Oecologia (2011) 167:599–611
DOI 10.1007/s00442-011-2107-9

CONCEPTS, REVIEWS AND SYNTHESES

A primer for data assimilation with ecological models using Markov Chain Monte Carlo (MCMC)

J. M. Zobitz · A. R. Desai · D. J. P. Moore ·
M. A. Chadwick

Received: 11 January 2011 / Accepted: 5 August 2011 / Published online: 27 August 2011
© Springer-Verlag 2011

Abstract Data assimilation, or the fusion of a mathematical model with ecological data, is rapidly expanding knowledge of ecological systems across multiple spatial and temporal scales. As the amount of ecological data available to a broader audience increases, quantitative proficiency with data assimilation tools and techniques will be an essential skill for ecological analysis in this data-rich era. We provide a data assimilation primer for the novice user by (1) reviewing data assimilation terminology and methodology, (2) showcasing a variety of data assimilation studies across the ecological, environmental, and atmospheric sciences with the aim of gaining an understanding of potential applications of data assimilation, and (3) applying data assimilation in specific ecological examples to determine the components of net ecosystem carbon uptake in a forest and also the population dynamics of the mayfly (*Hexagenia limbata*, Serville). The review and examples are then used to provide guiding principles to newly proficient data assimilation practitioners.

Keywords Data assimilation · Markov Chain Monte Carlo · NEE · Aquatic insects · Ecological models

Communicated by Russell Monson.

Electronic supplementary material The online version of this article (doi:10.1007/s00442-011-2107-9) contains supplementary material, which is available to authorized users.

J. M. Zobitz (✉)
Department of Mathematics, Augsburg College,
2211 Riverside Avenue, Minneapolis, MN 55454, USA
e-mail: zobitz@augsburg.edu

A. R. Desai
Department of Atmospheric and Oceanic Sciences,
University of Wisconsin-Madison,
1225 W Dayton St, Madison, WI 53706, USA
e-mail: desai@aos.wisc.edu

D. J. P. Moore · M. A. Chadwick
Department of Geography, King's College London,
Strand, London WC2R 2LS, UK
e-mail: dave.moore@kcl.ac.uk

M. A. Chadwick
e-mail: michael.chadwick@kcl.ac.uk

Present Address:
D. J. P. Moore
School of Natural Resources and Environment,
University of Arizona, 1955 E. Sixth Street, Suite 205,
Tucson, AZ 85721, USA
e-mail: davidjpmoore@email.arizona.edu

Introduction

Ecological data are varied and complex because the relationships between living organisms and their environment are numerous and non-linear. To illuminate these complicated relationships, ecologists tend to focus their attention on model systems and on specific processes and response variables or to represent systems using mathematical models. The fusion of models with data, or data assimilation, has improved the inference that we can draw from data (Canadell et al. 2004; Doney and Ducklow 2006; Fox et al. 2009; Friend et al. 2007; Mathieu and O'Neill 2008; Olden et al. 2008; Raupach et al. 2005; Williams et al. 2009).

Data assimilation is a general term for methods that systematically combine measurements with a model, and in recent years it has been applied at a breadth of ecologically relevant scales (Table 1). These applications have included the (1) gap filling of missing measurements, (2) estimation of process model parameters or other non-observed model-derived quantities, or (3) forecasting of future measurements. While data assimilation has been extensively used to corroborate and synthesize data from large-scale ecological networks (Williams et al. 2009), it also is a tool appropriate for the analysis of smaller scale, plot-level studies (Cable et al. 2011; Ogle and Barber 2008). Table 1 in the Electronic Supplementary Material (ESM) represents a broad range of questions where data assimilation can be applied, serving as a good entry point into its use for new studies or users of data assimilation.

Four catalysts drive the expansion of data assimilation in the ecological sciences. First, the maturation of ecological datasets available from long-term investigations and emerging high-volume datasets (e.g., LTER network, FLUXNET, Baldocchi 2008; Peters et al. 2008) poses a new set of ecological questions which can be explored computationally. Second, the collation of these data into online datasets (e.g., National Center for Ecological Analysis and Synthesis, NCEAS) makes data accessible to a wide group of researchers who were not involved in the data collection process. Third, data assimilation algorithms can account for non-linearity in the model structure, reducing the need for simplifying model assumptions and making more complex synthetic questions tractable. Fourth, computational advances have allowed any ecologist with a desktop computer to efficiently process large amounts of data, making sophisticated data assimilation algorithms computationally feasible. It is likely that methods to extract information from large ecological datasets will become increasingly important for a new generation of ecologists with the increasing availability of ecological data and the establishment of the National Ecological Observatory Network (NEON; Keller et al. 2008). The network promises hundreds of new data streams measured at each of 60 ecological observatories over the next 30 years.

Data assimilation has been widespread in atmospheric sciences (Daley 1994), and researchers in hydrology have adopted similar techniques (e.g., Beven and Freer 2001). With this development, ecologists have been readily adopting data assimilation to address ecological questions (Clark 1998). While recent reviews and syntheses have highlighted early achievements and the utility of data assimilation for ecological forecasting and to address key ecological questions (Clark 2005a; Fox et al. 2009; Trudinger et al. 2007; Wang et al. 2009; Williams et al. 2009), there are still key limitations barring the widespread adoption of these techniques in ecology.

Data assimilation requires the mastery of mathematical and computational techniques coupled with a strong biological understanding of the mechanisms and processes represented in the data. New graduate students in biological sciences are not always well versed in mathematics (A'Brook and Weyers 1996; Metz 2008). Commensurate with this training gap, many reports from professional societies highlight the need to increased quantitative training and interdisciplinary collaboration between the mathematical and biological sciences (Ellison and Dennis

Table 1 Parameters, state variables, and driver variables for Example 1

Parameters, state variables, and driver variables	Description	ER or GEP model			
		Eq. 3	Eq. 4	Eq. 5	Eq. 6
Parameters					
B_R	Basal respiration rate at 10°C (μmol m^{-2} s^{-1})	•	•		
Q_{10}	Exponential temperature sensitivity	•	•		
R	Down-regulation of ER at high temperature (°C)		•		
A_{max}	Maximum photosynthetic capacity (μmol m^{-2} s^{-1})			•	•
LUE	Light use efficiency (μmol light $μmol^{-1}$ CO_2)			•	•
T_{min}	Lower limit on fractional reduction for LUE as a linear function of temperature (°C)				•
T_{max}	Upper limit on fractional reduction for LUE as a linear function of temperature (°C)				•
State variables					
ER	Whole ecosystem respiration (μmol m^{-2} s^{-1})				
GEP	Whole ecosystem primary productivity (μmol m^{-2} s^{-1})				
NEE	Net ecosystem exchange of CO_2 (μmol m^{-2} s^{-1})				
Driver variables					
T	Temperature				
I	Incoming light (μmol m^{-2} s^{-1})				

Equations 3–6 describe the functional forms utilized for the various types of gross ecosystem production and ecosystem respiration models

GEP gross ecosystem production, *ER* ecosystem respiration

2010; National Research Council 2003; Steen 2005). The presentation of data assimilation in textbooks may contain mathematical terminology, unfamiliar to ecologists, that lacks the necessary practical details to set up and apply data assimilation to an ecological problem. In addition, the data assimilation studies in the ecological literature have tended to succinctly synthesize mathematical details to allow for greater exposition of biological results. In our opinion, this discrepancy has created a gap between theory and practice, potentially stifling the application of data assimilation in ecological sciences for new users.

The objective of this paper is to describe the terminology and principles of data assimilation for an ecological audience and to provide ecologists with some basic data assimilation tools. The paper is primarily aimed at researchers relatively new to the application of data assimilation techniques; as such, it serves as a "Hitchhiker's Guide to Data Assimilation for Ecologists". We describe basic terminology commonly used in data assimilation and survey applications of data assimilation in the ecological and hydrological literature. We also illustrate how to apply data assimilation with basic examples familiar to many ecologists. In both of these examples, we highlight different aspects of data assimilation. While we exclusively use the Markov Chain Monte Carlo (MCMC) technique as our data assimilation method for both examples, the principles and analysis techniques described can be generalized to any data assimilation algorithm. The ESM provides helpful "Rules of Thumb" for data assimilation, a glossary of data assimilation terminology, a table outlining different data assimilation studies, and all data and codes used to generate the results.

Methods

Components of data assimilation

A model is a necessary component of data assimilation. Mathematically, a model describes a dependency between input and output. In this context, a model is a codified version of the conceptual relationships and dependencies within an ecological system. Inputs to a model could include *state variables*, *parameters*, and *driver variables*, whereas model outputs could be state variables or measurements. State variables are considered the model "building blocks". For models based on ecological processes, state variables could represent terrestrial pools of carbon (i.e., soil or wood), water (i.e., groundwater, snow, or precipitation), or population characteristics (i.e., number of individuals, density, mean height, etc.); these are variables that describe the state of the ecological system. Driver variables, such as temperature or the amount of incident radiation, influence the rate and nature of processes (e.g., growth, reproduction, evapotranspiration, photosynthesis, etc.) that ultimately change the state variables. Parameters control processes within a specific functional form articulated in the model (the model structure) or serve as unit conversions between different types of state variables or measurements.

The general model framework in Fig. 1 encompasses many different types of models which can be simple or complex. Ecological knowledge is included in the process primarily through the structure of the model itself. Empirical (or statistical) models have a simple structure and may not have state variables but rather relate measurements to each other. As such, statistical data fitting is one form of data assimilation. Process-based models may have more defined structures with several state variables, parameters, and measurements that may be dynamically linked through the time evolution of a differential equation. However, models are not limited to time evolution, as system state may also change through space or other dimensions. More complex models require a greater number of assumptions and typically a greater number of parameters and state variables that need to be defined, either through direct measurement or through data assimilation. Some parameters and state variables are difficult to measure directly and so must be estimated using an objective method.

Ecological information is also included in the process by incorporating prior knowledge on state variables and parameters in the data assimilation framework. Prior information may include reasonable maximum and minimum values for parameters and state variables, designed to ensure that the model cannot generate impossible ecological predictions, or the likely distribution of parameter or state variable values, which indicates some degree of understanding of the ecological system. The use of Bayesian approaches in ecology is driven by a need to systematically incorporate all possible information when pursuing a problem, and this is accomplished by including prior information directly into the data assimilation scheme

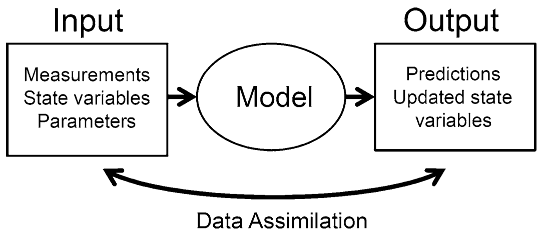

Fig. 1 Conceptual model framework for data assimilation. All models require inputs, including measurements of driver variables, estimates of state variables which define the starting state of the system, and parameters which govern how model processes respond to driver variables. Data assimilation is the systematic combination of measurements with the model to estimate parameters, forecast state variables, or gap-fill missing measurements

(Jaynes 2003; Tarantola 2005). This approach differs from frequentist data assimilation methods (e.g., least squares, maximum likelihood), which rely completely on measurements to estimate state variables and parameters. Generally speaking, as the uncertainty of the prior information increases (e.g., wide ranges for parameter values or non-informative prior values), parameter estimates derived from Bayesian results will match estimates derived from frequentist approaches. In other words, when we have a poor understanding of the likely parameter values, we can do no better than to estimate these from the data at hand. Whether wide or narrow priors are included, specific techniques or algorithms are applied to objectively compare the model with observations. Ultimately, what this implies is that data assimilation can be used to test complex hypotheses in a more nuanced way. Instead of merely falsifying a hypothesis (in the form of a model), a Bayesian approach can tell us how "likely" or "probable" a set of hypotheses are given data.

Information from the data is included in the simplest framework by the repeated estimation of the difference between the model and the data by means of a "cost function" or "merit function". Evaluating the cost function allows us to determine how well the current set of parameters and state variables derived from the model agree with the observations.

While such a framework could be solved analytically for simple models (e.g., maximum likelihood for linear models), it becomes prohibitive to do so for complex, non-linear models that are common to ecological systems. In these cases, computational approaches are taken which utilize concepts from probability, such as random walks, likelihood analysis, and probability density functions (PDFs), to estimate the final result uncertainty and confidence intervals (Clark 2005b). Here, the likelihood function is very similar to the cost function. Random samples are taken through parameter or state space, conditioned on prior probability, and tested against the cost function. Prior probabilities are typically assumed to be uniform within a range defined by previous literature or observations, although such assumptions are not required. Posterior probability is then built from the likelihood function. Mathematically, it can be shown that with sufficient sampling of the prior space, the posterior PDF will converge on the global optimum.

Non-linear or complex cost functions of models with many parameters require more computational time. The cost-function evaluation time can be directly proportional to the number of observations. Likewise, the computational time needed is exponentially dependent on the number of parameters. Large, complex models with many observations or parameters may require hours of days of computational time, although increasing access to clusters and high-performance computing in the ecological sciences are rapidly increasing our ability to use these algorithms in high-volume data and complex models that were previously intractable for data assimilation.

Two broad classes of data assimilation algorithms used to find the most likely solution to a problem or the best fit for a comparison between model predictions and measurements are sequential and non-sequential algorithms, respectively. Both algorithms make multiple comparisons between observations and model predictions to estimate the most likely values for parameters or state variables. First, *non-sequential* or *batch* data assimilation algorithms assimilate all measurements at the same time. These algorithms are typically chosen under the assumption that the measurements are independent of one another or there is spatial or temporal autocorrelation in the likelihood function. Such algorithms reduce the computational complexity because all of the data (as a batch) can be used at once to estimate parameters or state variables. Non-sequential algorithms use least squares (Clark 2005b; Sokal and Rohlf 1995) and the MCMC technique (Metropolis et al. 1953) to tune the model parameters or state variables based on comparisons between model predictions and *all* of the data for every iteration of the algorithm. Generally speaking, these algorithms are easier to computationally implement; the disadvantage is that they may require more model evaluations and computational time. *Sequential* data assimilation algorithms, such as the Kalman filter (Evensen 2009), compare a subset of the data sequentially during each comparison with the model until all of the data has been considered. Sequential techniques are used when measurements are not independent of one another (e.g., time series data), when parameter or state variables are allowed to vary through time, or when there is spatial or temporal correlation in the likelihood function.

Data assimilation algorithm utilized

As previously described, the computational time needed to sufficiently explore a parameter space can become prohibitive as the number of observations and parameters increase. For this primer, the data assimilation algorithm used is the MCMC technique, which is a variation of the Metropolis algorithm (Metropolis et al. 1953). The MCMC algorithm is one of a suite of data assimilation techniques. Because of the applicability of MCMC across the biological, mathematical, and physical sciences and its rich history, MCMC has become a universal algorithm (Richey 2010).

To briefly describe the algorithm, we assume that a given measurement x is modeled by a process with unknown parameters. As described in our examples below, the measurement might be ecosystem carbon uptake or the length–frequency distribution of a mayfly population.

Given a collection of N measurements, we define the cost or likelihood function, L:

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\eta_i)^2}{2\sigma^2}}. \qquad (1)$$

In Eq. 1, x_I represents the Ith of N measurements, η_I is the model-derived estimate of a measurement, and σ represents the uncertainty of the data about the model. The "Π" symbol denotes the product of all measurements (unlike "Σ" for summation), generated from assuming that all our measurements are independent. This likelihood function assumes that all errors are normally distributed and that the standard deviation σ follows a uniform distribution, although these assumptions can be modified to accommodate other distributions. In practice, to avoid issues with numerical precision and to increase computation speed, Eq. 1 is written as a sum function by taking the natural logarithm of both sides and using standard logarithm algebraic identities:

$$\mathrm{LL} = -N \cdot \ln(\sqrt{2\pi}) - N \cdot \ln(\sigma) - \sum_{i=1}^{N} \frac{(x_i - \eta_i)^2}{2\sigma^2}. \qquad (2)$$

Then the cost function is interpreted in terms of "log likelihood" or LL.

For parameter estimation, the MCMC technique iteratively changes the parameter values until the log likelihood is a global optimum, and the term Markov Chain implies that the current value of the parameters during the iteration is determined solely by the previous value of the parameter. Additional information about this algorithm can be found in Hurtt and Armstrong (1996) and Braswell et al. (2005). We include the sample computer code implementing this algorithm for all examples in the ESM.

An output of the MCMC method is the set of accepted parameter values and the PDF. These PDFs can be quite revealing in terms of the nature of the parameters and provide a way to estimate the uncertainty of the accepted parameter or state set and to generate confidence intervals of parameters and model output. An important point to make here is that the accepted "global optimum" and posterior "accepted" parameter sets represent the PDF of the most likely parameters or state *consistent* with the prior, observations, model structure, and cost function provided. Presumably, different priors, observations, cost function, or model structure are likely to lead to different accepted states, and hypotheses can often be tested by examining the sensitivity of the data assimilation to these choices (e.g., model selection, observational system design, or sampling biases).

The Metropolis–Hastings MCMC algorithm uses three techniques to increase the speed of optimization, described more fully in Braswell et al. (2005). First, the speed can be increased through local optimization by random walk. The parameter estimation proceeds by randomly changing a parameter within the current parameter space such that the new parameter space is *near* this original point, based on a distance criterion (usually changing one parameter by some random amount within an arbitrary range), and re-evaluating the cost function. If the new point produces a higher likelihood, then it is accepted and this "chain" moves to that point. A random walk through parameter space is then accomplished by extending the chain until both a sufficient number of samples have been made so that the percentage of accepted parameter sets reaches a threshold. Given these criteria, a local optimum has likely been found, but there is no guarantee that it is the global optimum.

To solve this problem of local versus global optima, multiple chains are initialized from different points in the parameter or state space, and the chain with the best final likelihood is selected as the best chain. The posterior PDF of accepted values is then built by evaluating the cost function at random points in the neighborhood of the chain end. For complex models, a "burn-in" period is sometimes applied, which rejects a certain fraction of these neighborhood explorations before "accepting" points for the posterior PDF as a way to ensure that the optimum found is a true global optimum.

A second approach to increase optimization speed is the random relaxation of the acceptance criteria. In this situation, the algorithm will randomly accept worse model–data fits in proportion to the evaluated likelihood. This approach also helps to avoid the problem of the algorithm converging on a local optimum (which would prevent the discovery of the "true" optimum).

Third, the distance or "jump" that a parameter changes in the random walk can also be modified. This approach decreases the number of cost-function evaluations away from the optimum value, reducing computational time. When the random walk is initially started, the jump should be relatively large to allow for large departures from the current randomly selected point. The jump distance should decrease further into the chain in proportion to the fraction of accepted parameter sets since the algorithm is more likely to be close to the "true" optimum value. This process is called tempering, in reference to the process of glass making, where large changes occur at a "high" temperature, and smaller changes are made as the temperature cools.

Collectively, these three techniques (random walk, relaxation of acceptance criteria, and adjustment of the jump distance) improve the speed at which MCMC converges on the global optimum and limits false convergence to local optima. The MCMC algorithm is one of many Bayesian parameter or state estimation algorithms. Various flavors of MCMC have different acceptance criteria.

MCMC without the acceptance of worse fits is sometimes called Gibbs sampling.

The advantages of the MCMC technique are that it does not require computation of derivatives of the model with respect to state or parameters, as is needed for some gradient descent methods, nor does it require computation of a backwards-in-time or adjoint model, as is needed for variational assimilation. Also, MCMC is easy to implement and works with virtually any model. However, a particular issue with MCMC is the need for a significant number of model evaluations, often on the order of 10^5–10^6 depending on the number of parameters. This high number of evaluations makes MCMC difficult to efficiently parallelize with computers. MCMC is not an appropriate data assimilation technique for models that have an evaluation time on the order of a second or a minute.

Other estimators are better designed to work with models where such a number of evaluations would be prohibitive. Models sometimes can only be evaluated up to a certain time (e.g., a forecast model) or exhibit extreme sensitivity to initial conditions (such as chaos in Lorenz 1963), both of which require careful evaluation of state space. It is these instances where sequential assimilation, such as Ensemble Kalman filtering or variational assimilation (both commonly used in atmospheric sciences), or other forms of batch assimilation, such as stratified random sampling, and hierarchical modeling can be a more proper assimilation technique (Ogle and Barber 2008). The appropriate choice of data assimilation framework is as important as the choice of model or model inputs. Here, we focus on two ecological examples where MCMC would be an appropriate choice.

Results and discussion

Example 1: determining the temperature sensitivity of ecosystem carbon respiration

The global carbon cycle is driven by transformations and exchanges of carbon between the ocean, atmosphere, and terrestrial ecosystems. Terrestrial carbon cycling is driven by processes of gross ecosystem photosynthesis (GEP) and total ecosystem respiration (ER). These two key components of the terrestrial carbon cycle have a demonstrated latitudinal variation and response to climate variation (Janssens et al. 2001; Litton et al. 2007; Nemani et al. 2003; Piao et al. 2008; Piovesan and Adams 2000; Schimel et al. 2001; Solomon et al. 2007; Valentini et al. 2000). The net ecosystem exchange of carbon (NEE) is an observation that represents the sum of both GEP and ER. Long-term, continuous, half-hourly measurements of NEE (the eddy–covariance technique; see Wofsy et al. 1993) have been ongoing and part of a global monitoring network, with 500 sites and approximately 40 million observations of half-hourly biosphere–atmosphere carbon exchange, with a few sites providing over 15 years of continuous data, and many nearly a decade (Eugster et al. 2000; Saigusa et al. 2002; Wofsy et al. 1993).

Flux partitioning refers to a suite of techniques that estimate GEP and ER from NEE measurements (Desai et al. 2008). A commonly used technique is statistical parameter estimation of NEE using climatic variables, such as temperature, light, and moisture. (Reichstein et al. 2005; Yi et al. 2004). Because photosynthesis is a light-dependent reaction, the measurement of NEE during nighttime periods should represent only ecosystem respiration, which is assumed to be exponentially dependent on air temperature, as in Lloyd and Taylor (1994):

$$\mathrm{ER} = B_{\mathrm{R}} Q_{10}^{(T-10)/10} \tag{3}$$

where B_{R} is the basal respiration rate at 10°C (μmol m^{-2} s^{-1}), Q_{10} is the exponential temperature sensitivity (unitless), and T is the measured temperature in degrees Celsius. This equation assumes a constant temperature sensitivity of respiration independent of temperature. Additional work has shown that this "activation energy" of respiration decreases with increasing temperature (Lloyd and Taylor 1994). A modification of this equation, rewritten in a form similar to the above equation, is:

$$\mathrm{ER} = B_{\mathrm{R}} Q_{10}^{309\left(\frac{1}{R}-\frac{1}{T-227}\right)}. \tag{4}$$

In Eq. 4, B_{R} and Q_{10} remain the same, but we have added a third parameter, R, to represent down-regulation of ER at high temperature. Using these empirical relationships and fitted parameters, then for a given half-hour during the day, ER is determined from air temperature measurements, and GEP is determined as the difference between measured daytime NEE and ER.

So a simple question might be: which one of these models is more appropriate given observations of NEE, and given so, what is the temperature sensitivity of respiration? However, it would be an oversimplification to simply fit nocturnal NEE to these two models, because information on the diurnal cycle of ER may also be present in the daytime, and the climate sensitivity of ER during the day may not be the same as that during the night (e.g., autotrophic respiration is known to vary with GEP; see Davidson et al. 2006; Zobitz et al. 2007). So a model of GEP could be added, and once again, there are two competing models to best explain the light-response of GEP. The first is a hyperbolic Michaelis–Menten light use equation:

$$\mathrm{GEP} = -A_{\max}\frac{I\cdot \mathrm{LUE}}{I\cdot \mathrm{LUE} + A_{\max}} \tag{5}$$

where I is incoming light (μmol light m^{-2} s^{-1}), $A_{\max}$ is a parameter that represents maximum photosynthetic capacity (μmol CO_2 m^{-2} s^{-1}), and LUE is the light use efficiency (μmol CO_2 μmol^{-1} light). The sign conventions are made to be consistent with our observations of NEE and are negative for carbon uptake and positive for emission into the atmosphere. This model only contains the light-limitation of photosynthesis. Temperature is also known to limit photosynthesis, which can simply be included by modifying LUE by the fractional multiplier $g(T)$ in the prior equation, based on the Monteith relationship (Heinsch et al. 2006):

$$g(T) = \begin{cases} 0 & T < T_{\min} \\ \frac{T - T_{\min}}{T_{\max} - T_{\min}} & T_{\min} \le T \le T_{\max} \\ 1 & T_{\max} < T \end{cases} \tag{6}$$

where two new parameters $T_{\max}$ and $T_{\min}$ reflect a fractional reduction for LUE as a linear function of temperature T. Arguably, the models represented in Eqs. 3–6 are simplistic, but they will serve as an example here, with applications to more complex models covered in the literature (e.g., Sacks et al. 2006).

Bayesian data assimilation is well suited to model selection, especially among those models with different numbers of parameters. Here, we have two ER models with two or three parameters (B_R, Q_{10}, and R), and GEP models with two or four parameters ($A_{\max}$, LUE, $T_{\max}$, $T_{\min}$). Therefore, we could construct four NEE models that have four, five, six or seven parameters (Table 1). All things being equal, a model with more parameters generally will be a better fit to data than one with fewer (e.g., the canonical case of polynomial fitting). However, with the interest here being the nature of these parameters (e.g., the value of Q_{10}), we want to find the most parsimonious model that explains observations. MCMC can be used to estimate parameters and confidence intervals for all combinations of the models, and a Bayesian "information criteria" metric (Schwartz 1978) can help select the best set of models among the four, namely the one which is most consistent with the data, while using the minimal number of parameters. We can then be satisfied that the parameters derived from this model can be analyzed without concern of overfitting.

Given the large amount of NEE data available, we can also use another technique to minimize overfitting, which is to stratify the data into fitting and prediction sets. Once the parameters are estimated via MCMC from one set of the data, the final likelihood and information criteria calculations can be performed with the unassimilated data. Here, we will use a 10-year half-hourly dataset of NEE from a flux tower in the Rocky Mountains (Monson et al. 2002). In this example, we will use the one growing season (May–September 2007) of data for model fitting and the following growing season (May–September 2008) for prediction. We will rely on investigator screening flags and only incorporate "good" observed data, not gap-filled or inferred data, nor observations that are contaminated by other processes (e.g., cold air drainage). More sophisticated methods involving adding a weighting term to the cost function based on data quality, or directly using an estimate of observational uncertainty (Gaussian or otherwise) instead of the model-observed residual variance in the cost function shown in Eq. 1. For the purpose of this example, we will assume errors are Gaussian and in proportion to the variance of the model-observed residual. In this case, this is a decent assumption given the large amount of data (high N); however, in real experiments, this assumption must be examined for all data assimilation problems.

For priors, we will assume uniform weighting and rely on literature estimates of the likely range of these values (ESM Table 2). Generally, the literature estimates are further broadened within limits of reality. Broad, uniform priors are the best course of attack when no other information is available. The disadvantage to using broad priors is that the number of reliably estimated parameters decreases. This trade-off occurs because the broader the priors, the greater the chance of finding multiple parameter set combinations that optimize the cost function, a problem known as equifinality (Luo et al. 2009). The example here converges relatively quickly using six chains started randomly within the prior ranges, assuming uniform distribution, checking for convergence after every 10,000 steps, with a final stage with 70,000 burn-in steps, and up to 80,000 accepted parameter sets.

The results reveal that all four combinations of the equations (Table 2) are able to adequately capture a significant fraction of the observed variability in the data ($p < 0.001$). The correlation coefficient of model NEE compared to measured NEE, r^2, ranges from 0.73 to 0.80 against the withheld data ($N = 3{,}515$). Graphically, all four models generally replicate most of the pattern of daily NEE in the prediction data (Fig. 2), although Models 3 and 4 are better able to capture the variability of NEE in the early part of the data record than the other two models. A number of high nocturnal NEE values, reflecting high respiration, are not simulated by any model, suggesting that other mechanisms (including observational error) explain these.

In parameter space (ESM Table 2), posterior estimates of LUE and $A_{\max}$ are essentially the same given posterior ranges of accepted parameters for all four models, although in the temperature-limited GEP model (Model 4), both are slightly higher. The major difference is on posterior Q_{10} and

Table 2 Results for data assimilation of the net ecosystem exchange of carbon

Model[a]	Eqs.	p	r^2	LL	BIC
1	3 and 5	4	0.74	−10,564	21,160
2	4 and 5	5	0.73	−10,603	21,246
3	**3 and 6**	**6**	**0.80**	**−10,338**	**20,726**
4	4 and 6	7	0.80	−10,362	20,782

p, Number of parameters for the model (see Table 1 for a description of the parameters); r^2, correlation coefficient from 1:1 model to data comparisons of net ecosystem exchange of carbon (NEE); LL, log likelihood defined by Eq. 2; BIC, Bayesian Information Criterion (defined by Eq. 7)

[a] The model with the lowest BIC is said to have the greatest support from the data and is given in Table 2 in bold

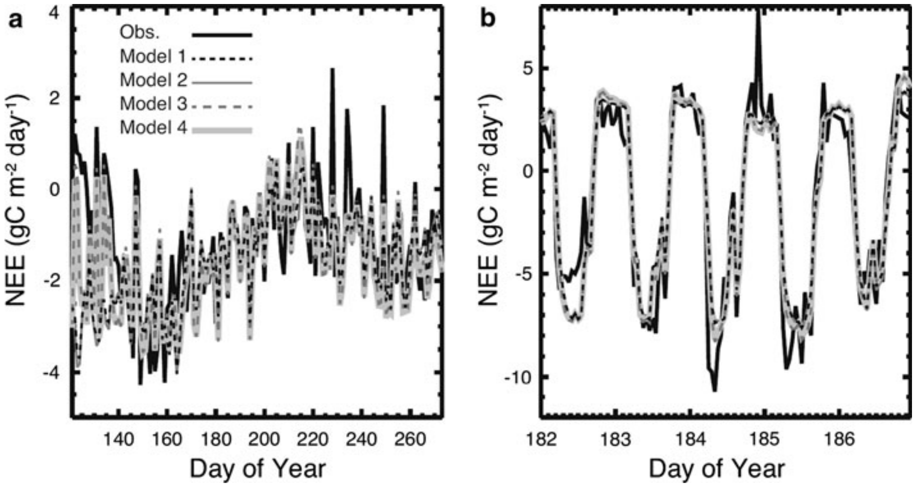

Fig. 2 Results from Markov Chain Monte Carlo (MCMC) optimization of Example 1. **a** Daily net ecosystem exchange of carbon (*NEE*) from observations and the four models utilized from Eqs. 3–6 (see Table 1), **b** half-hourly NEE for a 5-day period in July. The same legend is used in both plots

B_R, with posterior Q_{10} being significantly larger in both ER models when temperature limitation is applied to GEP. This finding illustrates the nature of model parameters being closely tied to model structural formulation, and, as such, the definition of "Q_{10}" in one model is not necessarily the same as that in another, calling into question traditional methods of parameterizing models against field data.

Given the range of accepted posterior parameter sets, it is reasonable to ask if correlations between parameters explain some of the spread in optimal values of Q_{10} (Fig. 3). Some amount of equifinality is present in all four models for Q_{10} and B_R (Fig. 3), showing that estimations of the temperature sensitivity of respiration will vary depending on the expectation of B_R, and this range can be quite large. The models show a range of correlations: over-constrained (Model 1, whose posterior ranges are very small), unconstrained for one parameter (Model 2), strongly correlated parameters (Model 3), and well constrained (Model 4). Parameter correlation plots are a powerful feature of Bayesian data assimilation and suggest here that fixing B_R (or formulating a model for B_R), especially in the case of a weak constraint or strong correlation, would improve representation (reduce posterior range) for Q_{10}. Another approach would be to use tighter prior ranges on Q_{10} and B_R, with other field data or a literature review.

Finally, as previously mentioned, Bayesian metrics can be used to select the "optimal" model among the four. Here the likelihood values calculated by Eq. 1 can be used to define "information criteria" that are derived from theoretical arguments on information entropy and data compression in computer sciences (Akaike 1974; Burnham and Anderson 2002; Konishi and Kitagawa 2008; Schwartz 1978). These criteria assess the tradeoff between number of parameters (p) and goodness of fit. It should be noted that these are not absolute quantitative metrics, but rather a way to weight the relative value of different models. There is no one correct information criterion, and there are a number of justifications for choosing one over the other, depending on the question being asked—a topic beyond the scope of this paper (see Burnham and Anderson 2002; Johnson and Omland 2004; Konishi and Kitagawa 2008). Some of these include the Bayesian Information Criterion (BIC), Deviance Information Criterion (DIC), and the Aikake Information Criterion (AIC). Here we will use the BIC, which is defined as:

$$\mathrm{BIC} = -2 \cdot \mathrm{LL} + p \cdot \ln(N). \qquad (7)$$

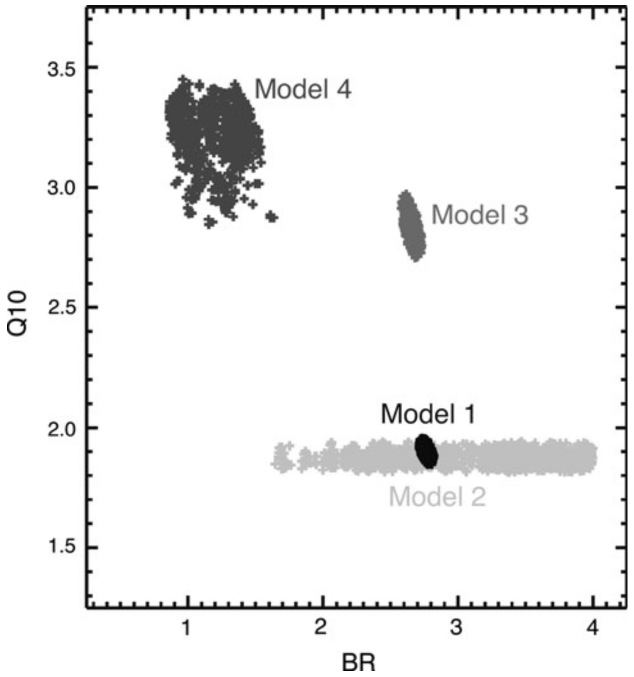

Fig. 3 Parameter space plots of basal respiration rate at 10°C (B_R; μmol m^{-2} s^{-1}) versus exponential temperature sensitivity (Q_{10}; no units) for the four models utilized in Example 1

Thus, given the estimate of log likelihood LL from the MCMC algorithm, the number of free parameters in the model (which ranges from 4 to 7), and the number of observed data points N in the likelihood function (3,515, although in reality, this should be corrected for temporal autocorrelation of half-hourly flux data), we can compute the BIC and compare the relative values. The model with the lowest BIC is the one that is most parsimonious with the data with respect to explaining variance with the minimal number of parameters. In our example, Model 3 (Eqs. 3 and 6) has the lowest BIC, suggesting that high temperature limitation for ER is not supported by the data (Eq. 4), while temperature limitation to photosynthesis (Eq. 5) is supported by the data given the observations. Of course, for a more realistic case, multiple sets of models and larger datasets should be tested. Nevertheless, both the issues of caution (parameter equifinality) and experimental design (data stratification, multiple model testing) are well described by this simple example.

A benefit of MCMC with this example is the ability to investigate predicted values of GEP, ER, and NEE and their respective uncertainties arising from parameter uncertainties. Figure 4 shows the predicted GEP and ER and NEE derived from Model 3, the model with the lowest BIC. We represent the results as cumulative GEP and ER and NEE for the 2008 growing season. To investigate prediction uncertainties, we used the entire set of accepted parameter values derived from MCMC (the number of accepted parameter sets for this model was 34,731). At a given timestep there is a distribution of GEP, ER, and NEE values from which we can determine the mean (solid lines in Fig. 4) and 95% confidence intervals (shaded regions in Fig. 4) of the predicted quantities. Figure 4 shows that the confidence interval increases throughout the growing season, attributable to the range in our accepted parameters. The computed uncertainties on NEE, GEP, and ER, which reflect both errors in observation and model parameters, provides a quantitative way to test for statistical difference against other estimates of these values (e.g., biometric upscaling, other sites).

Example 2: predicting mayfly emergence with an age-structured model

A more sophisticated example of data assimilation can be used in the case of population forecasting. Growth rates and, in turn, time to emergence for aquatic insects depends in part on the thermal conditions experienced at the larval stage (Sweeney and Vannote 1978). In North America, the burrowing mayfly, *Hexagenia limbata*, Serville, is found from Manitoba to Florida, with adult emergence occurring in late spring and summer. Along this latitudinal gradient, time to emergence, ranging from 1 to 4 years, is strongly

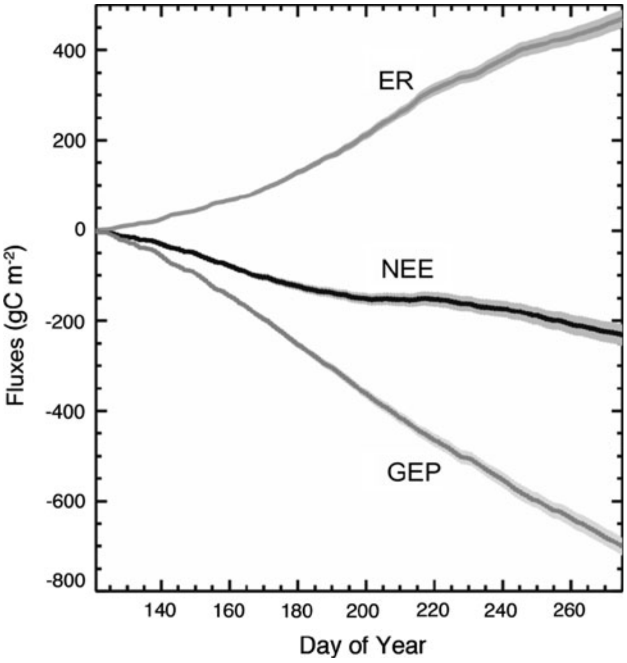

Fig. 4 Cumulative gross ecosystem production (*GEP*), ecosystem respiration (*ER*), and NEE for Model 3 using the entire accepted parameter set from the MCMC algorithm for the 2008 growing season at the Niwot Ridge AmeriFlux site. *Lines* Mean of fluxes when Model 3 is run forward with each set of accepted parameters (the number of accepted parameter sets was 34,731), *shading* 95% confidence intervals for GEP, NEE, or ER

correlated with latitude and the accumulation of degree days between minimum and maximum temperature thresholds (Chadwick and Feminella 2001). By that assuming growth rates are mainly driven by temperature, we can predict the likelihood of adult emergence via estimates of larval size driven by the accumulation of growing degree days.

We present a model for the length distribution of a mayfly population. We utilized temperature and mayfly length data collected at the Lower Mobile River, Alabama, USA from a study done by Chadwick and Feminella (2001). The time measurement is months. The dataset collected is frequency of observed mayflies by size class. To model how this distribution changes over time and its sensitivity to climate, we construct an individual population growth and mortality model using a randomly derived population of mayflies conditioned on the observed initial frequency. The starting state for this model is 10,000 mayflies of random size, such that the size distribution matches the first month's observation. From this state, the model predicts growth and mortality for each mayfly over the ensuing 11 months of the year and constructs frequency distributions for each month, as described in more detail below. These distributions are compared to the observed distribution beyond the first month (since this was used for the initial state) in the likelihood function.

We generated a representative mayfly population from length distribution results reported in Chadwick and Feminella (2001) (see Fig. 5 in Chadwick and Feminella 2001). Assuming an initial mayfly population of 10,000 individuals, we assigned a mayfly to a length class by comparing a randomly generated number to the probability of a mayfly being in that size class. We added random noise to each mayfly's length to account for length variation other than the size class mean. Using this approach, we computationally generated a mayfly population (Fig. 5) that follows a length distribution approximately equal to results reported by Chadwick and Feminella (2001).

A population of mayflies at a given time will have a distribution of lengths because of different emergence times and different sizes at emergence. However, for each mayfly we assume that mayfly length is a function of temperature, given by the following differential equation:

$$\frac{dl}{dt} = \alpha \cdot I(\mathrm{MDT}) \tag{8}$$

where l (mm) is the length of the mayfly, α is the mayfly growth rate (mm/day), and I(MDT) is a length increment function dependent on the MDT (mean daily temperature in degrees Celsius) given by the following step function:

$$I(\mathrm{MDT}) = \begin{cases} 1 & T_1 \leq \mathrm{MDT} \leq T_2 \\ 0 & \text{otherwise} \end{cases}. \tag{9}$$

Parameters T_1 and T_2 (both degrees Celsius) are the lower and upper temperature thresholds for growth, respectively.

If we assume that for a given month m there are D days where the length increment function is non-zero, then the length for a single mayfly is the following:

$$l_m = l_{m-1} + \alpha \cdot D. \tag{10}$$

Finally we assume that the mayfly population has a relative mortality rate δ (% mayflies month^{-1}). In our model this is implemented stochastically by randomly removing no greater than δ percentage of the total mayfly population each month.

The two temperature thresholds (T_1 and T_2), the growth rate (α), and the relative mortality rate (δ) are parameters which may be estimated by constraining the model using field observations of nymph length and measurements of MDT (°C) (see Table 3). ESM Fig. 1 shows mean daily temperature for 9 months starting from October 1995 in the Lower Mobile River Alabama, from data presented in Chadwick and Feminella (2001). The temperature data are used as driver variables for the mayfly length model in Eqs. 8–10. As described above, we simulated 80 length measurements (ten length classes across 8 months of reported temperature data). The MCMC routine maximizes the likelihood for model parameters against the length distribution presented in Fig. 5.

We estimated parameters for two cases, one with mortality estimated by MCMC ($\delta \neq 0$, ranging from 0 to 1) and one with mortality fixed to zero ($\delta = 0$), in addition to the other three parameters (Table 4). In both cases, the model is able to explain a major portion (>80%) of the variance in observed frequency distribution. Both models have a similar likelihood and BIC, although the mortality model has a slightly higher fit, likelihood, and lower BIC, indicating that the model with mortality is more appropriate, as would be expected. Graphically (Fig. 5), both models (gray bars) are able to simulate well the frequency distribution observed (black bars), although the mortality model (Fig. 5a) is better able to simulate the size distribution in the later months, during mayfly emergence, than the no mortality model (Fig. 5b).

However, the mortality model is clearly the more "reasonable" model when the accepted parameters are examined (ESM Table 3); the no mortality model has a very broad posterior parameter set in α, ranging from 0.1 to 1, indicating that the actual value for α for this model is highly uncertain. This model also shows a strong correlation between α and T_1, reflected in the broad posterior range for the latter. Finally, the range between T_1 and T_2 is exceptionally narrow (<1°C). When $\delta = 0$, we do not explicitly model mayfly mortality. In order to maximize the log likelihood function, the MCMC method forces the

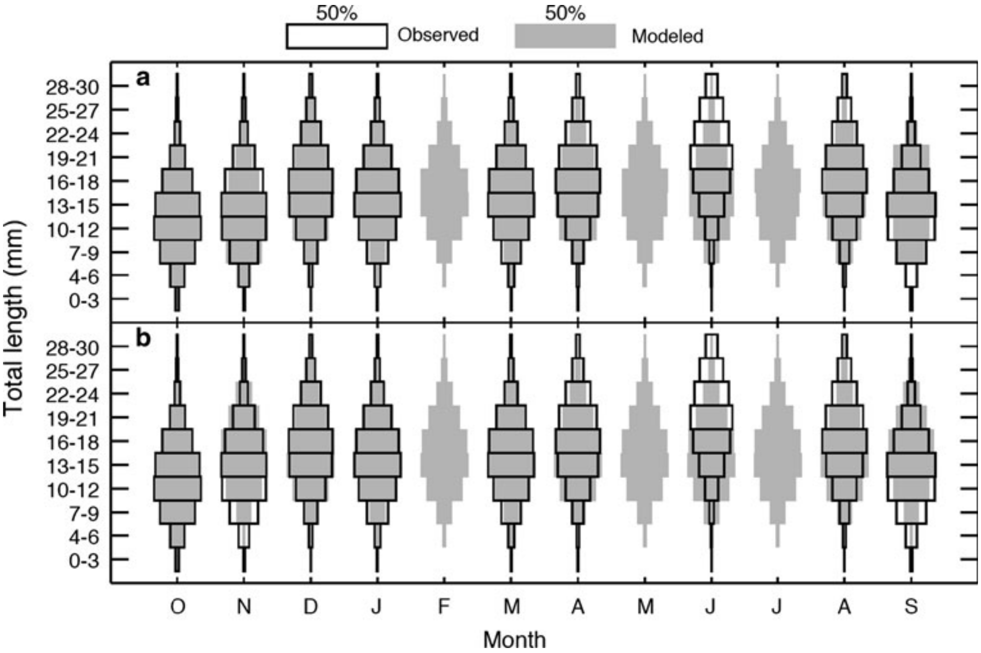

Fig. 5 Frequency distribution of mayfly population by 3-mm bin size class of observations (*black bars*) versus model (*gray bar*) run with mortality (**a**) and without mortality (**b**)

Table 3 Parameters, state variables, and driver variables for the mayfly population model described in Example 2

Parameters, state variables, and driver variables	Description
Parameters	
T_1	Lower limit on mayfly temperature range for growth (°C)
T_2	Upper limit on mayfly temperature range for growth (°C)
α	Mayfly growth rate (mm day^{-1})
δ	Relative mayfly mortality rate (% mayflies month^{-1})
State variables	
l	Mayfly length (mm)
Driver variables	
MDT	Mean daily temperature (°C)

temperature parameters to a biologically unreasonable narrow range. The lesson here is that when a model has a clear structural flaw (e.g., that mayflies do not die), the other parameters being optimized "compensate" for this deficiency. In contrast, the mortality model has parameters that are better constrained, have less correlation, and reflect reasonable values.

More complex models that utilize a wider range of state variables, parameters, and driver variables (e.g., recruitment, sex-biased mortality, size-specific growth rates, water and food quality, etc.) could be easily envisaged; however, this temperature-based model is a useful tool for simulating mayfly length-frequency and predicting emergence. The extension of this model to forecast or predict the climate sensitivity of mayfly distribution and emergence is discussed in the ESM. From these results, we suggest a new hypothesis to test, namely, that there is a latitudinally specific, optimal temperature for mayfly growth, and that when conditions are beyond these thresholds, mayfly emergence would be predicted to decline. Additional observations or manipulative experiments could then be used to confirm or falsify this hypothesis.

Table 4 Results for data assimilation for the mayfly population models (Example 2)

Experiment[a]	p	r^2	LL	BIC
Mortality	**4**	**0.85**	**47**	**−77**
No mortality	3	0.82	36	−59

r^2, Correlation coefficient from 1:1 model to data comparisons of monthly mean mayfly length; all other abbreviations are as given in the footnote to Table 2

[a] The first experiment is with mortality being a free parameter (δ ranging from 0 to 1), and the second experiment is with mortality fixed to zero ($\delta = 0$), The model with the lowest BIC is said to have the greatest support from the data and is given in Table 4 bold

Acknowledgments The authors would like to thank T. Quaife for helpful discussions, students of the Annual Summer Course in Flux Measurements and Modeling (NSF 0634649), and anonymous reviewers. ARD acknowledges funding support from Department of Energy (DOE) Office of Biological and Environmental Research (BER) National Institute for Climatic Change Research (NICCR) Midwestern Region Subagreement 050516Z19.

References

A'Brook R, Weyers J (1996) Teaching of statistics to UK undergraduate biology students in 1995. J Biol Educ 30(4):281–288

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19(6):716–723. doi:10.1109/TAC.1974.1100705

Baldocchi D (2008) "Breathing" of the terrestrial biosphere: lessons learned from a global network of carbon dioxide flux measurement systems. Aust J Bot 56(1):1–26

Beven K, Freer J (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J Hydrol 249(1–4):11–29. doi:10.1016/S0022-1694(01)00421-8

Braswell BH, Sacks WJ, Linder E, Schimel DS (2005) Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. Glob Change Biol 11(2):335–355. doi:10.1111/j.1365-2486.2005.00897.x

Burnham KP, Anderson DR (eds) (2002) Model selection and multimodel inference. Springer, New York

Cable JM, Ogle K, Lucas RW, Huxman TE, Loik ME, Smith SD, Tissue DT, Ewers BE, Pendall E, Welker JM, Charlet TN, Cleary M, Griffith A, Nowak RS, Rogers M, Steltzer H, Sullivan PF, van Gestel NC (2011) The temperature responses of soil respiration in deserts: a seven desert synthesis. Biogeochemistry 103:71–90. doi:10.1007/s10533-010-9448-z

Canadell J, Ciais P, Cox P, Heimann M (2004) Quantifying, understanding and managing the carbon cycle in the next decades. Clim Change 67(2):147–160. doi:10.1007/s10584-004-3765-y

Chadwick Ma, Feminella JW (2001) Influence of salinity and temperature on the growth and production of a freshwater mayfly in the Lower Mobile River, Alabama. Limnol Oceanogr 46(3):532–542

Clark JS (1998) Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. Am Nat 152(2):204–224

Clark J (2005a) Why environmental scientists are becoming Bayesians. Ecol Lett 8:2–14. doi:10.1111/j.1461-0248.2004.00702.x

Clark JS (2005b) Models for ecological data: statistical computation for classical and Bayesian approaches. Princeton University Press, Princeton

Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century, National Research Council (2003) BIO2010: transforming undergraduate education for future research biologists. The National Academies Press, Washington, DC

Daley R (1994) Atmospheric data analysis, cambridge atmospheric and space science series. Cambridge University Press, New York

Davidson EA, Janssens IA, Luo Y (2006) On the variability of respiration in terrestrial ecosystems: moving beyond Q10. Glob Change Biol 12:154–164. doi:10.1111/j.1365-2486.2005.01065.x

Desai AR, Richardson AD, Moffat AM, Kattge J, Hollinger DY, Barr A, Falge E, Noormets A, Papale D, Reichstein M, Stauch VJ (2008) Cross-site evaluation of eddy covariance GPP and RE decomposition techniques. Agric For Meteorol 148(6–7):821–838. doi:10.1016/j.agrformet.2007.11.012

Doney S, Ducklow H (2006) A decade of synthesis and modeling in the US Joint Global Ocean Flux Study. Deep Sea Res (2 Top Stud Oceanogr) 53(5–7):451–458. doi:10.1016/j.dsr2.2006.01.019

Ellison AM, Dennis B (2010) Paths to statistical fluency for ecologists. Front Ecol Environ 8(7):362–370. doi:10.1890/080209

Eugster W, Rouse WR, Pielke RA Sr, Mcfadden JP, Baldocchi DD, Kittel TGF, Chapin FS III, Liston GE, Vidale PL, Vaganov E, Chambers S (2000) Land-atmosphere energy exchange in Arctic tundra and boreal forest: available data and feedbacks to climate. Glob Change Biol 6:84–115. doi:10.1046/j.1365-2486.2000.06015.x

Evensen G (2009) Data assimilation: the ensemble Kalman filter, 2nd edn. Springer, New York

Fox A, Williams M, Richardson AD, Cameron D, Gove JH, QuaifeT, Ricciuto D, Reichstein M, Tomelleri E, Trudinger CM, Van Wijk MT (2009) The REFLEX project: comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data. Agric For Meteorol 149(10):1597–1615. doi:10.1016/j.agrformet.2009.05.002

Friend AD, Arneth A, Kiang NY, Lomas M, Ogée J, Rödenbeck C, Running SW, Santaren JD, Sitch S, Viovy N, Woodward FI, Zaehle S (2007) FLUXNET and modelling the global carbon cycle. Glob Change Biol 13(3):610–633. doi:10.1111/j.1365-2486.2006.01223.x

Heinsch FA, Zhao M, Running SW, Kimball JS, Nemani RR, Davis KJ, Bolstad PV, Cook BD, Desai AR, Ricciuto DM, Law BE, Oechel WC, Kwon H, Luo H, Wofsy SC, Dunn AL, Munger JW, Baldocchi DD, Xu L, Hollinger DY, Richardson AD, Stoy PC, Siqueira MBS, Monson RK, Burns SP, Flanagan LB (2006) Evaluation of remote sensing based terrestrial productivity from MODIS using regional tower eddy flux network observations. IEEE Trans Geosci Remote Sens 44(7):1908–1925. doi:10.1109/TGRS.2005.853936

Hurtt GC, Armstrong RA (1996) A pelagic ecosystem model calibrated with BATS data. Deep Sea Res (2 Top Stud Oceanogr) 43:653–683

Janssens IA, Lankreijer H, Matteucci G, Kowalski AS, Buchmann N, Epron D, Pilegaard K, Kutsch W, Longdoz B, Grünwald T, Montagnani L, Dore S, Rebmann C, Moors EJ, Grelle A, Rannik Ü, Morgenstern K, Oltchev S, Clement R, Guðmundsson J, Minerbi S, Berbigier P, Ibrom A, Moncrieff J, Aubinet M, Bernhofer C, Jensen NO, Vesala T, Granier A, Schulze ED, Lindroth A, Dolman AJ, Jarvis PG, Ceulemans R, Valentini R (2001) Productivity overshadows temperature in determining soil and ecosystem respiration across European forests. Glob Change Biol 7:269–278

Jaynes ET (2003) Probability theory: the logic of science. Cambridge University Press, Cambridge

Johnson JB, Omland KS (2004) Model selection in ecology and evolution. Trends Ecol Evol 19(2):101–108. doi:10.1016/j.tree.2003.10.013

Keller M, Schimel DS, Hargrove WW, Hoffman FM (2008) A continental strategy for the National Ecological Observatory Network. Front Ecol Environ 6(5):282–284. doi:10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2

Konishi S, Kitagawa G (2008) Information criteria and statistical modeling. Springer, New York

Litton CM, Raich JW, Ryan MG (2007) Carbon allocation in forest ecosystems. Glob Change Biol 13(10):2089–2109. doi:10.1111/j.1365-2486.2007.01420.x

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

Lorenz E (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141

Luo Y, Weng E, Wu X, Gao C, Zhou X, Zhang L (2009) Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. Ecol Appl 19(3):571–574. doi:10.1890/08-0561.1

Mathieu P, O'Neill A (2008) Data assimilation: from photon counts to Earth System forecasts. Remote Sens Environ 112(4):1258–1267. doi:10.1016/j.rse.2007.02.040

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21(6):1087–1092. doi:10.1063/1.1699114

Metz AM (2008) Teaching statistics in biology: using inquiry-based learning to strengthen understanding of statistical analysis in biology laboratory courses. Cell Biol Educ 7(3):317–326. doi:10.1187/cbe.07-07-0046

Monson RK, Turnipseed AA, Sparks JP, Harley PC, Scott-Denton LE, Sparks K, Huxman TE (2002) Carbon sequestration in a high-elevation, subalpine forest. Glob Change Biol 8:459–478

Nemani RR, Keeling CD, Hashimoto H, Jolly WM, Piper SC, Tucker CJ, Myneni RB, Running SW (2003) Climate-driven increases in global terrestrial net primary production from 1982 to 1999. Science 300(5625):1560–1563. doi:10.1126/science.1082750

Ogle K, Barber JJ (2008) Bayesian data—model integration in plant physiological and ecosystem ecology. Prog Bot 69:281–311. doi:10.1007/978-3-540-72954-9_12

Olden JD, Lawler JJ, Poff NL (2008) Machine learning methods without tears: a primer for ecologists. Q Rev Biol 83(2):171–193

Peters DP, Groffman PM, Nadelhoffer KJ, Grimm NB, Collins SL, Michener WK, Huston MA (2008) Living in an increasingly connected world: a framework for continental-scale environmental science. Front Ecol Environ 6(5):229–237. doi:10.1890/070098

Piao S, Ciais P, Friedlingstein P, Peylin P, Reichstein M, Luyssaert S, Margolis H, Fang J, Barr A, Chen A, Grelle A, Hollinger DY, Laurila T, Lindroth A, Richardson AD, Vesala T (2008) Net carbon dioxide losses of northern ecosystems in response to autumn warming. Nature 451(7174):49–52. doi:10.1038/nature06444

Piovesan G, Adams JM (2000) Carbon balance gradient in European forests: interpreting EUROFLUX. J Veg Sci 11(6):923–926. doi:10.2307/3236563

Raupach MR, Rayner PJ, Barrett DJ, DeFries RS, Heimann M, Ojima DS, Quegan S, Schmullius CC (2005) Model-data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications. Glob Change Biol 11(3):378–397. doi:10.1111/j.1365-2486.2005.00917.x

Reichstein M, Falge E, Baldocchi D, Papale D, Aubinet M, Berbigier P, Bernhofer C, Buchmann N, Gilmanov T, Granier A, Grünwald T, Havránková K, Ilvesniemi H, Janous D, Knohl A, Laurila T, Lohila A, Loustau D, Matteucci G, Meyers T, Miglietta F, Ourcival JM, Pumpanen J, Rambal S, Rotenberg E, Sanz M, Tenhunen J, Seufert G, Vaccari F, Vesala T, Yakir D, Valentini R (2005) On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. Glob Change Biol 11:1424–1439. doi:10.1111/j.1365-2486.2005.001002.x

Richey M (2010) The evolution of Markov Chain Monte Carlo methods. Am Math Monthly 117(5):343–383. doi:10.4169/000298910X485923

Sacks WJ, Schimel DS, Monson RK, Braswell BH (2006) Model-data synthesis of diurnal and seasonal CO_2 fluxes at Niwot Ridge, Colorado. Glob Change Biol 12:240–259. doi:10.1111/j.1365-2486.2005.01059.x

Saigusa N, Yamamoto S, Murayama S, Kondo H, Nishimura N (2002) Gross primary production and net ecosystem exchange of a cool-temperate deciduous forest estimated by the eddy covariance method. Agric For Meteorol 112:203–215. doi:10.1016/S0168-1923(02)00082-5

Schimel DS, House JI, Hibbard KA, Bousquet P, Ciais P, Peylin P, Braswell BH, Apps MJ, Baker D, Bondeau A, Canadell J, Churkina G, Cramer W, Denning AS, Field CB, Friedlingstein P, Goodale C, Heimann M, Houghton RA, Melillo JM, Moore B III, Murdiyarso D, Noble I, Pacala SW, Prentice IC, Raupach MR, Rayner PJ, Scholes RJ, Steffen WL, Wirth C (2001) Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. Nature 414(6860):169Gǘ172. doi:10.1038/35102500

Schwartz G (1978) Estimating the dimensions of a model. Ann Stat 6(2):461–464

Sokal R, Rohlf J (1995) Biometry. W. H. Freeman & Co, New York

Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (2007) 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

Steen LA (2005) Math & bio 2010: linking undergraduate disciplines. Mathematical Association of America (MAA), Washington DC

Sweeney BW, Vannote RL (1978) Size variation and the distribution of hemimetabolous aquatic insects: two thermal equilibrium hypotheses. Science 200(4340):444–446. doi:10.1126/science.200.4340.444

Tarantola A (2005) Inverse problem theory and model parameter estimation. SIAM Books, Philadelphia

Trudinger CM, Raupach MR, Rayner PJ, Kattge J, Liu Q, Pak B, Reichstein M, Renzullo L, Richardson AD, Roxburgh SH, Styles J, Ping Wang YP, Briggs P, Barrett D, Nikolova S (2007) OptIC project: an intercomparison of optimization techniques for parameter estimation in terrestrial biogeochemical models. J Geophys Res 112(G2). doi:10.1029/2006JG000367

Valentini R, Matteucci G, Dolman AJ, Schulze ED, Rebmann C, Moors EJ, Granier A, Gross P, Jensen NO, Pilegaard K, Lindroth A, Grelle A, Bernhofer C, Grünwald T, Aubinet M, Ceulemans R, Kowalski AS, Vesala T, Rannik Ü, Berbigier P, Loustau D, Guethmundsson J, Thorgeirsson H, Ibrom A, Morgenstern K, Clement R, Moncrieff J, Montagnani L, Minerbi S, Jarvis PG (2000) Respiration as the main determinant of carbon balance in European forests. Nature 404:861–865. doi:10.1038/35009084

Wang Y-P, Trudinger CM, Enting IG (2009) A review of applications of model-data fusion to studies of terrestrial carbon fluxes at different scales. Agric For Meteorol 149(11):1829–1842. doi:10.1016/j.agrformet.2009.07.009

Williams M, Richardson AD, Reichstein M, Stoy PC, Peylin P, Verbeeck H, Carvalhais N, Jung M, Hollinger DY, Kattge J, Leuning R, Luo Y, Tomelleri E, Trudinger CM, Wang YP (2009) Improving land surface models with FLUXNET data. Biogeosciences 6(7):1341–1359. doi:10.5194/bg-6-1341-2009

Wofsy SC, Goulden ML, Munger JW, Fan SM, Bakwin PS, Daube BC, Bassow SL, Bazzaz FA (1993) Net exchange of CO_2 in a mid-latitude forest. Science 260:1314–1317

Yi C, Li R, Bakwin PS, Desai A, Ricciuto DM et al (2004) A non-parametric method for separating photosynthesis and respiration components in CO_2 flux measurements. Geophys Res Lett 31. doi:10.1029/2004GL020490

Zobitz J, Burns S, Ogee J, Reichstein M, Bowling R (2007) Partitioning net ecosystem exchange of CO_2: a comparison of a Bayesian/isotope approach to environmental regression methods. J Geophys Res–Biogeosciences 112(G3). doi:10.1029/2006JG000282