Plant, Cell & Environment

Plant, Cell and Environment (2013) **36**, 1575–1585 doi: 10.1111/pce.12043

On improving the communication between models and data

MICHAEL C. DIETZE[1], DAVID S. LEBAUER[2] & ROB KOOPER[3]

[1]Department of Earth and Environment, Boston University, 675 Commonwealth Ave., Rm. 130, Boston, MA 02215, USA, [2]Energy Biosciences Institute, University of Illinois, 1206 W. Gregory Dr., Urbana, IL 61801, USA and [3]National Center for Supercomputing Applications, 1205 W. Clark St., Urbana, IL 61801, USA

ABSTRACT

The potential for model–data synthesis is growing in importance as we enter an era of 'big data', greater connectivity and faster computation. Realizing this potential requires that the research community broaden its perspective about how and why they interact with models. Models can be viewed as scaffolds that allow data at different scales to inform each other through our understanding of underlying processes. Perceptions of relevance, accessibility and informatics are presented as the primary barriers to broader adoption of models by the community, while an inability to fully utilize the breadth of expertise and data from the community is a primary barrier to model improvement. Overall, we promote a community-based paradigm to model–data synthesis and highlight some of the tools and techniques that facilitate this approach. Scientific workflows address critical informatics issues in transparency, repeatability and automation, while intuitive, flexible web-based interfaces make running and visualizing models more accessible. Bayesian statistics provides powerful tools for assimilating a diversity of data types and for the analysis of uncertainty. Uncertainty analyses enable new measurements to target those processes most limiting our predictive ability. Moving forward, tools for information management and data assimilation need to be improved and made more accessible.

Key-words: accessibility; Bayesian statistics; data assimilation; informatics; PEcAn; provenance; uncertainty; workflow.

INTRODUCTION

Recent scientific advances, as well as the underlying currents of technology in society, are changing the way that we think about and interact with models and data. New technologies such as smartphones and tablets, cellular data networks, cloud computing and virtual machines are increasingly making information available at our fingertips and moving computing away from the desktop. Parallel computing, once the territory of supercomputing centres, can now occur on a handheld device, while the demands for faster graphics by gamers have led to new approaches to graphics processing unit (GPU)-based scientific computing. Data are increasingly accessible and interconnected, and advances in scientific technologies, from gene sequencing to satellite remote sensing, have led to an explosion in data quantity (Baraniuk 2011). Some have argued that we are entering a fourth era of data-intensive 'eScience' (Hey, Tansley & Tolle 2009), and whether or not this is true it is clear that scientific synthesis is more important and more attainable than ever before, and that many researchers are shifting beyond the 'my single data set' approach to science.

While other reviews in this special issue have highlighted the state-of-the-art in plant models across a wide range of disciplines, our objective is to discuss more generally new approaches to how models interact with data and how researchers interact with models. Our aim is to broaden the perspective of the general research community about how and why they interact with models, to promote the idea of a community approach model–data synthesis, and to highlight some of the tools and techniques that can make this happen. We will be upfront that this is not meant to just be a review, but also a call for a new way of viewing the interactions between models and data, and between empiricists and modellers. We focus primarily on terrestrial ecosystem models, crossing scales from individual-level ecophysiology to global vegetation, as this is our area of expertise. However, most of the model–data concepts presented translate across the different levels of biological organization. Because of the large spatial and temporal scales involved, and the diversity of biological, chemical and physical processes represented, ecosystem models represent a particularly complex class of models and model/data issues (Dietze & Latimer 2011). Therefore, ecosystem modellers have had to confront many of these issues earlier than other subdisciplines. That said, there are specific subtopics, such as the archiving and informatics of genomic data, where other disciplines are more advanced.

As mentioned earlier, one of the great opportunities in modern science is the explosion in the volume and diversity of data that are being generated and made available. However, for many scientific questions no single data source provides a complete picture of the processes we are interested in. For example, biometric inventory data, eddy-covariance towers, soil respiration chambers and numerous remote-sensing technologies each provide abundant but partial information about the terrestrial carbon cycle. Traditional research focused on a small number of data types at one site, but at present even large syntheses only make use of a subset of the available data. For many critical global change and carbon cycle questions, we are more limited by our ability to use the data that have already been collected than by the need for new data. Many modern data sources, from next generation sequencing to remote sensing, are highly

Correspondence: M. C. Dietze. Fax: +1 617 358 8399; e-mail: dietze@bu.edu

automated and generate unprecedented volumes of data. Beyond automated data, there also exists critical information scattered across smaller data sets, often in non-standardized file formats and in the hands of individual investigators, that make up the 'long tail' of data. These data represent the majority of research effort but are near impossible to synthesize without a bottom-up 'community' effort. In addition to synthesizing data, there is also growing recognition of the importance of characterizing the uncertainties in our data (Clark 2007). Not only do we need to be able to characterize our confidence in models and data for them to be useful (Clark *et al*. 2001), but how variance is partitioned can qualitatively change model predictions (Clark *et al*. 2007) and estimates of model parameters (Trudinger *et al*. 2007). Finally, data collection is currently driven largely by intuition (we measure what we feel to be important) or by technology (we measure what is easy to measure) rather than a quantitative understanding of what data will most reduce uncertainties and how much data are required to do so. In the following sections, we will discuss the current state of approaches for model–data synthesis and strategies for moving forward on improving tools and the way that users interact with them.

MODELS AS A SCAFFOLD

The traditional approach to the interface between models and data has focused on calibration and validation, and as typified by the oft-repeated expression 'confronting models with data' (Hilborn & Mangel 1997). However, societal demands have driven a push for quantitative forecasts (Clark *et al*. 2001) and for ecology to become a more predictive science (Moorcroft 2006). Even outside concerns over global change in ecology, the era of 'big data' opens up the possibility of syntheses that were not previously possible. One of the major challenges of data synthesis and prediction is that data sets that operate on different scales cannot 'talk' to one another directly (e.g. leaf-level gas exchange, remote sensing and palaeoecological proxies), but all provide us with partial information about the underlying biological processes. However, models can represent processes at a hierarchy of spatial, temporal, phylogenetic and organizational scales, and encapsulate our current understanding of a system. Here we present the paradigm of using *models as a scaffold* for fusing different data sets. In this context, our process understanding provides the capacity to move across scales. Models thus serve as the natural extension of synthetic efforts that have evolved from qualitative reviews to numerical meta-analysis.

As an example, consider the terrestrial biosphere, which remains one of the largest sources of uncertainty in climate change projections (Friedlingstein *et al*. 2006; Purves & Pacala 2008). Despite an abundance of data, no one data source provides a complete picture of the carbon cycle, and therefore multiple data sources must be integrated in a sensible manner. Process-based ecosystem models represent an ideal scaffold for integrating these data streams because they represent multiple processes in ways that capture our current understanding of the causal connections across scales and among data types. However, current terrestrial ecosystem models only make use of a subset of the available data and remain inaccessible to much of the scientific community (see *Barriers to Modelling* below). The lack of integration of data in models is a major hindrance to reducing uncertainty in climate change projections, and separates the information we have gathered from the understanding required to inform policy and management. Addressing this need requires the development of tools able to accommodate a diverse array of data operating across a large range of spatial and temporal scales and with different levels of associated uncertainty.

There are a number of significant model–data challenges associated with improving synthesis and prediction (McMahon *et al*. 2009). Process models are complex, nonlinear and computationally demanding. They suffer from issues of equifinality, whereby many combinations of parameters can predict the same net outcome (Luo *et al*. 2009; Williams *et al*. 2009), and latent variables, where models infer the dynamics of internal state variables and processes that are not directly observable. Model–data synthesis requires that we be able to combine multiple types of data operating on different scales and of different types (literature versus observation versus experiment). Both models and data possess multiple types of error, and correct inference can depend strongly on how errors are treated and partitioned (Clark *et al*. 2007; Trudinger *et al*. 2007).

There are a number of statistical techniques available to synthesize models and data, and to account for multiple types of data and uncertainty. These methods are predominantly Bayesian in their approach because the Bayesian perspective offers a number of conceptual advantages over traditional methods when it comes to fitting complex models (Clark 2005). Firstly, Bayesian methods provide probability distributions as their output, rather than point estimates, which makes it straightforward to characterize uncertainty and transfer it into further analyses and forecasts. Secondly, because Bayesian analyses are based on conditional probabilities, they provide a flexible framework for fitting complex models with multiple data constraints and multiple sources of uncertainty. Each small part of a complex analysis is expressed as a probability conditioned on all other variables in the analysis, allowing complex analyses to be built up by combining these conditional probabilities. Thirdly, Bayesian statistics is an inherently sequential approach to inference, whereby we are able to incorporate earlier information and update estimates based on new information. Returning to our analogy of model as scaffold, this means that as we build up our analysis, adding new study sites, new experiments, new types of data or even just the latest year of data from a long-term study, we do not need to start our analysis from scratch each time.

BARRIERS TO MODELLING

Given the importance of models to science, one could ask the question, 'Why don't all researchers use models?' In this section, we will discuss some of the barriers to modelling and how they might be reduced. In our experience, there are three major barriers to modelling: accessibility, relevance and

Table 1. Barriers to modelling

Accessibility	
	Perceptions of model complexity
	Need for more intuitive interfaces
	Need for 'driver' training
Relevance	
	Ability to use models for non-trivial hypothesis testing
	Experimental design: what to measure, where, when and how much
Informatics	
	Data to run models
	Data to evaluate and improve models
	Visualization and analysis of model outputs

informatics (Table 1). By *accessibility*, we refer both to the perception that models are too complicated or that they require expert training to use. Many biologists believe that models are too complicated, which may be related to more general patterns of poor communication between theory and experimentation in biology (Fawcett & Higginson 2012). However, from our experience, once a model is explained to a subject-matter expert they frequently assert that the same model, which was previously perceived as too complicated, is actually too simple. Indeed, we firmly believe that an expert's conceptual model of how their system works is much more complicated and nuanced than most mathematical models.

There are multiple reasons for the simplicity of models. While simulation models are different from theoretical models, in that they frequently seek predictability rather than pure abstraction, they still require generality. The result of this is that a novel finding, which is frequently first demonstrated at one place in one system, is not easy to incorporate into a model until sufficient data have accumulated to calibrate the process more generally. However, even after taking this into account, models lag behind our current understanding of a system for two important and interrelated reasons. The first, which we dub the *assimilation challenge*, is that, as stated earlier, models are not yet taking full advantage of the available data and insight. The second, which we dub the *community challenge*, is that in biology modellers are greatly outnumbered by empiricists and lack the time and resources to assimilate the vast array of current and historical experiments and observations. Furthermore, the nature of synthetic work requires that modellers be academic generalists rather than the expert in any particular subdiscipline. These two challenges overlap considerably in the 'long tail' data problem, whereby modellers lack both the expert's knowledge and their data. We believe that solving these challenges requires not only new tools to make models and data more accessible, but *a new interdisciplinary commitment to community involvement in model–data synthesis and the evaluation of models by subject-matter experts*.

The second half of the accessibility problem is the perception that models require expert training to use. This perception is absolutely true. However, to make an analogy, driving a car requires training and it is dangerous to drive without experience or an understanding of the rules of the road. Still, the training required to drive a car is very different from the training required by a mechanic to repair a car or by an engineer to design a car. Right now, the primary people 'driving' models are their mechanics and engineers. Part of the solution lies in the need for more opportunities for multidisciplinary training that integrates data collection, statistics, modelling and data assimilation. Successful examples of such training are the 'Summer Course in Flux Measurements and Advanced Modeling' (http://www.fluxcourse.org) and the PalEON summer course 'Assimilating long-term data into ecosystem models' (http://www.paleonproject.org), which integrate modelling and data assimilation with carbon flux and palaeoecological measurements, respectively (Fig. 1). Another large part of the solution is the need to make modelling tools more accessible – to make models easier to drive.

The second barrier to modelling is *relevance*. Many researchers understand the importance of models but, given the multitude of other demands on their time, do not see how learning to use models helps them understand their system better or how models can help them in the field or the laboratory. Here it is useful to view models as working hypotheses – as a quantitative expression of what we currently know about how a system works. As such, models allow us to ask non-trivial hypotheses (Hilborn & Mangel 1997; Anderson, Burnham & Thompson 2000). For example, we know that net primary production (NPP) responds positively to nitrogen addition in almost all systems (LeBauer & Treseder 2008). Yet, nitrogen addition experiments continue to use 'no response to N' as a trivial null hypothesis. The interesting question is not whether there is a response to N, but whether the response is different from our expectation given our current understanding of plant physiology and biogeochemistry. Not only can models provide sensible competing alternative hypotheses, but also Bayesian credible intervals on model projections provide a means of formally testing these hypotheses. In addition to testing hypotheses, uncertainty

Figure 1. Students at the PalEON summer course 'Assimilating long-term data into ecosystem models' spend an intensive week learning about palaeoecological data and measurements, Bayesian statistics, ecosystem models and data assimilation. Photo credit by John W. Williams.

analyses of models can be used to quantitatively determine what processes drive system responses, where the gaps are in our scientific knowledge of these processes, and how variable these processes are and at what scale (LeBauer *et al*. 2012). This information can be formally incorporated into power analyses and economic design optimization to help inform what measurements would most efficiently reduce overall uncertainties about a process, what sample size is required, and how the sampling should be done (e.g. how many study sites versus how many samples per site) (M. Dietze *et al*., unpublished data).

The final barrier to modelling is *informatics*. Many models require massive amounts of information about the systems they describe in order to operate because computational models, unlike theory, aim to describe real plants in a real environment. For example, a terrestrial ecosystem model might require that the user have information about a dozen meteorological variables at an hourly time step, soil texture profiles, soil carbon and nitrogen pools, vegetation composition and structure, topography, land use trajectories, nitrogen deposition, ozone concentration and disturbance frequencies. Each input may come from a different source, in a different file format, and may require gap filling, interpolation or other forms of synthesis and processing so that the model can read it. All these data and effort are required just to run the model forward, and need to exist at whatever spatial grain (pixel size) and extent the user is interested in. If one is interested in model–data fusion, then in addition there may be a score or more of different data types being assimilated, each available at a specific spatial and temporal resolution and in its own file format, as well as the need to run the model anywhere from dozens to hundreds of thousands of times depending upon what method is being used for data assimilation. Each of these model simulations may generate dozens of output variables, each written out numerous times a day for each pixel across the whole spatial and temporal domain of the run. These outputs need to be visualized, compared with data, and otherwise interpreted. Trying to manage these flows of information into and out of models is like drinking from a fire hose. From our experience, of the three barriers to modelling it is teaching students how to manage information that is the major bottleneck for training and research. Fortunately, thanks to advances in scientific workflow software, data archiving and interoperability, this problem is surmountable. Even more progress is anticipated with future advances in informatics and cyberinfrastructure. The remainder of this review focuses on recent and ongoing work that addresses these three barriers to modelling and the idea of a model as a scaffold.

TRANSPARENCY, QUALITY ASSURANCE AND REPEATABILITY

If there is one truth that came out of the 2009 University of East Anglia Climatic Research Unit (CRU) email controversy, also known as 'Climategate', it is that the bar has been raised on the need for transparency and repeatability with data processing and models. Unlike running an experiment in the physical world, computer outputs should be easy to recreate by any reasonably skilled user, allowing one to verify results, check assumptions and build upon past research. In practice, this is rarely the case. Even with a task as simple as reporting a statistical test, the exact software used is often not reported, the software is not always still available, and backward compatibility is not always guaranteed (Ellison 2010). For proprietary software, the exact computation being done is typically inaccessible and unverifiable. Therefore, for more complex models the archiving of computer codes is a necessary condition for repeatability and transparency. However, archiving code is not sufficient. The reason is that this does not capture the informatics of where the inputs came from, how they were processed, how sets of model runs were completed, and how the model output was post-processed and visualized. In other words, transparency and repeatability in data processing and modelling require that we capture the full *workflow*. Fortunately, scientific workflow management software has been an active area of development over recent years and there are a number of options to choose from (Curcin & Ghanem 2008). These systems generally give the user the capacity to interact with workflows using graphical analytic webs that depict different modules and actions, with arrows between them representing flows of information (Boose *et al*. 2007). Popular workflows in plant biology include Kepler (https://kepler-project.org), SciWalker (Ellison *et al*. 2006), Taverna (http://www.taverna.org.uk) and Cyberintegrator (http://isda.ncsa.uiuc.edu/cyberintegrator/). Kepler is particularly popular among ecologists since its development originated at the National Center for Ecological Analysis and Synthesis (NCEAS). Taverna is a UK project that tends to be more widely used for bioinformatics.

Another critical function of workflows is provenance tracking (Reichman, Jones & Schildhauer 2011). Data provenance refers to the tracking of data from its origins through all the processing steps and analyses. Part of provenance is carrying the appropriate metadata forward with the data. Another important part is knowing exactly which version of a data set was used in an analysis or model. When data are curated by hand it is extremely easy to end up with multiple versions of files and analyses and it can be very difficult to later reconstruct what precisely was done. These problems are multiplied with models where larger volumes of outputs can be generated quickly and generating numerous model runs is frequently part of an analysis.

ACCESSIBILITY

The above discussion of workflows leads naturally into discussions of accessibility, as graphical workflows make data and models more transparent and repeatable, and are generally more accessible to the non-expert than reams of computer code. They make it easier to see the 'big picture' and compartmentalize a lot of the details of the model–data process within modules. More generally, there is a clear need for tools that make models more accessible. These tools need to deal with the challenges in informatics and statistical inference surrounding models as much as they do with running

the models themselves. One thing that has become apparent from simple statistical software like JMP (http://www.jmp.com) is that we make the most use of the tools that are easy to use [e.g. analysis of variance (ANOVA)]. However, such tools need to accommodate the need for flexibility that is part of doing novel science and the need for automation required among expert users. Traditional software engineering approaches of designing graphical user interfaces (GUIs) may succeed in making models more accessible at the cost of flexibility and automation. Indeed, from personal experience, GUIs designed by biologists often fail on all fronts, being hard to use by the novice, hard to maintain by the developer, and perpetually lagging behind the newest features in the model. Web-based interfaces, on the other hand, are often more intuitive to current users, easier to maintain, and require less modification of the underlying model. Overall, there is a need for modellers to be conscious of design issues, and to work with those who have greater expertise in making interfaces that are intuitive, provide user feedback, and project the correct conceptual model of how the system works (Norman 2002).

In addition to designing better interfaces, there is often a significant challenge to novice modellers in learning how to compile and install software, and porting code from one system to another can challenge even expert users. These challenges are multiplied when one is dealing not just with a model, but with the workflows for analysis and data assimilation being advocated here. Virtualization provides the potential to share not just software with users, but whole integrated and fully functional systems of models, data, workflows and databases. Because a virtual machine brings its operating system with it, the user sees the exact same environment regardless of the host operating system. Finally, a number of cloud-computing services, such as Amazon EC2 and Google Compute Engine, let you run virtual applications in a scalable environment without having to invest in the hardware or information technology (IT) of running your own compute cluster.

DATA ASSIMILATION

The relationship between models and data in biology is changing rapidly as the statistical methods for data assimilation, which have been commonplace in other scientific disciplines such as numerical weather prediction, are being picked up by biologists and adapted to deal with the differences between the physical and natural sciences. Past approaches to model parameterization (the process of putting hard numbers on all the coefficients in a model) at times lacked transparency. The process of model tuning and informal calibration based on the literature and expert opinion gave rise to inverse modelling approaches and numerical optimization (e.g. Medvigy *et al*. 2009). These approaches have been followed by Bayesian approaches to model–data fusion capable of a richer accounting of uncertainties in model parameters and data. Interestingly, Bayesian approaches have now come full circle in acknowledging the importance of the scientific literature and expert opinion, but now formalize the process through meta-analysis and expert elicitation for defining priors (LeBauer *et al*. 2012).

From the perspective of ecosystem models, there have been a number of recent reviews and perspectives published on data assimilation techniques and their potential (Luo *et al*. 2009, 2011; Williams *et al*. 2009; Keenan *et al*. 2011; Peng *et al*. 2011; Zobitz *et al*. 2011; Hartig *et al*. 2012). These techniques can be broadly grouped into two approaches: *batch* methods that assimilate a whole data set at once, and *sequential* methods that assimilate time series data in chronological order. The most common batch method is to employ the same Markov Chain Monte Carlo (MCMC) techniques that are used in fitting Bayesian statistical models, only treating the computer model as a 'black box' (Braswell *et al*. 2005; Sacks *et al*. 2006; Xu *et al*. 2006; Zobitz *et al*. 2008; Keenan *et al*. 2012). The general approach here involves iteratively proposing a new set of model parameters, running the model with those parameters and comparing model output to data. The algorithm then accepts or rejects the new parameters with some probability that depends upon the *likelihood* of generating the observed data under these parameters and the prior probability that those parameters are biologically realistic. In this way, the MCMC algorithm generates random samples from the parameter distributions that can be used to approximate the full distribution.

Applications of the sequential approach to data assimilation generally employ some flavours of Kalman Filter, most commonly the ensemble Kalman filter (EnKF) (Williams *et al*. 2005; Gove & Hollinger 2006; Quaife *et al*. 2008; Gao *et al*. 2011). With sequential approaches, the probability distribution of model predictions from the previous time point to the current one is treated as the prior distribution. The prior distribution is then updated based on the likelihood of generating the observed data, and the resulting posterior distribution serves as the initial conditions for the next forecast step.

For those seeking greater technical detail beyond the aforementioned ecological reviews, there are a number of more general reviews (Wikle & Berliner 2007; Evensen 2009a) and textbooks (Lewis, Lakshmivarahan & Dhall 2006; Evensen 2009b) on the subject; however, there are not yet any detailed texts that take a biological perspective. This is significant because there are some non-trivial differences between the data assimilation in the physical and natural sciences. Foremost is that data assimilation in the physical sciences benefits from models that encapsulate physical laws where the equations and parameters are known, but where the predictions may be highly divergent with time, or even chaotic, such as with weather forecasting. By contrast, most biological systems appear to have stabilizing feedbacks that make them predictable at certain scales in space and time (e.g. succession). In the physical sciences, data assimilation is primarily focused on estimating the *state* of the system, in essence nudging models to keep on track with observations. This state-variable data assimilation problem is different from the typical biological application, where the equations and parameters are statistical approximations of diverse, complex phenomena, and the most common concern is

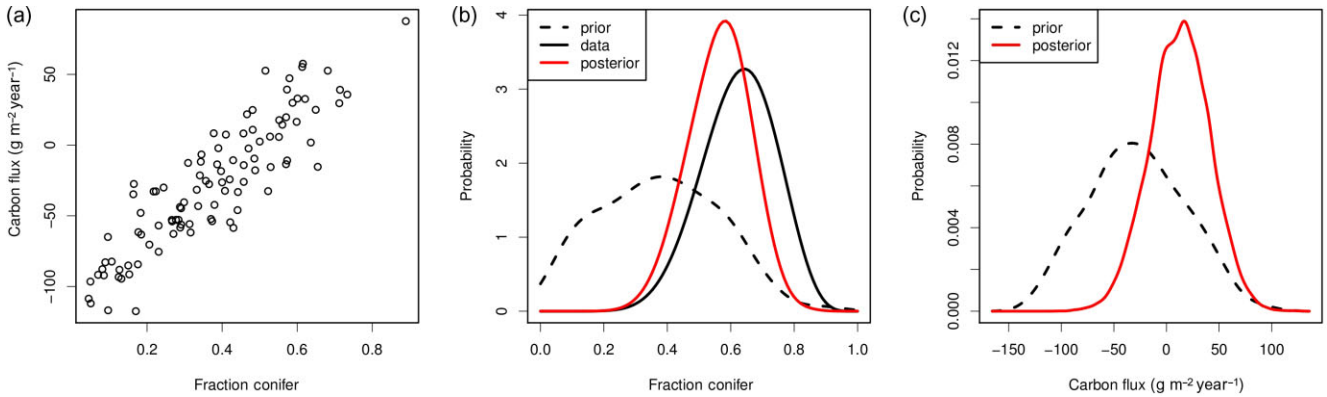

Figure 2. Conceptual example of state-variable data assimilation. Imagine an ensemble of model runs that predict the ecosystem dynamics for time *t* and incorporate uncertainty in drivers, state variables and model parameters. Panel (a) gives a hypothetical scatter plot of such an ensemble containing one output variable that maps to an observable quantity (e.g. fraction of the community that is conifer) and one that is unobservable (carbon flux). In data assimilation, the model output is treated as the prior and panels (b) and (c) give the model-ensemble priors for the two variables (dashed lines) that are just the marginal distributions from panel (a). Consider now the case where observations suggest a higher conifer abundance than the model (panel b solid line) but there is still uncertainty in this estimate. The data assimilation posterior estimate of conifer abundance (panel b, red) combines both the model and the data. Furthermore, because of the covariance between composition and carbon flux the composition data also constrain the carbon flux estimate (panel c, red). The posterior distributions from this time step then serve to generate the ensemble of model inputs for the next time step.

reducing parameter uncertainty. That said, there is a large potential utility for state-variable data assimilation in biology for those problems where the primary interest lies in estimating the state of a system subject to multiple or indirect data constraints (Fig. 2). Overall, while the physical data assimilation literature is an invaluable starting point, and there are excellent tools available to assist with the process, such as the NCAR DART (Data Assimilation Research Testbed, http://www.image.ucar.edu/DAReS/DART), there remain significant data assimilation challenges that are unique to biology.

One of the major strengths of formal data assimilation techniques is that multiple data types can be assimilated simultaneously. Indeed, because of issues of parameter identifiability and equifinality (many parameter combinations giving the same net prediction), the assimilation of multiple data sets is often necessary to ensure that models are not getting the right answer for the wrong reasons (Luo *et al.* 2009; Williams *et al.* 2009). There are a number of examples of recent work with ecosystem models that have assimilated some combination of field measurements, eddy covariance and remote sensing (Xu *et al.* 2006; Quaife *et al.* 2008; Gao *et al.* 2011; Weng & Luo 2011; Keenan *et al.* 2012), although to our knowledge, a robust synthesis across all three data types has not yet occurred. It is also noteworthy that the majority of data assimilation work with ecosystem models occurs using simple models, as data assimilation with larger and more complex terrestrial biosphere models presents non-trivial computational challenges. Another challenge when synthesizing multiple data streams is how to prevent abundant, automated data sources from swamping out the signal from other less abundant, but often very valuable, data sources. We suggest elsewhere (M. Dietze, unpublished data) that much of the solution lies in the details of how variance is partitioned. For example, many automated data streams such as eddy covariance are highly autocorrelated on multiple temporal scales. Failure to account for this can also lead to qualitative mistakes in identifying the processes that drive model error (Dietze *et al.* 2011). Indeed, even when just assimilating a single data stream, the partitioning of variance has a larger impact on data assimilation than the choice of assimilation method (Trudinger *et al.* 2007). Finally, it is important to take to heart the warning that 'data-model integration is not magic' (Classen & Langley 2005), but just one tool that needs to be part of the larger process of evaluating models and bringing modellers and experimentalists closer together.

REPRESENTATION OF UNCERTAINTY

As discussed in the *Data Assimilation* and *Models as a Scaffold* sections, the way that uncertainty is represented in models can have a large impact on inference and prediction. There are numerous sources of uncertainty in process-based models and to date most modelling exercises only accommodate a small fraction of these processes, if incorporating any at all. It is still not uncommon to see model projections that lack any confidence interval or standard error around the model output. However, it is easy to be fooled into believing that an *accurate* model is a *precise* one. We have generated model predictions that passed beautifully through both calibration and validation data, only to be disappointed when further computation revealed a confidence interval that spanned two orders of magnitude. Complex models often contain numerous parameters – many more than typically used in statistical models such as multiple regressions – and thus parameter uncertainty is frequently the dominant source of uncertainty. Because Bayesian approaches to data assimilation generate posterior parameter distributions as their output, it is conceptually straightforward to propagate parameter error by sampling many sets of parameter values

from their joint posterior distribution, running the model many times with these parameter sets, and then using the distribution of model outputs to estimate the mean, confidence interval, variance, etc. This *ensemble* approach to error propagation is a simple and statistically valid approximation to the transformation of parameter error into model output, with the only real cost being computation, as the accuracy of the approximation clearly increases with the size of the ensemble. Alternatively, there are analytical methods for doing these transformations – these in fact are the origin of the error estimates and confidence intervals in classical statistical models such as regression – but they require the ability to calculate the full Jacobian matrix (the matrix of derivatives between every model parameter and every model output), which are not generally available or easy to generate for complex computer models (Casella & Berger 2001).

A second major source of uncertainty is the choice of model structure. The typical way of estimating this error is to use an ensemble of different models to make predictions. A very common phenomenon with multi-model ensembles is for the mean across models to provide a more accurate prediction than most, if not all, of the individual models (Schwalm *et al*. 2010; Dietze *et al*. 2011). In some sense, the errors in modelling assumptions tend to cancel each other out; however, one thing that is generally not accounted for is the fact that models are frequently not independent of one another and therefore multi-model ensembles likely underestimate the true structural uncertainty. This lack of independence arises from the fact that modellers clearly learn from the successes and failures of one another, borrowing approaches and sometimes even large chunks of code. From a scientific perspective this is a good thing, even if it makes the statistics a bit more complicated.

Parameter and structural errors are the most common sources of errors currently being addressed, but by no means the only sources (McMahon *et al*. 2009). In addition, there is process error in the models and measurement error in the data. With classical methods these are often lumped together into a 'residual' error, but it is important to distinguish them because process error propagates forward into predictions while measurement error should not. The incorporation of process error is currently rare in models, but in a regression context this is analogous to the distinction between a confidence interval, which accounts for parameter uncertainty, and a predictive interval, which includes parameter and residual error. This distinction is important because as the volume of data increases the width of the confidence interval goes to zero but the predictive interval does not. In addition, model parameters may vary with space, time, plot, taxa, individual, etc. in a consistent fashion – what in a regression context would be either correlated or uncorrelated random effects. Accounting for these sources of variability, which are quite different from random noise in the process, can qualitatively change the outcome of a model prediction. For example, within a forest community model accounting for random effects in time and at the individual tree level changed the coexistence dynamics of two species from competitive exclusion to coexistence with the previously excluded species becoming the more dominant (Clark *et al*. 2007). In addition to uncertainty in the model itself, there can also be considerable uncertainties in the inputs to the models. Here we need to distinguish between uncertainties in the initial conditions, which are often large but typically diminish in influence over time, from uncertainties in the model drivers, such at the meteorology driving an ecosystem model. For drivers, it is important to get not just the mean, but also the variance structure, correct as the variability can have as much impact on model projections as the mean (Medvigy *et al*. 2010). The uncertainty in meteorology increases when moving from a study site, where the main source of error is gaps in the data, to regional scales where one has to work with derived data products (Spadavecchia & Williams 2011). It also increases when one moves back in time before instrumental records or forward with different climate change scenarios.

In addition to propagating uncertainty through models, it is important to understand and explain the sources of error and variability in models. This helps us better understand where errors can be reduced, which sources of error need to be propagated into predictions, and where systems are unpredictable. Unpredictability can arise in situations where stochasticity dominates the process, such as long-distance seed dispersal (Clark *et al*. 2003), or it can arise if the dominant source of variability is one that we have no capacity to reduce. Some systems can also be inherently unpredictable due to the chaotic nature of the system, or computationally irreducible, whereby the dynamics of the system needs to be represented with great detail and cannot be approximated with simple models (Beckage, Gross & Kauffman 2011).

For those problems that do have some degree of predictability, attributing uncertainty to specific processes is a critical part of the feedback loop between models and measurements. Uncertainty analyses identify which processes need further study or additional data constraints, and can be used to prioritize field and laboratory work (LeBauer *et al*. 2012). The information from uncertainty analyses can be formally incorporated into experimental design in order to determine what sample sizes and sampling designs would most efficiently reduce uncertainty. There are a suite of methods that can be used for sensitivity and uncertainty analyses (Saltelli *et al*. 2008). One common approach is the use of Monte Carlo sensitivity analyses that are very similar to ensemble approaches to error propagation discussed earlier. Specific methods vary based on whether parameters are varied individually or all at once. In the latter, variance is usually partitioned statistically, such as with an ANOVA approach or by using flexible statistical models such as splines, GAMS, or Gaussian process models to estimate the response surface (Petropoulos *et al*. 2009). For the former, the uncertainty contribution is assessed by the incremental change in either the model output (for a sensitivity analysis) or the model predictive uncertainty (for an uncertainty analysis) as each term is varied either individually (holding all others constant) or sequentially (incrementally adding a new term). In all cases, the bottleneck for uncertainty

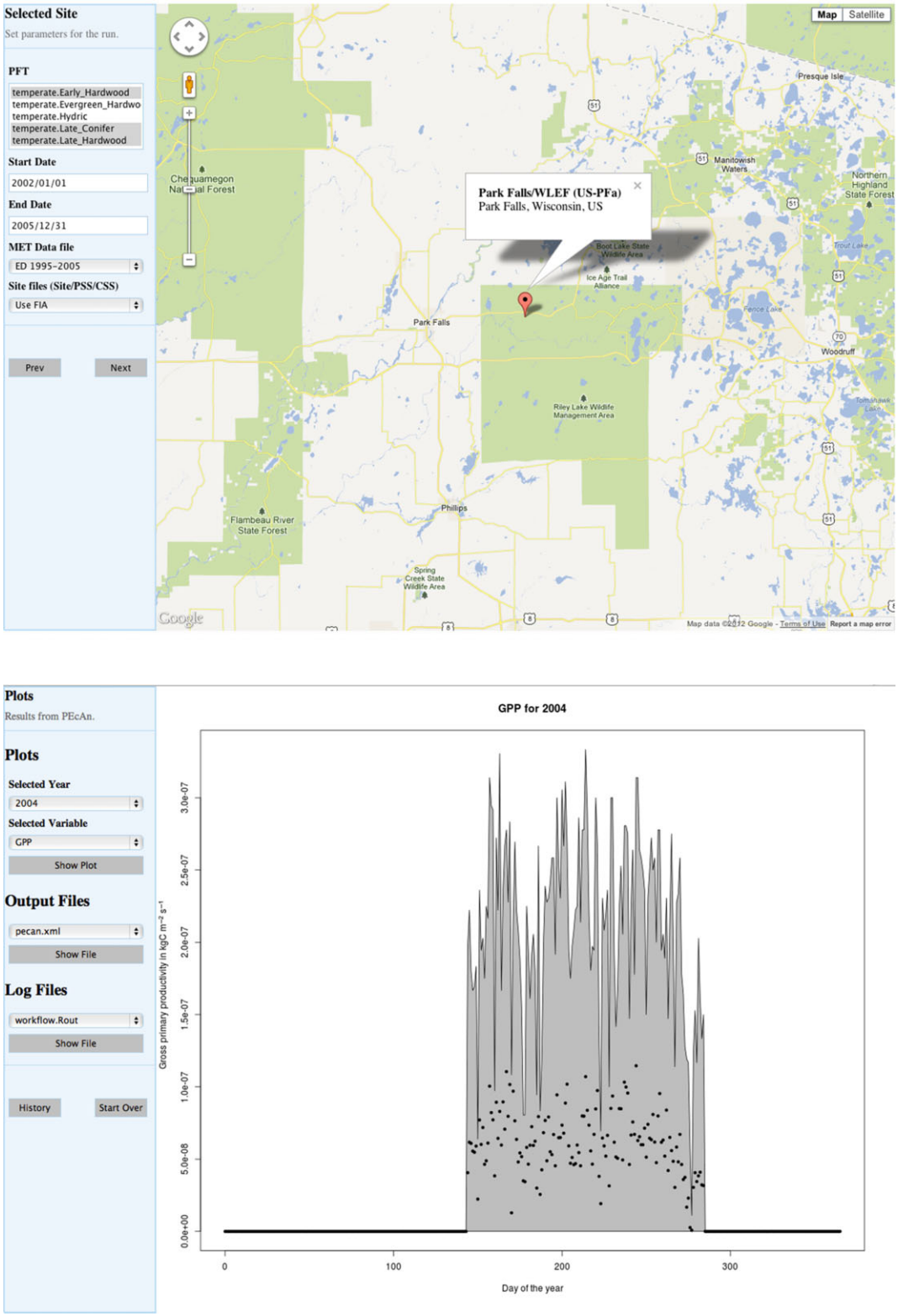

Figure 3. Sample screenshots of the PEcAn interface. The upper panel demonstrates map-based tools for setting up a model run, while the lower panel shows tools for visualization and evaluation of model performance. In this figure, the annual cycle of gross primary productivity is plotted with the grey interval showing the diurnal range and the black points the daily mean.

analyses is generally in *knowing* the probability distributions associated with each source of uncertainty (Verbeeck *et al*. 2006).

PUTTING THINGS TOGETHER: THE PREDICTIVE ECOSYSTEM ANALYSER (PEcAn)

To conclude, we will consider a case study that provides an example of how the various threads of discussion can be pulled together within an integrated system.

Understanding climate change responses requires tighter integration between modelling and data collection. However, achieving this integration requires the development of accessible tools that put the power of models in the hands of the broader research community. We initiated the PEcAn project (http://pecanproject.org) to meet the need for ecosystem modelling in general, and data assimilation in particular, to be more accessible, transparent and repeatable. PEcAn is also an attempt to begin to address and automate many of the informatics challenges that slow or impede synthesis and prediction. PEcAn is not a model itself, it is instead an ecoinformatics toolbox and a set of workflows that wrap around an ecosystem model and manage the flow of information in and out of regional-scale terrestrial biosphere models (LeBauer *et al*. 2012). PEcAn provides the user with intuitive web-based interfaces that run multiple ecosystem models, visualize model–data comparisons and interact with databases and workflows (Fig. 3). The system employs an iterative Bayesian approach to model calibration that updates prior distributions with a meta-analysis of compiled trait data. These distributions are further updated through multiple rounds of parameter data assimilation against observational data, such as eddy-covariance fluxes of carbon and water, soil respiration and forest inventories (Davidson 2012; Wang, LeBauer & Dietze 2012). PEcAn automates analyses aimed at understanding and propagating uncertainties through these models. Ensemble forecasts propagate parameter uncertainties, while sensitivity analyses and variance decomposition attribute the sources of model uncertainty to specific model processes and parameters. Power analyses and optimal design tools formally estimate the quantity and types of new data that need to be collected to most efficiently achieve a given reduction in uncertainty, including within- versus across-site sample sizes. These tools enable more effective feedbacks between models and field research. In addition to making the PEcAn source code open source, the system is also available as a fully functional virtual application that runs on a wide range of operating systems, meaning users can be up on running quickly and whole projects with their full provenance can easily be moved around, shared with colleagues and archived. The system can also interact with remote high-performance computing environments, allowing model runs to be done in parallel on remote clusters.

Looking forward, the ongoing development of PEcAn seeks to build upon the iterative nature of the Bayesian approach, and combine it with distributed computing approaches, in order to create an interconnected environment that 'learns' each time an individual user uses PEcAn to make a comparison between models and data. In keeping with the idea of model as scaffold, there has been a steady increase in the diversity of data types that can be assimilated, and moving forward there is a particular focus on remotely sensed data and the 'long tail' data provided by individual researchers. We are also steadily increasing the number of model inputs and data constraints that are managed automatically, pulling in and processing resources off the web. More ecosystem models are slowly but steadily being added to the list of models supported by the system. Ultimately, PEcAn aims to make ecosystem modelling and data assimilation routine tools for answering scientific questions and informing policy and management.

DIRECTIONS AND CONCLUSIONS

As stated in the Introduction, the goal in reviewing recent advances in model–data informatics and assimilation was to present a new perspective about how researchers *are* interacting with models, and how and why we as a research community could further reimagine this relationship. A lot of nitty-gritty work remains to be done in the process of building tools that are more accessible and capable of assimilating a wider range of data and broader range of uncertainties. There are also conceptual advances that need to occur in the repurposing of data assimilation techniques for biological problems. However, perhaps the most essential part of the work to be done is building a community approach to model–data synthesis. Biologists have long been more sceptical of models than their physical science peers, but model-driven synthesis is a critical component of conducting science and applying science to real-world problems. To build a community approach requires that empiricists play a more active role in confronting models with data, but this new role has multiple benefits in helping to refine hypotheses, support experimental design, and shorten the loop between measurement and synthesis. For this to occur, there will be a need for significant changes in how all students are trained in informatics, statistics and modelling (Fig. 1). However, it bears repeating that the training required to use these tools is different from that required for those who build them.

ACKNOWLEDGMENTS

This project was supported by funding from the National Science Foundation (Advances in Biological Informatics #10-62547 & Emerging Frontiers #10-65848) and the Energy Biosciences Institute. Ankur Desai, Joshua Mantooth, Kenton McHenry, Dan Wang, Jack Williams and an anonymous reviewer provided useful feedback for this manuscript.

REFERENCES

Anderson D., Burnham K. & Thompson W. (2000) Null hypothesis testing: problems, prevalence, and an alternative. *The Journal of Wildlife Management* **64,** 912–923.

Baraniuk R.G. (2011) More is less: signal processing and the data deluge. *Science* **331,** 717–719.

Beckage B., Gross L.J. & Kauffman S. (2011) The limits to prediction in ecological systems. *Ecosphere* **2,** art125.

Boose E.R., Ellison A.M., Osterweil L.J., Clarke L.A., Podorozhny R., Hadley J.L., Wise A. & Foster D.R. (2007) Ensuring reliable datasets for environmental models and forecasts. *Ecological Informatics* **2,** 237–247.

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

Casella G. & Berger R.L. (2001) *Statistical Inference* 2nd edn, Duxbury Press, Pacific Grove, CA.

Clark J.S. (2005) Why environmental scientists are becoming Bayesians. *Ecology Letters* **8,** 2–14.

Clark J.S. (2007) *Models for Ecological Data: An Introduction*. Princeton University Press, Princeton, NJ.

Clark J.S., Carpenter S.R., Barber M., *et al*. (2001) Ecological forecasts: an emerging imperative. *Science* **293,** 657–660.

Clark J.S., Lewis M., McLachlan J.S. & HilleRisLambers J. (2003) Estimating population spread: what can we forecast and how well? *Ecology* **84,** 1979–1988.

Clark J.S., Dietze M.C., Chakraborty S., Agarwal P.K., Wolosin M.S., Ibanez I. & LaDeau S. (2007) Resolving the biodiversity paradox. *Ecology Letters* **10,** 647–659.

Classen A.T. & Langley J.A. (2005) Data-model integration is not magic. *New Phytologist* **166,** 367–369.

Curcin V. & Ghanem M. (2008) Scientific workflow systems – can one size fit all? *2008 Cairo International Biomedical Engineering Conference* 1–9.

Davidson C.D. (2012) The modeled effects of fire on carbon balance and vegetation abundance in Alaskan Tundra. M.S. thesis, University of Illinois Urbana-Champaign.

Dietze M., Vargas R., Richardson, A.D. *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.

Dietze M.C. & Latimer A.M. (2011) Forest simulators. In *Encyclopedia of Theoretical Ecology* (eds A. Hastings & L. Gross), pp. 307–316. University of California Press, Berkeley, CA.

Ellison A., Osterweil L., Clarke L. & Hadley J. (2006) Analytic webs support the synthesis of ecological data sets. *Ecology* **87,** 1345–1358.

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

Evensen G. (2009a) The ensemble Kalman filter for combined state and parameter estimation. *IEEE Control Systems Magazine* 83–104.

Evensen G. (2009b) *Data Assimilation: The Ensemble Kalman Filter*, 2nd edn, Springer, New York.

Fawcett T.W. & Higginson A.D. (2012) Heavy use of equations impedes communication among biologists. *Proceedings of the National Academy of Sciences of the United States of America* **2012,** 1–5.

Friedlingstein P., Cox P., Betts R., Bopp L. & Von W. (2006) Carbon-cycle feedback analysis: results from the C4MIP model intercomparison. *Journal of Climate* **19,** 3337–3353.

Gao C., Wang H., Weng E., Lakshmivarahan S., Zhang Y. & Luo Y. (2011) Assimilation of multiple data sets with ensemble Kalman filter for parameter estimation and forecasts of forest carbon dynamics. *Ecological Applications* **21,** 1461–1473.

Gove J.H. & Hollinger D.Y. (2006) Application of a dual unscented Kalman filter for simultaneous state and parameter estimation in problems of surface-atmosphere exchange. *Journal of Geophysical Research* **111,** 1–21.

Hartig F., Dyke J., Hickler T., Higgins S.I., O'Hara R.B., Scheiter S. & Huth A. (2012) Connecting dynamic vegetation models to data – an inverse perspective. *Journal of Biogeography* **39,** 2240–2252.

Hey T., Tansley S. & Tolle K. (eds) (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, WA.

Hilborn R. & Mangel M. (1997) *The Ecological Detective: Confronting Models with Data*, 1st edn, Princeton University Press, Princeton, NJ.

Keenan T.F., Carbone M.S., Reichstein M. & Richardson A.D. (2011) The model–data fusion pitfall: assuming certainty in an uncertain world. *Oecologia* **167,** 587–597.

Keenan T.F., Davidson E., Moffat A., Munger W. & Richardson A.D. (2012) Using model-data fusion to interpret past trends, and quantify uncertainties in future projections, of terrestrial ecosystem carbon cycling. *Global Change Biology* **18,** 2555–2569.

LeBauer D.S. & Treseder K.K. (2008) Nitrogen limitation of net primary productivity in terrestrial ecosystems is globally distributed. *Ecology* **89,** 371–379.

LeBauer D.S., Wang D., Richter K.T., Davidson C.C. & Dietze M.C. (2012) Facilitating feedbacks between field measurements and ecosystem models. *Ecological Monographs*. doi: 10.1890/12-0137.1.

Lewis J.M., Lakshmivarahan S. & Dhall S.K. (2006) *Dynamic Data Assimilation: A Least Squares Approach*. Cambridge University Press, New York.

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

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

McMahon S.M., Dietze M.C., Hersh M.H., Moran E.V. & Clark J.S. (2009) A predictive framework to understand forest responses to global change. The year in ecology and conservation biology. *Annals of the New York Academy of Sciences* **1162,** 221–236.

Medvigy D.M., Wofsy S.C., Munger J.W., Hollinger D.Y. & Moorcroft P.R. (2009) Mechanistic scaling of ecosystem function and dynamics in space and time: ecosystem demography model version 2. *Journal of Geophysical Research* **114,** 1–21.

Medvigy D.M., Wofsy S.C., Munger J.W. & Moorcroft P.R. (2010) Responses of terrestrial ecosystems and carbon budgets to current and future environmental variability. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 8275–8280.

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

Norman D. (2002) *The Design of Everyday Things*, 2nd edn, Basic Books, New York.

Peng C., Guiot J., Wu H., Jiang H. & Luo Y. (2011) Integrating models with data in ecology and palaeoecology: advances towards a model-data fusion approach. *Ecology Letters* **14,** 522–536.

Petropoulos G., Wooster M.J., Carlson T.N., Kennedy M.C. & Scholze M. (2009) A global Bayesian sensitivity analysis of the 1d SimSphere soil–vegetation–atmospheric transfer (SVAT) model using Gaussian model emulation. *Ecological Modelling* **220,** 2427–2440.

Purves D. & Pacala S. (2008) Predictive models of forest dynamics. *Science* **320,** 1452–1453.

Quaife T., Lewis P., Dekauwe M., Williams M., Law B., Disney M. & Bowyer P. (2008) Assimilating canopy reflectance data into an ecosystem model with an ensemble Kalman filter. *Remote Sensing of Environment* **112,** 1347–1364.

Reichman O.J., Jones M.B. & Schildhauer M.P. (2011) Challenges and opportunities of open data in ecology. *Science* **331,** 703–705.

Sacks W.J., Schimel D.S., Monson R.K. & Braswell B.H. (2006) Model-data synthesis of diurnal and seasonal CO2 fluxes at Niwot Ridge, Colorado. *Global Change Biology* **12,** 240–259.

Saltelli A., Ratto M., Andres T., Campolongo F., Cariboni J., Gatelli D., Saisana M. & Tarantola S. (2008) *Global Sensitivity Analysis. The Primer*, 1st edn, John Wiley & Sons, West Sussex, UK.

Schwalm C.R., Williams C.A., Schaefer K., *et al*. (2010) A model-data intercomparison of CO2 exchange across North America: results from the North American Carbon Program site synthesis. *Journal of Geophysical Research* **115,** G00H05.

Spadavecchia L. & Williams M. (2011) Uncertainty in predictions of forest carbon dynamics: separating driver error from model error. *Ecological Applications* **21,** 1506–1522.

Trudinger C.M., Raupach M.R., Rayner P.J., *et al*. (2007) OptIC project: an intercomparison of optimization techniques for parameter estimation in terrestrial biogeochemical models. *Journal of Geophysical Research* **112,** G02027.

Verbeeck H., Samson R., Verdonck F. & Lemeur R. (2006) Parameter sensitivity and uncertainty of the forest carbon flux model FORUG: a Monte Carlo analysis. *Tree Physiology* **26,** 807–817.

Wang D., LeBauer D.S. & Dietze M.C. (2012) Predicting yields of short-rotation hybrid poplar (Populus spp.) for the contiguous US. *Ecological Applications*. doi: 10.1890/12-0854.1.

Weng E. & Luo Y. (2011) Relative information contribution of model vs. data to constraints of short- and long-term forecasts of forest carbon dynamics. *Ecological Applications* **21,** 1490–1505.

Wikle C. & Berliner L. (2007) A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena* **230,** 1–16.

Williams M., Schwarz P.A., Law B.E., Irvine J. & Kurpius M.R. (2005) An improved analysis of forest carbon dynamics using data assimilation. *Global Change Biology* **11,** 89–105.

Williams M., Richardson A.D., Reichstein M., *et al*. (2009) Improving land surface models with FLUXNET data. *Biogeosciences* **6,** 1341–1359.

Xu T., White L., Hui D. & Luo Y. (2006) Probabilistic inversion of a terrestrial ecosystem model: analysis of uncertainty in parameter estimation and model prediction. *Global Biogeochemical Cycles* **20,** 1–15.

Zobitz J.M., Moore D.J.P., Sacks W.J., Monson R.K., Bowling D.R. & Schimel D.S. (2008) Integration of process-based soil respiration models with whole-ecosystem CO2 measurements. *Ecosystems* **11,** 250–269.

Zobitz J.M., Desai A.R., Moore D.J.P. & Chadwick M.A. (2011) A primer for data assimilation with ecological models using Markov Chain Monte Carlo (MCMC). *Oecologia* **167,** 599–611.

Received 29 September 2012; received in revised form 15 November 2012; accepted for publication 18 November 2012