spread of the posterior (black) compared to the prior (gray) parameter distributions. The influence of the prior information on the posterior distribution increased when the prior was more constrained or when fewer data were available for use in the meta-analysis. For example, data substantially constrained the uncertainty in the $V_{\text{c,max}}$ and SLA posteriors relative to the priors. By contrast, there was little effect of additional data on the parameter estimates for fine-root to leaf allocation and root respiration rate; these parameters had relatively well constrained priors and limited species-specific data.

Model analysis

Ensemble.—Within the model ensemble analysis (Fig. 5), both the prior and posterior parameterizations produced yield estimates that were consistent with yields observed at the Urbana site for which the model was run (Heaton et al. 2008) and with 1902 previously reported yields of switchgrass (Wang et al. 2010). In both the prior and posterior ensembles, the predicted aboveground biomass was clearly bimodal. These two modes had little overlap and a distinct break at 2 Mg/ha. We inferred that the first peak represents nonviable plants generated by unrealistic parameter sets; thus plants with aboveground biomass <2 Mg/ha were considered "nonviable." When summarizing the model output, we consider viable and nonviable ensemble members separately; all runs are considered in the sensitivity analysis and variance decomposition. A greater percentage of runs in the prior ensemble fell below this threshold (53.4 vs. 36.6; $P \simeq 0$).

Compared to the prior ensemble prediction, the data-constrained posterior runs had lower median yields and a more constrained 95% credible interval (16.5 [7.2, 37] Mg/ha vs. 25 [7.7, 56] Mg/ha). This reflects the substantial shrinkage of the posterior relative to the prior SD estimates of model output uncertainty (from $\sigma = 19.7$ to $\sigma = 11.9$). In particular, the upper tail of the modeled yield was reduced toward the observed yields.

Despite the reduction in ensemble uncertainty, the ensemble posterior yield was still relatively imprecise and had much greater uncertainty than the field trial ($\sigma = 4.1$; Heaton et al. 2008) or the meta-analysis of all observations ($\sigma = 5.4$; Wang et al. 2010). The spline ensemble viable plants had a median 18.8 [2.9, 48] Mg/ha and $\sigma = 13$.

Sensitivity analysis.—Sensitivity analysis demonstrated that traits varied in their effect on aboveground biomass (Fig. 6). For example, parameters associated with photosynthesis and carbon allocation (including $V_{\text{c,max}}$, SLA, growth respiration, and root allocation) were particularly sensitive. For particularly sensitive parameters, the sensitivity functions had coverage of unrealistic yields >30 Mg/ha. Constraining SLA and $V_{\text{c,max}}$ parameters with data resulted in a more realistic range of yields. On the other hand, aboveground biomass was largely insensitive to leaf width, seed dispersal, and mortality rate.

Variance decomposition.—The variance decomposition showed that data-constrained parameters substantially reduced their contribution to overall model variance (Fig. 7). Prior to including species-specific trait data, SLA contributed the most to model uncertainty, followed by growth respiration, fine-root allocation, $V_{\text{c,max}}$, seedling mortality, and stomatal slope (Fig. 7c, gray bars). Incorporating species-level data substantially reduced the contributions of SLA, $V_{\text{c,max}}$, seedling mortality, and stomatal slope to model uncertainty.

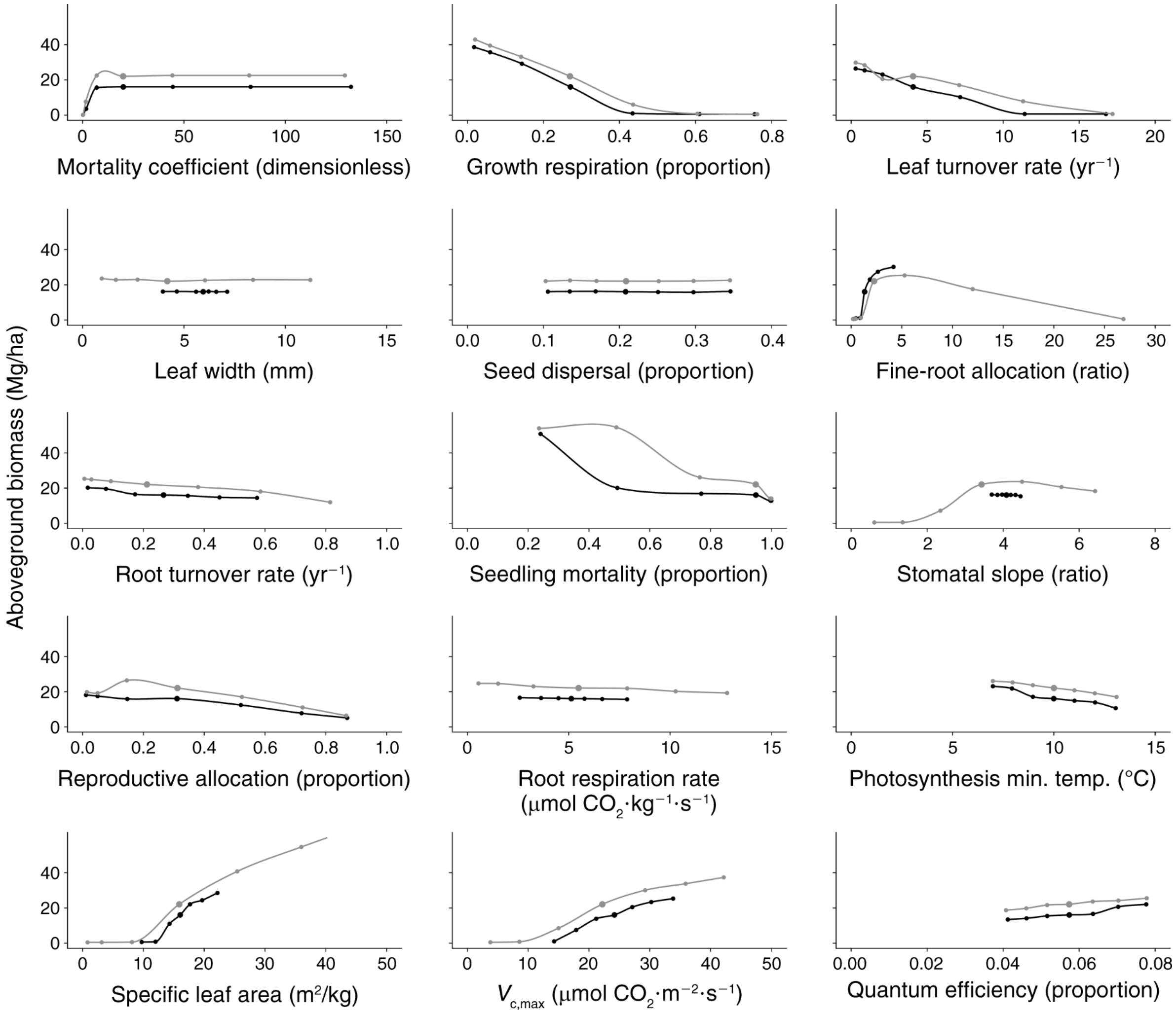

FIG. 6. Sensitivity of aboveground biomass to 15 plant traits: univariate relationships between parameters and 2004–2006 average modeled yield. Parameter values are on the *x*-axis, and biomass is on the *y*-axis; runs centered around the prior median are in gray, and those centered around the posterior median are in black. The univariate responses were estimated using a cubic spline to fit model output at the median and ±[1, 2, 3]σ quantiles of each parameter while holding other parameters to the median value.

For example, SLA fell from first to fourth and stomatal slope fell from sixth to 14th in rank contribution to ensemble variance. Although the addition of data reduced parameter uncertainty for fine-root to leaf allocation, aboveground biomass was more sensitive to this parameter at the posterior median. These changes cancelled each other out, and as a result the contribution of the fine-root allocation parameter to ensemble variance remained constant.

There was no effect of increasing the sample size from 500 to 10 000 on the variance estimates (Table 3). The variance of the ensemble was less than the variance calculated in the variance decomposition, and this difference is the closure term, ω. The closure term accounted for ~28% of the variance decomposition estimate on the standard deviation scale (Table 4).

DISCUSSION

Switchgrass trait meta-analysis

When species-level data were available, the meta-analysis constrained estimates of the trait mean parameter (Fig. 4) and provided insight into the sources of parameter uncertainty (Table 2). In the context of constraining model parameters, we were interested in accounting for, but not directly investigating, the specific effects of site, treatment, or greenhouse effects. However, we can use the meta-analysis results to identify sources and scales of parameter variability. This insight into parameter variability helps to inform future sampling designs, development of the ecosystem model, and improvement of methods used to parameterize the ecosystem model.

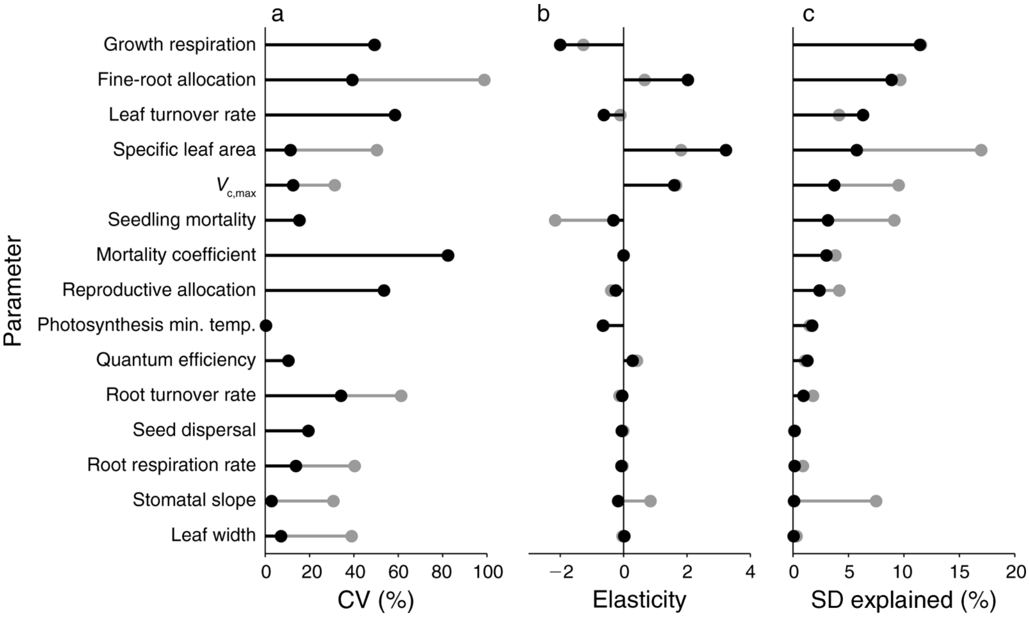

FIG. 7. Partitioning of variance by parameter results from variance decomposition conducted before (gray) and after (black) updating parameter estimates with species-level data in the meta-analysis. (a) Uncertainty associated with each parameter (coefficient of variation, CV = σ/μ). The degree to which some parameters have been constrained by species-level data is indicated by the reduction in CV in the black compared to the gray bars; sample sizes are indicated in Table 2. (b) The sensitivity of modeled aboveground biomass to each parameter presented as elasticity (normalized sensitivity; an elasticity of 1 indicates that model output will double when the parameter value doubles). Sensitivity is the slope of the line at the median in Fig. 6. Parameters with larger bars have greater influence on model output. (c) SD is the contribution of each parameter to model uncertainty. This is a function of both the parameter variance and sensitivity. Parameters with both large CV and elasticity have the highest SD, indicating that they explain the most uncertainty in model output.

Where data from multiple sites were available, we could evaluate the relative importance of within- vs. among-site variance for the range of sites represented in the data (Table 2). Recent studies demonstrate important effects of intraspecific trait variability on ecosystem functioning (Albert et al. 2011, Breza et al. 2012, Violle et al. 2012).

TABLE 3. Comparison of sample variance estimates (s, on standard deviation scale) for the aboveground biomass (all values Mg/ha) estimated from data-constrained parameters calculated from model ensemble, spline-emulated model ensembles, and variance decomposition.

Sample size, n	Model ensemble $s_{f(\boldsymbol{\beta}_0)}$	Spline ensemble $s_{\mathbf{g}(\boldsymbol{\beta}_0)}$	Variance decomposition $\sum s_{g_t(\beta_{0_t})}$
500	13 (14)	13.8 (13)	18.2 (6)
10 000	†	14.1 (2.8)	18.1 (1.2)

Notes: Values in parentheses are estimates of uncertainty in the sample estimate of variance. Sample size, n, refers to the ensemble size of the sample from the posterior parameter distribution. Terms are: $\boldsymbol{\beta}_0$, vector of parameter values; f, ED2 model; $\mathbf{g}$, spline approximation of the ED2 model (Eq. 7); t, the tth of m parameters (Eq. 6).

† Analysis of the closure term is based on estimates with $n =$ 10 000 parameter sets, except in the case of the model ensemble because evaluation of the model ensemble at n = 10 000 is computationally prohibitive.

Therefore, for traits that do exhibit greater variability across than within sites, explicit inclusion of spatial, environmental, and even genetic information into the meta-analytical model would be justified. This approach would enable the estimation of site-specific parameters for use in the ecosystem model and will be investigated in future development of the meta-analysis module.

For the other parameters ($V_{c,max}$, fine-root allocation, root respiration rate, and root turnover rate) data came from one site, so it is not possible to estimate the across-site variability. For these traits, obtaining data from additional sites would better constrain both the global

TABLE 4. Components of the variance closure term, ω (see Eq. 6), the difference between the variance decomposition and model ensemble estimates of σ.

	Calculation	Mg/ha
ω_{total}	$\sum s_{g_t} - s_f$	5.2
$\omega_{\text{covariance}}$	$s_{\mathbf{g}} - s_f$	1.1
$\omega_{\text{truncation}}$	$\sum s_{g_t} - s_{\mathbf{g}}$	4.1

Notes: The closure due to parameter interactions is estimated as the difference between the spline ensemble and the model ensemble; the closure due to the absence of a lower bound of zero on the spline functions is estimated as the difference between the variance decomposition and the spline ensemble estimates.

PLATE 1. Switchgrass stand in winter, Urbana, Illinois, USA. Photo credit: D. S. LeBauer.

mean and the across-site variance. This additional data collection is particularly justified for traits that contribute most to the uncertainty in the model ensemble.

Model ensemble

Despite the large reduction in model uncertainty from the prior to the posterior model ensemble, the uncertainty in projected yield is substantial (Fig. 5) and further constraint would increase the utility of this model output. However, the explicit accounting of parameter uncertainty is an important first step toward producing more informative model output. If model parameters had been treated as fixed constants, we would have no estimate of model uncertainty; without an estimate of uncertainty, the similarity between the modeled (16.5 Mg/ha) and observed (12.0 Mg/ha) median yields would be difficult to interpret; a naive interpretation could create false confidence in the model. Including the nonviable plants would have made the model mean more accurate but less precise (Fig. 5). The 90% CI would have contained the possibility that switchgrass would not grow in Champaign County, Illinois, even though extensive research (Heaton et al. 2008, VanLoocke et al. 2012; D. LeBauer, *personal observation*) demonstrates that it does grow very well in this area.

The reduction in median modeled yield in the posterior relative to the prior model ensemble is consistent with the reduced probability of high SLA and $V_{\mathrm{c,max}}$ values in the posterior relative to the prior distributions. As expected, the use of switchgrass trait data to inform model parameters succeeded in both reducing total uncertainty and bringing modeled yield in line with observations of switchgrass yields, both at this site (Heaton et al. 2008) and globally (Wang et al. 2010). Reducing uncertainty in model outputs (in this case, yield) is key to increasing the value of ecological forecasts (Clark et al. 2001).

Although reducing uncertainty does not necessarily increase model accuracy, an estimate of model uncertainty is a first step toward generating meaningful statistical inference from the model itself. Without an estimate of model uncertainty, it is not possible to make such a basic inference as the probability that the model predictions overlap with observed data; this limits the capacity of models to inform research and applied problems (Clark et al. 2001). Instead, comparisons of ecosystem models with observations have focused on differences and correlations between model output and data (Bellocchi et al. 2010, Schwalm et al. 2010, Dietze et al. 2011) without providing a confidence interval around the model output itself. The ability to identify, with confidence, the conditions under which a model produces valid output helps to determine appropriate applications of the model and to identify targets for further model development (Williams et al. 2009). Although parameter uncertainty is clearly just one of many sources of uncertainty in models (McMahon et al.

2009), and constraining model parameters by no means guarantees that a model will match reality, it is difficult to assess the accuracy of a model if it has low precision. In deterministic models, such as most ecosystem models, parameter uncertainty is a major driver of the precision of a model.

In this study, we can state with 90% confidence that the mean switchgrass yield during the Heaton et al. (2008) study should have been between 7.2 and 37 Mg/ha, and if we had made this prediction in advance, we could have said that we were correct because the mean did fall within this range. We can also see that the model uncertainty contains the 90% CI for observed switchgrass yields globally (Wang et al. 2010), even though we know that important drivers of variability in the global meta-analysis (e.g., climate, soil) are different from the source of uncertainty in our model predictions (e.g., parameters). The model ensemble left open the possibility that the yields could have been much more or much less than those actually observed, and we conclude that much of this variability could be constrained with additional trait-level data or data assimilation. Wang et al (2013) provide an example of combining the PEcAn meta-analysis and variance decomposition with data assimilation of biomass to constrain uncertainty in parameter estimates and improve the accuracy and precision of model output. Once the model can make more precise predictions, it will be possible to begin investigation of other sources of uncertainty, such as model structure and state variables (e.g., climate, soil).

Although the present analysis focuses on modeled aboveground biomass, PEcAn can analyze any model output, including pools and fluxes of water, energy, and carbon.

Indeed, PEcAn's approach to the synthesis of data and mechanistic models is independent of the system being modeled, and thus has potential applications far beyond the scope of its current development to support ecosystem modeling.

Variance decomposition

Variance decomposition quantified the contribution of each parameter to model uncertainty, helping to identify a subset of parameters for further constraint. SLA, $V_{\mathrm{c,max}}$, fine-root to leaf ratio, and leaf turnover rate dominated uncertainty in yield prior to incorporating species-level data. Therefore, SLA, which can be measured quickly and at low cost, would make a good first target for reducing uncertainty when a new species is evaluated. SLA also correlates strongly with other important model parameters, such as photosynthetic rate, leaf life span, and nitrogen content (Wright et al. 2004). The ranking of parameters based on variance contribution depends on the response variable of choice (in this case, aboveground biomass) as well as the conditions of the run (duration, soil, climate), and the species or PFT being evaluated. In general, for a given species and model output, overall patterns of parameter importance are consistent across a broad range of climates (Wang et al. 2013).

Variance decomposition (Eq. 6) demonstrates that it is not parameter uncertainty or model sensitivity alone, but the combination of the two, that determines the importance of a variable. For example, despite the high uncertainty in seed dispersal, switchgrass yield is insensitive to this parameter (Figs. 6 and 7); therefore a better understanding of switchgrass seed dispersal would not reduce model uncertainty. By contrast, although uncertainty in the growth respiration is not particularly large, switchgrass yield was very sensitive to growth respiration, and for this reason growth respiration is the greatest contributor to model uncertainty. In addition, although no seedling mortality data were available, model sensitivity to this parameter was much lower in the posterior compared to prior runs. Using the sensitivity analysis or parameter uncertainties alone would thus lead to incorrect conclusions about what parameters are most important and would be an inefficient approach to reducing predictive uncertainties.

This analysis only represents the first step toward more comprehensive accounting of known sources of uncertainty. The next step in reducing uncertainty would be to use the results of the variance decomposition to target the most influential model parameters for further constraint through data collection. We have demonstrated how the use of species-level data to constrain parameter uncertainty reduced ensemble variance, resulting in a new set of targets for additional field observations and refined literature surveys.

Processes that are difficult to measure, such as belowground carbon cycling, can be indirectly constrained with ecosystem-level observations using data assimilation (Luo et al. 2009, 2011). Integrating data assimilation into PEcAn will allow ecosystem-level observations to further constrain parameters for which trait-level observations are difficult or impossible to obtain. To date most Bayesian data assimilation approaches applied by ecologists have employed flat, uninformative priors (assigning equal probability to values over many orders of magnitude), which has lead to the problems of parameter identifiably and the criticism that model parameters are allowed to take on biologically unrealistic values. The use of the meta-analysis posteriors as priors in the data assimilation step ensures that any parameter estimates are consistent with what is known about different plant traits. In this way, Bayesian methods are, in effect, updating the literature-derived estimates with new data and providing a coherent and rigorous framework for combining multiple different types of data.

In addition, by effectively restricting parameter space based on observed values, the use of informed priors in data assimilation reduces problems of equi-finality and identifiability. This is consistent with the argument by Beven and Freer (2001) that only the feasible parameter range should be sampled.

To a first order, the spline interpolations of the univariate relationships between parameters and aboveground biomass (Fig. 6) provide a good estimate of the total model variance. The closure term (Table 4) accounted for ~5.2 Mg/ha or 28% of the variance decomposition estimate (18.1 Mg/ha; Table 3), suggesting that although parameter interactions are important, univariate parameter uncertainty drives overall model variance. One key concern of parameter interactions is that the combination of the variance decomposition terms would result in the prediction of negative yields, which is clearly biologically impossible. By comparing the spline ensemble, where this term is truncated, to the spline-based variance decomposition, we can conclude that this truncation effect accounts for 4.1 Mg/ha or 80% of the closure term in the variance decomposition.

By contrast, evaluating the spline ensemble for different ensemble sizes shows that ensemble size had negligible effect on the mean variance estimate, although it does improve the precision of this estimate (Table 3). Finally, the difference between the model and spline ensembles (Table 4) suggests that the absence of parameter interactions in the variance decomposition account for the remaining 20% of the closure term (<6% of the variance decomposition calculation), which could be improved by a multivariate meta-analysis and sensitivity analysis, both of which are planned for future development of PEcAn. Overall, the closure term is relatively well constrained, even when the parameter interactions are assumed to be linear.

Model–fieldwork feedback

Variance decomposition can be used to inform data collection by identifying candidate parameters for further refinement based on their contributions to model variance. Recall that this variance contribution is a function of parameter sensitivity and the parameters' probability density (Eq. 6; Fig. 7). Sensitivity is a function of the model and so there is no direct way to reduce sensitivity. However, because $\mathrm{Var}(f) \propto \mathrm{Var}(\beta_0)$, it is possible to reduce the model uncertainty by reducing parameter variances.

Through simple power analyses, one can explicitly estimate the relationship between an increase in sample size and the reduction in posterior variance. Not only can we calculate the reduction in parameter uncertainty that would be expected for a given sample size, we can also use Eq. 6 to estimate the resulting decrease in the variance of the model output. This then allows us to directly compare the value of different data types in a common currency.

Quantitatively evaluating the relationship between data and model uncertainty provides a path of communication between field research and modeling, opening the door for a new framework in which modeling and fieldwork can be mutually informative. Given the current data and model uncertainties, it is possible to identify effective data acquisition strategies based on this analysis. For example, data could be ranked by the ratio of reduction in model uncertainty to the cost of acquiring each sample in terms of dollars and/or man hours. In this way, data collection could be optimized in terms of the efficiency with which new information is gained.

These approaches close the model–data loop by enabling models to inform data collection, and data to inform models. Such a shift has the potential to put field ecologists and modelers in closer connection. It also gives us the tools to answer the long-standing question among many field ecologists about what exactly modelers need to know. Indeed, this shift highlights a need for greater accessibility to models by the general research community so that field ecologists can drive this loop directly. This is exactly the objective of PEcAn: to encapsulate these tasks in a way that makes the analysis of models transparent, repeatable, and accessible.

In addition to informing sample size, the parameter meta-analysis can inform experimental design by providing valuable information on the scales of variability. For example, when data from multiple sites are available, the meta-analysis partitions among-site and within-site variance. This information can be used to construct optimal sampling designs that balance intensive vs. extensive sampling, and may help to identify environmental covariates that should be measured in order to explain parameter variability.

Based on our switchgrass example, variance partitioning also highlights broad data needs and the discrepancy between the relative ease of parameterizing aboveground processes compared to belowground processes. Indeed, one of the greatest challenges in ecosystem ecology is the ability to constrain belowground processes including allocation, respiration, and turnover. These parameters are uncertain precisely because measurement is difficult, often indirect, and data may reflect the diverse methods used to indirectly estimate a pool or flux. Many parameters in the ED2 model correspond to processes that are not directly observable. For example, the root respiration parameter in ED2 is not total root respiration, but just maintenance respiration, whereas measurements typically cannot separate growth, maintenance, and rhizosphere respiration. Whole-plant growth respiration, which is currently the most important model parameter, is also difficult to estimate directly from measurements (Amthor 2000). In this case, data assimilation is likely to be the most efficient route to constrain this parameter; data assimilation would effectively use mass balance of ecosystem carbon exchange to estimate this respiration parameter once other parameters are more directly constrained by data.

Future directions

The analyses presented here represent the first phase in the development of the PEcAn project. In the near term, we intend to expand the existing analyses to include a multivariate meta-analysis and sensitivity analysis to reduce model uncertainty by accounting for parameter covariances. In addition, we plan to implement the power analyses just discussed to more quantitatively inform data collection. A data assimilation module is in progress for PEcAn that will allow the use of ecosystem-level data, including plot-level inventory data, eddy covariance fluxes, and remote-sensing imagery, to enter the analysis and provide additional constraint on uncertainty in both parameters and output. The basic concept of variance decomposition will also be expanded to investigate other sources of variability, such as uncertainty in initial conditions or in driver data. We are implementing ecosystem models other than ED2 within the PEcAn workflow. This will provide opportunities for multi-model ensemble forecasting and assessing data requirements across models.

Integrating modeling into a workflow system has distinct advantages not just for model analysis, but also for managing the flows of information coming in and out of the model. In this sense we also envision PEcAn as an informatics and data management tool.

Finally, it is our hope that other researchers will find PEcAn useful and will contribute modules that extend the functionality of the system in creative and exciting ways.

Conclusion

In this paper, we demonstrate an approach to the parameterization of a terrestrial ecosystem model that synthesizes available data while providing a robust accounting of parameter uncertainty. We also present a scientific workflow that enables more efficient constraint of this uncertainty by identifying the key areas requiring data collection and model refinement. By quantifying the effect that each parameter has on model output uncertainty, this analysis identifies additional data that, once obtained, would improve model precision. In addition, the analysis calculates probabilities of alternate potential outcomes, resulting in more useful assessments.

Acknowledgments

D. S LeBauer and M. C. Dietze developed the concept, statistical analyses, and writing; D. S LeBauer implemented analyses and visualization; M. C. Dietze provided scientific and statistical guidance; D. Wang provided feedback on design and implementation of BETY, PEcAn, and this manuscript; K. T. Richter provided stomatal slope data and revisions to the manuscript; C. D. Davidson assisted in code design and implementation. Michael Sterling Burns assisted with the writing process. Patrick Mulrooney assisted with database design and implementation. John Skillman provided raw data; John Skillman, Peter Reich, Ian Wright, and Stan Wullschleger provided assistance with interpretation of their published data. D. S LeBauer, D. Wang, David Bettinardi, Andy Tu, Xiaohui Feng, and others entered previously published data into the database and contributed to the design of the web-based data entry workflow interface. Andrew Leakey, Evan Delucia, and Don Ort assisted with prior estimation. Juan Alvarez provided feedback on statistical calculations. Statistical and programming advice was obtained from contributors to crossvalidated.com, stackoverflow.com, and other stackexchange websites. We appreciate all individuals, including but not limited to the authors of publications cited in this paper, whose time, efforts, and intellect designed and produced the data used here, and whose recognition of the importance of archiving and sharing large data sets contributes to the advancement of science.

Funding for this research was provided by the Energy Biosciences Institute and an NSF Advances in Bioinformatics grant #1062547 to M. C. Dietze.

Literature Cited

Albert, C. H., F. Grassein, F. M. Schurr, G. Vieilledent, and C. Violle. 2011. When and how should intraspecific variability be considered in trait-based plant ecology? Perspectives in Plant Ecology, Evolution and Systematics 13:217–225.

Amthor, J. 2000. The McCree-de Wit-Penning de Vries-Thornley respiration paradigms: 30 years later. Annals of Botany 86:1–20.

Angel, J. 2010. Historical climate data. Illinois State Climatologist data for Station 118749 (Urbana). Illinois State Water Survey, Urbana, Illinois, USA. http://www.isws.illinois.edu/atmos/statecli/Summary/118740.htm

Ball, J., I. Woodrow, and J. Berry. 1987. A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions. Pages 221–224 *in* J. Biggins, editor. Progress in photosynthesis research. Martinus Nijhoff Publishers, Leiden, The Netherlands.

Bellocchi, G., M. Rivington, M. Donatelli, and K. Matthews. 2010. Validation of biophysical models: issues and methodologies. A review. Agronomy for Sustainable Development 30:109–130.

Beven, K. 2006. A manifesto for the equifinality thesis. Journal of Hydrology 320:18–36.

Beven, K., and J. Freer. 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Journal of Hydrology 249:11–29.

Breza, L. C., L. Souza, N. J. Sanders, and A. T. Classen. 2012. Within and between population variation in plant traits predicts ecosystem functions associated with a dominant plant species. Ecology and Evolution 2:1151–1161.

Cariboni, J., D. Gatelli, R. Liska, and A. Saltelli. 2007. The role of sensitivity analysis in ecological modelling. Ecological Modelling 203:167–182.

Casella, G., and R. L. Berger. 2001. Statistical inference. Second edition. Duxbury Press, Pacific Grove, California, USA.

Clark, J. S., et al. 2001. Ecological forecasts: an emerging imperative. Science 293:657–660.

Clark, J. S., M. Silman, R. Kern, E. Macklin, and J. HilleRisLambers. 1999. Seed dispersal near and far: patterns across temperate and tropical forests. Ecology 80:1475–1494.

Collatz, G., M. Ribas-Carbo, and J. Berry. 1992. Coupled photosynthesis-stomatal conductance model for leaves of C_4 plants. Functional Plant Biology 19:519–538.

Denman, K., G. et al. 2007. Couplings between changes in the climate system and biogeochemistry. Pages 499–587 *in* Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, New York, New York, USA.

Dietze, M. C., and A. M. Latimer. 2011. Forest simulators. Pages 307–316 *in* A. Hastings and L. Gross, editors. Sourcebook in theoretical ecology. University of California Press, Berkeley, California, USA.

Dietze, M. C., D. S. LeBauer. and R. Kooper. 2013. On improving the communication betweeen models and data. Plant, Cell and Environment, *in press*. http://dx.doi.org/10.1111/pce.12043

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. Journal of Geophysical Research 116:G04029.

Ellison, A. M. 2010. Repeatability and transparency in ecological research. Ecology 91:2536–2539.

Ernst, W. H. O., E. M. Veenendaal, and M. M. Kebakile. 1992. Possibilities for dispersal in annual and perennial grasses in a savanna in Botswana. Vegetatio 102:1–11.

Farquhar, G. D., and T. D. Sharkey. 1982. Stomatal conductance and photosynthesis. Annual Review of Plant Physiology 33:317–345.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7:457–472.

Gelman, A., and K. Shirley. 2011. Inference from simulations and monitoring convergence. Pages 163–174 *in* S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors. Handbook of Markov chain Monte Carlo. CRC Press, Boca Raton, Florida, USA.

Gill, R. A., and R. B. Jackson. 2000. Global patterns of root turnover for terrestrial ecosystems. New Phytologist 147:13–31.

Heaton, E. A., F. G. Dohleman, and S. P. Long. 2008. Meeting US biofuel goals with less land: the potential of Miscanthus. Global Change Biology 14:2000–2014.

Jackson, R. B., H. A. Mooney, and E. D. Schulze. 1997. A global budget for fine root biomass, surface area, and nutrient contents. Proceedings of the National Academy of Sciences USA 94:7362–7366.

Jones, M. B., M. P. Schildhauer, O. Reichman, and S. Bowers. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics 37:519–544.

Jongejans, E., and P. Schippers. 1999. Modeling seed dispersal by wind in herbaceous species. Oikos 87:362–372.

Justus, C., W. Hargraves, A. Mikhail, and D. Graber. 1978. Methods for estimating wind speed frequency distributions. Journal of Applied Meteorology 17:350–353.

Kruschke, J. 2010. Doing Bayesian data analysis. First edition. Elsevier/Academic Press, Boston, Massachusetts, USA.

Kubien, D. S., and R. F. Sage. 2004. Low-temperature photosynthetic performance of a C4 grass and a co-occurring C3 grass native to high latitudes. Plant, Cell and Environment 27:907–916.

Larocque, G. R., J. S. Bhatti, R. Boutin, and O. Chertov. 2008. Uncertainty analysis in carbon cycle models of forest ecosystems: research needs and development of a theoretical framework to estimate error propagation. Ecological Modelling 219:400–412.

Leuning, R. 1995. A critical appraisal of a combined stomatal-photosynthesis model for C_3 plants. Plant, Cell and Environment 18:339–355.

Luo, Y., K. Ogle, C. Tucker, S. Fei, C. Gao, S. LaDeau, J. S. Clark, and D. S. Schimel. 2011. Ecological forecasting and data assimilation in a data-rich era. Ecological Applications 21:1429–1442.

Luo, Y., E. Weng, X. Wu, C. Gao, X. Zhou, and L. Zhang. 2009. Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. Ecological Applications 19:571–574.

Massad, R.-S., A. Tuzet, and O. Bethenod. 2007. The effect of temperature on C_4-type leaf photosynthesis parameters. Plant, Cell and Environment 30:1191–1204.

McLaughlin, S., and L. Kszos. 2005. Development of switchgrass (*Panicum virgatum*) as a bioenergy feedstock in the United States. Biomass and Bioenergy 28:515–535.

McMahon, S. M., M. C. Dietze, M. H. Hersh, E. V. Moran, and J. S. Clark. 2009. A predictive framework to understand forest responses to global change. Annals of the New York Academy of Sciences 1162:221–236.

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. Journal of Geophysical Research 114:1–21.

Mesinger, F., et al. 2006. North American Regional Reanalysis. Bulletin of the American Meteorological Society 87:343–360.

Michener, W. K., and M. B. Jones. 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology and Evolution 27:85–93.

Miguez, F., X. Zhu, S. Humphries, G. Bollero, and S. Long. 2009. A semimechanistic model predicting the growth and production of the bioenergy crop *Miscanthus giganteus*: description, parameterization and validation. GCB Bioenergy 1:282–296.

Mood, A. M., F. A. Graybill, and D. C. Boes. 1974. Introduction to the theory of statistics. Third edition. McGraw-Hill, New York, New York, USA.

Moorcroft, P. R., G. C. Hurtt, and S. W. Pacala. 2001. A method for scaling vegetation dynamics: the Ecosystem Demography model (ED). Ecological Monographs 71:557–586.

Moore, D. J., J. Hu, W. J. Sacks, D. S. Schimel, and R. K. Monson. 2008. Estimating transpiration and the sensitivity of carbon uptake to water availability in a subalpine forest using a simple ecosystem process model informed by measured net CO_2 and H_2O fluxes. Agricultural and Forest Meteorology 148:1467–1477.

Oehlert, G. 1992. A note on the delta method. American Statistician 46:27–29.

Oyarzabal, M., J. M. Paruelo, F. Pino, M. Oesterheld, and W. K. Lauenroth. 2008. Trait differences between grass species along a climatic gradient in South and North America. Journal of Vegetation Science 19:183–192.

Plummer, M. 2010. JAGS (Just Another Gibbs Sampler). Version 2.2.0 user manual. http://ftp.jaist.ac.jp/pub/sourceforge/m/project/mc/mcmc-jags/Manuals/2.x/jags_user_manual.pdf

Reich, P. B., and J. Oleksyn. 2004. Global patterns of plant leaf N and P in relation to temperature and latitude. Proceedings of the National Academy of Sciences USA 101:11001–11006.

Reichstein, M., J. Tenhunen, O. Roupsard, J.-M. Ourcival, S. Rambal, F. Miglietta, A. Peressotti, M. Pecchiari, G. Tirone, and R. Valentini. 2003. Inverse modeling of seasonal drought effects on canopy CO_2/H_2O exchange in three Mediterranean ecosystems. Journal of Geophysical Research 108:4726.

Richardson, A., and D. Hollinger. 2005. Statistical modeling of ecosystem respiration using eddy covariance data: Maximum likelihood parameter estimation, and Monte Carlo simulation of model and parameter uncertainty, applied to three simple models. Agricultural and Forest Meteorology 131:191–208.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. 2008. Global sensitivity analysis. John Wiley, Chichester, West Sussex, UK.

Saugier, B., J. Roy, and H. Mooney. 2001. Estimations of global terrestrial productivity: Converging toward a single number? Pages 543–557 *in* B. Saugier and H. Mooney, editors. Terrestrial global productivity. Academic Press, San Diego, California, USA.

Schwalm, C. R., et al. 2010. A model–data intercomparison of CO_2 exchange across North America: Results from the North American Carbon Program site synthesis. Journal of Geophysical Research 115:G00H05.

Skillman, J. B. 2008. Quantum yield variation across the three pathways of photosynthesis: not yet out of the dark. Journal of Experimental Botany 59:1647–1661.

Stodden, V., et al. 2010. Reproducible Research 12(5):8–13. IEEE Computer Society, CS Digital Library. http://doi.ieeecomputersociety.org/10.1109/MCSE.2010.113

Thompson, S., and G. Katul. 2008. Plant propagation fronts and wind dispersal: an analytical model to upscale from seconds to decades using superstatistics. American Naturalist 171:468–479.

Tjoelker, M. G., J. M. Craine, D. Wedin, P. B. Reich, and D. Tilman. 2005. Linking leaf and root trait syndromes among 39 grassland and savannah species. New Phytologist 167:493–508.

USDA NRCS. 2011. The PLANTS Database. U.S. Department of Agriculture, Natural Resources Conservation Service. http://plants.usda.gov/

VanLoocke, A., T. E. Twine, M. Zeri, and C. J. Bernacchi. 2012. A regional comparison of water use efficiency for miscanthus, switchgrass and maize. Agricultural and Forest Meteorology 164:82–95.

Violle, C., B. J. Enquist, B. J. McGill, L. Jiang, C. H. Albert, C. Hulshof, V. Jung, and J. Messier. 2012. The return of the variance: intraspecific variability in community ecology. Trends in Ecology and Evolution 27:244–252.

Wang, D., D. S. LeBauer, and M. C. Dietze. 2010. A quantitative review comparing the yield of switchgrass in monocultures and mixtures in relation to climate and management factors. GCB Bioenergy 2:16–25.

Wang, D., D. S. LeBauer, and M. C. Dietze. 2013. Predicting yields of short-rotation hybrid poplar (*Populus* spp.) for the contiguous USA through model–data synthesis. Ecological Applications 23:944–958.

Wang, D., M. W. Maughan, J. Sun, X. Feng, F. E. Miguez, D. K. Lee, and M. C. Dietze. 2011. Impacts of canopy position and nitrogen on nitrogen allocation and photosynthesis of switchgrass (*Panicum virgatum* L.). Aspects of Applied Biology 112:341–351.

Williams, M., et al. 2009. Improving land surface models with FLUXNET data. Biogeosciences Discussions 6:2785–2835.

Wright, I. J., et al. 2004. The worldwide leaf economics spectrum. Nature 428:821–827.

Wullschleger, S. D. 1993. Biochemical limitations to carbon assimilation in C_3 plants—a retrospective analysis of the A/Ci curves from 109 species. Journal of Experimental Botany 44:907–920.

Supplemental Material

Appendix

Data transformations used by PEcAn, including the Arrhenius correction and equations used to estimate SE from various statistics reported in the literature (*Ecological Archives* M083-006-A1).

Supplement 1

Data used in the present analysis, including new stomatal slope data for five species and previously published switchgrass observations of $V_{c,max}$, SLA, leaf width, and fine-root : leaf ratio (*Ecological Archives* M083-006-S1).

Supplement 2

Documentation of database (BETYdb) used in the present study, including schema and contents (*Ecological Archives* M083-006-S2).

Data Availability

Data associated with this paper have been deposited in the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS): http://www.ideals.illinois.edu/handle/2142/34655