Biogeosciences Discuss., 10, 10937–10995, 2013
www.biogeosciences-discuss.net/10/10937/2013/
doi:10.5194/bgd-10-10937-2013

Biogeosciences
Discussions
Open Access

Evaluation of biospheric components in Earth system models using modern and palaeo observations: the state-of-the-art

A. M. Foley[1,*], D. Dalmonech[2], A. D. Friend[1], F. Aires[3], A. Archibald[4], P. Bartlein[5], L. Bopp[6], J. Chappellaz[7], P. Cox[8], N. R. Edwards[9], G. Feulner[10], P. Friedlingstein[8], S. P. Harrison[11], P. O. Hopcroft[12], C. D. Jones[13], J. Kolassa[3], J. G. Levine[14], I. C. Prentice[15], J. Pyle[4], N. Vázquez Riveiros[16], E. W. Wolff[14], and S. Zaehle[2]

[1]Department of Geography, University of Cambridge, Cambridge, UK
[2]Max Planck Institute for Biogeochemistry, Jena, Germany
[3]Estellus, Paris, France
[4]Centre for Atmospheric Science, University of Cambridge, Cambridge, UK
[5]Department of Geography, University of Oregon, Eugene, Oregon, USA
[6]Laboratoire des Sciences du Climat et de l'Environnement, Gif sur Yvette, France
[7]UJF – Grenoble I and CNRS Laboratoire de Glaciologie et Géophysique de l'Environnement, Grenoble, France
[8]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK
[9]Environment, Earth and Ecosystems, The Open University, Milton Keynes, UK
[10]Potsdam Institute for Climate Impact Research, Potsdam, Germany

[11]Department of Biological Sciences, Macquarie University, Sydney, Australia and Geography and Environmental Sciences, School of Human and Environmental Sciences, Reading University, Reading, UK
[12]School of Geographical Science, University of Bristol, Bristol, UK
[13]Met Office Hadley Centre, Exeter, UK
[14]British Antarctic Survey, Cambridge, UK
[15]Department of Life Sciences and Grantham Institute for Climate Change, Imperial College, Silwood Park, UK and Department of Biological Sciences, Macquarie University, Sydney, Australia
[16]Department of Earth Sciences, University of Cambridge, Cambridge, UK
[*]now at: Cambridge Centre for Climate Change Mitigation Research, Department of Land Economy, University of Cambridge, Cambridge, UK

Received: 3 June 2013 – Accepted: 10 June 2013 – Published: 4 July 2013

Correspondence to: A. M. Foley (amf62@cam.ac.uk)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Earth system models are increasing in complexity and incorporating more processes than their predecessors, making them important tools for studying the global carbon cycle. However, their coupled behaviour has only recently been examined in any detail, and has yielded a very wide range of outcomes, with coupled climate-carbon cycle models that represent land-use change simulating total land carbon stores by 2100 that vary by as much as 600 Pg C given the same emissions scenario. This large uncertainty is associated with differences in how key processes are simulated in different models, and illustrates the necessity of determining which models are most realistic using rigorous model evaluation methodologies. Here we assess the state-of-the-art with respect to evaluation of Earth system models, with a particular emphasis on the simulation of the carbon cycle and associated biospheric processes. We examine some of the new advances and remaining uncertainties relating to (i) modern and palaeo data and (ii) metrics for evaluation, and discuss a range of strategies, such as the inclusion of pre-calibration, combined process- and system-level evaluation, and the use of emergent constraints, that can contribute towards the development of more robust evaluation schemes. An increasingly data-rich environment offers more opportunities for model evaluation, but it is also a challenge, as more knowledge about data uncertainties is required in order to determine robust evaluation methodologies that move the field of ESM evaluation from "beauty contest" toward the development of useful constraints on model behaviour.

1 Introduction

Earth system models (ESMs), which use sets of equations to represent atmospheric, oceanic, cryospheric and biospheric processes and interactions (Claussen et al., 2002; Le Treut et al., 2007; Lohmann et al., 2008), are important tools for the study of the Earth system. The current generation of ESMs are more complex than their predecessors in terms of land and ocean biogeochemistry, and can also account for land cover change, which is an important driver of the climate system through biophysical and biogeochemical feedbacks. Yet, their coupled behaviour has only recently been explored in detail.

In the context of coupled behaviour, the carbon cycle is a particular relevant feature of ESMs, and one which is associated with a lot of biotic feedbacks. Interestingly, ESM simulation results submitted to the Coupled Model Intercomparison Project Phase 5 (CMIP5) simulate total land carbon stores in 2100 that vary by as much as 600 Pg C across models which represent land-use change even when forced with the same anthropogenic emissions (Jones et al., 2013), signaling that there are large uncertainties associated with how key carbon cycle processes are represented in different models. As such, robust evaluation of a model's ability to simulate key carbon cycle processes is essential, as this provides an important measure of the confidence that can be placed in a model's abilities to project future behaviours and states of the system.

Evaluation is further complicated by the fact that not all ESMs have the same level of complexity. To take the example of land cover, while some models only account for the biophysical effects, i.e. related to changes in surface albedo, some ESMs also account for the biogeochemical effects, e.g. the emission of CO_2 following conversion. Similarly, not all ESMs include nutrient cycles, but current model projections that do include terrestrial carbon-nitrogen (-phosphorus) cycles show that taking nutrient limitation into account attenuates the carbon cycle responses to disturbance because of a feedback between nutrient availability from soils and plant productivity. This attenuation generally leads to a stronger accumulation of CO_2 in the atmosphere by the end of the 21st century than projected by carbon cycle models (Sokolov et al., 2008; Thornton et al., 2009; Zaehle et al., 2010). Nitrogen dynamics have been shown to reduce potential terrestrial carbon storage over the 21st century by up to 50 % (Zaehle et al., 2010), therefore ESMs that do not include nutrient interactions likely substantially overestimate carbon sequestration. The representation, or lack thereof, of such interactions in the current generation of models raises a key point that ESMs encompass a wide range of complexity, necessitating robust, useful and comparable evaluation across the range of ESMs.

In climate modelling, recent evaluation studies have highlighted how choice of methodology can significantly impact the conclusions reached about model skill (e.g. Radić and Clarke, 2011; Foley et al., 2013). Several studies have found that the mean of an ensemble of models outperforms all or most single models of that ensemble (e.g. Evans, 2008; Pincus et al., 2008). However, Schaller et al. (2011) have demonstrated that the multi-model mean outperforms individual models when the ability to reproduce global fields of climate variables is evaluated, but performs averagely when the ability to simulate regional climatic features is tested. This outcome highlights the need for more robust assessments of model skill, as model evaluations which use inappropriate metrics or fail to consider key aspects of the system have the potential to lead to overconfidence in model projections.

Developing robust approaches to model evaluation is challenging for a variety of reasons, which are not exclusive to carbon cycle modelling but applicable across all aspects of Earth system modelling. Datasets may lack uncertainty estimates, rendering them sub-optimal for model evaluation. Critical analysis may also be required to reconcile differences between datasets intended to describe similar phenomena (e.g. ice core global CO_2 record versus plant stomatal local CO_2 record, van Hoof et al., 2005). Furthermore, there are many different metrics in use in model evaluation and often, the rationale for applying a specific metric is unclear. These issues, along with strategies for improvement, will be discussed within this paper.

Overview of this paper

In this paper, some key uncertainties and challenges relating to model evaluation are discussed, along with new approaches and strategies for more robust evaluation. These issues, though illustrated here with examples relating to global carbon cycle modelling, are relevant across many branches of Earth system modelling.

Given that knowledge of the system under observation is essential for the assessment of model performance (Oreskes et al., 1994), we begin with a discussion of some challenges associated with the use of *modern and palaeo data* in model evaluation. An appreciation of the advantages, uncertainties and limitations of each type of data is a crucial starting point. Data validity (Sargent, 2010) is a crucial aspect. Key issues include uncertainties associated with our understanding of the changes captured in each type of record, mismatches between available data and what is needed for evaluation, and the challenges of using data collected at a specific spatial or temporal scale to develop larger-scale constraints for model evaluation.

Next, we assess the state-of-the-art with respect to metrics for model evaluation. Whether using classical metrics (such as root mean square error, correlation or model efficiency) or advanced statistical techniques (such as applied neural networks) to compare models with data and quantify model skill, there is a need to be aware of the statistical properties of metrics and statistical techniques, as well as the properties of the model variables under assessment and the corresponding evaluation datasets, in order to select an appropriate methodology. There is significant potential to draw false conclusions about model skill by using an inappropriate metric and furthermore, the complexity of the metric may render such errors difficult to detect. Recent attempts to provide a benchmarking framework for land surface model evaluation indicate a move toward setting community-accepted standards (Randerson et al., 2009; Luo et al., 2012; Kelley et al., 2012). However, different levels of complexity in ESMs, different parameterization procedures and modelling approaches, the validity of data, and an unavoidable level of subjectivity makes the task of identifying universally applicable procedures a challenge.

Finally, *recommendations for more robust evaluation* are discussed. We consider that evaluation can be process-based (“bottom-up”) or system-level (“top-down”) (Fig. 1). It can also utilize pre-calibration, or emergent constraints across multiple models, although the development of emergent constraints is a community-wide activity rather than an evaluation methodology that can be applied to an individual model. A combination of approaches can greatly increase our understanding of a model’s ability to realistically simulate processes across multiple temporal and spatial scales. For example, both locally and globally, the terrestrial carbon budget is a fine balance between large uptake (photosynthesis) and release (respiration) terms. Even if each process could be modelled with very high precision, the net effect could still be poorly constrained because of the difference in the magnitude of the fluxes. Hence, single, process-based tests are necessary but may not be sufficient. Conversely, observations of the seasonal cycle or interannual variability of carbon balance give the overall balance, but not detail about the component processes. It is possible to simulate the overall balance with a number of different combinations of the components, and therefore there is significant potential to get the right answer for the wrong reasons. Furthermore, some of the most accurate features of climate simulations (such as the pattern of near-surface temperatures) are poor predictors of the sensitivity of a model to increasing CO_2, because it is possible to get a skilful simulation of the present through cancellation of multiple errors.

Using a combination of “bottom-up” constraints on the processes and “top-down” constraints on the balance between them is required to give confidence that the model gives the right behaviour for the right reason. Consideration will also be given to how key questions arising in the paper could potentially be resolved through coordinated research activities.

2 The role of datasets in ESM evaluation

Due to intensive, but necessary, parameterization of processes, there is potential for compensating error when modelling the interactive system. Different parameter combinations are potentially able to recreate the historical record (Sitch et al., 2008; Booth et al., 2012) but structurally, erroneous parameter combinations, or even model subcomponent combinations, might compensate for each other. Another consideration is that as the models aim to simulate a complex system, due to non-linearity, even a small change in one of the components might unexpectedly influence another component of the system (Roe and Baker, 2007). As such, robust model evaluation is critical, to assist in understanding the behaviour of ESMs and the limitations of what we can and cannot represent quantitatively.

Modern and palaeo data are both used for model evaluation, although each kind of data has advantages and limitations (Table 1). Experimental data provides benchmarks for a range of carbon cycle-relevant processes (e.g. physiological behaviour in the terrestrial carbon cycle) that cannot be tested in other ways. However, for processes that are biome specific, the limited geographical scope of the relatively few existing records is problematic. Datasets also exist with more global coverage, documenting changes in the recent past (last 30–50 yr), but an inherent limitation of these modern datasets is that they sample the carbon cycle response to a limited range of climate variability. Palaeoclimate evaluation is therefore an important test of how well ESMs reproduce climate changes (e.g. Braconnot et al., 2012). The past does not provide direct analogues for the future, but does offer the opportunity to examine climate changes that are as large as those anticipated during the 21st century, and to evaluate climate system feedbacks with response times longer than the instrumental period (e.g. cryosphere, ocean circulation, some components of the carbon cycle).

2.1 Modern datasets

Evaluation analysis can benefit from modern datasets, to test and constrain components within ESMs in a hierarchical approach to evaluation (Leffelaar, 1990; Wu and David, 2002). Recent initiatives of land and ocean model evaluation and benchmarking (Randerson et al., 2009; Luo et al., 2012; Kelley et al., 2012; Dalmonech and Zaehle, 2012) gave examples of suitable modern datasets for model evaluation and their use in diagnosing model inconsistencies with respect to behaviour of the carbon cycle. These include instrumental data, such as direct measurements of e.g. CO_2, CH_4 spanning the last 30–50 yr, measurements from carbon flux monitoring networks and satellite-based data (Table 1).

Due to their excellent spatial and/or temporal resolution, satellite datasets offer the potential to explore structural and parameter uncertainties in detail, and to reveal compensating errors in ESMs. However, the lack of full independency of the data and the model is an issue that often affects such data. Satellite data is model-derived (type 1, Table 1), with some sort of model used to transform the direct measurements of the satellite into other parameters of interest. If a radiative transfer model is used for this transformation, then there will be similarities between the functions used in the retrieval and those used in a climate model, resulting in a lack of model-data independency. However, the functions are used in opposite direction; in climate models, radiative transfer functions are generally used to derive radiative forcing from the atmospheric or surface state, while during satellite retrieval, these functions are used to estimate the atmospheric or surface state from measured radiation. So, even if the functions are similar, a retrieved product can still help to evaluate whether the initial assumptions of the atmospheric or surface state of a model are correct. Statistical and change detection retrievals rely not on physical models but on statistical links between variables or on a modulation of the satellite signal. These two types of retrievals sometimes use model data for calibration but are otherwise independent of models. Statistical models in particular are not only useful to evaluate specific parameters in a model, but can also be used to perform process-based evaluations.

In addition, uncertainty estimates are not always provided or propagated during the retrieval process. This affects the evaluation methodology, and the selection of metric. For example, satellite-based datasets are an advantageous choice for evaluation of specific modelled processes primarily due to their global coverage, high spatial resolution, and consistently repeated measurements. However, one of the main concerns is the lack of full consistency between what we can observe with different satellite sensors (e.g. top of the atmosphere reflectance) and what models actually simulate (e.g. net primary productivity). As such, the choice of a particular dataset and selection of a proper methodology for robust model evaluation analysis are vital.

Despite uncertainties, modern datasets are a very rich data source with a number of useful applications. For example, robust spatial and temporal information emerging from data can be used to rule out unreasonable simulations and diagnose model weaknesses. Satellite-based datasets of vegetation activity clearly depict ecosystem response to climate variability at seasonal and interannual time scales and return patterns of forced variability that can be useful for model evaluation (Beck et al., 2011; Dahlke et al., 2012), even if bias within the dataset is greater than data-model differences (e.g. Fig. 2).

Ecosystem observations, such as FLUXNET, and ecosystem manipulation studies, such as drought treatments and free air CO_2 enrichment (FACE) experiments provide a unique source of information to evaluate process formulations in the land component of ESMs (Friend et al., 2007; Bonan et al., 2012; de Kauwe et al., 2013). Manipulation experiments (e.g. FACE experiments: Nowak et al., 2004; Ainsworth and Long, 2005; Norby and Zak, 2011) are a particularly powerful test of key processes in ESMs and their constituent components, as shown by Zaehle et al. (2010) in relation to C-N cycle coupling, and de Kauwe et al. (2013) for carbon-water cycling. It should be a key expectation that models would be able to reproduce experimental results involving manipulations of global change drivers such as CO_2, temperature, rainfall and N addition, but there are limitations.

The application of such data for the evaluation of ESMs is challenging because the limited spatial representativeness of the observations, resulting from the lack of a coherent global monitoring strategy and the high costs of running these observational sites. Upscaling monitoring data using data-mining techniques and ancillary data (such as remote-sensing and climate data) provide a means to bridge the spatial gap between the scale of observation and ESMs (Jung et al., 2011), however, at the cost of introducing model assumptions and uncertainties that are difficult to quantify and include in metric calculations. Such upscaling is near impossible for ecosystem manipulation experiments, as they are scarce, and rarely performed following a similar protocol. More and better coordinated manipulation studies are needed to better constrain

ESM prediction (Batterman and Larsen, 2011; Vicca et al., 2012). Hickler et al. (2008), for example, showed that the LPJ-GUESS model produced quantitatively realistic enhancement of NPP due to CO_2 elevation in temperate forests, but also showed greatly different responses in boreal and tropical forests, for which no adequate manipulation studies exist. Thus, these predictions remain to be tested.

Another limitation is that the interpretation of experiments is not unambiguous because it is seldom that just one external variable can be altered at a time. For example, Bauerle et al. (2012) showed that the widely observed decline of Rubisco capacity (Vcmax) in leaves during the transition from summer to autumn could be abated by supplementary lighting designed to maintain the summer photoperiod, and concluded that Vcmax is under photoperiodic control. However, their treatment also inevitably increased total daily photosynthetically active radiation (PAR) in autumn. The results can therefore be interpreted as showing that seasonal variations in Vcmax are simply related to daily total PAR.

This example illustrates a key challenge for the modelling community. One modelling response to new experimental studies is to increase model complexity by adding new processes based on the supposed advances in knowledge. We advise a more critical and cautious approach, employing case-by-case comparison of model results and experiments, rather than general interpretation of experiments, to reduce the potential for ambiguities and avoid unnecessary complexity in models (Crout et al., 2009).

2.2 Palaeo data

The key purpose of palaeo-evaluation is to establish whether the model has the correct sensitivity for large-scale processes. Models are developed using modern observations (i.e. under a limited range of climate conditions and behaviours) and we need to determine how well they simulate a large climate change, to assess whether they can capture the behaviour of the system outside the modern range. If our understanding of the physics and biology of the system is correct, we should be able to predict past changes as well as present behaviour.

Reconstructions of global temperature changes over the last 1500 yr (e.g. Mann et al., 2009) are primarily derived from tree-ring and isotopic records, while reconstructions of climates over the last deglaciation and the Holocene (e.g. Davis et al., 2003; Viau et al., 2008; Seppä et al., 2009) are primarily derived from pollen data, although other biotic assemblages and geochemical data have been used at individual sites (e.g. Larocque and Bigler, 2004; Hou et al., 2006; Millet et al., 2009). Marine sediment cores have been used extensively to generate sea surface temperature reconstructions (e.g. Marcott et al., 2013), and to reconstruct different past climate variables (see review in Henderson, 2002) related to ocean conditions, but the interpretation of these data is often not straightforward, since the measured indicators are frequently influenced by more than one climatic variable (e.g., the benthic δ^{18}O measured in foraminiferal shells contains information on both global sea level and deep water temperature). Errors associated with the data and their interpretation also need to be stated, as while analytical errors on the measurements are often small, errors in the calibrations used to obtain reconstructions tend to be much bigger. As such, the incorporation of palaeoproxy data such as marine carbonate concentrations (e.g. Ridgwell et al., 2007), δ^{18}O (e.g. Roche et al., 2004) or δ^{13}C (e.g. Crucifix, 2005;) in models is an important advance, as it allows comparison of model outputs directly with proxy data, increasing understanding of the proxies themselves, and the different climatic variables that affect them.

Ice cores provide a polar contribution to climate response reconstruction, as well as crucial information on a range of climate-relevant factors; e.g. forcing by solar (through ^{10}Be), volcanic (through sulphate spikes) and greenhouse gas (e.g. CO_2, CH_4, N_2O) and dust changes. CH_4 can be measured from both Greenland and Antarctic ice cores, while CO_2 measurements require Antarctic cores, due to high concentrations of impurities in Greenland samples in-situ producing extra CO_2 (Tschumi and Stauffer, 2000; Stauffer et al., 2002). For the recent past (i.e. last few millennia), choosing sites with the highest snow accumulation rates yields decadal resolution. The highest resolution records to date are from Law Dome (MacFarling Meure et al., 2006), making this data more reliable for model evaluation (e.g. Frank et al., 2010), although further work at high accumulation sites would provide further reassurance on this point. Over longer time periods, sites with progressively lower snow accumulation rates, and therefore lower intrinsic time resolution, have to be used. Through the Holocene (last ~ 11 000 yr) (Elsig et al., 2009) and the termination out of the last glacial maximum (LGM) into the Holocene (Lourantou et al., 2010; Schmidt et al., 2012), there are now high quality $^{13}C/^{12}C$ of CO_2 data available, as well as much improved information about the phasing between the change in Antarctic temperature and CO_2 (Pedro et al., 2012; Parrenin et al., 2013), and between CO_2 and the global mean temperature (Shakun et al., 2012) across the termination.

Compared to the amount of effort spent on reconstructing past climates and atmospheric composition, there are comparatively few datasets that provide information on different components of the terrestrial carbon cycle. Nevertheless, there are datasets that provide information on changes in vegetation distribution (e.g. Prentice et al., 2000; Bigelow et al., 2003; Harrison and Sanchez Goñi, 2010; Prentice et al., 2011a), biomass burning (Power et al., 2008; Daniau et al., 2012), and peat accumulation (e.g. Yu et al., 2010; Charman et al., 2012). These datasets are important because they can be used to test the response of individual components of ESMs to changes in forcing.

The major advantage of evaluating models using the palaeo record is that it is possible to focus on times when the signal is large compared to the noise. The change in forcing at the Last Glacial Maximum (LGM) relative to the pre-industrial control is of comparable magnitude, though opposite direction, to the change in forcing from quadrupling CO_2 relative to that same control (Izumi et al., 2013). Thus, comparisons of palaeoclimatic simulations and observations since the LGM can provide a measure of individual model performance, discriminating between models, and allowing diagnosis of the sources of model error for a range of climate states similar in scope to those expected in the future (Harrison et al., 2013). However, as is the case for many modern observational datasets (e.g. Kelley et al., 2012), not all palaeo-reconstructions provide adequate documentation of errors and uncertainties and there is a lack of standard-

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◂◂ | ▸▸
◂ | ▸
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

ization between datasets where such estimates are provided (e.g. Leduc et al., 2010; Bartlein et al., 2010). Reconstructions based on ice or sediment cores are intrinsically site-specific (except for the globally significant greenhouse gas records), therefore many records are ideally required to synthesise regional or global distribution patterns and estimates (Fig. 3). Community efforts to provide high quality compilations of already available data (e.g., Waelbroeck et al., 2009; Bartlein et al., 2010) make it possible to use palaeo data for model evaluation, but an increase in the coverage of palaeo-reconstructions is still required to address many regional signals.

Most attempts to compare simulations and reconstructions in palaeo-mode focus on qualitative agreement of simulated and observed spatial patterns (e.g. Otto-Bliesner et al., 2007; Miller et al., 2010). There has been surprisingly little use of metrics for data-model comparisons (for exceptions see e.g. Guiot et al., 1999; Paul and Schäfer-Neth, 2004; Harrison et al., 2013). This probably reflects problems in developing meaningful ways of taking uncertainties into account in these comparisons. Quantitative assessments have generally focused on individual large-scale features of the climate system, for example the magnitude of insolation-induced increase in precipitation over northern Africa during the mid-Holocene (see e.g. Joussaume et al., 1999; Jansen et al., 2007), of zonal cooling in the tropics at the LGM (Otto-Bleisner et al., 2009), or of the amplification of cooling over Antarctica relative to the tropical oceans at the LGM (Masson-Delmotte et al., 2006; Braconnot et al., 2012). Comparisons of simulated vegetation changes have been based on assessments of the number of matches to site-based observations from a region (e.g. Harrison and Prentice, 2003; Wohlfahrt et al., 2004, 2008). Observational uncertainty is represented visually in such comparisons, and only used explicitly to identify extreme behaviour amongst the models. Nevertheless, the recent trend is towards explicit incorporation of uncertainties and systematic model benchmarking (see e.g. Harrison et al., 2013; Izumi et al., 2013).

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
|◀ | ▶|
◀ | ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

3 Key metrics for ESM evaluation

“Metrics” are simple formulae or mathematical procedures that measure the similarity or difference between two datasets. Many different metrics have been proposed in the literature (Tables 2–4), and the choice of an appropriate metric in model evaluation is crucial because the use of inappropriate metrics can lead to overconfidence in model skill. The choice should be based on the properties of the metric, the properties of the datasets, and the specific objectives of the evaluation analysis. Metric formalism – that is, the treatment of metrics as well-defined mathematical and statistical concepts – can help the interpretation of metrics, their analysis, or their combination into a “skill-score” (Taylor, 2001) in an objective way.

The use of metrics draws on the mathematical concept of “distance” ($d(x,y)$), expressed in terms of three characteristics: separation: $d(x,y) = 0 \leftrightarrow x = y$, symmetry: $d(x,y) = d(y,x)$, and triangle inequality $d(x,z) \leqq d(x,y) + d(y,z)$. The two datasets could be two model outputs, where the metric is used to measure how similar the two models are, or one model output and one reference observation dataset, where the metric is used to evaluate the model against real measurements. Three levels of metric complexity can be identified, relating to the state-space on which to apply the distance:

- Level 1 – “comparisons of raw bio-geophysical variables”. Here the distance generally reflects errors and provides assessment of model performance where there is a reasonable degree of similarity between the model and reference dataset (such as climate variables in weather models).

- Level 2 – “comparisons of statistics on bio-geophysical variables”. Here the distance is measured on a statistical property of the datasets. This is particularly useful for models that are expected to characterise the statistical behaviour of a system (e.g. climate models). This level is appropriate for most of the biophysical variables simulated by ESMs.

- Level 3 – “comparisons of relationships among bio-geophysical variables”. Here, the distance is diagnostic of relationships related to physical and/or biological processes and this level of comparison is therefore useful for understanding the behaviour of two datasets.

At all levels of metric complexity, the metric needs to be synthetic enough to aid in understanding the similarities and differences between the two datasets, and be understandable by non-specialists in order to facilitate its use by other communities. Next, the particular uses, advantages and limitations of metrics in each level of metric complexity will be discussed.

3.1 Metrics on raw bio-geophysical variables

Level 1 metrics are probably the most widely used. The distance measures the discrepancies between two datasets of a key bio-geophysical variable. Discrepancies can be measured at site level or at pixel level for gridded datasets, and thus such comparisons can be used for model evaluation against sparse data, e.g. site-based NPP data (e.g. Zaehle and Friend, 2010), eddy-covariance data (e.g. Blyth et al., 2011), or atmospheric CO_2 concentration records at remote monitoring stations (e.g. Cadule et al., 2010; Dalmonech and Zaehle, 2012). Where there is sufficient data to make the calculation meaningful, comparisons can be made against spatial averages or global means of the bio-geophysical variables. Comparisons can also be made in the time domain because climate change and climate variability act on Earth system components across a wide range of temporal scales. The distance can thus be measured on instantaneous variables or on time-averaged variables, such as annual means. Many distances, summarized in Table 2, can be considered to measure these discrepancies.

The Euclidean distance (Eq. 2) is the most commonly used distance. It is more sensitive to outliers compared to the Manhattan distance (Eq. 1). Both of these distances assume that direct comparisons of the data can be made. Some examples are reported in Joliff et al. (2009), where the Euclidean distance is used to evaluate three ocean surface bio-optical fields.

In the case of the weighted Euclidean distance (Eq. 3), a weight is associated with each variable. This is useful for various reasons: (1) normalization against a mean value provides a dimensionless metric and allows comparisons to be drawn between datasets with different orders of magnitude; (2) the weighting can take account of uncertainties in the reference dataset (e.g. instrumental errors in an observational dataset, or uncertainty in a model ensemble); (3) this type of metric can be useful when the data have a different dynamical range. For example, in a time series of Northern Hemisphere monthly surface temperature, the variability is different for summer and winter, and it makes sense to normalize the differences by the variance.

The Chi-Squared "distance" (Eq. 4) is related to the Pearson chi-squared test or goodness-of-fit, and differs from previous distances discussed here as it measures the similarity between two Probability Density Functions (PDFs), rather than between data points. It is particularly useful if the focus of the analysis is at the population level. Distances on PDFs are defined, in this paper, to be Level 2 metrics, but the Chi-square distance can be used when the geophysical variables are supposed to have a particular shape (e.g. an atmospheric profile of temperature). Equation (5) can also be used, in particular to facilitate the symmetry property of distances.

The Tchebychev distance (Eq. 6) can be used, for example, to identify the maximum annual discrepancy in a climatic run. It can be useful if the focus is on extreme events.

The Mahalanobis distance (Eq. 7) is particularly suitable if variables have very different units, as each one will be normalized by its variance, and if they are correlated with each other, since the distance takes these correlations into account. High correlation between two datasets has no impact on the distance computed, compared to two independent datasets. This distance is directly related to the quality criterion of the variational assimilation and Bayesian formalism that optimally combines weather forecast and real observations. This criterion needs to take into account the covariance matrices and the uncertainties of the state variables.

Interesting links can be established between metrics and the operational developments of the numerical weather prediction centres. The Mahalanobis distance is well suited for Gaussian distributions (meaning here that the data/model misfit distribution follows a Gaussian distribution with covariance matrix A, e.g. Min and Hense, 2007). General Bayesian formalism can be used to generalize this distance to more complex distributions. The Mahalanobis distance and the more general Bayesian framework are particularly suitable to treat several evaluation issues at once, such as the quantification of multiple sources of error and uncertainty in models or the combination of multiple sources of information (including the acquisition of new information). For instance, Rowlands et al. (2012), use a goodness-of-fit statistic similar to the Mahalanobis distance applied to surface temperature datasets.

We present here distances between two points, possibly multivariate. Some metrics use these distances and have been defined over the two whole datasets D_1 and D_2. For example, the Normalized Mean Error (NME) is a normalization of the bias between the two datasets (Eq. 8). Similarly, the Normalized Mean Square Error (NMSE) is a normalization of the Euclidean distance used in Eq. (2) The Nash–Sutcliffe Model Efficiency Coefficient (NSMEC) is again the Euclidean distance, normalized by the variance in dataset D_1.

Several other distances exist in the literature that have been applied in different scientific fields and that are not listed here (e.g. Deza and Deza, 2006). However most of these distances are particular cases or an extension of the preceding distances.

3.2 Metrics on statistical properties

Level 2 metrics, summarised in Table 3, use statistical quantities estimated for two datasets D_1 and D_2. Some of the metrics presented in the previous section can then be applied to the selected statistics. For instance, the PDF can be estimated for both datasets and the Chi-squared distance can be used to measure their discrepancy. For example, Anav et al. (2013) compared the PDFs of GPP and LAI from the CMIP5 model simulations with two selected datasets.

The Kullback–Leibler divergence (Eq. 9) is based on information theory and can also be used to measure the similarity of two PDFs. The Kolmogorov–Smirnov distance can be used when it is of interest to measure the maximum difference between the cumulative distributions. Tchebychev or other distances acting on estimated seasons are also considered here to be level 2 metrics, since the seasons are statistical quantities estimated on D_1 and D_2 (although very close to level 1 raw geophysical variables). Similarly the distance can operate on derived variables from the original time series as decomposed signals in the frequency domain. Cadule et al. (2010), for example, analysed model performance in terms of representing the long-term trend and the seasonal signal of the atmospheric CO_2 record.

The variance of data and model is often used to formulate metrics for the quantification of the data-model similarity. In coupled systems, the use of a metric based on distance can become inadequate; the metric no longer facilitates definite conclusions on the model error, because it includes an unknown parameter in the form of the unforced variability. Furthermore, when applied to spatial fields, as variance is strongly location-dependent, a global spatial variance can be misleading. Gleckler et al. (2008) proposed a more suitable model variability index which has been applied to climatic variables, but is also highly applicable to several of the biophysical variables simulated by land and ocean coupled models, and thus relevant to the carbon cycle.

A distance can also operate on derived variables such as decomposed signals in the frequency domain from the original time series. The metric can also focus on extreme events, with the distance acting on the percentile, assuming that the length of the records is sufficient to characterize these extremes.

3.3 Metrics on relationships

Level 3 metrics, summarized in Table 4, focus on relationships. The aim here is to diagnose a physical or a biophysical process that is particularly important; e.g. the link between two variables in the climate system. Various “relationship diagnostics” have been used, summarized in Table 4.

The correlation between two variables is a very simple and widely used metric; it satisfies the need to compare the data-model phase correspondence of a particular bio-geophysical variable. In this case parametric statistics such as the Pearson correlation coefficient (Eq. 10), or non-parametric statistics such as the Spearman correlation coefficient, are directly used as a metric. This is particularly used to evaluate the correspondence of the mean seasonal cycle of several variables, from precipitation (Taylor, 2001), to LAI, GPP (Anav et al., 2013), and atmospheric CO_2 (Dalmonech and Zaehle, 2012).

The sensitivity of one variable to another can be estimated using simple to very complex techniques (Aires and Rossow, 2003). It can be obtained by dividing concomitant perturbations of the two variables using spatial or temporal differences (Eq. 11), or by perturbing a model and measuring the impact when reaching equilibrium. The first approach can be used to evaluate, for example, site-level manipulative experiments to estimate carbon sensitivity to soil temperature or nitrogen deposition in terrestrial ecosystem models (e.g. Luo et al., 2012).

From the linear regression of two variables the slope or bias can be compared for D_1 or D_2 (Eq. 12). The slope is very close to the concept of sensitivity, but sensitivities are very dependent on the way they are measured. For example, sensitivity of the atmospheric CO_2 to climatic fluctuations may depend on the timescales they are calculated on (Cadule et al., 2010). An alternative, when more than two variables are involved in the physical or biophysical relationship under study, is a multiple linear regression (Eq. 13), or any other linear or nonlinear regression model such as neural networks. See, for example, the results obtained at site-level by Moffat et al. (2010).

Pattern-oriented approaches use graphs to identify particular patterns in the dataset. These graphs aim at capturing relationships of more than two variables. For example, in Bony and Dufresne (2005), the tropical circulation is first decomposed into dynamical regimes using mid-tropospheric vertical velocity and then the sensitivity of the cloud forcing to a change in local sea surface temperature (SST) is examined for each dynamical regime. Moise and Delage (2011) proposed a metric that assesses the similarity of field structure of rainfall over the South Pacific Convergence Zone in terms of errors in replacement, rotation, volume, and pattern. The same metric could be applied to ocean Sea-viewing Wide Field-of-view Sensor (SeaWiFS) satellite-based fields in areas where particular spatial structures emerge.

Clustering algorithms have been used to obtain weather regimes based only on the samples of a dataset. For example, Jakob (2003) and Chéruy and Aires (2009) obtained cloud regimes based on cloud properties (optical thickness, cloud top pressure). The same methodology can be used in D_1 and D_2 and the two sets of regimes can be compared. The regimes can also be obtained on one dataset and only the regime frequencies of the two datasets are compared. Abramowitz and Gupta (2008) applied a distance metric to compare several density functions of modelled net ecosystem exchange (NEE) clustered using the “self-organizing map” technique.

It is often difficult to use a real mathematical distance to measure the discrepancy between the two “relationship diagnostics”. Although very useful for understanding differences in the physical behaviour, the simple comparison of two graphs (for D_1 and D_2) is not entirely satisfactory since it does not allow combination of multiple metrics or definition of scoring systems. In this paper, it is not possible to list all the ways to define a rigorous distance on each one of the relationship diagnostics that have been presented: Euclidean distance can be used on the regression parameters or the sensitivity coefficients, or two weather regime frequencies can be measured using confusion matrices (e.g. Aires et al., 2011). The distance needs to be adapted to the relationship diagnostic. The most limiting factor to this type of approach for ESM evaluation is that the relationship obtained might be not robust enough (i.e. statistically significant) or not easily framed within a process-based context.

4 Recommendations for robust model evaluation

4.1 A framework for robust model evaluation

Robust model evaluation relies on a combination of approaches, each informed using appropriate data and metrics (Fig. 4). Calibration and, ideally, pre-calibration must first be employed to rule out implausible outcomes, using data independent to that which may subsequently be used in model evaluation. Then, evaluation approaches must be a combination of process-based and system-wide, to ensure that both the representation of processes and the balance between them are realistic in the model. Optionally, the results of different model evaluation tests can be combined into a single model score, perhaps for the purposes of weighting future projections. When employed as part of a multi-model ensemble, the simulation can also contribute to the calculation of emergent constraints, which can then be used in subsequent model development.

4.2 Recommendations for improved data availability and usage

The increasingly data-rich environment is both an opportunity and a challenge, in that it offers more opportunities for model validation but requires more knowledge about the generation of datasets and their uncertainties in order to determine the best dataset for validating specific processes. While improved documentation of datasets would go some way to alleviating the latter problem, there is scope for improved collaboration between the modelling and observational communities to develop an appropriate benchmarking system, that evolves to reflect new model developments (such as representing ecosystem-scale responses to combined environmental drivers) not addressed by existing benchmarks.

4.2.1 Coordinate data collection efforts

A key question for both the modelling and data communities to address together is how well model evaluation requirements and data availability are reconciled. There is an on-going and critical need for new and better datasets for model evaluation, datasets which are appropriately documented and for which useful information about errors and uncertainties are provided. The temporal and spatial coverage of datasets also needs to be sufficient to capture potential climatic perturbations, a point that is illustrated in the evaluation of marine productivity, a key variable in controlling marine carbon fluxes and exchanges of carbon with the atmosphere. Modelling studies offer conflicting evidence of its behaviour under a changing climate (e.g. Sarmiento et al., 2004; Steinacher et al., 2010; Taucher et al., 2012; Laufkötter et al., 2013), therefore model evaluation is essential. Recent compilations of observations of marine-productivity proxies give us a reasonably well-documented picture of qualitative changes in productivity over the last glacial-interglacial transition (e.g. Kohfeld et al., 2005), and in response to Heinrich events (e.g. Mariotti et al., 2012). These datasets are being used to evaluate the same ESMs used to predict changes in NPP in response to climate change (e.g. Bopp et al., 2003; Mariotti et al., 2012), and these studies show reasonable agreement. On more recent timescales, remote sensing observations of ocean colour have been used to infer decadal changes in marine NPP. Studies show an increase in the extent of oligotrophic gyres over 1997–2008 with the SeaWiFS data (Polovina et al., 2008). However, on longer time-scales, and using Coastal Zone Color Scanner (CZCS) and SeaWiFS datasets, analysis yields contrasting results of increase or decrease of NPP from 1979–1985 to 1998–2002 (Gregg et al., 2003; Antoine et al., 2005). Henson et al. (2010) have shown that at least 20–30 yr of uninterrupted data would be needed to detect any global warming-induced changes in NPP, highlighting that the need for continued, focused data collection efforts.

4.2.2 Maximise usefulness of current data in modelling studies

Modelling studies should be designed in a manner that makes the best use of the available data. For example, equilibrium model simulations of the distant past require time slice reconstructions for testing issues relating to the carbon cycle. These reconstructions rely on synchronisation of records from ice cores, marine sediments and terrestrial sequences, to take account of differences between forcings and responses in different archives, which is a significant effort even within a particular palaeo archive, let alone across multiple archives. Yet the strength of palaeo data is precisely that it offers information about rates of change, and such information is discarded in a time slice simulation. For that reason, the increasing use of transient model runs is particularly important.

There is an also an increasing need for forward modelling to simulate the parameters that are actually measured and observed during data collection, such as isotopes in ice cores and pollen. Ice core gas concentration measurements are unusual because what is measured is what we want to know, and is a variable that most climate models yield as direct outputs. This is not generally the case, nor are all model setups easily able to simulate even the trace gas isotopic data that are available from ice. A corollary of this is that we need to recognise the difficulty of trying to reconstruct with palaeo data quantities that are essentially model constructs, e.g inferring the strength of the meridional overturning circulation (MOC) from the ^{231}Pa/^{230}Th ratio in marine sediment cores (McManus et al., 2004). In the latter context, direct simulation of the ^{231}Pa/^{230}Th ratio is necessary to properly deconvolute the multiple competing processes (Siddall et al., 2005, 2007).

4.2.3 Use data availability to inform model development

Model development should also focus on incorporating processes that, at least collectively, are constrained by a wealth of data. Notable examples are processes such as those governing methane emissions (e.g. from wetlands and permafrost) and the removal of methane from the atmosphere (e.g. via oxidation by the hydroxyl radical and atomic chlorine). There are four main observational constraints on the methane (CH_4) budget with which we can evaluate the performance of ESMs: the concentration of CH_4, [CH_4]; its isotopic composition with respect to carbon and deuterium, $\delta^{13}CH_4$ and δD (CH_4); and CH_4 fluxes at specific sites. We have no natural record of CH_4 fluxes so their use in ESM evaluation is limited to the relatively recent period in which they have been measured, though recent measurements of CH_4 fluxes at specific sites can be used to verify spatial and seasonal distributions of CH_4 emissions inferred from tall tower and satellite measurements of [CH_4], by inverse modelling. However, a range of [CH_4], $\delta^{13}CH_4$, and $\delta D(CH_4)$ records are available, spanning up to 800 000 yr in the case of polar ice cores, which can be used to evaluate the ability of ESMs to capture changes to the CH_4 budget in response to past changes in climate. The variety of climatic changes we can probe, from large glacial-interglacial changes spanning thousands of years to substantial changes over just a few tens of years at the beginning of Dansgaard–Oeschger events, and still more rapid, subtle changes following volcanic eruptions, enables us to evaluate the ability of ESMs to capture both the observed size and speed of changes known to have taken place. The complementary nature of the [CH_4], $\delta^{13}CH_4$, and $\delta D(CH_4)$ constraints is key to ESM evaluation. Each CH_4 source and sink affects these three constraints in different ways. As such, scenarios that can explain only one set of observations can be eliminated. For instance, an increase in CH_4 emissions from tropical wetlands, biomass burning or methane hydrates could explain an increase in [CH_4], but of these, only an increase in biomass burning emissions could explain an accompanying enrichment in $\delta^{13}CH_4$. Of course, more than one factor can change at a time, but the key point is that the most rigorous test of ESM performance utilizes all three constraints and, therefore, ESMs should track the influence of each CH_4 source and sink on each of these constraints.

4.3 Recommendations for model calibration

4.3.1 Key principles of model calibration

Model evaluation is closely linked to model calibration (or tuning). ESMs contain a large number of poorly constrained parameters, resulting from incomplete knowledge of certain processes or from the simplification of complex computations, which can be calibrated in order to improve model behaviour. In general, model calibration should follow a number of fundamental guiding principles. The principles detailed here are mostly based on the discussion in Petoukhov et al. (2000) for the CLIMBER-2 model.

First, parameters which are well constrained from observations or from theory must not be used for model calibration. Normally it would be physically inappropriate to modify the values of fundamental constants, for example, or use a value for a parameter which is different from the accepted empirical measurement just to improve the performance of the model.

Second, whenever possible, parameterizations and sub-modules should be tuned separately against observations rather than in the coupled system. In the case of parameterizations, this ensures that they represent the physical behaviour of the process described rather than their effect on the coupled system. The same principle should be applied as far as possible to the individual sub-modules of any Earth-system model to make sure that their behaviour is self-consistent and to facilitate calibration of the much more complex fully-coupled system.

Third, parameters must describe physical processes rather than unexplained differences between geographic regions. It is preferable for the model to represent the physical behaviour of the system rather than apply hidden flux corrections.

Fourth, the number of tuning parameters must be smaller than the predicted degrees of freedom, which is usually large for the case of Earth-system models.

Finally, one of the key challenging points relating to data used in ESM evaluation is to what extent ESM development and evaluation data are independent. In principle, the same observational data should not be used for calibration and evaluation. This is difficult to enforce in practice, however. Even if the observational data are divided into two parts, with one part used for calibration and the other for evaluation, for example, any mismatch in the evaluation will likely lead to a readjustment of model tuning parameters, making the evaluation not completely independent of the calibration procedure (Oreskes et al., 1994). Standard leave-one-out cross-validation techniques divide calibration datasets into multiple subsets, sequentially testing the calibration on each left-out subset (in the limit each data point) in turn but in Earth system modelling the subsets are unlikely to be fully independent.

4.3.2 Pre-calibration

The essence of pre-calibration is to apply weak constraints to model inputs in the initial ensemble design and weak constraints on the model outputs to rule out implausible regions of input and output spaces (Edwards et al., 2011).

The pre-calibration approach is based on relatively simple statistical modelling tools and robust scientific judgements but avoids the formidable challenges of applying full Bayesian calibration to a complex model (Rougier, 2007). A large set of model experiments sampling the variability in multiple input parameter values with the full simulator, here the Earth system model, is used to derive a statistical surrogate model or "emulator" of the dependence of key model outputs on uncertain model inputs. The choice of sampling points must be highly efficient to span the input space and is usually based on Latin hypercube designs. The resulting emulator is computationally many orders of magnitude faster than the original model and can therefore be used for extensive, multi-dimensional sensitivity analyses to understand the behaviour of the model. Holden et al. (2009, 2013a,b) demonstrated the approach in constraining glacial and future terrestrial carbon storage.

The process is usually iterative, in that a large proportion of the initial parameter space may be deemed implausible, but one or more subsequent simulated ensembles can be designed by rejection sampling from the emulator to locate the not-implausible region of parameter space. The resulting simulated ensembles are then used to refine the emulator and the definition of the implausible space. The final output is an emulator of model behaviour and an ensemble of simulations, corresponding to a subset of parameter space that is deemed "plausible" in the sense that simulations from the identified parameter region do not disagree with a set of observational metrics by more than is deemed reasonable for the given simulator. The level of agreement is therefore dependent on the model and represents an assessment of the expected magnitude of its structural error. The plausible ensemble, however, is a general result for the model that can be applied to any relevant prediction problems and embodies an estimate of the structural and parametric error inherent in the model predictions.

Ideally, pre-calibration is a first step in a full Bayesian calibration analysis. The advantage of the logistic mapping or pure rejection sampling approach used is that, because no weighting is applied, a subsequent Bayesian calibration can be applied to refine the evaluation without any need to unravel convolution effects or the multiple use of constraints. In practice, however, the pre-calibration step can be sufficient to extract all the information that is readily available from top-down constraints given the magnitude of uncertainties in inputs and of structural errors in intermediate complexity Earth system models.

4.4 Recommendations for model evaluation methodologies

Both bottom-up and top-down evaluation is required for evaluating ESMs: the first approach can give process-by-process information but not the balance between them; the second will give the balance but not the single terms. When bottom-up, process-based improvements can be shown to have top-down, system-level benefits, then we know our multi-pronged evaluation has worked.

4.4.1 Process-based evaluation (bottom-up)

Bottom-up, process-based evaluation will often require combinations of data to create the appropriate metrics as it is more likely to focus on the sensitivity of one output variable to changes in a single input. For example to assess if a model has the right sensitivity of NPP to precipitation a test could be to compute the partial derivative of NPP with respect to precipitation at constant values of temperature, radiation etc. for the model and observations (Randerson et al., 2009). This approach requires processing a dataset of, in this case, NPP to combine it with precipitation data to derive a relationship. The same NPP data could be combined with temperature data to derive a similar NPP(T) relationship. This is much more likely to isolate at least a small number of processes than simply comparing simulated NPP to a map or timeseries.

It is also widespread for model development to focus on specific features or aspects of the model required in order to have faith in the model's ability to make projections. For example climate modelling centres may focus on the ability of their GCMs to represent coupled phenomena such as ENSO, or the timing and intensity of monsoon systems. In this way, bottom-up evaluation pinpoints important model processes, and helps to confirm that the model is a sufficiently accurate representation of the real system, giving the right results for the right reasons. However, a key limitation of this approach is that the relevant observations needed to assess a particular process may not exist.

Process-based evaluation requires metrics based on process-based sensitivities, as described in Sect. 3.2. Sensitivity analysis (e.g. Saltelli et al., 2000; Zaehle et al., 2005) may be useful to determine the parameters and processes to focus on in a bottom-up evaluation. In this approach, a simple statistical model is used to represent the physical relationships in the reference dataset. A similar model is calibrated on the model simulations and the complex multivariate and non-linear relationships can then be compared. Measuring these sensitivities allows prioritization of the important parameters to validate in the model and isolate processes not well simulated in the model.

For example, Aires et al. (2013) used neural networks to develop a reliable statistical model for the analysis of land-atmosphere interactions over the continental US in the North American Regional Reanalysis (NARR) dataset. Such sensitivity analyses enable identification of key factors in the system and in this example, characterization of rainfall frequency and intensity according to three factors: cloud triggering potential, low-level humidity deficit, and evaporative fraction.

4.4.2 System-level evaluation (top-down)

Top-down constraints tend to focus on whole-system behaviour and are more likely to involve evaluation of spatial or timeseries data. Typical quantities used for top-down evaluations include surface temperature, pressure, precipitation, and wind speed maps. Observational datasets exist for many of these quantities throughout the atmosphere, so zonal-mean, or 3-dimensional comparisons are also possible (Randall et al., 2007). Anav et al. (2013) extend this approach to assess new biogeochemical outputs of CMIP5 ESMs, such as distribution and time evolution of carbon stores and fluxes.

The appropriate choice of metrics is important, as discussed in Sect. 3. A correlation coefficient might seem an obvious choice to assess the seasonal cycle of a given variable, but a model with the right phase of seasonal cycle but a magnitude 5 times too big/small would score a high correlation coefficient, while a model with the correct magnitude but lagged by just one month would score poorly. To overcome these limitations of correlation-based metrics, additional metrics such as mean error should be included in the analysis to aid interpretation of the correlation, while lag errors could first be corrected so that the correlation gives a more meaningful result. There are also many studies which have attempted to overcome this issue by presenting summary statistical metrics for multiple components across multiple models.

Taylor (2001) is one of the examples where a metric based on correlation and a distance metric have been developed as a skill score. Gleckler et al. (2008) use Taylor diagrams to compare the performance of models in terms of both the magnitude and phase of the seasonal cycle. Reichler and Kim (2008) normalise model error variance on a gridpoint basis to come up with a composite score and measure progress in model skill between generations of IPCC reports. Such scoring systems can be useful to synthesize the results of numerous metric comparisons, but should be used with caution as they can be hard to interpret – it is not always clear what model failing has led to a low score. The choice of which observations to use in the weighting is also subjective.

Model errors will inevitably evolve in time, affecting the reliability of simulations of future Earth system states. Measuring this type of uncertainty is an extremely difficult challenge. Presently, the best approach is to use expert judgement to estimate the growth of errors beyond the known forcing space, and this logic underpins the large, subjective choice of input ranges in the precalibration technique. Palaeoclimate analysis expands the space of forcings applied to the Earth system, such that possible future states might be more likely to occur inside the envelope of testable simulations. With sufficient high-quality data, it would be possible to cross-validate predictions against extreme past states that stretched the envelope in the most appropriate way. The Paleocene–Eocene Thermal Maximum offers perhaps the best opportunity for this, due to the large difference from current climate and atmospheric CO_2 conditions. An advanced theoretical approach is the "Reification" technique of Goldstein and Rougier (2009), which allows the error in a given model to be successively related to more and more accurate models, but its implementation is very much under development (see Williamson et al., 2012).

4.4.3 Emergent constraints

ESM evaluation suffers from a timescale problem; the observational data that we have is on short timescales and therefore, does not relate directly to the Earth System sensitivities of interest (e.g. climate sensitivity to CO_2, or carbon cycle sensitivity to climate) that are most important for the 21st and 22nd centuries. Analogue approaches, which evaluate ESM sensitivity against known changes in the past, are also limited by observational data, as the analogue events in the Earth's past are obviously not as well characterised as the contemporary period.

An innovative way to convert the copious short-timescale information available for the contemporary period into longer-timescale constraints on Earth system sensitivity is through "emergent constraints" (Table 5), which relate some observable aspect of the contemporary Earth system to a key system sensitivity, using an ensemble of Earth system simulations (Collins et al., 2012). The archetypal example of this relates the magnitude of the snow-albedo feedback to the size of the seasonal cycle in snow-cover in the Northern Hemisphere, across more than twenty GCMs (Hall and Qu, 2006). Since the seasonal cycle of snow-cover can be estimated from observations, this model-derived relationship provides a means to convert observations to a constraint on the size of the snow-albedo feedback in the real climate system, for which there is no direct reliable measurement. A similar emergent constraint has been used that relates the sensitivity of the interannual variability in atmospheric CO_2 to the loss of carbon from tropical land under climate change (Cox et al., 2013).

In general terms, such emergent constraint methods build on the realisation that analysis of short-time fluctuations in a system can assist in determining the sensitivity of that system to external forcing (Leith, 1975). Conversely, valuable information is unnecessarily lost when taking long-term trends and ignoring the shorter-timescale variations about these trends. Such emergent constraints utilize the large differences amongst Earth System model projections to reduce uncertainties in the sensitivities of the real Earth System to anthropogenic forcing.

5 Outlook

Although the current generation of ESMs encompass a wide range of processes, they are likely to become increasingly complex as processes that are currently being explored in, for example, dynamic global vegetation models such as better representation of nutrient cycles (e.g. Gotangco Castillo et al., 2012), fire (e.g. Thonicke et al., 2010; Prentice et al., 2011b; Pfeiffer et al., 2013), permafrost (e.g. Lawrence et al., 2012) and wetland (e.g. Collins et al., 2011) dynamics, or dust- (e.g. Shannon and Lunt, 2010), vegetation-climate interactions (Quillet et al., 2010) and aerosol-climate interactions (Woodage et al., 2010; Bellouin et al., 2011) are incorporated. This growing complexity has the potential to mask model errors, making robust evaluation of the model and its components essential.

Common to any dynamical system that needs to be evaluated, key challenges include choosing the most important variables in the system, identifying the fundamental relationships, estimating non-linear and multivariate sensitivities, and analysing the interaction between processes. We have outlined how approaches such as pre-calibration and robust calibration, along with a combination of process- and system-level evaluation with relevant data, can be used to characterise model skill. We have also illustrated the usefulness of emergent constraints to further refine model outcomes

A key limitation of current model evaluation approaches is that the widely used statistical measures of sensitivities are based on "coincident increments", such as correlations, not on causality. A very interesting extension of sensitivities would investigate causal links among the important parameters in the system. Some tentative studies have investigated measures such as Granger causality; see Notaro et al. (2006) for an application to vegetation patterns. However, a more complete framework needs to be used (Pearl, 2009). Due to the complexity of this type of work, a close collaboration of climate-carbon cycle scientists and statisticians would be required.

Model complexity and structure has to be kept in mind when making comparisons of skill with respect to any given metric across a range of ESMs. Comparing models of different complexity could create an artificially large model spread, which does not reflect current process knowledge. However, comparing only models of similar complexity could lead to underestimation of the true uncertainty in model projections due to structural similarities between models and restricted sample size.

Benchmarking models against a set of well-chosen observations (Sect. 2) and using appropriate metrics (Sect. 3) should be considered a vital step in model evaluation. However, while it may be tempting to simply evaluate the performance of the model against every dataset that can be found – and indeed a "perfect" model should be able to withstand such a test – if this comes at the expense of being able to interpret the results then it may be more beneficial to focus on a smaller set of tests which may be more important for the model to perform well in. This level of discrimination is inevitably an expert judgement, but is required if the field of ESM evaluation is to move from "beauty contest" to constraint.

Acknowledgements. This paper emerged from the GREENCYCLESII mini-conference "Evaluation of Earth system models using modern and palaeo observations" held in Cambridge, UK, in September 2012. We would like to thank the Marie Curie FP7 Research and Training Network GREENCYCLESII for providing funding which made this meeting possible. Research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7 2007–2013) under grant agreement no. 238366. The work of C. D. Jones was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). N. R. Edwards acknowledges support from FP7 grant no. 265170 (ERMITAGE).

References

Abramowitz, G. and Gupta, H.: Toward a model space and model independence metric, Geophys. Res. Lett., 35, L05705, doi:10.1029/2007GL032834, 2008.

Ainsworth, E. A. and Long, S. P.: What have we learned from 15 years of free-air CO_2 enrichment (FACE)? A meta-analytic review of the responses of photosynthesis, canopy properties and plant production to rising CO_2, New Phytol., 165, 351–371, doi:10.1111/j.1469-8137.2004.01224.x, 2005.

Aires, F. and Rossow, W. B.: Inferring instantaneous, multivariate and nonlinear sensitivities for the analysis of feedback processes in a dynamical system: Lorenz model case-study, Q. J. Roy. Meteor. Soc., 129, 239–275, doi:10.1256/qj.01.174, 2003.

Aires, F., Marquisseau, F., Prigent, C., and Sèze, G.: A Land and Ocean Microwave Cloud Classification Algorithm Derived from AMSU-A and -B, Trained Using MSG-SEVIRI Infrared and Visible Observations, Mon. Weather Rev., 139, 2347–2366, doi:10.1175/MWR-D-10-05012.1, 2011.

Aires, F., Gentine, P., Findell, K., Lintner, B. R., and Kerr, C.: Neural network-based sensitivity analysis of summertime convection over continental US, submitted, 2013.

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R., and Zhu, Z.: Evaluating the land and ocean components of the global carbon cycle in the CMIP5 Earth System Models, J. Climate, doi:10.1175/JCLI-D-12-00417.1, in press., 2013.

Antoine, D., Morel, A., Gordon, H. R., Banzon, V. F., and Evans, R. H.: Bridging ocean color observations of the 1980s and 2000s in search of long-term trends, J. Geophys. Res.-Oceans, 110, C06009, doi:10.1029/2004JC002620, 2005.

Bartlein, P. J., Harrison, S. P., Brewer, S., Connor, S., Davis, B. A. S., Gajewski, K., Guiot, J., Harrison-Prentice, T. I., Henderson, A., Peyron, O., Prentice, I. C., Scholze, M., Seppä, H., Shuman, B., Sugita, S., Thompson, R. S., Viau, A. E., Williams, J., and Wu, H.: Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis, Clim. Dynam., 37, 775–802, doi:10.1007/s00382-010-0904-1, 2010.

Batterman, S. A. and Larsen, K. S.: Integrating empirical studies and global models to improve climate change predictions, Eos. T. Am. Geophys. Un., 92, 353–353, doi:10.1029/2011EO410011, 2011.

Bauerle, W. L., Oren, R., Way, D. A., Qian, S. S., Stoy, P. C., Thornton, P. E., Bowden, J. D., Hoffman, F. M., and Reynolds, R. F.: Photoperiodic regulation of the seasonal pattern of photosynthetic capacity and the implications for carbon cycling, P. Natl. Acad. Sci. USA, 109, 8612–8617, doi:10.1073/pnas.1119131109, 2012.

Beck, H. E., McVicar, T. R., Van Dijk, A. I. J. M., Schellekens, J., De Jeu, R. A. M., and Bruijnzeel, L. A.: Global evaluation of four AVHRR–NDVI data sets: Intercomparison and assessment against Landsat imagery, Remote Sens. Environ., 115, 2547–2563, doi:10.1016/j.rse.2011.05.012, 2011.

Bellouin, N., Rae, J., Jones, A., Johnson, C., Haywood, J., and Boucher, O.: Aerosol forcing in the Climate Model Intercomparison Project (CMIP5) simulations by HadGEM2-ES and the role of ammonium nitrate, J. Geophys. Res., 116, D20206, doi:10.1029/2011JD016074, 2011.

Bigelow, N. H., Brubaker, L. B., Edwards, M. E., Harrison, S. P., Prentice, I. C., Anderson, P. M., Andreev, A. A., Bartlein, P. J., Christensen, T. R., Cramer, W., Kaplan, J. O., Lozhkin, A. V., Matveyeva, N. V., Murray, D. F., McGuire, A. D., Razzhivin, V. Y., Ritchie, J. C., Smith, B., Walker, D. A., Gajewski, K., Wolf, V., Holmqvist, B. H., Igarashi, Y., Kremenetskii, K., Paus, A., Pisaric, M. F. J., and Volkova, V. S.: Climate change and Arctic ecosystems: 1. vegetation changes north of 55° N between the last glacial maximum, mid-Holocene, and present, J. Geophys. Res., 108, 8170, doi:10.1029/2002JD002558, 2003.

Blyth, E., Clark, D. B., Ellis, R., Huntingford, C., Los, S., Pryor, M., Best, M., and Sitch, S.: A comprehensive set of benchmark tests for a land surface model of simultaneous fluxes of water and carbon at both the global and seasonal scale, Geosci. Model Dev., 4, 255–269, doi:10.5194/gmd-4-255-2011, 2011.

Bonan, G. B., Oleson, K. W., Fisher, R. A., Lasslop, G., and Reichstein, M.: Reconciling leaf physiological traits and canopy flux data: Use of the TRY and FLUXNET databases in the Community Land Model version 4, J. Geophys. Res.-Biogeo., 117, G02026, doi:10.1029/2011JG001913, 2012.

Bony, S. and Dufresne, J.-L.: Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models, Geophys. Res. Lett., 32, L20806, doi:10.1029/2005GL023851, 2005.

Booth, B. B. B., Jones, C. D., Collins, M., Totterdell, I. J., Cox, P. M., Sitch, S., Huntingford, C., Betts, R. A., Harris, G. R., and Lloyd, J.: High sensitivity of future global warming to land carbon cycle processes, Environ. Res. Lett., 7, 024002, doi:10.1088/1748-9326/7/2/024002, 2012.

Bopp, L., Kohfeld, K. E., Le Quéré, C., and Aumont, O.: Dust impact on marine biota and atmospheric CO_2 during glacial periods, Paleoceanography, 18, 1046, doi:10.1029/2002PA000810, 2003.

Braconnot, P., Harrison, S. P., Kageyama, M., Bartlein, P. J., Masson-Delmotte, V., Abe-Ouchi, A., Otto-Bliesner, B., and Zhao, Y.: Evaluation of climate models using palaeoclimatic data, Nat. Clim. Change., 2, 417–424, doi:10.1038/nclimate1456, 2012.

Cadule, P., Friedlingstein, P., Bopp, L., Sitch, S., Jones, C. D., Ciais, P., Piao, S. L., and Peylin, P.: Benchmarking coupled climate-carbon models against long-term atmospheric CO_2 measurements, Global Biogeochem. Cy., 24, GB2016, doi:10.1029/2009GB003556, 2010.

Charman, D. J., Beilman, D. W., Blaauw, M., Booth, R. K., Brewer, S., Chambers, F. M., Christen, J. A., Gallego-Sala, A., Harrison, S. P., Hughes, P. D. M., Jackson, S. T., Korhola, A., Mauquoy, D., Mitchell, F. J. G., Prentice, I. C., van der Linden, M., De Vleeschouwer, F., Yu, Z. C., Alm, J., Bauer, I. E., Corish, Y. M. C., Garneau, M., Hohl, V., Huang, Y., Karofeld, E., Le Roux, G., Loisel, J., Moschen, R., Nichols, J. E., Nieminen, T. M., MacDonald, G. M., Phadtare, N. R., Rausch, N., Sillasoo, Ü., Swindles, G. T., Tuittila, E.-S., Ukonmaanaho, L., Väliranta, M., van Bellen, S., van Geel, B., Vitt, D. H., and Zhao, Y.: Climate-related changes in peatland carbon accumulation during the last millennium, Biogeosciences, 10, 929–944, doi:10.5194/bg-10-929-2013, 2013.

Chéruy, F. and Aires, F.: Cluster analysis of cloud properties over the southern European mediterranean area in observations and a model, Mon. Weather Rev., 137, 3161–3176, doi:10.1175/2009MWR2882.1, 2009.

Claussen, M., Mysak, L., Weaver, A., Crucifix, M., Fichefet, T., Loutre, M.-F., Weber, S., Alcamo, J., Alexeev, V., Berger, A., Calov, R., Ganopolski, A., Goosse, H., Lohmann, G., Lunkeit, F., Mokhov, I., Petoukhov, V., Stone, P., and Wang, Z.: Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models, Clim. Dynam., 18, 579–586, doi:10.1007/s00382-001-0200-1, 2002.

Collins, M., Chandler, R. E., Cox, P. M., Huthnance, J. M., Rougier, J., and Stephenson, D. B.: Quantifying future climate change, Nat. Clim. Change, 2, 403–409, doi:10.1038/nclimate1414, 2012.

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C. D., Joshi, M., Liddicoat, S., Martin, G., O'Connor, F., Rae, J., Senior, C., Sitch, S., Totterdell, I., Wiltshire, A., and Woodward, S.: Development and evaluation of an Earth-System model – HadGEM2, Geosci. Model Dev., 4, 1051–1075, doi:10.5194/gmd-4-1051-2011, 2011.

Cox, P. M., Pearson, D., Booth, B. B., Friedlingstein, P., Huntingford, C., Jones, C. D., and Luke, C. M.: Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability, Nature, 494, 341–344, doi:10.1038/nature11882, 2013.

Crout, N. M. J., Tarsitano, D., and Wood, A. T.: Is my model too complex? Evaluating model formulation using model reduction, Environ. Modell. Softw., 24, 1–7, doi:10.1016/j.envsoft.2008.06.004, 2009.

Crucifix, M.: Distribution of carbon isotopes in the glacial ocean: a model study, Paleoceanography, 20, PA4020, doi:10.1029/2005PA001131, 2005.

Dahlke, C., Loew, A., and Reick, C.: Robust identification of global greening phase patterns from remote sensing vegetation products, J. Climate, 25, 8289–8307, doi:10.1175/JCLI-D-11-00319.1, 2012.

Dalmonech, D. and Zaehle, S.: Constraints from atmospheric CO_2 and satellite-based vegetation activity observations on current land carbon cycle trends, Biogeosciences Discuss., 9, 16087–16138, doi:10.5194/bgd-9-16087-2012, 2012.

Daniau, A.-L., Harrison, S. P., and Bartlein, P. J.: Fire regimes during the Last Glacial, Quaternary Sci. Rev., 29, 2918–2930, doi:10.1016/j.quascirev.2009.11.008, 2010.

Davis, B. A. S., Brewer, S., Stevenson, A. C., and Guiot, J.: The temperature of Europe during the Holocene reconstructed from pollen data, Quaternary Sci. Rev., 22, 1701–1716, doi:10.1016/S0277-3791(03)00173-2, 2003.

De Kauwe, M. G., Medlyn, B. E., Zaehle, S., Walker, A. P., Dietze, M. C., Hickler, T., Jain, A. K., Luo, Y., Parton, W. J., Prentice, I. C., Smith, B., Thornton, P. E., Wang, S., Wang, Y.-P., Wårlind, D., Weng, E., Crous, K. Y., Ellsworth, D. S., Hanson, P. J., Seok Kim, H.-, Warren, J. M., Oren, R., and Norby, R. J.: Forest water use and water use efficiency at elevated CO_2?: a model-data intercomparison at two contrasting temperate forest FACE sites, Glob. Change Biol., 19, 1759–1779, doi:10.1111/gcb.12164, 2013.

Deza, E. and Deza, M.-M.: Chapter 17 – Distances and similarities in data analysis, in: Dictionary of Distances, Elsevier, Amsterdam, 217–229, 2006.

Elsig, J., Schmitt, J., Leuenberger, D., Schneider, R., Eyer, M., Leuenberger, M., Joos, F., Fischer, H., and Stocker, T. F.: Stable isotope constraints on Holocene carbon cycle changes from an Antarctic ice core, Nature, 461, 507–510, doi:10.1038/nature08393, 2009.

Edwards, N. R., Cameron, D., and Rougier, J.: Precalibrating an intermediate complexity climate model, Clim. Dynam., 37, 1469–1482, doi:10.1007/s00382-010-0921-0, 2011.

Evans, J. P.: 21st century climate change in the Middle East, Climatic Change, 92, 417–432, doi:10.1007/s10584-008-9438-5, 2008.

Foley, A., Fealy, R., and Sweeney, J.: Model skill measures in probabilistic regional climate projections for Ireland, Clim. Res., 56, 33–49, doi:10.3354/cr01140, 2013.

Frank, D. C., Esper, J., Raible, C. C., Büntgen, U., Trouet, V., Stocker, B., and Joos, F.: Ensemble reconstruction constraints on the global carbon cycle sensitivity to climate, Nature, 463, 527–530, doi:10.1038/nature08769, 2010.

Friend, A. D., Arneth, A., Kiang, N. Y., Lomas, M., Ogée, J., Rödenbeck, C., Running, S. W., Santaren, J.-D., Sitch, S., Viovy, N., Ian Woodward, F., and Zaehle, S.: FLUXNET and modelling the global carbon cycle, Glob. Change Biol., 13, 610–633, doi:10.1111/j.1365-2486.2006.01223.x, 2007.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res., 113, D06104, doi:10.1029/2007JD008972, 2008.

Goldstein, M. and Rougier, J.: Reified Bayesian modelling and inference for physical systems, J. Stat. Plan. Infer., 139, 1221–1239, doi:10.1016/j.jspi.2008.07.019, 2009.

Gotangco Castillo, C. K., Levis, S., and Thornton, P.: Evaluation of the new CNDV option of the Community Land Model: effects of dynamic vegetation and interactive nitrogen on CLM4 means and variability*, J. Climate, 25, 3702–3714, doi:10.1175/JCLI-D-11-00372.1, 2012.

Gregg, W. W., Conkright, M. E., Ginoux, P., O'Reilly, J. E., and Casey, N. W.: Ocean primary production and climate: global decadal changes, Geophys. Res. Lett., 30, 1809, doi:10.1029/2003GL016889, 2003.

Guiot, J., Boreux, J. J., Braconnot, P., and Torre, F.: Data-model comparison using fuzzy logic in paleoclimatology, Clim. Dynam., 15, 569–581, doi:10.1007/s003820050301, 1999.

Hall, A. and Qu, X.: Using the current seasonal cycle to constrain snow albedo feedback in future climate change, Geophys. Res. Lett., 33, L03502, doi:10.1029/2005GL025127, 2006.

Harrison, S. P. and Bartlein, P.: Records from the past, lessons for the future, in: The Future of the World's Climate, edited by: Henderson-Sellars, A. and McGuffie, K. J., Elsevier, 403–436, 2012.

Harrison, S. P. and Prentice, C. I.: Climate and CO_2 controls on global vegetation distribution at the last glacial maximum: analysis based on palaeovegetation data, biome modelling and palaeoclimate simulations, Glob. Change Biol., 9, 983–1004, doi:10.1046/j.1365-2486.2003.00640.x, 2003.

Harrison, S. P. and Sanchez Goñi, M. F.: Global patterns of vegetation response to millennial-scale variability and rapid climate change during the last glacial period, Quaternary Sci. Rev., 29, 2957–2980, doi:10.1016/j.quascirev.2010.07.016, 2010.

Harrison, S. P., Bartlein, P. J., Brewer, S., Prentice, I. C., Boyd, M., Hessler, I., Holmgren, K., Izumi, K., and Willis, K.: Model benchmarking with glacial and mid-Holocene climates, submitted, 2013.

Henderson, G. M.: New oceanic proxies for paleoclimate, Earth Planet. Sc. Lett., 203, 1–13, doi:10.1016/S0012-821X(02)00809-9, 2002.

Henson, S. A., Sarmiento, J. L., Dunne, J. P., Bopp, L., Lima, I., Doney, S. C., John, J., and Beaulieu, C.: Detection of anthropogenic climate change in satellite records of ocean chlorophyll and productivity, Biogeosciences, 7, 621–640, doi:10.5194/bg-7-621-2010, 2010.

Hickler, T., Smith, B., Prentice, I. C., MjöFors, K., Miller, P., Arneth, A., and Sykes, M. T.: CO_2 fertilization in temperate FACE experiments not representative of boreal and tropical forests, Glob. Change Biol., 14, 1531–1542, doi:10.1111/j.1365-2486.2008.01598.x, 2008.

Holden, P. B., Edwards, N. R., Oliver, K. I. C., Lenton, T. M., and Wilkinson, R. D.: A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1, Clim. Dynam., 35, 785–806, doi:10.1007/s00382-009-0630-8, 2009.

Holden, P. B., Edwards, N. R., Gerten, D., and Schaphoff, S.: A model-based constraint on CO_2 fertilisation, Biogeosciences, 10, 339–355, doi:10.5194/bg-10-339-2013, 2013a.

Holden, P. B., Edwards, N. R., Müller, S. A., Oliver, K. I. C., Death, R. M., and Ridgwell, A.: Controls on the spatial distribution of oceanic $\delta^{13}C_{DIC}$, Biogeosciences, 10, 1815–1833, doi:10.5194/bg-10-1815-2013, 2013b.

Hou, J., Huang, Y., Wang, Y., Shuman, B., Oswald, W. W., Faison, E., and Foster, D. R.: Postglacial climate reconstruction based on compound-specific D/H ratios of fatty acids from Blood Pond, New England, Geochem. Geophy. Geosy., 7, Q03008, doi:10.1029/2005GC001076, 2006.

Izumi, K., Bartlein, P. J., and Harrison, S. P.: Consistent large-scale temperature responses in warm and cold climates, Geophys. Res. Lett., 40, 1817–1823, doi:10.1002/grl.50350, 2013.

Jansen, E., Overpeck, J., Briffa, K. R., Duplessy, J.-C., Joos, F., Masson-Delmotte, V., Olago, D., Otto-Bliesner, B., Peltier, W. R., Rahmstorf, S., Ramesh, R., Raynaud, D., Rind, D., Solomina, O., Villalba, R., and Zhang, D.: Palaeoclimate, in: Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., Cambridge University Press, Cambridge, UK and New York, NY, USA, 2007.

Jakob, C.: An improved strategy for the evaluation of cloud parameterizations in GCMS, B. Am. Meteorol. Soc., 84, 1387–1401, doi:10.1175/BAMS-84-10-1387, 2003.

Jolliff, J. K., Kindle, J. C., Shulman, I., Penta, B., Friedrichs, M. A. M., Helber, R., and Arnone, R. A.: Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, J. Marine Syst., 76, 64–82, doi:10.1016/j.jmarsys.2008.05.014, 2009.

Jones, C., Gregory, J., Thorpe, R., Cox, P., Murphy, J., Sexton, D., and Valdes, P.: Systematic optimisation and climate simulation of FAMOUS, a fast version of HadCM3, Clim. Dynam., 25, 189–204, doi:10.1007/s00382-005-0027-2, 2005.

Jones, C., Robertson, E., Arora, V., Friedlingstein, P., Shevliakova, E., Bopp, L., Brovkin, V., Hajima, T., Kato, E., Kawamiya, M., Liddicoat, S., Lindsay, K., Reick, C. H., Roelandt, C., Segschneider, J., and Tjiputra, J.: 21st Century compatible CO_2 emissions and airborne fraction simulated by CMIP5 Earth System models under 4 representative concentration pathways, J. Climate, doi:10.1175/JCLI-D-12-00554.1, in press, 2013.

Joussaume, S., Taylor, K. E., Braconnot, P., Mitchell, J. F. B., Kutzbach, J. E., Harrison, S. P., Prentice, I. C., Broccoli, A. J., Abe-Ouchi, A., Bartlein, P. J., Bonfils, C., Dong, B., Guiot, J., Herterich, K., Hewitt, C. D., Jolly, D., Kim, J. W., Kislov, A., Kitoh, A., Loutre, M. F., Masson, V., McAvaney, B., McFarlane, N., de Noblet, N., Peltier, W. R., Peterschmitt, J. Y., Pollard, D., Rind, D., Royer, J. F., Schlesinger, M. E., Syktus, J., Thompson, S., Valdes, P., Vettoretti, G., Webb, R. S., and Wyputta, U.: Monsoon changes for 6000 years ago: results of 18 simulations from the Paleoclimate Modeling Intercomparison Project (PMIP), Geophys. Res. Lett., 26, 859–862, doi:10.1029/1999GL900126, 1999.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, J. Geophys. Res.-Biogeo., 116, G00J07, doi:10.1029/2010JG001566, 2011.

Kelley, D. I., Colin Prentice, I., Harrison, S. P., Wang, H., Simard, M., Fisher, J. B., and Willis, K. O.: A comprehensive benchmarking system for evaluating global vegetation models, Biogeosciences Discuss., 9, 15723–15785, doi:10.5194/bgd-9-15723-2012, 2012.

Kohfeld, K. E., Quéré, C. L., Harrison, S. P., and Anderson, R. F.: Role of marine biology in glacial–interglacial CO_2 cycles, Science, 308, 74–78, doi:10.1126/science.1105375, 2005.

Larocque, I. and Bigler, C.: Similarities and discrepancies between chironomid- and diatom-inferred temperature reconstructions through the Holocene at Lake 850, northern Sweden, Quatern. Int., 122, 109–121, doi:10.1016/j.quaint.2004.01.033, 2004.

Laufkötter, C., Vogt, M., and Gruber, N.: Long-term trends in ocean plankton production and particle export between 1960–2006, Biogeosciences Discuss., 10, 5923–5975, doi:10.5194/bgd-10-5923-2013, 2013.

Lawrence, D. M., Slater, A. G., and Swenson, S. C.: Simulation of present-day and future permafrost and seasonally frozen ground conditions in CCSM4, J. Climate, 25, 2207–2225, doi:10.1175/JCLI-D-11-00334.1, 2012.

Leduc, G., Schneider, R., Kim, J.-H., and Lohmann, G.: Holocene and Eemian sea surface temperature trends as revealed by alkenone and Mg/Ca paleothermometry, Quaternary Sci. Rev., 29, 989–1004, doi:10.1016/j.quascirev.2010.01.004, 2010.

Leffelaar, P. A.: On scale problems in modelling: an example from soil ecology, in: Theoretical Production Ecology?: Reflections and Prospects, edited by: Rabbinge, J. G. R., Pudoc, Wageningen, 57–73, available at: http://edepot.wur.nl/171931 (last access: 24 June 2013), 1990.

Leith, C. E.: Climate response and fluctuation dissipation, J. Atmos. Sci., 32, 2022–2026, doi:10.1175/1520-0469(1975)032<2022:CRAFD>2.0.CO;2, 1975.

Le Treut, H., Somerville, R., Cubasch, U., Ding, Y., Mauritzen, C., Mokssit, A., Peterson, T., and Prather, M.: Historical overview of climate change, in: Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., Cambridge University Press, Cambridge, UK and New York, NY, USA, 2007.

Lohmann, G., Schulz, M., and Otto-Bliesner, B.: Earth System Models: Their Use and Reliability in Past Climate Reconstructions and Future Predictions, available at: http://pages-142.unibe.ch/cgi-bin/WebObjects/products.woa/wa/product?id=300 (last access: 3 April 2013), 2008.

Lourantou, A., Lavrič, J. V., Köhler, P., Barnola, J.-M., Paillard, D., Michel, E., Raynaud, D., and Chappellaz, J.: Constraint of the CO_2 rise by new atmospheric carbon isotopic measurements during the last deglaciation, Global Biogeochem. Cy., 24, GB2015, doi:10.1029/2009GB003545, 2010.

Luo, Y. Q., Randerson, J. T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., Ciais, P., Dalmonech, D., Fisher, J. B., Fisher, R., Friedlingstein, P., Hibbard, K., Hoffman, F., Huntzinger, D., Jones, C. D., Koven, C., Lawrence, D., Li, D. J., Mahecha, M., Niu, S. L., Norby, R., Piao, S. L., Qi, X., Peylin, P., Prentice, I. C., Riley, W., Reichstein, M., Schwalm, C., Wang, Y. P., Xia, J. Y., Zaehle, S., and Zhou, X. H.: A framework for benchmarking land models, Biogeosciences, 9, 3857–3874, doi:10.5194/bg-9-3857-2012, 2012.

MacFarling Meure, C., Etheridge, D., Trudinger, C., Steele, P., Langenfelds, R., Van Ommen, T., Smith, A., and Elkins, J.: Law dome CO_2, CH_4 and N_2O ice core records extended to 2000 years BP, Geophys. Res. Lett., 33, L14810, doi:10.1029/2006GL026152, 2006.

McManus, J. F., Francois, R., Gherardi, J.-M., Keigwin, L. D., and Brown-Leger, S.: Collapse and rapid resumption of Atlantic meridional circulation linked to deglacial climate changes, Nature, 428, 834–837, doi:10.1038/nature02494, 2004.

Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., Ammann, C., Faluvegi, G., and Ni, F.: Global signatures and dynamical origins of the little ice age and medieval climate anomaly, Science, 326, 1256–1260, doi:10.1126/science.1177303, 2009.

Marcott, S. A., Shakun, J. D., Clark, P. U., and Mix, A. C.: A reconstruction of regional and global temperature for the past 11 300 years, Science, 339, 1198–1201, doi:10.1126/science.1228026, 2013.

Mariotti, V., Bopp, L., Tagliabue, A., Kageyama, M., and Swingedouw, D.: Marine productivity response to Heinrich events: a model-data comparison, Clim. Past, 8, 1581–1598, doi:10.5194/cp-8-1581-2012, 2012.

Masson-Delmotte, V., Kageyama, M., Braconnot, P., Charbit, S., Krinner, G., Ritz, C., Guilyardi, E., Jouzel, J., Abe-Ouchi, A., Crucifix, M., Gladstone, R. M., Hewitt, C. D., Kitoh, A., LeGrande, A. N., Marti, O., Merkel, U., Motoi, T., Ohgaito, R., Otto-Bliesner, B., Peltier, W. R., Ross, I., Valdes, P. J., Vettoretti, G., Weber, S. L., Wolk, F., and Yu, Y.: Past and future polar amplification of climate change: climate model intercomparisons and ice-core constraints, Clim. Dynam., 27, 437–440, doi:10.1007/s00382-006-0149-1, 2006.

Miller, G. H., Alley, R. B., Brigham-Grette, J., Fitzpatrick, J. J., Polyak, L., Serreze, M. C., and White, J. W. C.: Arctic amplification: can the past constrain the future?, Quaternary Sci. Rev., 29, 1779–1790, doi:10.1016/j.quascirev.2010.02.008, 2010.

Millet, L., Arnaud, F., Heiri, O., Magny, M., Verneaux, V., and Desmet, M.: Late-Holocene summer temperature reconstruction from chironomid assemblages of Lake Anterne, northern French Alps, Holocene, 19, 317–328, doi:10.1177/0959683608100576, 2009.

Min, S.-K. and Hense, A.: Hierarchical evaluation of IPCC AR4 coupled climate models with systematic consideration of model uncertainties, Clim. Dynam., 29, 853–868, doi:10.1007/s00382-007-0269-2, 2007.

Moffat, A. M., Beckstein, C., Churkina, G., Mund, M., and Heimann, M.: Characterization of ecosystem responses to climatic controls using artificial neural networks, Glob. Change Biol., 16, 2737–2749, doi:10.1111/j.1365-2486.2010.02171.x, 2010.

Moise, A. F. and Delage, F. P.: New climate model metrics based on object-orientated pattern matching of rainfall, J. Geophys. Res., 116, D12108, doi:10.1029/2010JD015318, 2011.

Norby, R. J. and Zak, D. R.: Ecological lessons from Free-Air CO_2 Enrichment (FACE) experiments, Annu. Rev. Ecol. Evol. S., 42, 181–203, doi:10.1146/annurev-ecolsys-102209-144647, 2011.

Norby, R. J., DeLucia, E. H., Gielen, B., Calfapietra, C., Giardina, C. P., King, J. S., Ledford, J., McCarthy, H. R., Moore, D. J. P., Ceulemans, R., De Angelis, P., Finzi, A. C., Karnosky, D. F., Kubiske, M. E., Lukac, M., Pregitzer, K. S., Scarascia-Mugnozza, G. E., Schlesinger, W. H., and Oren, R.: Forest response to elevated CO_2 is conserved across a broad range of productivity, P. Natl. Acad. Sci. USA, 102, 18052–18056, doi:10.1073/pnas.0509478102, 2005.

Notaro, M., Liu, Z., and Williams, J. W.: Observed vegetation–climate feedbacks in the United States, J. Climate, 19, 763–786, doi:10.1175/JCLI3657.1, 2006.

Nowak, R. S., Ellsworth, D. S., and Smith, S. D.: Functional responses of plants to elevated atmospheric CO_2 – do photosynthetic and productivity data from FACE experiments support early predictions?, New Phytol., 162, 253–280, doi:10.1111/j.1469-8137.2004.01033.x, 2004.

Oreskes, N., Shrader-Frechette, K., and Belitz, K.: Verification, validation, and confirmation of numerical models in the Earth sciences, Science, 263, 641–646, doi:10.1126/science.263.5147.641, 1994.

Otto-Bliesner, B. L., Hewitt, C. D., Marchitto, T. M., Brady, E., Abe-Ouchi, A., Crucifix, M., Murakami, S., and Weber, S. L.: Last Glacial Maximum ocean thermohaline circulation: PMIP2 model intercomparisons and data constraints, Geophys. Res. Lett., 34, L12706, doi:10.1029/2007GL029475, 2007.

Otto-Bliesner, B. L., Schneider, R., Brady, E. C., Kucera, M., Abe-Ouchi, A., Bard, E., Braconnot, P., Crucifix, M., Hewitt, C. D., Kageyama, M., Marti, O., Paul, A., Rosell-Melé, A., Waelbroeck, C., Weber, S. L., Weinelt, M., and Yu, Y.: A comparison of PMIP2 model simulations and the MARGO proxy reconstruction for tropical sea surface temperatures at last glacial maximum, Clim. Dynam., 32, 799–815, doi:10.1007/s00382-008-0509-0, 2009.

Parrenin, F., Masson-Delmotte, V., Köhler, P., Raynaud, D., Paillard, D., Schwander, J., Barbante, C., Landais, A., Wegner, A., and Jouzel, J.: Synchronous change of atmospheric CO_2 and Antarctic temperature during the last deglacial warming, Science, 339, 1060–1063, 2013.

Paul, A. and Schäfer-Neth, C.: How to combine sparse proxy data and coupled climate models, Quaternary Sci. Rev., 24, 1095–1107, doi:10.1016/j.quascirev.2004.05.010, 2005.

Pearl, J.: Causality: Models, Reasoning, and Inference, Cambridge University Press, 2000.

Pedro, J. B., Rasmussen, S. O., and van Ommen, T. D.: Tightened constraints on the time-lag between Antarctic temperature and CO_2 during the last deglaciation, Clim. Past, 8, 1213–1221, doi:10.5194/cp-8-1213-2012, 2012.

Petoukhov, V., Ganopolski, A., Brovkin, V., Claussen, M., Eliseev, A., Kubatzki, C., and Rahmstorf, S.: CLIMBER-2: a climate system model of intermediate complexity, Part I: model description and performance for present climate, Clim. Dynam., 16, 1–17, doi:10.1007/PL00007919, 2000.

Pfeiffer, M., Spessa, A., and Kaplan, J. O.: A model for global biomass burning in preindustrial time: LPJ-LMfire (v1.0), Geosci. Model Dev., 6, 643–685, doi:10.5194/gmd-6-643-2013, 2013.

Pincus, R., Batstone, C. P., Hofmann, R. J. P., Taylor, K. E., and Glecker, P. J.: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models, J. Geophys. Res., 113, D14209, doi:10.1029/2007JD009334, 2008.

Polovina, J. J., Chai, F., Howell, E. A., Kobayashi, D. R., Shi, L., and Chao, Y.: Ecosystem dynamics at a productivity gradient: a study of the lower trophic dynamics around the northern atolls in the Hawaiian Archipelago, Prog. Oceanogr., 77, 217–224, doi:10.1016/j.pocean.2008.03.011, 2008.

Power, M., Marlon, J., Ortiz, N., Bartlein, P., Harrison, S., Mayle, F., Ballouche, A., Bradshaw, R., Carcaillet, C., Cordova, C., Mooney, S., Moreno, P., Prentice, I., Thonicke, K., Tinner, W., Whitlock, C., Zhang, Y., Zhao, Y., Ali, A., Anderson, R., Beer, R., Behling, H., Briles, C., Brown, K., Brunelle, A., Bush, M., Camill, P., Chu, G., Clark, J., Colombaroli, D., Connor, S., Daniau, A.-L., Daniels, M., Dodson, J., Doughty, E., Edwards, M., Finsinger, W., Foster, D., Frechette, J., Gaillard, M.-J., Gavin, D., Gobet, E., Haberle, S., Hallett, D., Higuera, P., Hope, G., Horn, S., Inoue, J., Kaltenrieder, P., Kennedy, L., Kong, Z., Larsen, C., Long, C., Lynch, J., Lynch, E., McGlone, M., Meeks, S., Mensing, S., Meyer, G., Minckley, T., Mohr, J., Nelson, D., New, J., Newnham, R., Noti, R., Oswald, W., Pierce, J., Richard, P., Rowe, C., Sanchez Goñi, M., Shuman, B., Takahara, H., Toney, J., Turney, C., Urrego-Sanchez, D., Umbanhowar, C., Vandergoes, M., Vanniere, B., Vescovi, E., Walsh, M., Wang, X., Williams, N., Wilmshurst, J., and Zhang, J.: Changes in fire regimes since the Last Glacial Maximum: an assessment based on a global synthesis and analysis of charcoal data, Clim. Dynam., 30, 887–907, doi:10.1007/s00382-007-0334-x, 2008.

Prentice, I. C., Jolly, D., et al.: Mid-Holocene and Glacial-Maximum vegetation geography of the northern continents and Africa, J. Biogeogr., 27, 507–519, 2000.

Prentice, I. C., Harrison, S. P., and Bartlein, P. J.: Global vegetation and terrestrial carbon cycle changes after the last ice age, New Phytol., 189, 988–998, doi:10.1111/j.1469-8137.2010.03620.x, 2011a.

Prentice, I. C., Kelley, D. I., Foster, P. N., Friedlingstein, P., Harrison, S. P., and Bartlein, P. J.: Modeling fire and the terrestrial carbon balance, Global Biogeochem. Cy., 25, GB3005, doi:10.1029/2010GB003906, 2011b.

Quillet, A., Peng, C., and Garneau, M.: Toward dynamic global vegetation models for simulating vegetation–climate interactions and feedbacks: recent developments, limitations, and future challenges, Environ. Rev., 18, 333–353, doi:10.1139/A10-016, 2010.

Radić, V. and Clarke, G. K. C.: Evaluation of IPCC models' performance in simulating late-twentieth-century climatologies and weather patterns over North America, J. Climate, 24, 5257–5274, doi:10.1175/JCLI-D-11-00011.1, 2011.

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., Stouffer, R. J., Sumi, A., and Taylor, K. E.: Climate models and their evaluation, in: Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., Cambridge University Press, Cambridge, UK and New York, NY, USA, 2007.

Randerson, J. T., Hoffman, F. M., Thornton, P. E., Mahowald, N. M., Lindsay, K., Lee, Y.-H., Nevison, C. D., Doney, S. C., Bonan, G., Stöckli, R., Covey, C., Running, S. W., and Fung, I. Y: Systematic assessment of terrestrial biogeochemistry in coupled climate–carbon models, Glob. Change Biol., 15, 2462–2484, doi:10.1111/j.1365-2486.2009.01912.x, 2009.

Reichler, T. and Kim, J.: Uncertainties in the climate mean state of global observations, reanalyses, and the GFDL climate model, J. Geophys. Res.-Atmos., 113, D05106, doi:10.1029/2007JD009278, 2008.

Ridgwell, A.: Interpreting transient carbonate compensation depth changes by marine sediment core modeling, Paleoceanography, 22, PA4102, doi:10.1029/2006PA001372, 2007.

Roche, D., Paillard, D., and Cortijo, E.: Constraints on the duration and freshwater release of Heinrich event 4 through isotope modelling, Nature, 432, 379–382, doi:10.1038/nature03059, 2004.

Roe, G. H. and Baker, M. B.: Why is climate sensitivity so unpredictable?, Science, 318, 629–632, doi:10.1126/science.1144735, 2007.

Rougier, J.: Probabilistic inference for future climate using an ensemble of climate model evaluations, Climatic Change, 81, 247–264, doi:10.1007/s10584-006-9156-9, 2007.

Rowlands, D. J., Frame, D. J., Ackerley, D., Aina, T., Booth, B. B. B., Christensen, C., Collins, M., Faull, N., Forest, C. E., Grandey, B. S., Gryspeerdt, E., Highwood, E. J., Ingram, W. J., Knight, S., Lopez, A., Massey, N., McNamara, F., Meinshausen, N., Piani, C., Rosier, S. M., Sanderson, B. M., Smith, L. A., Stone, D. A., Thurston, M., Yamazaki, K., Yamazaki, Y. H., and Allen, M. R.: Broad range of 2050 warming from an observationally constrained large climate model ensemble, Nat. Geosci., 5, 256–260, doi:10.1038/ngeo1430, 2012.

Saltelli, A., Chan, K., and Scott, E. M.: Sensitivity Analysis: Gauging the Worth of Scientific Models, Wiley, 2000.

Sargent, R. G.: Verification and validation of simulation models, in: Proceedings of the 2010 Winter Simulation Conference (WSC), 166–183, 2010.

Sarmiento, J. L., Slater, R., Barber, R., Bopp, L., Doney, S. C., Hirst, A. C., Kleypas, J., Matear, R., Mikolajewicz, U., Monfray, P., Soldatov, V., Spall, S. A., and Stouffer, R.: Response of ocean ecosystems to climate warming, Global Biogeochem. Cy., 18, GB3003, doi:10.1029/2003GB002134, 2004.

Schaller, N., Mahlstein, I., Cermak, J., and Knutti, R.: Analyzing precipitation projections: a comparison of different approaches to climate model evaluation, J. Geophys. Res., 116, D10118, doi:10.1029/2010JD014963, 2011.

Schmidt, G. A., Jungclaus, J. H., Ammann, C. M., Bard, E., Braconnot, P., Crowley, T. J., Delaygue, G., Joos, F., Krivova, N. A., Muscheler, R., Otto-Bliesner, B. L., Pongratz, J., Shindell, D. T., Solanki, S. K., Steinhilber, F., and Vieira, L. E. A.: Climate forcing reconstructions for use in PMIP simulations of the Last Millennium (v1.1), Geosci. Model Dev., 5, 185–191, doi:10.5194/gmd-5-185-2012, 2012.

Seppä, H., Bjune, A. E., Telford, R. J., Birks, H. J. B., and Veski, S.: Last nine-thousand years of temperature variability in Northern Europe, Clim. Past, 5, 523–535, doi:10.5194/cp-5-523-2009, 2009.

Shakun, J. D., Clark, P. U., He, F., Marcott, S. A., Mix, A. C., Liu, Z., Otto-Bliesner, B., Schmittner, A., and Bard, E.: Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation, Nature, 484, 49–54, doi:10.1038/nature10915, 2012.

Shannon, S. and Lunt, D. J.: A new dust cycle model with dynamic vegetation: LPJ-dust version 1.0, Geosci. Model Dev., 4, 85–105, doi:10.5194/gmd-4-85-2011, 2011.

Siddall, M., Henderson, G. M., Edwards, N. R., Frank, M., Müller, S. A., Stocker, T. F., and Joos, F.: ^{231}Pa/^{230}Th fractionation by ocean transport, biogenic particle flux and particle type, Earth Planet. Sc. Lett., 237, 135–155, doi:10.1016/j.epsl.2005.05.031, 2005.

Siddall, M., Stocker, T. F., Henderson, G. M., Joos, F., Frank, M., Edwards, N. R., Ritz, S. P., and Müller, S. A.: Modeling the relationship between ^{231}Pa/^{230}Th distribution in North Atlantic sediment and Atlantic meridional overturning circulation, Paleoceanography, 22, 1–14, doi:10.1029/2006PA001358, 2007.

Sitch, S., Huntingford, C., Gedney, N., Levy, P. E., Lomas, M., Piao, S. L., Betts, R., Ciais, P., Cox, P., Friedlingstein, P., Jones, C. D., Prentice, I. C., and Woodward, F. I.: Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five Dynamic Global Vegetation Models (DGVMs), Glob. Change Biol., 14, 2015–2039, doi:10.1111/j.1365-2486.2008.01626.x, 2008.

Sokolov, A. P., Kicklighter, D. W., Melillo, J. M., Felzer, B. S., Schlosser, C. A., and Cronin, T. W.: Consequences of considering carbon–nitrogen interactions on the feedbacks between climate and the terrestrial carbon cycle, J. Climate, 21, 3776–3796, doi:10.1175/2008JCLI2038.1, 2008.

Stauffer, B., Fluckiger, J., Monnin, E., Schwander, J., Barnola, J.-M., and Chappellaz, J.: Atmospheric CO_2, CH_4 and N_2O records over the past 60 000 years based on the comparison of different polar ice cores, Ann. Glaciol., 35, 202–208, doi:10.3189/172756402781816861, 2002.

Steinacher, M., Joos, F., Frölicher, T. L., Bopp, L., Cadule, P., Cocco, V., Doney, S. C., Gehlen, M., Lindsay, K., Moore, J. K., Schneider, B., and Segschneider, J.: Projected 21st century decrease in marine productivity: a multi-model analysis, Biogeosciences, 7, 979–1005, doi:10.5194/bg-7-979-2010, 2010.

Taucher, J., Schulz, K. G., Dittmar, T., Sommer, U., Oschlies, A., and Riebesell, U.: Enhanced carbon overconsumption in response to increasing temperatures during a mesocosm experiment, Biogeosciences, 9, 3531–3545, doi:10.5194/bg-9-3531-2012, 2012.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res., 106, 7183–7192, doi:10.1029/2000JD900719, 2001.

Thonicke, K., Spessa, A., Prentice, I. C., Harrison, S. P., Dong, L., and Carmona-Moreno, C.: The influence of vegetation, fire spread and fire behaviour on biomass burning and trace gas emissions: results from a process-based model, Biogeosciences, 7, 1991–2011, doi:10.5194/bg-7-1991-2010, 2010.

Thornton, P. E., Doney, S. C., Lindsay, K., Moore, J. K., Mahowald, N., Randerson, J. T., Fung, I., Lamarque, J.-F., Feddema, J. J., and Lee, Y.-H.: Carbon-nitrogen interactions regulate climate-carbon cycle feedbacks: results from an atmosphere-ocean general circulation model, Biogeosciences, 6, 2099–2120, doi:10.5194/bg-6-2099-2009, 2009.

Tschumi, J. and Stauffer, B.: Reconstructing past atmospheric CO_2 concentration based on ice-core analyses: open questions due to in situ production of CO_2 in the ice, J. Glaciol., 46, 45–53, doi:10.3189/172756500781833359, 2000.

Van Hoof, T. B., Kaspers, K. A., Wagner, F., Van De Wal, R. S. W., Kürschner, W. M., and Visscher, H.: Atmospheric CO_2 during the 13th century AD: reconciliation of data from ice core measurements and stomatal frequency analysis, Tellus B, 57, 351–355, doi:10.1111/j.1600-0889.2005.00154.x, 2005.

Viau, A. E., Gajewski, K., Sawada, M. C., and Bunbury, J.: Low- and high-frequency climate variability in eastern Beringia during the past 25 000 years, Can. J. Earth Sci., 45, 1435–1453, 2008.

Vicca, S., Gilgen, A. K., Camino Serrano, M., Dreesen, F. E., Dukes, J. S., Estiarte, M., Gray, S. B., Guidolotti, G., Hoeppner, S. S., Leakey, A. D. B., Ogaya, R., Ort, D. R., Ostrogovic, M. Z., Rambal, S., Sardans, J., Schmitt, M., Siebers, M., van der Linden, L., van Straaten, O., and Granier, A.: Urgent need for a common metric to make precipitation manipulation experiments comparable, New Phytol., 195, 518–522, doi:10.1111/j.1469-8137.2012.04224.x, 2012.

Waelbroeck, C., Paul, A., Kucera, M., Rosell-Melé, A., Weinelt, M., Schneider, R., Mix, A. C., Abelmann, A., Armand, L., Bard, E., Barker, S., Barrows, T. T., Benway, H., Cacho, I., Chen, M.-T., Cortijo, E., Crosta, X., de Vernal, A., Dokken, T., Duprat, J., Elderfield, H., Eynaud, F., Gersonde, R., Hayes, A., Henry, M., Hillaire-Marcel, C., Huang, C.-C., Jansen, E., Juggins, S., Kallel, N., Kiefer, T., Kienast, M., Labeyrie, L., Leclaire, H., Londeix, L., Mangin, S., Matthiessen, J., Marret, F., Meland, M., Morey, A. E., Mulitza, S., Pflaumann, U., Pisias, N. G., Radi, T., Rochon, A., Rohling, E. J., Sbaffi, L., Schäfer-Neth, C., Solignac, S., Spero, H., Tachikawa, K., and Turon, J.-L.: Constraints on the magnitude and patterns of ocean cooling at the Last Glacial Maximum, Nat. Geosci., 2, 127–132, doi:10.1038/ngeo411, 2009.

Williamson, D., Goldstein, M., and Blaker, A.: Fast linked analyses for scenario-based hierarchies, J. Roy. Stat. Soc. C-App., 61, 665–691, doi:10.1111/j.1467-9876.2012.01042.x, 2012.

Wohlfahrt, J., Harrison, S. P., and Braconnot, P.: Synergistic feedbacks between ocean and vegetation on mid- and high-latitude climates during the mid-Holocene, Clim. Dynam., 22, 223–238, doi:10.1007/s00382-003-0379-4, 2004.

Wohlfahrt, J., Harrison, S. P., Braconnot, P., Hewitt, C. D., Kitoh, A., Mikolajewicz, U., Otto-Bliesner, B. L., and Weber, S. L.: Evaluation of coupled ocean–atmosphere simulations of the mid-Holocene using palaeovegetation data from the Northern Hemisphere extratropics, Clim. Dynam., 31, 871–890, doi:10.1007/s00382-008-0415-5, 2008.

Woodage, M. J., Slingo, A., Woodward, S., and Comer, R. E.: U. K. HiGEM: simulations of desert dust and biomass burning aerosols with a high-resolution atmospheric GCM, J. Climate, 23, 1636–1659, doi:10.1175/2009JCLI2994.1, 2010.

Wu, J. and David, J. L.: A spatially explicit hierarchical approach to modeling complex ecological systems: theory and applications, Ecol. Model., 153, 7–26, doi:10.1016/S0304-3800(01)00499-9, 2002.

Yu, Z., Loisel, J., Brosseau, D. P., Beilman, D. W., and Hunt, S. J.: Global peatland dynamics since the Last Glacial Maximum, Geophys. Res. Lett., 37, L13402, doi:10.1029/2010GL043584, 2010.

Zaehle, S. and Friend, A. D.: Carbon and nitrogen cycle dynamics in the O-CN land surface model: 1. Model description, site-scale evaluation, and sensitivity to parameter estimates, Global Biogeochem. Cy., 24, GB1005, doi:10.1029/2009GB003521, 2010.

Zaehle, S., Sitch, S., Smith, B., and Hatterman, F.: Effects of parameter uncertainties on the modeling of terrestrial biosphere dynamics, Global Biogeochem. Cy., 19, GB3020, doi:10.1029/2004GB002395, 2005.

Zaehle, S., Friedlingstein, P., and Friend, A. D.: Terrestrial nitrogen feedbacks may accelerate future climate change, Geophys. Res. Lett., 37, L01401, doi:10.1029/2009GL041345, 2010.

Table 1. Summary of key data types for evaluation.

Type of data	Description	Examples	Advantages	Limitations
Modern last 30–50 yr	In-situ instrumental data	Atmospheric CO_2, CH_4	Direct observations of key variables, known uncertainties	Local observation
	Experimental data	Contolled environments (e.g. Phytotrons and glasshouses) Field experiments (e.g. Free Air Carbon dioxide Enrichment (FACE))	Provides new situations against which to test model behaviour, such as representing ecosystem-scale responses to combined environmental drivers	Large-scale field experiments generally do not provide information across all biomes Interpretation of experiments may be ambiguous
	Model-derived type I	Satellite-based data	Excellent spatial and/or temporal resolution	Lack full data-model independency, as data is model-derived (radiative transfer model converts radiation measurements into the parameter of interest) Inconsistent documentation of errors and uncertainties
	Model-derived type II	C-fluxes up-scaled data (e.g. MTE-MPI dataset)	“Data-model” conceptual correspondence	Lack fully data-model independency, as data is model-derived Inconsistent documentation of errors and uncertainties
Palaeo	Reconstructions based on interpretation of biological or geochemical records Measurements of concentrations and isotopic ratios from ice cores	Tree-ring datasets Pollen and plant macrofossil data (e.g. BIOME 6000) Ice cores, e.g. Law Dome, EPICA	Tests ability to capture behaviour of the system outside modern range Signal is large compared to noise	Site-specific records (except for long-lived greenhouse gases), synthesis required to produce global estimates Variable temporal resolution necessitates appropriate selection of data to address e.g. rapid changes Inconsistent documentation of errors and uncertainties

Table 2. Summary of Level 1 metrics (x, y represent points while D_1, D_2 are datasets).

Metric	Equation		Suitability
Manhattan distance	$\mathrm{d}(x,y) = \sum_{i=1}^{n} x_i - y_i$	(1)	Implicitly supposes that x and y are comparable, so is not suited to mixed variables
Euclidean distance	$\mathrm{d}(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$	(2)	More sensitive to outliers compared to the Manhattan distance Like the Manhattan distance, it also supposes that direct comparisons of the variables can be made
Weighted Euclidean distance	$\mathrm{d}(x,y) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}$	(3)	Uncertainty in the reference dataset, such as instrumental errors in an observation dataset, or model uncertainty in a model ensemble, can be accounted for using a model efficiency metric: E.g. $w_i = \frac{1}{\sigma_i^2}$ where σ_i = uncertainty
Chi-squared "distance"	$\mathrm{d}(x,y) = \sqrt{\sum_{i=1}^{n} \frac{1}{2} \frac{(x_i - y_i)^2}{y_i}}$	(4)	Measures similarity of two PDFs (Probability Distribution Functions) Particularly useful if the focus of the analysis is at population level
	$\mathrm{d}(x,y) = \sqrt{\sum_{i=1}^{n} \frac{1}{2} \frac{(x_i - y_i)^2}{x_i + y_i}}$	(5)	Alternative formulation Eq. (5) can be used to facilitate the symmetry property of distances
Tchebychev distance	$\mathrm{d}(x,y) = \underset{i}{max} (\lvert x_i - y_i \rvert)$	(6)	Useful for extreme events, or maximum annual discrepancy in a climatic run
Mahalanobis distance	$\mathrm{d}(x,y) = \sqrt{(x-y)^T A^{-1} (x-y)}$ where A^{-1} = covariance matrix of x or y	(7)	Particularly useful if x or y include coordinates with very different units (each one will be normalised by its variance), if they are correlated one with each other (since the distance takes into account these correlations), and for combination of multiple sources of information
Normalized mean error	$\frac{1}{E} \sum_e \frac{(D_{1,e} - D_{2,e})}{\bar{D}_1 \bar{D}_2}$ where E is the total number of samples in D_1 and D_2	(8)	Applies the distance over the entirety of two datasets D_1 and D_2

Table 3. Summary of Level 2 metrics.

Metric	Equation		Suitability
Chi-squared distance	See Eq. (4)		For PDFs of two datasets (e.g. observed and modelled data)
Kullback–Leibler divergence	$d(p \parallel q) = \int p(x) \frac{p(x)}{q(x)}$	(9)	For PDFs of two datasets, p and q
Variance	Depends on application		Suitable if long observational record is available Use the diagnostic that best suits the application

Table 4. Summary of Level 3 metrics.

Metric	Equation or method		Suitability
Pearson correlation coefficient	$\rho_{x,y} = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}$	(10)	Very simple – measures only correlation, not the causality Assumes data are normally distributed
Sensitivity	$\frac{\Delta x}{\Delta y}$ From empirical increments or model experiments	(11)	Simple technique – univariate only Does not consider possible interactions and non-linearities
Linear regression	For two variables a and b, $a = c \cdot b + d$ where c = slope and d = bias	(12)	Very simple – can become multivariate but has linear limitation Provides "coincidence" information, not causality information
Sensitivity (more advanced)	e.g. multiple linear regression $a = C \cdot B + d$ where B is now a vector of biogeophysical variables e.g. nonlinear model such as neural network	(13)	Nonlinear model that provides access to threshold, interactions and saturation behaviours The metric can then be defined as the percentage of variance of a explained by B in the data and in the model Still not causal
Pattern-oriented approaches	Various methods		Very process oriented, but requires a good understanding a priori of what needs to be examined
Clustering algorithms	e.g. K-means, self-organizing maps Uses a similarity distance, similar to level 1 metrics		Ideal for obtaining a limited set of prototypes, describing the variability of the datasets as much as possible

Table 5. Summary of evaluation methodologies.

Type of evaluation	Description	Examples	Advantages	Limitations
Process-based, "bottom-up"	Looks at relationships between variables in a way that isolates a single process, or small number of processes	NPP vs. precip. (Randerson et al., 2009) Magnitude of seasonal cycle of T_{air} vs. T_{surf} to evaluate insulation by snow pack	Pinpoints important model processes "Right answer for right reason" Easy to interpret, e.g. can see if response is too big or too small for a given input	Only targets a small part of the model Relevant observations may not exist Even when process representation is close to perfect, this does not ensure overall balance between them is right
System-level, "top-down"	Compares large scale model outputs that emerge from interactions between many processes within the model with relevant observations	Global patterns of temperature, precipitation, etc. Seasonal cycle of carbon fluxes	Evaluates end-result, i.e. quantities that we actually want the model to predict Assesses overall balance between many (possibly finely balanced) processes	Compensating errors: "Right answer for wrong reason" Hard to interpret as offers no indication of what is causing an error and how to fix it
Multi-model emergent constraints	Robust (across models) relationship between a quantity we can observe and a future change we want to predict	Hall and Qu (2006): seasonal cycle of snow albedo Cox et al. (2013): IAV of tropical carbon fluxes	No requirement for models to be right – models might be wrong on individual basis regarding magnitude of response but the relationship may be robust Guides where we want observational effort	Relies on "bad" models more than "good" ones to derive regression May get false confidence if models systematically wrong (e.g. all lack long-term carbon release from permafrost)

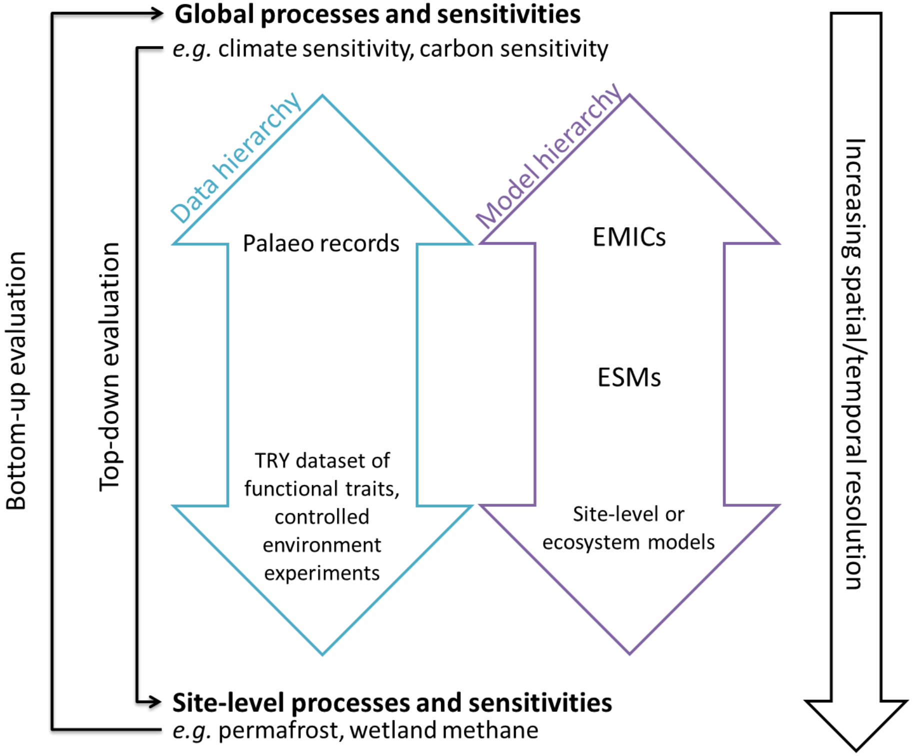
Global processes and sensitivities
e.g. climate sensitivity, carbon sensitivity
Data hierarchy
Model hierarchy
Palaeo records
EMICs
ESMs
TRY dataset of functional traits, controlled environment experiments
Site-level or ecosystem models
Bottom-up evaluation
Top-down evaluation
Increasing spatial/temporal resolution
Site-level processes and sensitivities
e.g. permafrost, wetland methane

Fig. 1. Conceptual diagram of hierarchical approach to model evaluation on different spatial and temporal scales.

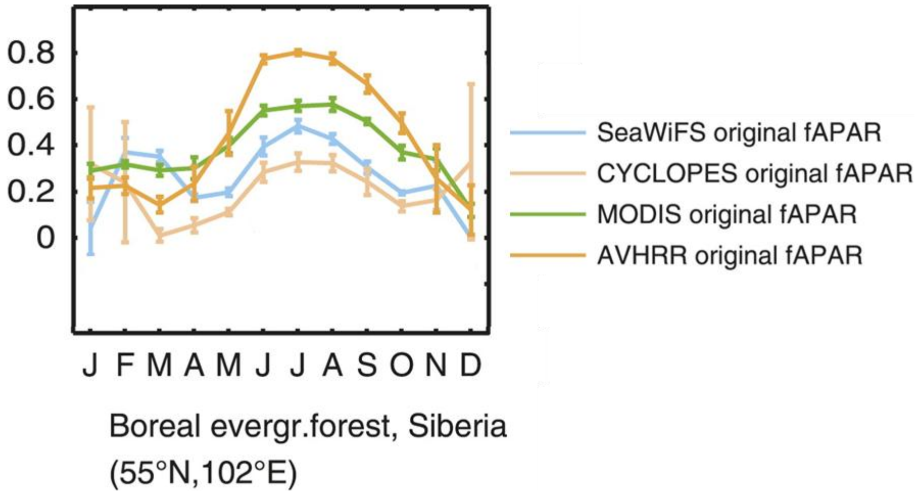

Fig. 2. Original fAPAR time series from a selected region (after Dahkle et al., 2012). © American Meteorological Society. Used with permission.

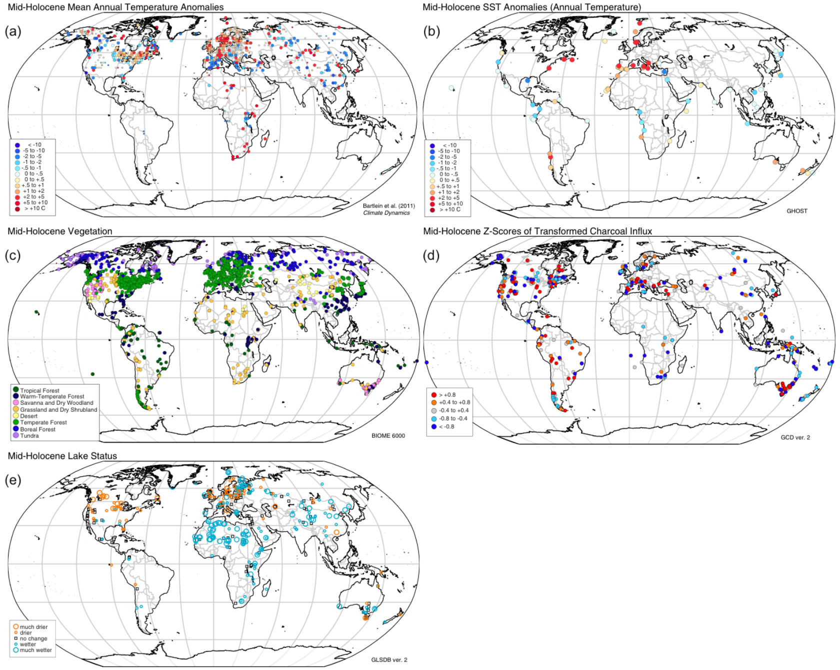

Fig. 3. Examples of global datasets documenting environmental conditions during the mid-Holocene (ca 6000 yr ago) that can be used for benchmarking ESM simulations. In general, these are expressed as anomalies, i.e. the difference between the mid-Holocene and modern conditions: **(a)** pollen-based reconstructions of anomalies in mean annual temperature, **(b)** reconstructions of anomalies in sea-surface temperatures based on marine biological and chemical records, **(c)** pollen and plant macrofossil reconstructions of vegetation during the mid-Holocene, **(d)** charcoal records of the anomalies in biomass burning, and **(e)** anomalies of changes in the hydrological cycle based on lake-level records of the balance between precipitation and evaporation (after Harrison and Bartlein, 2012). Reprinted from Harrison, S. P. and Bartlein, P.: Records from the Past, Lessons for the Future, in A. Henderson-Sellars and K.J. McGuffie eds.,The Future of the World's Climate, pp. 403–436. Copyright © 2012, with permission from Elsevier.

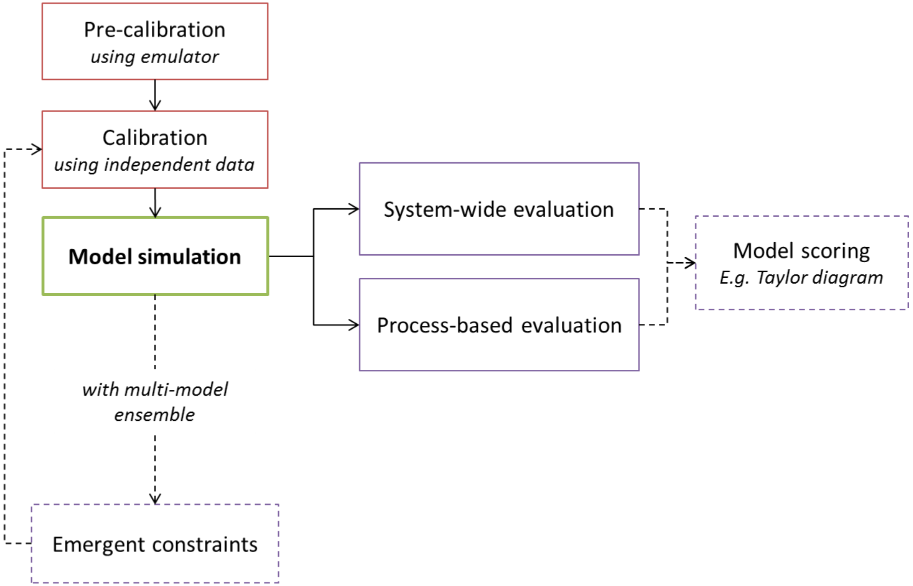

Fig. 4. Schematic diagram of model evaluation approaches, with optional approaches indicated by dashed lines.