ECOLOGY LETTERS

Ecology Letters, (2011) **14**: 522–536 doi: 10.1111/j.1461-0248.2011.01603.x

REVIEW AND SYNTHESIS

Integrating models with data in ecology and palaeoecology: advances towards a model–data fusion approach

Changhui Peng,[1,2,3]* Joel Guiot,[2] Haibin Wu,[4] Hong Jiang[5,6] and Yiqi Luo[7]

Abstract

It is increasingly being recognized that global ecological research requires novel methods and strategies in which to combine process-based ecological models and data in cohesive, systematic ways. Model–data fusion (MDF) is an emerging area of research in ecology and palaeoecology. It provides a new quantitative approach that offers a high level of empirical constraint over model predictions based on observations using inverse modelling and data assimilation (DA) techniques. Increasing demands to integrate model and data methods in the past decade has led to MDF utilization in palaeoecology, ecology and earth system sciences. This paper reviews key features and principles of MDF and highlights different approaches with regards to DA. After providing a critical evaluation of the numerous benefits of MDF and its current applications in palaeoecology (i.e. palaeoclimatic reconstruction, palaeovegetation and palaeocarbon storage) and ecology (i.e. parameter and uncertainty estimation, model error identification, remote sensing and ecological forecasting), the paper discusses method limitations, current challenges and future research direction. In the ongoing data-rich era of today's world, MDF could become an important diagnostic and prognostic tool in which to improve our understanding of ecological processes while testing ecological theory and hypotheses and forecasting changes in ecosystem structure, function and services.

Keywords

Carbon cycle, data assimilation, earth system modelling, ecological forecasting, global climate change, inverse modelling, palaeoclimatic reconstruction, sequential data assimilation, variational data assimilation.

Ecology Letters (2011) **14**: 522–536

INTRODUCTION

Ecology and palaeoecology are two fields of study that have become data-rich enterprises due to the rapid development of numerous global research networks (e.g. FLUXNET and BIOME6000), multi-sensor remote-sensing data (e.g. MODIS and Landsat) and the long-term accumulation of data through research network project initiatives (LTER). The enormous amount of available data offers tremendous opportunity to improve ecological model simulations by applying data assimilation (DA) and inverse modelling techniques. Data assimilation is the process of incorporating observations into a forecast model over a period of time to create an estimate of the system state (such as the state of the atmosphere or the biosphere). Data assimilation techniques have undergone continual development in the last 50 years. This is particularly true with respect to meteorology (Appendix S1) where numerical weather prediction (NWP) models have been constructed from the assimilation of satellite, atmospheric and surface observational data, which has led to dramatic improvements in forecasting techniques (Cressman 1959; Gandin 1963; Lorenc 1981; Evensen 2007; Lorenc & Payne 2007). Ecological models have concurrently undergone a conscientious development towards a more mechanistic, comprehensive and complex structure to prepare them for DA application (Williams *et al.* 2009; Wu *et al.* 2009). Inverse modelling (based on estimation theory) is a statistical technique that can be used to estimate parameters that are directly or indirectly related to the measured quantity. An inverse model differs from simulation (or forward) modelling in that it uses observed properties (e.g. carbon fluxes) to constrain physical or biological processes rather than using physics or biology to predict property distribution. Model–data fusion (MDF) (also called 'model and data integration' or 'model–data synthesis') includes both model inversion techniques and DA. The aim of the MDF approach is to improve the observational constraint of a model

[1]*Laboratory for Ecological Forecasting and Global Change, College of Forestry, Northwest A & F University, Yangling, Shaanxi 712100, China*
[2]*ECCOREV FR 3098, CNRS/Aix-Marseille Université, BP 80, 13545 Aix-en-Provence Cedex 4, France*
[3]*Department of Biology Sciences, Institute of Environment Science, University of Quebec at Montreal, Montreal, QC H3C 3P8, Canada*
[4]*Key Laboratory of Cenozoic Geology and Environment, Institute of Geology and Geophysics, Chinese Academy of Sciences, PO Box 9825, Beijing 100029, China*
[5]*State Key Laboratory of Subtropical Forest Science & Zhejiang Provincial Key Laboratory of Carbon Cycling in Forest Ecosystems and Carbon Sequestration, Zhejiang Agriculture and Forestry University, Hangzhou, 311300 Zhejiang, China*
[6]*International Institute for Earth System Science, Nanjing University, Nanjing 210093, Jiangsu, China*
[7]*Department of Botany and Microbiology, University of Oklahoma, Norman, OK 73019, USA*
**Correspondence: E-mail: peng.changhui@uqam.ca*

over that afforded by any single data set on its own (Raupach *et al.* 2005).

Despite three decades of advancements in ecological data acquisition, remote-sensing technology and inverse modelling there exists no single definitive data set or ecological model that can comprehensively generate all ecological products (or modelled state variables) needed for the reliable assessment of an environment and its monitoring and forecasting capacity (Clark *et al.* 2001; Luo *et al.* 2011). However, an emerging awareness is taking place in the ecology and palaeoecology communities where the integration of multiple observed data sets is starting to be perceived as a necessary step in overcoming the limitations imposed by uncertainties contained within any single data set (e.g. data gaps, biases, inaccurate processing algorithms, nonlinear dynamics and model error). Improved model forecasts require the development of innovative observational, statistical and computational techniques that optimally combine observation data sets and ecological models. Increasing demands to integrate model and data methods in the past decade has led to the development of MDF within ecology. MDF allows for the integration of multiple and different types of data (including associated uncertainties) as well as for the inclusion of prior knowledge of model parameters and/or initial state variables. It holds great promise as a tool for partitioning the photosynthetic and respiratory components of ecosystems while studying their separate response to environmental control (Wang *et al.* 2009). In addition, MDF can combine data sources into models that explicitly acknowledge sources of uncertainty. This is critical to the advancement of ecological forecasting (Clark *et al.* 2001; Luo *et al.* 2011).

Model–data fusion has played an increasingly important role not only in NWP (Lorenc 1981; Daley 1991; Kalnay 2003), oceanic sciences (Evensen 2003) and hydrologic modelling (Liu & Gupta 2007) but also in palaeoclimate, palaeovegetation and palaeocarbon reconstructions (Guiot *et al.* 2000, 2009; Wu *et al.* 2009), carbon cycle modelling (Wang *et al.* 2009), global carbon sink/source quantification through the application of atmospheric inversion (Bousquet *et al.* 2000; Gurney *et al.* 2002), earth system modelling (Mathieu & O'Neill 2008; Williams *et al.* 2009) and ecological estimation and forecasting (Cosby 1984; Luo *et al.* 2003, 2011). With increasing exigency to integrate models and data together, MDF has become more and more utilized in the fields of palaeoecology, ecology and earth system sciences. Figure 1 provides the number of publications and citations that applied MDF for global climate change ecological research from 1990 to 2009. It shows an exponentially increasing trend (almost by a factor of seven over the past 10 years) in terms of both publication and citation (Fig. 1).

This paper is not intended to be an exhaustive review of all methods and issues related to MDF reported in literature; instead, the aim was to provide neophyte readers with an illustrative and integrated overall picture of the MDF approach and how it is breaking new ground with regards to its current applications and benefits as well as its future potential within the global ecological community. Key features and principles of MDF and its various approaches are first assessed in relation to DA. After a critical evaluation of the variety of benefits of MDF and its current applications in palaeoecology (i.e. palaeoclimate, palaeovegetation and palaeocarbon storage reconstruction) and ecology (i.e. parameters and uncertainty estimations, identification of model error, remote sensing and ecological forecasting), the paper discusses its limitations, current challenges and directions for future research.

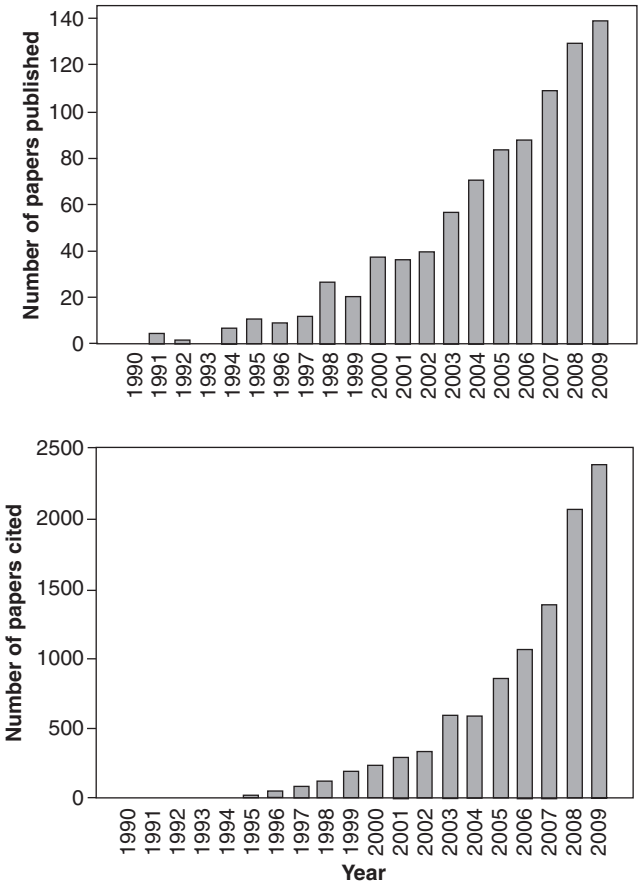

Figure 1 Number of publications and citations that used model–data fusion (MDF) for global change ecology research from 1990 to 2009. A Web of Science search was carried out using the following Boolean input keywords: ('Inversion' OR 'Data assimilation' AND 'Ecology' AND 'Global change').

KEY FEATURES, PRINCIPALS AND OPTIMIZATION METHODS OF THE MDF APPROACH

Key features and principles

Basically, MDF encompasses both DA and model inversion techniques (Raupach *et al.* 2005; Tarantola 2005; Evensen 2007). The key objective of MDF is to improve the performance of a model by either optimizing/refining the values of unknown parameters and initial state variables or by enhancing the predictive capacity of a model (state variable) according to a given data set. Notwithstanding a number of MDF approaches that have recently been reported in literature, common features among the different MDF approaches are: (1) a forward model structure that describes the temporal evolution of state variables (e.g. surface temperature, soil moisture and carbon stocks), (2) observational data that can be correlated to model results, (3) objective functions that combine model estimates and observations to any associated prior information and error structure and (4) optimization techniques that adjust forward model parameters or state variables to minimize the discrepancy between model estimates and field (or satellite) observations.

The general principle of MDF is to find an 'optimal match' between model observations by varying the properties of the model (e.g. structure, state and parameter). The 'optimal match' is a selection of model properties that minimize the gap between model system

representation and real systems based on observational and prior data. A cost function is then constructed to quantify the mismatch between model predictions and observations. To accomplish this, a common procedure (termed an M-estimator in statistics) is to have the cost function expressed as a weighted sum of the model–data mismatch. The cost function may also reflect the statistical characteristics of errors within the observations themselves (Wang *et al.* 2009). Discovering optimal parameters can help to improve predictions or test alternative hypotheses embedded into models. The Bayesian approach (a statistical inference method in which certain types of evidence or observations is used to calculate a prior probability over hypotheses) is the ultimate refinement as it provides a way to estimate uncertainties. A key feature of DA schemes is how they incorporate information that pertains to uncertainty for both the model and the observations, providing a best estimate of the true state of a system. In addition, estimates of model output uncertainty are calculated in compliance with the observed data so that model predictions are correlated to associated probabilistic density functions derived from the MDF approach. This is crucial in assessing the utility of model forecasting capacity.

Main optimization methods

The choice of optimization technique (Fig. 2) can be either batch or sequential DA functions that are built into the applications, which minimizes model and data differences throughout (or a subset of) all observations simultaneously. The search for an optimal solution as well as the estimation of uncertainties can be implemented by both batch and sequential techniques (Table 1). It all depends on whether the data are processed all at once (batch), in groups, or, potentially, even one at a time (sequential). Batch methods that include gradient-based methods as well as global search and variational DA methods (Fig. 2; Table 1) simultaneously process all data and observations. For batch techniques, the cost function is treated as a single function to be minimized.

In contrast to batch methods, sequential DA methods process data sequentially. One of the most popular examples of a sequential method is the Kalman filter (KF) (see Appendix S2) that was first introduced by Rudolf E. Kalman (1960). KF is a recursive algorithm that estimates the state of a system at each repetition using a state-space model in combination with (noisy) measurements. The objective of KF is to reduce the influence of noise that occurs in measurements. It provides a convenient representation of model error, data error and parameter error (Williams *et al.* 2005; Wang *et al.* 2009).

Kalman filter has two distinct phases: the prediction phase and the update phase. The prediction phase uses the state estimate from the previous time step to produce an estimate of the state at the point of the current time step. For the update phase, the current *a priori* prediction is combined with current observational data to refine the state estimate. This improved estimate is termed the '*a posteriori* state estimate'. Typically, the two phases alternate with the prediction, advancing the state until the next scheduled observation and the update that incorporates the observational data. A comparison between three major sequential DA methods is provided in Table 1.

The Ensemble Kalman filter (EnKF) is an extension of the traditional KF that in itself is based on Monte Carlo sampling and recursive data processing. Easy implementation of the EnKF method and its applicability to nonlinear problems has led to its extensive application in meteorology and hydrology as well as in other fields (Burgers *et al.* 1998; Reichle *et al.* 2002; Vrugt *et al.* 2005). Recently, ecologists (Williams *et al.* 2005; Fox *et al.* 2009) have begun to use EnKF to investigate problems concerning: (1) model parameter estimation (Quaife *et al.* 2008), (2) climate reconstruction (Wu *et al.* 2007a,b), (3) how to assimilate measured eddy fluxes and carbon pools into carbon cycle models (Williams *et al.* 2005; Mo *et al.* 2008), (4) wildland fire simulation and prediction (Mandel *et al.* 2008) and (5) how to estimate terrestrial water cycles on a regional scale by means of multiple satellite remote-sensing data (Pan *et al.* 2008).

RECENT DEVELOPMENTS AND APPLICATIONS OF MDF IN GLOBAL CHANGE RESEARCH

It is increasingly being recognized that global ecological research requires novel methods and strategies in which to combine process-

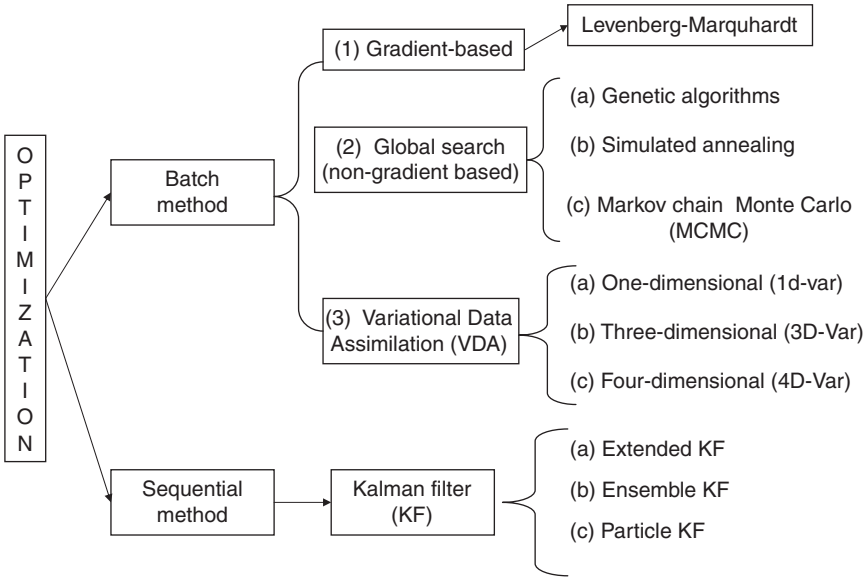

Figure 2 Optimization techniques (methods) used in model–data fusion (MDF).

Table 1 Key features and advantages of primary optimization techniques used in the MDF approach

Methods	Key features	Advantages and applications	Key references
1. Batch methods			
Gradient-based method	Levenberg-Marquardt (LM) is a gradient-descent method used for parameter estimation in nonlinear models. It provides a numerical solution to the problem of minimizing a function that is generally nonlinear over a space of parameters of the function. These minimization problems arise especially in least squares curve fittings and nonlinear programming. Gradient-based algorithms follow identified directions within the parameter space.	These methods are highly efficient but are not best suited for highly dimensional nonlinear models as they may end up discovering local rather than global minima. LM is often combined with a quasi-Monte Carlo algorithm to search for global optimal values (Luo *et al.* 2003). They require the calculation of model output sensitivity to model parameters to determine posterior uncertainty (Santaren *et al.* 2007).	Levenberg (1944), Marquardt (1963)
Non-gradient-based method	Non-gradient-based methods [also called 'global search' methods, e.g. genetic algorithms (GA), simulated annealing (SA) or Markov chain Monte Carlo (MCMC) methods] are often based on a random number generator (Braswell *et al.* 2005; Sacks *et al.* 2006), while GA and SA are typically applied to observations from a limited number of target variables. MCMC methods are a family of techniques that use Monte Carlo sampling to generate a discrete approximation of the posterior probability distribution of the parameter(s) that is/are sought for estimation. The Metropolis-Hastings sampler (Hastings 1970) is one of the more popular MCMC sampling algorithms in use.	The major advantage of global search methods is that they are able to treat all data simultaneously and, therefore, are more likely to discover the global minimum for the cost functions that possess multiple minima compared to gradient-based methods. The primary advantage of the Metropolis algorithm is to provide complete information concerning posterior distributions of parameters that can be used to generate standard errors and confidence intervals for both individual parameters and correlations between parameters. Markov chain Monte Carlo techniques iteratively produce a sample from a Bayesian posterior distribution of parameters until certain convergence criteria are met (Wang *et al.* 2009; Williams *et al.* 2009; Appendix S3).	Metropolis *et al.* (1953), Hastings (1970), Braswell *et al.* (2005), Sacks *et al.* (2006)
Variational data assimilation (VDA)	VDA methods operate in a batch processing manner over a given time window that contains a sequence of observational time points. In weather forecasting, depending on the spatial and temporal dimensions of the state variables, VDA methods can be primarily classified into three categories: one-dimensional (1D-Var), three-dimensional (3D-Var) and four-dimensional (4D-Var). For 3D-Var, only those observations available at the time of analysis were used. For 4D-Var, past observations were included and, thus, time dimension was added. 4D-Var uses the tangent-linear and adjoint versions of the forecast model to estimate the 4D atmospheric states that best fits assimilated observations distributed over a specified time window (Gauthier *et al.* 2007).	VDA methods are much less expensive to produce computationally than are KF and EKF methods. In light of this, they are preferable for data assimilation for use with realistic, complex systems (e.g. a numerical weather prediction framework). In addition, by simultaneously using observations inside the assimilation interval, VDA methods are also more optimal than KF and EKF methods inside (within) the interval (at the end of the interval). However, the VDA method itself does not provide any estimate of predictive uncertainty. The adjoint method is able to calculate exact gradient information of the objective function that is to be optimized. Important scientific advances such as 4D-Var and improvements in error specifications in combination with a large increase in available observations has led to considerable improvements in overall forecasting performance (Table S1).	Daley (1991), Kalnay (2003), Gauthier *et al.* (2007), Lorenc & Payne (2007)
2. Sequential methods [Kalman Filter (KF)]			
Extended KF	The KF is a sequential method for estimating the state of a system. EKF is the nonlinear version of the Kalman filter (Evensen 1992).	Unlike its linear counterpart, EKF is not ordinarily an optimal estimator. Another problem with EKF is that the estimated covariance matrix tends to underestimate the true covariance matrix and, therefore, risks becoming inconsistent in a statistical sense without the addition of 'stabilizing noise'. EFK can yield unstable results when the nonlinearity in a complex model is strong (Evensen 1994). The application can result in unbounded error growth as soon as a system enters an unstable regime. In addition, the enormous computational time required is a serious disadvantage of EKF. Ensemble Kalman Filter (EnKF) was introduced to overcome the drawbacks of EKF (Evensen 1994).	Kalman (1960), Evensen (1992, 1994)

Table 1 (*Continued*)

Methods	Key features	Advantages and applications	Key references
Ensemble KF	It results in the optimal estimation of strongly nonlinear dynamical systems with Gaussian probabilities (Evensen 1994).	EnKF is suitable for problems that possess a large number of variables such as the discretization of partial differential equations in geophysical models (Evensen 1994, 2007). One advantage of EnKF is that advancing the probability density function in time is achieved simply by advancing each member of the ensemble.	Evensen (2003, 2007)
Particle filtering (PF)	PF, also known as sequential Monte Carlo method and bootstrap filtering, is another commonly used data assimilation algorithm for the recursive estimation of model states (Arulampalam *et al.* 2002). It is typically used to estimate Bayesian models and is the sequential (online) analogue of the Markov chain Monte Carlo (MCMC) batch methods often similar in importance to sampling methods.	A well-designed PF can often operate much faster than MCMC. PF is typically used as an alternative to EKF with the advantage that (with sufficient numbers of samples) it approaches the Bayesian optimal estimate. It, therefore, can achieve greater accuracy compared to EKF. PF carries out updates on particle weights instead of state variables. In addition, PF has the desirable characteristic of being applicable to any state-space model in any format whether linear or nonlinear or Gaussian or non-Gaussian.	Arulampalam *et al.* (2002)

MDF, model–data fusion.

based ecological models and data in a cohesive, systematic manner. Increasing demands to integrate model and data methods in the past decade has led to MDF utilization in palaeoecology, ecology and earth system sciences (Fig. 3).

MDF applications in palaeoecology

Although palaeoecological data provides the means to understand past ecosystem dynamics and long-term climate changes by applying statistical methods (Guiot & de Vernal 2007), these methods are based on modern distributions of one or more species that reside within a climatic domain or on a statistical calibration of modern assemblages in relation to climatic variables. MDF appears to be an approach better suited to this contingency than traditional statistical methods due to its mechanistic, processed-base relationships compared to the latter's focus on simple statistical correlations. In recent years, MDF has emerged to combine environmental proxies (such as pollen, tree-rings, isotopes, etc.) and process-based dynamic vegetation models together to reconstruct palaeoenvironments.

Reconstructing palaeoclimates and palaeovegetation

It is important to note that statistical methods for reconstructing palaeoclimates are based on the assumption that plant–climate interactions remain constant through time and implicitly assume that these interactions are independent from climate forcing such as changes in atmospheric CO_2. These statistical methods are only valid for climatic niches presently realized, and their extrapolation to past conditions could prove problematic when they do not reflect present conditions (Guiot *et al.* 2009). Moreover, the methods primarily used were based on modern distributions of one or several species that reside in climatic domains or on a statistical calibration of modern assemblages in relation to climatic variables. They do not take into account the impact of atmospheric CO_2 concentrations on vegetation. Inverse vegetation modelling offers a novel approach to reconstruct palaeoclimates (Guiot *et al.* 2000; Wu *et al.* 2007a,b). Moreover, a physiologically based vegetation model can be utilized inversely for palaeoclimatic and carbon reconstructions (Fig. 4).

Guiot *et al.* (2000) developed the first palaeoclimate inverse modelling application. They estimated certain climatic variables from observed ecological data or simulated outputs. This new method involved the use of pollen data and a process-based BIOME3 vegetation model coupled to an artificial neural network both inverted by Markov chain Monte Carlo (MCMC) sampling. It provided a strategic framework and insight into how to take into account the effects of low CO_2 atmospheric concentrations to improve palaeoclimatic reconstructions. Wu *et al.* (2007a) extended the methodology to include European, African and Asian data for two historical periods in which atmospheric CO_2 concentrations were considerably different from the present day. They showed that bias could be as high as 10 °C for winter temperatures in Europe during the Last Glacial Maximum. This method has been extended to multiple proxies by Hatté & Guiot (2005) and Hatté *et al.* (2009) for reconstructing palaeoprecipitation using pollen and $\delta^{13}C$ data, a proxy strongly related to precipitation (Fig. 5). Data obtained on the Grande Pile Eemian have demonstrated that multi-constraints ascertained by means of the joint usage of pollen and carbon isotopes also reduce uncertainty in relation to precipitation reconstruction.

The two studies discussed in the above paragraph used an equilibrium vegetation model (BIOME3; Appendix S4) that accounts

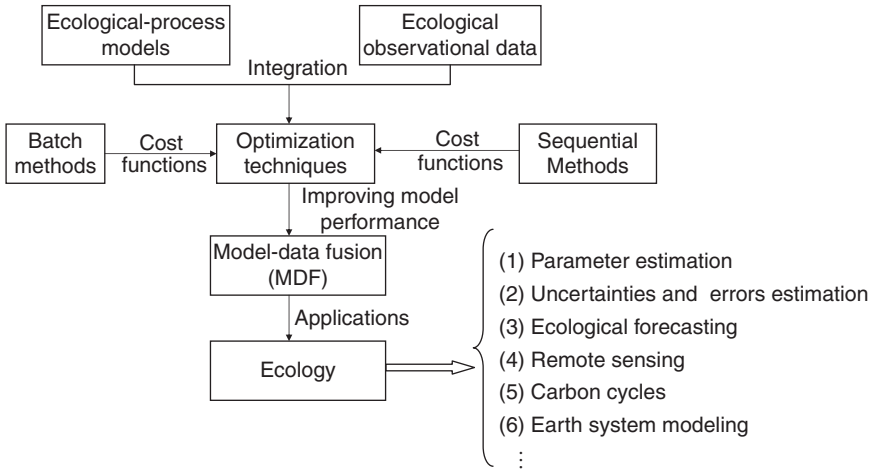

Figure 3 Overview of optimization techniques and model–data fusion (MDF) application in ecology.

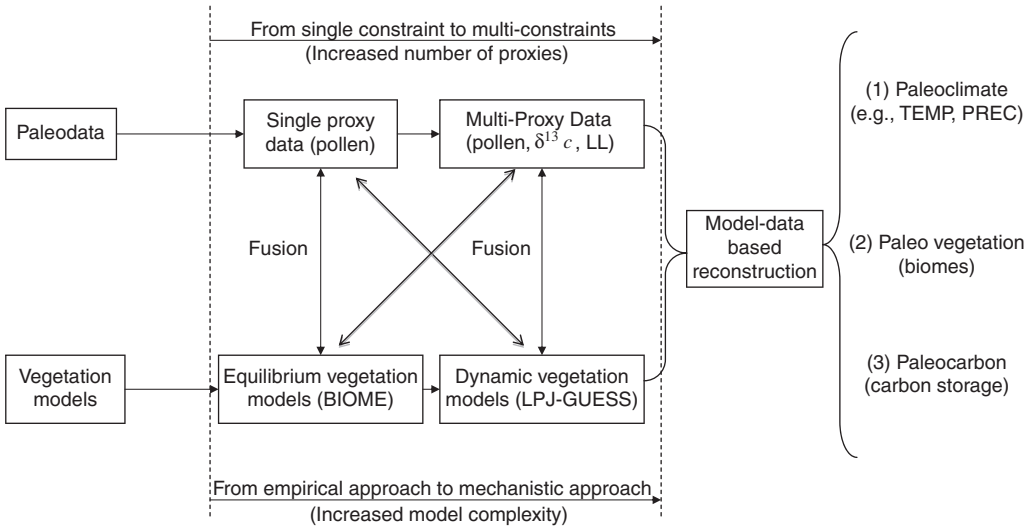

Figure 4 Development and evaluation of the model–data fusion (MDF) approach in palaeoecology fusion that includes inverse modelling and data assimilation. BIOME is an equilibrium vegetation model developed by Prentice *et al.* (1992). LPJ-GUESS is a dynamic vegetation model developed by Smith *et al.* (2001).

for processes related to carbon and water cycles but not to those related to plant competition and mortality. LPJ-GUESS (Smith *et al.* 2001), a more recent and sophisticated dynamic model, takes these processes into account. To carry out temporal inversion, Garreta *et al.* (2009), for example, developed a hierarchical Bayesian model and utilized a particle filter (PF) algorithm to the LPJ-GUESS dynamic vegetation model. The hierarchical Bayesian approach provides researchers with a classification model that takes into consideration uncertainty associated with measuring replicate samples while facilitating the probabilistic formalization of the inversion process. It offers attractive general features for palaeoclimatological research (see Haslett *et al.* 2006). This is possible due to it being a powerful tool that yields rich statistical models that more fully reflect a given problem in comparison to simple models. Guiot *et al.* (2009) recently proposed a number of prospective ideas regarding a shift from a statistical single proxy approach to a multi-proxy and dynamical approach with respect to palaeoclimate and palaeovegetation reconstructions. However, this inverse vegetation modelling approach is not a panacea. Since it is a model-based approach, it is highly dependent on the quality of the proxy model. Moreover, it requires a great deal of computational process time. Outputs of the model are not directly comparable to pollen data without first modelling pollen dispersion. Additional verification is also required by way of adapting this approach to other vegetation models. It remains important to use this approach in parallel with classical statistical methods (Guiot *et al.* 2009; Wu *et al.* 2009).

Reconstructing palaeocarbon

Methods used to reconstruct palaeocarbon storage with regards to glacial and interglacial environmental conditions can be generally classified into: (1) carbon density estimates, (2) vegetation model-based estimates that apply climate inputs and (3) inverse modelling that utilize palaeovegetation data (Peng *et al.* 1998; Wu *et al.* 2009). The carbon density method uses palaeoenvironmental proxy data (i.e. palynological, pedological and sedimentological data) to map the distribution of vegetation types and to estimate stocks by assuming that average carbon density in each biome is the same as observed conditions today (Adams *et al.* 1990; Van Campo *et al.* 1993). This

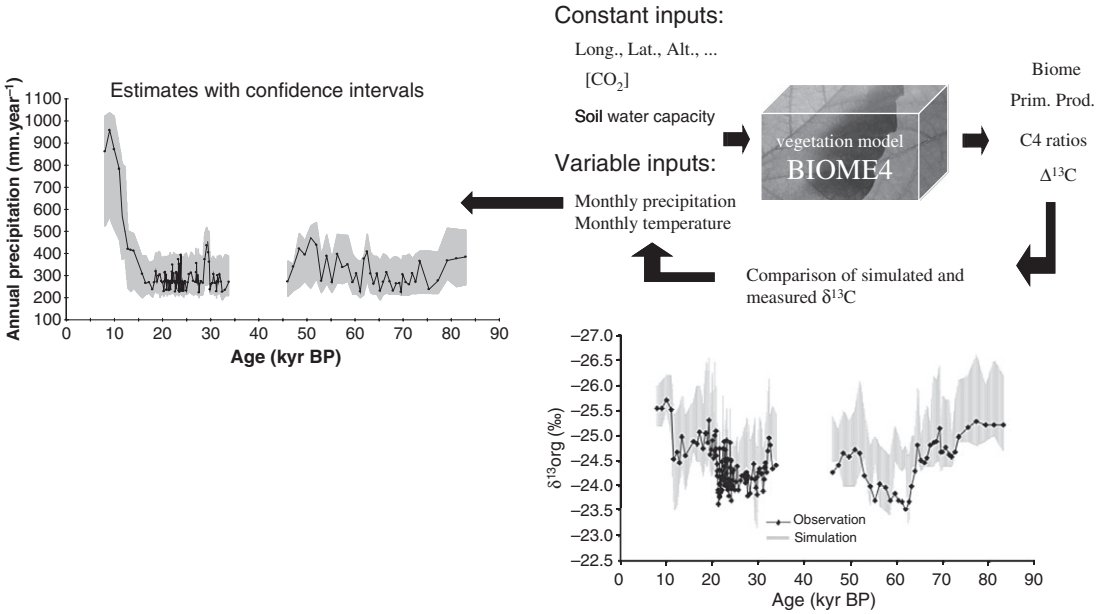

Figure 5 Estimated annual precipitation in Nussloch, Germany (Hatté & Guiot 2005), obtained by an inversion of the BIOME4 vegetation model for the 80 to 5 kyr BP period: (1) part of required inputs of the model was fixed to site values (coordinates, soil). (2) CO_2 is constant for a given time period but changes according to ice core measurements (from 190 p.p.m.v. during glacial periods to *c.* 280 p.p.m.v. during the Holocene epoch). (3) For each time period, a large number of climate scenarios were appraised and introduced into BIOME4. $\delta^{13}C$ was simulated and compared to measured values. Scenarios that provided simulated values close to the measured values were retained. They were used to define ranges of acceptable $\delta^{13}C$ (shown in grey). (4) Satisfactory scenarios were used to define reconstructed precipitation (median + 95% confidence intervals are represented by black lines and grey areas). The selection of climate scenarios is based on an Markov chain Monte Carlo (MCMC) algorithm. The number is typically a few thousand. Although the $\delta^{13}C$ output ranged from −23 to −26.5‰ (mean uncertainty range: ± 0.5‰) and was in agreement with observations ($r^2 = 0.75$, $n = 164$), certain discrepancies may be identified, particularly during the *c.* 62–58 kyr BP period in which simulations presented lower values than observations. The causes of the $\delta^{13}C$ 'mismatches' between the data and the model may arise either from the BIOME4 model itself (which assumes a constant value for incoming solar radiation) or from the isotopic data (which assumes a sampling problem as well as no apparent analogy between vegetation and soil types). The mismatch may also stem from parameters associated with fractionation.

estimate is solely dependent on vegetation or biome type and does not vary geographically for specific types under different climatic conditions and atmospheric CO_2 levels, which may lead to substantial errors during glacial–interglacial periods (Van Campo *et al.* 1993; Joos *et al.* 2004).

Wu *et al.* (2009) have recently developed a new MDF approach to estimate past biospheric carbon stocks based on the application of a new integrated ecosystem model [PaleoCarbon model (PCM)] that was built on top of a physiological process vegetation model (BIOME4) coupled with a process-based biospheric carbon model (DEMETER) (Fig. 6). The PCM was constrained to fit pollen data to obtain realistic estimates for both vegetation distribution and soil carbon storage. It was estimated that the probability distribution of climatic parameters (as simulated by BIOME4 utilizing an inverse process) was compatible with pollen data while DEMETER successfully simulated carbon storage values to corresponding outputs of BIOME4. The carbon model was validated to present day observations of vegetation biomes and soil carbon, and the inversion scheme was tested against 1491 surface pollen spectra sample sites in Africa and Eurasia. Results show that this method can successfully simulate biomes, terrestrial carbon variables and related climates for most of the selected pollen sites.

MDF application for use with parameter estimation

It has proven to be highly challenging to accurately estimate model parameters and their dynamical range regarding different spatiotemporal scales due to complex processes as well as the spatiotemporal variability found within terrestrial ecosystems. For example, terrestrial carbon cycle projections derived from multiple coupled ecosystem–climate models were at variance. Discrepancies ranged from a 10 Gt C $year^{-1}$ sink to a 6 Gt C $year^{-1}$ source by the year 2100 (Friedlingstein *et al.* 2006). The model discrepancy does not only arise from parameter uncertainties but also from the limited understanding of key ecological processes. In addition, by comparing nine process-based models applied to a Canadian boreal forest ecosystem, Potter *et al.* (2001) found that core parameter values [e.g. leaf area index (LAI), stomatal conductance and leaf nitrogen content] and their specific sensitivity to certain key environmental factors were inconsistent due to seasonal and locational variance in factors.

Parameter estimation is a major application of MDF wherein model parameters are adjusted so that model state(s) come into closer agreement with observations (Liu & Gupta 2007; Fox *et al.* 2009; Wang *et al.* 2009). Various published literature has show how inverse methods have been used to estimate and calibrate model parameters by assuming time-invariant parameters. For example, Braswell *et al.* (2005) estimated the Harvard Forest carbon cycle parameters using the Metropolis algorithm (a MCMC method). They found that most parameters were reasonably constrained and that estimated parameters can simultaneously fit both diurnal and seasonal variability patterns. With given *a priori* uncertainties, Knorr & Kattge (2005) reported that half-hourly CO_2 and water flux eddy covariance measurements could considerably reduce uncertainty in approximately five parameters in an ecosystem model. Luo *et al.* (2003) and Xu *et al.* (2006) added

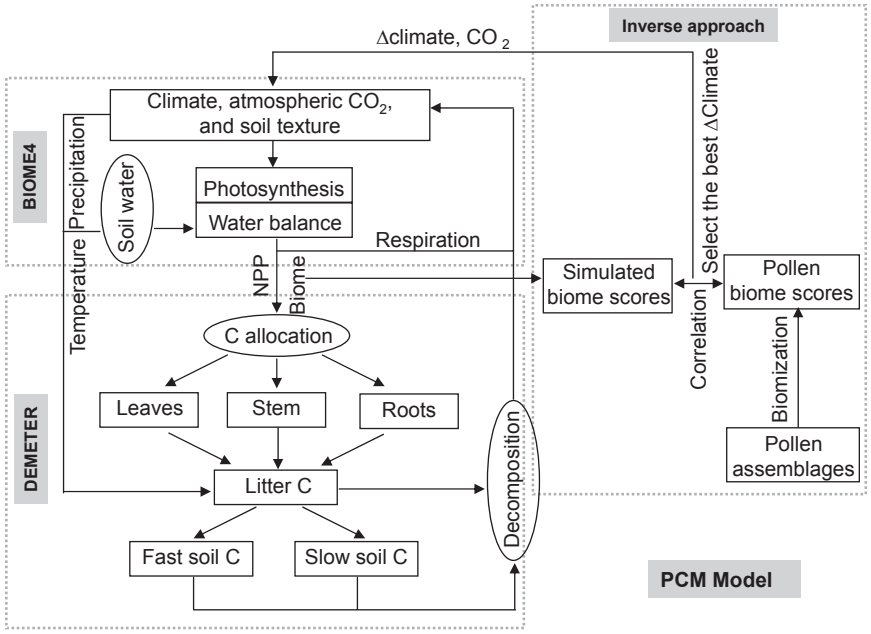

Figure 6 Schematic representation of the inverse vegetation modelling approach to past carbon storage reconstructions (adapted from Wu *et al.* 2009). PCM integrated the process-based biospheric models BIOME4 (Kaplan 2001) and DEMETER (Foley 1995) into an inverse model (Guiot *et al.* 2000). An innovative approach of the PCM model is the inversion of BIOME4 to estimate climatic variables that are introduced into DEMETER (which works in forward mode) to estimate terrestrial carbon storage. BIOME4 is a physiological process-based global model that is particularly useful when working with palaeovegetation as it operates with a limited number of inputs (monthly temperature, precipitation, sunshine, absolute minimum temperature, soil texture and atmospheric CO_2 concentrations) that are easily available. However, its equilibrium design makes it impossible to directly simulate terrestrial carbon stocks. DEMETER provides a better simulation of global and continental scale biospheric carbon storage for vegetation, litter and soil but offers only a simple potential vegetation submodel derived from the BIOME1 model (Prentice *et al.* 1992). To overcome the shortcomings of both approaches, these models were coupled with biome types and net primary production (NPP) that were calculated by BIOME4 and then used as inputs for the DEMETER model. PCM was constrained to fit pollen data to obtain realistic estimates. The principle behind the inversion method applied here was to estimate the input of BIOME4 (e.g. monthly climate) given the known data related to the output of the model, that is, the biome scores (Prentice *et al.* 1996) derived, in this case, from pollen. The reconstructed climate variables together with the vegetation parameters (NPP and biome type) were then used as an input for DEMETER to deduce the terrestrial vegetation carbon cycle.

photosynthate partitioning coefficients into plant pools (the initial values of pool size) and estimated parameters that describe carbon flow into receiving pools for inverse analysis by applying six data sets from a forest CO_2 experiment. These methods, however, do not take into account possible temporal variations that occur in model parameters, which may result in misrepresentation of certain processes. It is worth mentioning they also require a large historical data set. Moreover, it is important to point out that most MDF applications used for parameter estimation failed to further analyse and quantify systematic model error and uncertainty and their potential impacts on model parameters. The use of MDF to identify systematic model error and quantify uncertainty is discussed in the following section.

The assumption of time-invariant parameters is critical when using these inversions. It should be noted that this may not be the case for parameters that change over time. Sequential DA approaches such as recursive Bayesian estimation (Thiemann *et al.* 2001) as well as the KF and its various extensions (Trudinger *et al.* 2007; Mo *et al.* 2008; Gao *et al.* 2011) can be used to overcome these shortcomings, providing a general framework for the optimum consolidation of uncertainty in model prediction to observational data (Appendix S5). Sequential techniques have been used for the recursive estimation of time-varying parameters and predictive uncertainty in hydrological, climate and carbon cycle models. To account for seasonal parameter variation, for example, Mo *et al.* (2008) successfully used the sequential DA approach in combination with EnKF to optimize key parameters of the Boreal Ecosystem Productivity Simulator (BEPS) model, taking into account input error, parameters error and observational error. They found that sequential DA in combination with EnKF can be used to estimate the seasonal and interannual variation of parameters, and these optimized parameters can considerably improve ecosystem model accuracy. In addition, Chen *et al.* (2008) introduced a so-called smoothed ensemble Kalman filter (SEnKF) that combines EnKF with a kernel-smoothing technique to simultaneously estimate state variables and parameters of a forest carbon flux partition model (see Appendix S6). Furthermore, simultaneous parameter estimation can use near real time observations to improve the predictive capacity of dynamic models. Two recent studies compared a number of parameter estimation methods applied to common data sets and models. They highlighted the importance and difficulty in correctly representing uncertainties for both the model and parameter estimation measurements (Trudinger *et al.* 2007; Fox *et al.* 2009).

MDF application to estimate model parameter and prediction uncertainty and error

How to adequately address uncertainty associated with ecological prediction remains a critical and challenging issue despite considerable progress and recent developments in enhanced computational power

and ecological modelling sophistication. The most difficult technical challenges that remain are to understand, quantify and reduce uncertainty involved in palaeoecology and ecology modelling in a cohesive and systematic manner. Analysis of uncertainty and model performance are an integral part of any MDF application in quantifying and reducing model uncertainty (Wang *et al.* 2009; Williams *et al.* 2009). Estimating parameter uncertainty and corresponding model output uncertainty is an unrivaled benefit offered by most MDF approaches. Indeed, the outcome of the MDF approach is a set of parameter probability distribution functions that can be used to generate an assemblage of model runs using a time series of the climate forcing data. For example, Fox *et al.* (2009) compared various algorithms that estimate carbon model parameters that are consistent with both measured carbon flux and states and results acquired from a simpler carbon model. Their analysis indicates that the incorporation of additional constraints and the application of data to carbon pools (wood, soil and fine roots) can help to reduce uncertainties in model parameters poorly served by eddy covariance data. Moreover, a number of uncertainty analysis frameworks have been developed and reported on in atmospheric and oceanic sciences (e.g. Daley 1991) as well as in hydrological modelling (Liu & Gupta 2007). These methods and analysis frameworks, however, are relatively new in palaeoecology and ecology and have not yet been extensively tested and applied to global climate change ecological studies (Liu & Gupta 2007; Wang *et al.* 2009).

Two types of model errors exist: systematic and random (Appendix S7). MDF techniques can be used to identify systematic model errors. To separate systematic model error from errors specific to model parameters, an analysis of model errors can be carried out after model parameters have been optimized (Wang *et al.* 2009). This is accomplished by analysing model residuals and comparing different models together (Stöckli *et al.* 2008; Williams *et al.* 2009). Spectral analysis of model residuals provides a practical way in which to determine model error (Braswell *et al.* 2005; Williams *et al.* 2009). Based on wavelet analysis, Braswell *et al.* (2005) analysed model residuals and found that multi-year eddy covariance flux measurements collected from Harvard Forest provided good constraints on parameters related to the processes that control net carbon exchange (the difference between ecosystem production and ecosystem respiration) from daily to seasonal time scales. These measurements, however, did not provide good constraints regarding interannual time scales. The model's variance with regards to net carbon exchange was considerably lower than measurements with regards to variance. A more sophisticated method was recently developed by Abramowitz *et al.* (2007) to analyse systematic model errors in relation to complex global land surface models. For this study, a neural network-based flux correction technique was applied to three land surface models. It was used to show that the nature of systematic model error in simulations of latent heat, sensible heat and the net ecosystem exchange of CO_2 is shared between different vegetation types and, indeed, different models. The technique also offered insight into which land surface model process may be improved on to reduce systematic error. A comprehensive review regarding the application of MDF to quantify and reduce model uncertainty and error in hydrological modelling was reported by Liu & Gupta (2007).

MDF application to enhance remote-sensing products

Recent advances in biogeochemistry-based process models prove that in combination with regional scale remote-sensing products, promising approaches to test ecological hypotheses as well as to assess and forecast states of future landscapes can be achieved (Turner *et al.* 2004). Several studies have shown how remote sensing derived leaf area indices (Fang & Liang 2005; Hazarika *et al.* 2005), biomass (Schaepman *et al.* 2007) and canopy nitrogen (Ollinger & Smith 2005) can be adopted to constrain ecosystem models. Such approaches require spatially continuous inputs of the state of an ecosystem at the start point of a simulation and may profit from the assimilation of relevant remote sensing derived biophysical and biochemical state variables of ecosystems under consideration. MDF offers considerable promise in remote sensing with regards to improved state and parameter estimation, particularly when applied to multi-sensor image products (Liang 2007).

The MDF approach also allows for utilization of all available remote-sensing data within a time window to estimate various unknown parameters in land surface models (Liang & Qin 2008; Quaife *et al.* 2008). Hazarika *et al.* (2005) demonstrated the utility of combining satellite observations with ecosystem process models to achieve improved estimate accuracy as well as accuracy in monitoring global net primary production. Fang & Liang (2005) assimilated the MODIS LAI product into a crop growth model to estimate crop yield by determining critical parameters of the crop model. In addition, Renzullo *et al.* (2008) developed a 'multiple constraints' MDF (MCMDF) scheme that integrates multiple remote-sensing data sources (including the AMSR-E soil moisture content and MODIS land surface temperature data products) to a surface moisture and energy budget coupled biophysical model operating on a daily time scale for savannas in northern Australia. By utilizing EnKF (after Evensen 1994, 2003) with modifications made to joint state and parameter estimation, Stöckli *et al.* (2008) developed an MDF framework coupling the fraction of photosynthetically active radiation that is absorbed through vegetation (FPAR) and the LAI from MODIS data to constrain empirical temperature, light, moisture and structural vegetation parameters of a prognostic phenological model. They found that DA better constrains structural vegetation parameters than do climate control parameters. Furthermore, MDF effectively overcomes cloud and aerosol related deficiencies of satellite data sets in tropical regions.

MDF application for ecological forecasting

Ecological forecasting is a critical new tool to quantitatively characterize the most likely future states of ecological systems either under prevalent conditions or under different 'what-if' scenarios. Under prevalent conditions, short-term forecasts (i.e. days and months) generally can be made according to the system's own dynamics (Clark *et al.* 2001; Luo *et al.* 2011). Moreover, ecological forecasting uses the combined knowledge of physics, ecology and physiology to predict how ecosystems will diverge in the future in response to conditions of environmental fluctuation such as climate change (Clark *et al.* 2001). The ultimate goal of ecological forecasting is to provide resource managers and decision makers information that they can use to respond in advance to future changes. Ecological forecasting typically requires mechanistic knowledge of the processes being modelled. Forecasts are usually probabilistic and provide an estimate of the probability of a future state and not just a point estimate of its value.

The MDF approach helps improve ecological forecasting by using data to report on initial conditions and model parameters and, thereby,

constrains a model during simulation to yield results that approximate reality as close as possible. Specifically, MDF improves ecological forecasting through: (1) the estimation of model parameter and state variables, (2) the selection of the best model structure and (3) the quantification of uncertainties resulting from observations, models and their interactions (Luo *et al.* 2011).

In addition, MDF has been applied to improve model prediction of species distribution dynamics (Thuiller *et al.* 2009), recent epidemics of infectious disease (the outbreak of severe acute respiratory syndrome in Asia and the foot-and-mouth disease epidemic in the UK) and, because very few ecological examples exist up to this point, it has been applied to forecasting historical and future dynamics of terrestrial ecosystems under a changing environment as well as to carbon sink dynamics (Gao *et al.* 2011).

As an example, Kolomyts (2008) developed a landscape-ecological forecasting framework from computational models and palaeoreconstructions. Using the Volga basin as a case in point, this pilot study considered the mechanisms of local and regional responses to global warming that is expected to occur in the 21st century. Bayesian inversion and the MCMC technique were applied to a regional terrestrial ecosystem (TECO-R) model to quantify carbon residence times and assess their uncertainty in the conterminous USA (Zhou & Luo 2008). The proportion of carbon uptake in soil was found to be primarily regulated by carbon residence times. Therefore, the accurate estimation of spatial patterns with regards to carbon residence times is crucial for forecasting future change in soil carbon stocks. Results suggest that this MDF approach is an effective tool in which to estimate spatially distributed carbon residence times and assess uncertainty within the conterminous USA.

Today, the emerging ecological forecasting discipline is a data-driven scientific synthesis of physics, geology, chemistry and biology. Ecological forecasting through the application of the MDF approach requires accurate estimates of initial conditions and parameters before future states of an ecological system can be quantitatively estimated. This is true even with a perfect model structure. The feasibility and accuracy of ecological forecasting is directly dependant on the manner in which uncertainty is reduced in model prediction. The biggest challenge in ecological forecasting is to reduce model error and increase prediction accuracy.

Model–data fusion techniques can be used to systematically assess uncertainties in model prediction (Clark *et al.* 2001; Luo *et al.* 2009). Perhaps the most common in use are the KF and adjoint techniques that are often applied to numerical weather forecasting but have not yet been widely used in ecology (but Gao *et al.* 2011). Data assimilation, which merges multiple observations with numerical models, can advance ecological forecasting. For example, Scholze *et al.* (2007) used surface CO_2 concentration observational data from 1979 to 1999 to calibrate key model parameters of an ecosystem model and to forecast net CO_2 fluxes from 2000 to 2003. The REFLEX model-data fusion project (Fox *et al.* 2009) is another example of forecasting carbon dynamics by applying confidence intervals after the point in which model parameters were estimated. Iverson & Prasad (1998) used regression tree analysis to spatially forecast how tree distribution may change under a twofold CO_2 scenario estimated by several GCM models for 80 tree species in the eastern region of the USA. Thuiller W. (2003) and Thuiller *et al.* 2009) developed a new computer framework (called BIOMOD: BIOdiversity MODelling) to optimize predictions of species distribution as well as to forecast potential future shifts under global climate change.

MDF application to quantify global carbon sinks/sources

Inverse modelling has been used to deduce the spatial and temporal distribution of CO_2 sinks and sources to help identify the biogeochemical processes involved, providing a critical technique to refine our knowledge of global carbon budgets. Our growing understanding of carbon cycling provides the context for which the various developments of atmospheric inversions take place. Enting (2002) reviewed some key steps, noting discrepancies in flux estimates and controversies regarding a strong terrestrial sink located in North America. Much subsequent work has also been undertaken within the TRANSCOM intercomparison community (Baker *et al.* 2006). As greater time-resolution global data (with regards to additional species) become available (i.e. the atmospheric CO_2 $\delta^{13}C$ and $\delta^{18}O$ ratio and the atmospheric O_2/N_2 ratio), the use of synthesis inversion techniques in combination with atmospheric transport models will result in much more reliable estimates regarding the changing atmospheric global carbon budget. Using the TRANSCOM intercomparison project as an example, Denning *et al.* (1999) assessed the north–south transport of models by comparing simulations of SF6 – a relatively well-known inert anthropogenic tracer – to measured the recently investigated atmospheric concentrations of this tracer.

Overall, the application of MDF approaches has improved estimates of regional and global carbon fluxes over the last two decades. Robust estimates of decadal regional changes in carbon flux by Bousquet *et al.* (2000) as well as global carbon budgets for the 1990s and mean annual budgets from 1992 to 1996 were presented by Gurney *et al.* (2002). CARBONTRACKER is a recent advancement in the development and implementation of an operational tool that uses the KF to estimate surface flux and its inherent uncertainty by applying surface concentration measurements in near real time (Peters *et al.* 2007). The limitation of the inverse modelling approach, however, is that predicted CO_2 sinks/sources are sensitive to atmospheric CO_2 data. Thus, inborn small uncertainties that naturally occur in CO_2 concentration data and small errors that occur in the atmospheric transport models are magnified into large uncertainties for sink/source prediction. By enlarging the network of CO_2 sampling stations to cover continental regions and by careful analysis of noise amplification, it is expected that the atmospheric inverse modelling technique will resolve sink/source patterns in finer detail and will be applicable to other greenhouse gases in the near future.

CURRENT CHALLENGES, OPPORTUNITIES AND FUTURE DIRECTIONS

The application of MDF in the study of ecology is relatively new, and there is a lack of general guidance within global ecology literature on how to choose and implement a suitable MDF approach to address the four major types of MDF problems: system identification, parameter estimation, state estimation and uncertainty quantification. Although the MDF approach is still within the infancy stage in ecology and palaeoecology, it is rapidly becoming a more practical and powerful method to study global climate change issues due to increased data availability from global observational networks. New

DA methods taken from physical or meteorological science will need to be developed for ecological systems (Table S1). Moreover, an urgent need exists to develop dynamic simulation models that can explain past changes as well as forecast future ecosystem response and feedback under a changing global climate (Guiot *et al.* 2009). Uncertainty analysis in combination with DA has been carried out in a limited number of studies in the last few years (Xu *et al.* 2006; Verstraeten *et al.* 2008). However, parameter identifiability and equifinality in association with uncertainty analysis have not been fully addressed to date and must be explored in depth (Luo *et al.* 2009; Williams *et al.* 2009).

Key MDF challenges that both ecology and palaeoecology communities must first take into consideration include as follows:

The replacement of static vegetation inversion methods with dynamic vegetation model inversion in palaeoecological reconstructions

Garreta *et al.* (2009) proposed the use of LPJ-GUESS, a more complex and dynamic vegetation model (Smith *et al.* 2001), to replace equilibrium vegetation models used in combination with pollen assemblages. As the model is dynamic, the application takes into account temporal characteristics of data as suggested by Haslett *et al.* (2006). Vegetation is not merely assumed to be dependent on contemporaneous climates but also on historical vegetation. Time series autocorrelation is considered important information in LPJ-GUESS. Garreta *et al.* (2009) successfully used a PF technique better adapted to time series and stochastic processes. The use of a dynamic vegetation model allows for a better exploitation of the information available from the fossil record.

Parameter identifiability and effects of initial conditions

Parameter identifiability refers to parameters that can be constrained by a set of data with a given model structure. These parameters are identifiable when parameter maximum likelihood values are identified, which is the core to uncertainty analysis. The condition of equifinality exists in DA when different models or different parameter values of the same model fit data equally well without retaining the ability to distinguish which models or parameter values are better suited than others. Identifiability is therefore a reflection of parameter constraints and equifinality. Parameter identifiability is a critical but complex property that has not yet been extensively investigated. A careful assessment of the principle must be addressed in the near future (Luo *et al.* 2009). Global sensitivity analysis that examines parameters sensitive to available data sets may be an effective approach in which to select those identifiable parameters to be constrained by MDF (Tang & Zhuang 2009; Gao *et al.* 2011).

Problems related to initial values are not well addressed in MDF (Luo *et al.* 2011). Initial conditions (e.g. abundance and age distribution in demographical models and biomass and pool size in biogeochemical models) are critical and sometimes govern the subsequent trajectory of system performance. In a chaotic system, infinitesimal differences in initial conditions can lead to exponential divergence between trajectories. In carbon cycle modelling, the initial value of pool size determines the direction and magnitude of carbon sequestration (Carvalhais *et al.* 2008). Estimations of initial pool size by means of the MDF approach are essential when quantifying the rate of carbon sequestration in an ecosystem.

Assimilation of multi-data types and full quantification and reduction of multi-sources of uncertainties arising from data bias, model structure and initial condition estimates

Observations are always sparse and irregularly distributed throughout space and time. It is still not possible to measure all degrees of freedom of a model at a given time. Therefore, an efficient MDF method is required to combine irregular observations to generate a data set distributed on a regular model grid (Wang *et al.* 2009). For example, ground-based flux measurement networks are sparse (e.g. seen as patchy tower fluxes on land), especially in remote and inaccessible regions typical in the Southern Hemisphere. They can only provide a regional and incomplete picture of the fluxes taking place. In contrast, satellites provide global coverage. The issue, however, is that they cannot measure carbon flux directly but only related variables such as greenhouse gas concentrations. One powerful way in which to address this issue is to assimilate earth observational data into carbon cycle models to infer global estimates of carbon flux (Mathieu & O'Neill 2008).

Model development must involve the selection of appropriate model structure (e.g. concept models such as the PCM model provided for in Fig. 6) for any mathematical model (typically equations) to correctly represent the 'real' system (i.e. relationships between model inputs, parameters, states and outputs). Errors are the most important and also the most difficult to verify when identifying and quantifying model structure. In most cases, DA applications ignore parameter and model structure uncertainty (Liu & Gupta 2007). Data uncertainty, however, affects both the predicted uncertainty of outputs as well as the predicted best estimate. Due to this fact, an immediate need exists to improve our understanding of model structure error and to reduce parameter uncertainty relative to ecological systems.

Data assimilation has been widely used in meteorology, oceanography and hydrology, but more effort is needed to explore its potential for characterizing land surface environments in the remote-sensing community. Developing advanced inversion modelling and DA methods to solve multidimensional nonlinear inversion problems in remote sensing is critical as well as challenging as these inversion problems are typically affected by noise and measurement uncertainty. It is important to pay attention to model structure uncertainty, data uncertainty and initial conditions if MDF is to be attempted.

Advances towards 'ensemble forecasting' under a changing environment

The idea behind ensemble forecasting was formally developed by French mathematician P. Laplace in 1818 (Laplace 1818). It was not until the pioneering work of Bates & Granger (1969) that ensemble forecasting blossomed after they developed the idea of combining forecast data together. They observed that combining forecasts yielded lower mean errors more so than any constituent individual forecasts could do separately as long as individual forecasts contained some independent information. Since that time numerous studies have been reviewed (Clemen 1989; Cheung 2001) and applied to a variety of research fields, including economics (Gregory *et al.* 2001), managerial practices (Makridakis & Winkler 1983), systematics (Miyamoto 1985), meteorology (Sanders 1963) and climatology (Benestad 2004). These ideas have yet to be adapted and developed for ecological systems. Only recently has ensemble forecasting been

explicitly attempted on bioclimatic modelling of species distribution (Thuiller 2003; Araújo *et al.* 2005; Thuiller *et al.* 2009). More recently, Araújo & New (2006) reviewed various modelling techniques that incorporated elements of ensemble forecasting approaches. They proposed the use of multi-model simulations to construct ensemble forecasting. Ensemble forecasting has become one of the new MDF applications in global climate change ecology. It applies DA and model forecasting techniques to obtain future predictions through the running of multiple dynamical system simulations incorporating various initial conditions. The real advantage of ensemble forecasting is to help quantify the intrinsic error in each individual model.

Advances towards a 'hybrid data assimilation' approach in combination with global optimization and sequential data assimilation

Ensemble Kalman filter is able to provide a general framework in which to consider input, output, model structure uncertainty and a predefined filter for use with state estimations. However, no recursive parameter estimation procedure presently exists. A failure to explicitly take into account the effects of parameter uncertainty and interaction typically occurs when applying EnKF to recursively estimated state variables. It would be necessary to combine parameter estimation and state estimation to account for all kinds of uncertainty. For example, Vrugt *et al.* (2005) recently introduced a simultaneous optimization and data assimilation (SODA) method that uses EnKF to recursively update model states while estimating time-invariant values for model parameters that utilize the SCEM-UA optimization algorithm. A novel feature of SODA is its explicit treatment of error due to parameter uncertainty as well as its treatment of uncertainty in the initialization and propagation of state variables, model structure error and output measurement error.

In fact, state of the art assimilation techniques of 4D-Var are seldom used on global scale advanced biogeochemical modelling analysis frameworks partly because of the enormous numerical computation involved. A recent trend in DA is to combine the advantages of 4D-Var and KF techniques together. 4D-Var has proven itself to be an efficient analytical method when applied to a real time assimilation system that is run over a short time interval. Recent versions of the EnKF approach have applied ensemble estimates of error covariance of the 4D atmospheric state (Hunt *et al.* 2004). This feature was implemented into the Canadian NWP operational EnKF system on 10 July 2007 (Houtekamer *et al.* 2009). Similar to 4D-Var, this approach allows EnKF to estimate the 4D-Var atmospheric state that best fits assimilated observations distributed over time. A good compromise between 4D-Var and EnKF techniques and their constituent algorithms can be achieved. New hybrid DA methods also provide numerous byproducts that remain to be used (assessed) as diagnostic tools to improve assimilation and forecast systems.

CONCLUSION

Model–data fusion is an application comprised of statistically based inverse modelling and DA techniques that integrates and combines dynamic models and observed data sets in optimal ways to enhance the nowcasting, hindcasting and forecasting capacity of complex systems. A great variety of DA techniques and application domains exist within the various spheres of earth system sciences. Different MDF methods hold different assumptions and possess different strengths and weaknesses. The improvement in ecological forecasting requires the development of innovative mathematical, observational and computational techniques that optimally combine observational data sets and models. The intercomparison of optimization techniques and algorithms would aid in the selection of cost functions and quantify model errors in any optimization as well as assist in reducing uncertainties of model parameters poorly constrained by available observations. Ecological models are highly nonlinear and are not fully constrained by available observations. This may lead to problems in certain DA techniques. Moreover, variational methods (3/4 D-Var) may prove unfeasible for more sophisticated terrestrial ecosystem biogeochemical models. The hybrid application of 4D-Var and EnKF for global analysis and EnKF for regional analysis appears promising for future MDF utilization as well as the future direction of ecology. Overall, MDF approaches hold great potential in enhancing the capacity of vegetation and ecosystem carbon models in relation to their predictive response to terrestrial vegetation and carbon cycling in a changing climate.

ACKNOWLEDGEMENTS

This work was conducted in China and France during the sabbatical leave of C. Peng. For financial support, we would like to thank the China QianRen program, the Institute of Ecology and Environment (INEE) of CNRS (French National Center for Scientific Research), the National Science and Engineering Research Council of Canada (NSERC) Discover Grant, and the Canada Research Chair Program. C. Peng acknowledges the financial and scientific support he received during his sabbatical leave at ECCOREV, France, and at Northwest A&F University, China. We are grateful to Bin He, the editor and three anonymous referees for their constructive comments and suggestions concerning the paper. We also thank Wenhua Zhang for her technical assistance and Brian Doonan for his editorial help.

REFERENCES

Abramowitz, G., Pitman, A., Gupta, H., Kowalczyk, E. & Wang, Y. (2007). Systematic bias in land surface models. *J. Hydrometeorol.*, 8, 989–1001.

Adams, J.M., Faure, H., Faure-Denard, L., McGlade, J.M. & Woodward, F.I. (1990). Increase in terrestrial carbon storage from the Last Glacial Maximum to the present. *Nature*, 348, 711–714.

Araújo, M. & New, M. (2006). Ensemble forecasting of species distributions. *Trends Ecol. Evol.*, 22, 42–47.

Araújo, M., Whittaker, R., Ladle, R. & Erhard, M. (2005). Reducing uncertainty in projections of extinction risk from climate change. *Global Ecol. Biogeogr.*, 14, 529–538.

Arulampalam, M., Maskell, S., Gordon, N. & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.*, 50, 174–188.

Baker, D.F., Law, R.M., Gurney, K.R., Rayner, P., Peylin, P., Denning, A.S. *et al.* (2006). TransCom 3 inversion intercomparison: impact of transport model errors on the interannual variability of regional CO_2 fluxes, 1988-2003. *Global Biogeochem. Cycles*, 20, 1002–1017.

Bates, J. & Granger, C. (1969). The combination of forecasts. *OR*, 20, 451–468.

Benestad, R. (2004). Tentative probabilistic temperature scenarios for northern Europe. *Tellus A*, 56, 89–101.

Bousquet, P., Peylin, P., Ciais, P., Le Quéré, C., Friedlingstein, P. & Tans, P.P. (2000). Regional changes in carbon dioxide fluxes of land and oceans since 1980. *Science*, 290, 1342–1346.

Braswell, B., Sacks, W., Linder, E. & Schimel, D. (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, 335–355.

Burgers, G., van Leeuwen, P. & Evensen, G. (1998). Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, 126, 1719–1724.

Carvalhais, N., Reichstein, M., J., S., G.J., C., Santos Pereira, J., Berbigier, P. *et al.* (2008). Implications of carbon cycle steady state assumptions for biogeochemical modeling performance and inverse parameter retrieval. *Global Biogeochem. Cycles*, 22, GB2007, doi: 10.1029/2007GB003033.

Chen, M., Liu, S., Tieszen, L. & Hollinger, D. (2008). An improved state-parameter analysis of ecosystem models using data assimilation. *Ecol. Model.*, 219, 317–326.

Cheung, K. (2001). A review of ensemble forecasting techniques with a focus on tropical cyclone forecasting. *Meteorol. Appl.*, 8, 315–332.

Clark, J., Carpenter, S., Barber, M., Collins, S., Dobson, A., Foley, J. *et al.* (2001). Ecological forecasts: an emerging imperative. *Science*, 293, 657.

Clemen, R. (1989). Combining forecasts: a review and annotated bibliography. *Int. J. Forecast.*, 5, 559–583.

Cosby, B. (1984). Dissolved oxygen dynamics of a stream: model discrimination and estimation of parameter variability using an extended Kalman filter. *Water Sci. Technol.*, 16, 561–569.

Cressman, G. (1959). An operational objective analysis system. *Mon. Wea. Rev.*, 87, 367–374.

Daley, R. (1991). *Atmospheric Data Analysis*. Cambridge University Press, Cambridge, 471 pp.

Denning, A.S., Holzer, M., Gurney, K.R., Heimann, M., Law, R.M., Rayner, P.J. *et al.* (1999). Threedimensional transport and concentration of SF6: a model intercomparison study (TransCom 2). *Tellus*, 51B, 266–297.

Enting, I.G. 2002. *Inverse Problems in Atmospheric Constituent Transport*. Cambridge University Press, New York, 392 pp.

Evensen, G. (1992). Using the extended Kalman filter with a multilayer quasi-geostrophic ocean model. *J. Geophys. Res.*, 97, 17905–17924.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, 99, 10143–10162.

Evensen, G. (2003). The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynam.*, 53, 343–367.

Evensen, G. (2007). *Data Assimilation: The Ensemble Kalman Filter*. Springer Verlag, Berlin.

Fang, H. & Liang, S. (2005). A hybrid inversion method for mapping leaf area index from MODIS data: experiments and application to broadleaf and needleleaf canopies. *Remote Sens. Environ.*, 94, 405–424.

Foley, J.A. (1995). An equilibrium model of the terrestrial carbon budget. *Tellus*, 47B, 310–319.

Fox, A., Williams, M., Richardson, A., Cameron, D., Gove, J., Quaife, T. *et al.* (2009). The REFLEX project: comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data. *Agr. Forest Meteorol.*, 149, 1597–1615.

Friedlingstein, P., Bopp, L., Rayner, P., Cox, P., Betts, R., Jones, C. *et al.* (2006). Climate–carbon cycle feedback analysis: results from the C4MIP model intercomparison. *J. Climate*, 19, 3337–3353.

Gandin, L.S. (1963). Objective analysis of Meteorological field. Gidrometeorologicheskoe Izdatel'stvo, Leningrad, Translated From Russian in 1965 by Israel Program for Scientific Translations, Jerusalem, 242 pp.

Gao, G., Wang, H., Weng, E.S., Lakshmivarahan, S., Zhang, Y.F. & Luo, Y.Q. (2011). Assimilation of multiple data sets with ensemble Kalman filter for parameter estimation and forecasts of forest carbon dynamics. *Ecol. Appl.*, in press.

Garreta, V., Miller, P., Guiot, J., Hély, C., Brewer, S., Sykes, M. *et al.* (2009). A method for climate and vegetation reconstruction through the inversion of a dynamic vegetation model. *Clim. Dyn.*, 35, 371–389.

Gauthier, P., Tanguay, M., Laroche, S., Pellerin, S. & Morneau, J. (2007). Extension of 3DVAR to 4DVAR: implementation of 4DVAR at the Meteorological Service of Canada. *Mon. Wea. Rev.*, 135, 2339–2354.

Gregory, A., Smith, G. & Yetman, J. (2001). Testing for forecast consensus. *J. Bus. Econ. Stat.*, 19, 34–43.

Guiot, J. & de Vernal, A. (2007). Chapter thirteen transfer functions: methods for quantitative paleoceanography based on microfossils. *Dev. Mar. Geol.*, 1, 523–563.

Guiot, J., Torre, F., Jolly, D., Peyron, O., Boreux, J. & Cheddadi, R. (2000). Inverse vegetation modeling by Monte Carlo sampling to reconstruct palaeoclimates under changed precipitation seasonality and CO_2 conditions: application to glacial climate in Mediterranean region. *Ecol. Model.*, 127, 119–140.

Guiot, J., Wu, H., Garreta, V., Hatté, C. & Magny, M. (2009). A few prospective ideas on climate reconstruction: from a statistical single proxy approach towards a multi-proxy and dynamical approach. *Clim. Past.*, 5, 99–125.

Gurney, K.R., Law, R.M., Scott Denning, A., Rayner, P.J., Baker, D., Bousquet, P. *et al.* (2002). Towards robust regional estimates of CO_2 sources and sinks using atmospheric transport models. *Nature*, 415, 626–630.

Haslett, J., Whiley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S., Allen, J. *et al.* (2006). Bayesian palaeoclimate reconstruction. *J. R. Stat. Soc. A Stat.*, 169, 395–438.

Hastings, W.K. (1970). Monte Carlo sampling using Markov chains and their applications. *Biometrika*, 57, 97–109.

Hatté, C. & Guiot, J. (2005). Palaeoprecipitation reconstruction by inverse modelling using the isotopic signal of loess organic matter: application to the Nußloch loess sequence (Rhine Valley, Germany). *Clim. Dyn.*, 25, 315–327.

Hatté, C., Rousseau, D.-D. & Guiot, J. (2009). Climate reconstruction from pollen and *13C* records using inverse vegetation modelling: implication for past and future climates. *Clim. Past.*, 5, 147–156.

Hazarika, M.K., Yasuoka, Y., Ito, A. & Dye, D. (2005). Estimation of net primary productivity by integrating remote sensing data with an ecosystem model. *Remote Sens. Environ.*, 94, 298–310.

Houtekamer, P.L., Mitchell, H.L. & Deng, X. (2009). Model error representation in an operational ensemble Kalman filter. *Mon. Wea. Rev.*, 137, 2126–2143.

Hunt, B.R., Kalnay, E., Kostelich, E.J., Ott, E., Patil, D.J., Sauer, T. *et al.* (2004). Four-dimensional ensemble Kalman filtering. *Tellus A*, 56, 273–277.

Iverson, L.R. & Prasad, A.M. (1998). Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecol. Monogr.*, 68, 465–485.

Joos, F., Gerber, S., Prentice, I.C., Otto-Bliesner, B.L. & Valdes, P.J. (2004). Transient simulations of Holocene atmospheric carbon dioxide and terrestrial carbon since the Last Glacial Maximum. *Global Biogeochem. Cycles*, 18, GB2002.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng. (ASME)*, 32D, 35–45.

Kalnay, E. (2003). *Atmospheric Modelling, Data Assimilation and Predictability*. Cambridge University Press, UK.

Kaplan, J.O. (2001) *Geophysical Applications of Vegetation Modeling*. Lund University, Sweden, 210 pp.

Knorr, W. & Kattge, J. (2005). Inversion of terrestrial ecosystem model parameter values against eddy covariance measurements by Monte Carlo sampling. *Glob. Change Biol.*, 11, 1333–1351.

Kolomyts, E.G. (2008). Landscape-ecological forecasts from computational models and palaeoreconstructions (using the Volga basin as an example). *Geography and Nature Resources*, 29, 209–220.

Laplace, P.S. (1818). Deuxieme supplement a la theorie analytique des probabilites. *Courcier*, 7, 531–580.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Math.*, 2, 164–168.

Liang, S. (2007). Recent developments in estimating land surface biogeophysical variables from optical remote sensing. *Prog. Phys. Geog.*, 31, 501–516.

Liang, S. & Qin, J. (2008). Data assimilation methods for land surface variable estimation. In: *Advances in Land Remote Sensing* (ed. Liang, S.). Springer, Netherlands, pp. 313–339.

Liu, Y. & Gupta, H.V. (2007). Uncertainty in hydrologic modeling: toward an integrated data assimilation framework. *Water Resour. Res.*, 43, W07401.

Lorenc, A.C. (1981). A global three-dimensional multivariate statistical interpolation scheme. *Mon. Wea. Rev.*, 109, 701–721.

Lorenc, A. & Payne, T.J. (2007). The Met Office global four-dimensional variational data assimilation scheme. *Quart. J. R. Meteor. Soc.*, 133, 347–362.

Luo, Y., White, L.W., Canadell, J.G., DeLucia, E.H., Ellsworth, D.S., Finzi, A. *et al.* (2003). Sustainability of terrestrial carbon sequestration: a case study in Duke Forest with inversion approach. *Global Biogeochem. Cycles*, 17, 1021.

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, 571–574.

Luo, Y., Ogle, K., Tucker, C., Fei, S., Gao, C., LaDeau, S. *et al.* (2011). Ecological forecasting and data assimilation in a data-rich era. *Ecol. Appl.*, in press.

Makridakis, S. & Winkler, R.L. (1983). Averages of forecasts: some empirical results. *Manage. Sci.*, 29, 987–996.

Mandel, J., Bennethum, L.S., Beezley, J.D., Coen, J.L., Douglas, C.C., Kim, M. *et al.* (2008). A wildland fire model with data assimilation. *Math. Comput. Simulat.*, 79, 584–606.

Marquardt, D.W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.*, 11, 431–441.

Mathieu, P.-P. & O'Neill, A. (2008). Data assimilation: from photon counts to earth system forecasts. *Remote Sens. Environ.*, 112, 1258–1267.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21, 1087–1092.

Miyamoto, M.M. (1985). Consensus cladograms and general classifications. *Cladistics*, 1, 186–189.

Mo, X., Chen, J.M., Ju, W. & Black, T.A. (2008). Optimization of ecosystem model parameters through assimilating eddy covariance flux data with an ensemble Kalman filter. *Ecol. Model.*, 217, 157–173.

Ollinger, S. & Smith, M.-L. (2005). Net primary production and canopy nitrogen in a temperate forest landscape: an analysis using imaging spectroscopy, modeling and field data. *Ecosystems*, 8, 760–778.

Pan, M., Wood, E.F., Wójcik, R. & McCabe, M.F. (2008). Estimation of regional terrestrial water cycle using multi-sensor remote sensing observations and data assimilation. *Remote Sens. Environ.*, 112, 1282–1294.

Peng, C.H., Guiot, J. & Van Campo, E. (1998). Estimating changes in terrestrial vegetation and carbon storage: using palaeoecological data and models. *Quat. Sci. Rev.*, 17, 719–735.

Peters, W., Jacobson, A.R., Sweeney, C., Andrews, A.E., Conway, T.J., Masarie, K. *et al.* (2007). An atmospheric perspective on North American carbon dioxide exchange: CarbonTracker. *Proc. Natl. Acad. Sci. USA*, 104, 18925–18930.

Potter, C.S., Wang, S., Nikolov, N.T., McGuire, A.D., Liu, J., King, A.W. *et al.* (2001). Comparison of boreal ecosystem model sensitivity to variability in climate and forest site parameters. *J. Geophys. Res.*, 106, 33671–33687.

Prentice, I.C., Cramer, W., Harrison, S.P., Leemans, R., Monserud, R.A. & Solomon, A.M. (1992). A global biome model based on plant physiology and dominance, soil properties and climate. *J. Biogeogr.*, 19, 117–134.

Prentice, I.C., Guiot, J., Huntley, B., Jolly, D. & Cheddadi, R. (1996). Reconstructing biomes from palaeoecological data: a general method and its application to European pollen data at 0 and 6 ka. *Clim. Dyn.*, 12, 185–194.

Quaife, T., Lewis, P., De Kauwe, M., Williams, M., Law, B.E., Disney, M. *et al.* (2008). Assimilating canopy reflectance data into an ecosystem model with an Ensemble Kalman Filter. *Remote Sens. Environ.*, 112, 1347–1364.

Raupach, M.R., Rayner, P.J., Barrett, D.J., DeFries, R.S., Heimann, M., Ojima, D.S. *et al.* (2005). Model–data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications. *Glob. Change Biol.*, 11, 378–397.

Reichle, R.H., McLaughlin, D.B. & Entekhabi, D. (2002). Hydrologic data assimilation with the ensemble Kalman filter. *Mon. Wea. Rev.*, 130, 103–114.

Renzullo, L.J., Barrett, D.J., Marks, A.S., Hill, M.J., Guerschman, J.P., Mu, Q. *et al.* (2008). Multi-sensor model-data fusion for estimation of hydrologic and energy flux parameters. *Remote Sens. Environ.*, 112, 1306–1319.

Sacks, W.J., Schimel, D.S., Monson, R.K. & Braswell, B.H. (2006). Model-data synthesis of diurnal and seasonal CO_2 fluxes at Niwot Ridge, Colorado. *Glob. Change Biol.*, 12, 240–259.

Sanders, F. (1963). On subjective probability forecasting. *J. Appl. Meteorol.*, 2, 191–201.

Santaren, D., Peylin, P., Viovy, N. & Ciais, P. (2007). Optimizing a process-based ecosystem model with eddy-covariance flux measurements: a pine forest in southern France. *Global Biogeochem. Cycles*, 21, GB2013.

Schaepman, M.E., Wamelink, G.W.W., van Dobben, H., Gloor, M., Schaepman-Strub, G. & Kooistra, L. (2007). River floodplain vegetation scenario development using imaging spectroscopy and ecosystem models. *Photogramm. Eng. Rem. S.*, 73, 1179–1188.

Scholze, M., Kaminski, T., Rayner, P., Knorr, W. & Giering, R. (2007). Propagating uncertainty through prognostic carbon cycle data assimilation system simulations. *J. Geophys. Res.*, 112, D17305.

Smith, B., Prentice, I.C. & Sykes, M.T. (2001). Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within european climate space. *Glob. Ecol. Biogeogr.*, 10, 621–637.

Stöckli, R., Lawrence, D.M., Niu, G.Y., Oleson, K.W., Thornton, P.E., Yang, Z.L. *et al.* (2008). Use of FLUXNET in the community land model development. *J. Geophys. Res.*, 113, G01025.

Tang, J. & Zhuang, Q. (2009). A global sensitivity analysis and Bayesian inference framework for improving the parameter estimation and prediction of a process-based Terrestrial Ecosystem Model, *J. Geophys. Res.* 114: D15303, doi: 10.1029/2009JD011724.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, SIAM.

Thiemann, M., Trosset, M., Gupta, H.V. & Sorooshian, S. (2001). Bayesian recursive parameter estimation for hydrologic models. *Water Resour. Res.*, 37, 2521–2535.

Thuiller, W. (2003). BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. *Glob. Change Biol.*, 9, 1353–1362.

Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M.B. (2009). BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, 32, 369–373.

Trudinger, C.M., Raupach, M.R., Rayner, P.J., Kattge, J., Liu, Q., Pak, B. *et al.* (2007). OptIC project: an intercomparison of optimization techniques for parameter estimation in terrestrial biogeochemical models. *J. Geophys. Res.*, 112, G02027.

Turner, D.P., Ollinger, S.V. & Kimball, J.S. (2004). Integrating remote sensing and ecosystem process models for landscape- to regional-scale analysis of the carbon cycle. *Bioscience*, 54, 573–584.

Van Campo, E., Guiot, J. & Peng, C.H. (1993). A data-based re-appraisal of the terrestrial carbon budget at the Last Glacial Maximum. *Global Planet. Change*, 8, 189–201.

Verstraeten, W.W., Veroustraete, F., Heyns, W., Roey, T.V. & Feyen, J. (2008). On uncertainties in carbon flux modelling and remotely sensed data assimilation: the Brasschaat pixel case. *Adv. Space Res.*, 41, 20–35.

Vrugt, J.A., Diks, C.G.H., Gupta, H.V., Bouten, W. & Verstraten, J.M. (2005). Improved treatment of uncertainty in hydrologic modeling: combining the strengths of global optimization and data assimilation. *Water Resour. Res.*, 41, W01017.

Wang, Y.-P., Trudinger, C.M. & Enting, I.G. (2009). A review of applications of model-data fusion to studies of terrestrial carbon fluxes at different scales. *Agr. Forest Meteorol.*, 149, 1829–1842.

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

Williams, M., Richardson, A.D., Reichstein, M., Stoy, P.C., Peylin, P., Verbeeck, H. *et al.* (2009). Improving land surface models with FLUXNET data. *Biogeosciences*, 6, 2785–2835.

Wu, H., Guiot, J., Brewer, S. & Guo, Z. (2007a). Climatic changes in Eurasia and Africa at the last glacial maximum and mid-Holocene: reconstruction from pollen data using inverse vegetation modelling. *Clim. Dyn.*, 29, 211–229.

Wu, H., Guiot, J., Brewer, S., Guo, Z. & Peng, C. (2007b). Dominant factors controlling glacial and interglacial variations in the treeline elevation in tropical Africa. *Proc. Natl. Acad. Sci. USA*, 104, 9720–9724.

Wu, H.B., Guiot, J., Peng, C.H. & Guo, Z.T. (2009). A new coupled vegetation-carbon model used in inverse mode for reconstructing terrestrial carbon storage from pollen data: its validation using modern data. *Glob. Change Biol.*, 15, 82–95.

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 Biogeochem. Cycles*, 20, GB2007.

Zhou, T. & Luo, Y. (2008). Spatial patterns of ecosystem carbon residence time and NPP-driven carbon uptake in the conterminous United States. *Global Biogeochem. Cycles*, 22, GB3032.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Methodological development of data assimilation has gone through four primary phases: (1) simple analysis (e.g. Cressman algorithm), (2) statistical or optimum interpolation, (3) variational data assimilation (VDA) and (4) sequential data assimilation.
Appendix S2 The Kalman filter, named after Rudolf E. Kalman, is a mathematical method that uses measurements observed over time.
Appendix S3 Markov chain Monte Carlo (MCMC) techniques used to generate simulations from a probability distribution are a class of algorithms used in sampling from probability distributions based on constructing a Markov chain that has the desired distribution that it reflects its equilibrium distribution.
Appendix S4 BIOME3 (Haxeltine and Prentice 1996) is a process-based terrestrial biosphere model that includes a photosynthetic scheme that simulates the acclimation of plants to an altered state of atmospheric CO_2 by the optimization of nitrogen allocation to foliage and by accounting for the effects of CO_2 on net assimilation, stomatal conductance, leaf area index and the ecosystem water balance.
Appendix S5 The most important advantage of sequential methods is the ability of the optimal sate to differ from that embodied in the model equation.
Appendix S6 The study by Chen *et al.* (2008) is, in effect, a joint state-parameter approach that can integrate a kernel-smoothing algorithm into an ensemble Kalman filter to overcome the dramatic, sudden changes in parameter values through time and the loss of information between two consecutive points in time.
Appendix S7 Systematic errors are cumulative in nature.
Table S1 Comparison of main data assimilation (DA) methods used in numerical weather prediction (NWP)*.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Editor, John Arnone
Manuscript received 2 September 2010
First decision made 8 October 2010
Second decision made 22 December 2010
Manuscript accepted 30 January 2011
Revised MS by Editor Jon Chase