Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 4 (2011) 216–xxx

Procedia Environmental Sciences

Earth System Science 2010: Global Change, Climate and People

The evaluation of Earth System Models: discussion summary

Sönke Zaehle[a], Colin Prentice[b] and Sarah Cornell[c]*

[a]*Department for Biogeochemical Systems, Max Planck Institute for Biogeochemistry, Postfach 10 01 64, Hans-Knöll-Str. 10 07745 Jena, Germany*

[b]*Department of Biological Sciences, Faculty of Science, Macquarie University, NSW 2109, Australia*

[c]*School of Earth Sciences, University of Bristol, Wills Memorial Building, Queens Road, Bristol UK BS8 1RJ*

Abstract

Complex Earth system models, and their various sub-components, are not yet subject to rigorous evaluation against observations as much as they should be, despite the existence of hundreds of proposed diagnostics. A concerted process is urgently needed to make this the norm, not the exception. Earth Observation, field observations and palaeo data can be applied to contexts as diverse as wildfire, marine ecosystems, the land carbon cycle, and greenhouse gases. Model evaluation (by comparing models and benchmark data) and model weighting (defining the 'quality' of models on the basis of such a comparison) should be considered as separate issues. Systematic approaches to parameter optimization, such as the adjoint technique, allow structural differences between models to be identified and limitations to be addressed. Such methods are established in atmospheric tracer transport and carbon cycling; research carried out in the QUEST programme has demonstrated their applicability for climate modelling. Although it is impossible to devise a foolproof metric for the ability of models to predict the future, relevant metrics could be based on their ability to simulate the past. Furthermore, it should be possible to extend parameter optimization techniques to assimilate data from the past.

There are limits to what can be achieved by benchmarking against a mean state, when it is a change in state that is of greatest interest. It is useful to benchmark individual processes rather than aggregate properties. Coupling good components does not automatically result in a good Earth System model, so for complex models, a two-stage process is needed: first, benchmarking the components in stand-alone mode, and second, using the same benchmarks in coupled mode.

Selection under responsibility of C. Downy and S. Colston.

Keywords: benchmarking; Earth System Models;

* Corresponding author. Tel.: +44-(0)117-331-4468; fax: +44-(0)117-925-3385.
E-mail address: sarah.cornell@bristol.ac.uk.

1878-0296 © 2011 Published by Elsevier Open access under CC BY-NC-ND license.
doi:10.1016/j.proenv.2011.05.023

1. Background

As Earth System models become more complex, including more subsystems, processes and their interactions, it is imperative to develop strategies that can allow the different components of the models to be evaluated, both in stand-alone and fully coupled modes. Whereas climate modelling has established procedures and standards for benchmarking the performance of models in terms of contemporary climate, the evaluation of submodels for carbon cycling, atmospheric chemistry, and biogenic emissions has been much less systematic.

The Coupled Climate-Carbon Cycle Model Intercomparison Project (C4MIP), highlighted in the IPCC AR4, demonstrates the problem. It showed the existence of very large uncertainties in these models, stemming largely from the terrestrial ecosystem components. These components of most C4MIP models had never been subjected to rigorous evaluation, although targets had been established in the late 1990s, based on atmospheric CO_2 concentration measurements. Since then, other suitable benchmarking data sets have come on-stream, such as satellite-derived measurements of fractional absorbed photosynthetically active radiation (fAPAR) and the FLUXNET eddy covariance high - frequency measurements of CO_2 and energy fluxes between the atmosphere and the land.

Similarly, a variety of Earth System observations offer potentially powerful means to evaluate and better constrain the performance of models for different Earth System components, such as ocean biology and biogeochemistry, biogenic emissions and the uptake and deposition of trace gases at the land-atmosphere and land-ocean interfaces, and atmospheric aerosol and trace gas distributions. Major investments by the developed nations in ground- and space-based observing systems for multiple Earth System components have not yet been deployed adequately as aids to the development and quality assurance for complex Earth System models.

2. Data-Model Comparisons in QUEST

The issue of model evaluation has informed many aspects of the scientific strategy of the UK Natural Environment Research Council's programme for Earth System Science *Quantifying and Understanding the Earth System* (QUEST, http://quest.bris.ac.uk). Several QUEST projects and Working Groups contributed to the better integration of modelling and observations in Earth System Science. A selection of highlights from this work is given below:

- The MarQUEST project and the Green Ocean Modelling Working Group have developed innovative techniques using spectral ocean reflectance data to construct global seasonally-varying fields of the relative abundance of different phytoplankton functional types. These data are a key resource for the evaluation of the new generation of ocean biology/biogeochemistry model developed for the QUEST Earth System Model (QESM).
- The Carbon Cycle Modelling, Analysis and Prediction (CCMAP) project, in collaboration with the EU-funded ENSEMBLES project, has developed a series of benchmarks for terrestrial carbon cycle modelling based on seasonal cycles and interannual variability of FLUXNET measurements (CO_2 and latent heat exchanges with the atmosphere at selected eddy covariance sites), atmospheric CO_2 concentrations (selected measurement sites), fAPAR measurements (global fields) and streamflow (selected large catchments). The benchmarks have been piloted using the JULES model (as used in QESM), and also, for comparison, with the French ORCHIDEE model and three members of the offline LPJ/LPX family of models. The comparisons with CO_2 concentrations require the use of an atmospheric tracer transport model, but for this application it is possible to automate the process using the model's inverse. The comparison with streamflow

data requires the use of a hydrological routing model, which was applied by the Met Office to all of the models' gridded runoff fields.

- The QUEST Advanced Fellow (Dr Andrew Manning, UEA), has established a laboratory at UEA for high - precision atmospheric monitoring of key carbon cycle tracers, including O2 concentration. Atmospheric concentration measurements for long-lived atmospheric constituents including greenhouse gases had been a surprisingly neglected area in the UK, to the detriment of integrative science based around such measurements.
- The Fire into JULES project used remotely sensed data on burnt area, and its seasonal and interannual variations, to evaluate a fire prediction module for JULES. Subsequent work by the FireMAFS project has used remotely sensed fire data to improve parameter estimates in this model.
- The Global Palaeofire Working Group, in collaboration with the QUEST Deglaciation project, has produced a unique global data set of standardized charcoal records from terrestrial and marine sediments. This has been used to demonstrate the consistency between biomass burning variations and reconstructed northern hemisphere temperature variations during the past 2000 years (up to the Industrial Revolution), and the predictability of biomass burning trends on orbital (multimillennial) time scales over the past 20,000 years.
- QUEST funded a workshop launching a new international initiative, the iLAMB (international Land - Atmosphere Modelling and Benchmarking) project, which aims to set up universally accepted benchmarks for modelling land surface processes in the context of both numerical weather prediction and climate and Earth System modelling.

3. Discussion points

3.1. Model Weighting:

Given the variability of output from different climate and Earth system models, one approach to providing information to policy makers involves averaging results from several "independent" models (e.g. parallel model runs developed in different countries' contributions to IPCC reports). That the "average model" often performs better against benchmarks than individual models is an instance of the "wisdom of crowds". The problem is that this approach does not help improve any of the models.

Weighting the contribution of different models according to their performance against benchmarks introduces further problems. This could be done in many ways, but it has yet to be shown that it yields better results than simple averaging. More fundamentally, the model that performs best against a restricted set of benchmarks (for example, contemporary climate) is not necessarily the one that will provide the most accurate predictions when exercised outside the benchmarking domain (for example, future climate change). These considerations cast some doubt on the value of model weighting.

It was suggested that model evaluation and weighting should be treated as two separate problems: model evaluation is essential, whereas weighting is controversial.

Weighting (or selection, which is just a form of weighting in which the weights are one or zero) of models could in principle be done differently for different applications, depending on the models' ability to simulate phenomena relevant to that application. For example, prediction of dust storm frequencies might select models with good simulations of 3-dimensional dust fields. However for some applications, such as impacts of climate change on the hydrological cycle, it may be very difficult to identify which variables are key because so many different interacting processes are involved.

One well - founded approach to weighting was exemplified by the Met Office's Quantifying Uncertainties in Model Predictions (QUMP) project, where a climate model was systematically "detuned"

by creating an ensemble of model versions with parameter values drawn from a prior distribution determined by expert opinion (called a "perturbed physics ensemble"). Then, model versions were weighted according to a composite metric expressing the degree of agreement with observed climate.

3.2. Systematic parameter optimization of models:

The intercomparison and benchmarking of highly complex models is beset by the problem of suboptimal parameter estimation: many parameters are only known to within certain limits, and their values therefore may legitimately be "tuned" to improve benchmark performance. But this process is usually done in an ad hoc way.

The wider application of systematic parameter optimization methods would enable models to be compared under equal conditions. Pre-optimizing each model's performance against the benchmarks would leave residual data-model differences that must depend on structural inadequacies of the models.

The adjoint technique for parameter optimization is now well established for modelling of subsystems such as terrestrial and marine carbon cycling and atmospheric tracer transport. Work carried out within CCMAP has established that it also has promise for climate modelling with general circulation models.

3.3. Using past data:

One approach to developing greater confidence in the ability of models to make predictions outside their "comfort zone" is to extend the field of model evaluation to include the longer term past. This logic has been accepted for the IPCC Fifth Assessment Report, where the standard set of model runs for the Coupled Model Intercomparison Project (CMIP5) archive now includes the long-standing Palaeoclimate Modelling Intercomparison Project (PMIP) 'snapshot' climate simulations of the mid-Holocene (6000 years ago) and Last Glacial Maximum (LGM, 21,000 years ago), and an optional transient "millennium" simulation of changes in climate, forced by orbital, greenhouse gas, solar and volcanic variations, from 850 to 1850 AD.

'Palaeo benchmarking' was hampered until recently by the lack of global data sets on key Earth System variables to compare with model results. The evaluation of palaeoclimate simulations in the literature has tended to be lacking in quantitative rigour. QUEST made substantial investments and engaged in international initiatives to ensure this is no longer the case:

- The global data set on biomass burning changes over the past 20,000 years is mentioned above.
- The Quaternary Reconstructions group (an offshoot of the QUEST PMIP Working Group) has assembled a comprehensive data set of published pollen-based reconstructions of mean annual temperature and precipitation, effective moisture, growing degree days and mean coldest- and warmest-month temperatures. This work also demonstrated the high explained variance of these reconstruction methods when applied to surface pollen samples and confronted with modern observed climate.
- The Working Group on Abrupt Climate Changes has assembled a worldwide data set of long, high -resolution pollen and charcoal sequences extending back into the last glacial, and encompassing multiple rapid warming and cooling events. These events (Dansgaard-Oeschger and Heinrich events) are an under-exploited resource for the evaluation of "fast" Earth System behaviour and also of responses of species and ecosystems to large and rapid changes in climate.
- A QUEST workshop to launch the Palaeo Carbon Modelling Intercomparison Project (PCMIP) established a protocol for diagnosing atmospheric CO_2 variations from the millennium simulations, if run using a carbon-cycle enabled climate model. This allows the ice-core CO_2

record itself to be used as a benchmark for the carbon cycle component of the millennium simulations.

3.4. Specificity in benchmarking:

A useful approach to benchmarking looks at individual processes rather than aggregate outcomes. For example, it has proved more fruitful to compare modelled and observed soil drying rates during precipitation-free periods rather than monthly soil moisture values. This work indicated that land surface models tend to dry out soils too quickly after rain.

There are limits to what can be achieved by benchmarking against a mean state, when it is a change in state that is of greatest interest. Nevertheless, errors in the mean state have a habit of propagating into in the response. For example, the Hadley Centre model tends to dry the Amazon basin strongly in future climate-change simulations but this must be viewed against the background that the model produces too little precipitation there, even in the current state.

3.5. Quantitative metrics for data-model comparisons:

The utility, or otherwise, of expressing data-model similarities or differences using quantitative metrics was discussed in depth. Metrics have the advantage that they are easy to automate, and they can be combined across more than one indicator and multiple locations. But there is an inevitable arbitrariness in defining them, especially in multi-indicator composite metrics, and they are reductive by definition, so an over-reliance on metrics is undesirable. They can also lead to an unproductive 'beauty contest' among models. It has been shown that the best-established climate models generate the best statistics of agreement with indicators of present climate, but it is not entirely clear to what extent this simply reflects the resources available to different modelling groups. Well-funded groups could have just spent more effort on model tuning, especially as the target data sets are known in advance. Thus the metrics do not necessarily signal that these models have better predictive ability than others.

3.6. Benchmarking system characteristics:

There is no guarantee at all that selecting the 'best model components' and coupling them would lead to the 'best Earth System model'. There is a need for a two-stage process in evaluating coupled Earth System models. First, components needed to be evaluated 'offline' with other aspects prescribed (e.g., terrestrial carbon cycle models can be forced by observed climate data and CO_2 concentrations). This first step may already result in modifications and re-tuning of the model. Then, after coupling, the components all need to be evaluated again, using the same benchmarks. This second step may again generate a requirement for further modification and tuning of any or all of the components. This principle is well illustrated, for example, by the problem of simultaneously simulating the southern polar vortex and the stratospheric ozone hole, or wind speeds and dust sources. This type of problem is already familiar in climate modelling, but it becomes especially acute when highly non-linear processes are involved (e.g., the relation between dust ablation and wind speed).

This problem is most acute when the system shows threshold behaviour, as with the effect of freshwater forcing on the Atlantic Meridional Overturning Circulation (AMOC). The only credible way to evaluate the sensitivity of threshold behaviour in models involves attempting to simulate climate changes at times when the threshold has actually been crossed. This is the case for AMOC shutdown, which occurred repeatedly during rapid cooling events during the last glacial period. The most recent occurrence was the Younger Dryas period, a 'thousand-year chill' that interrupted the transition from the

LGM to the present interglacial. Work in the QUEST Deglaciation project, using the FAMOUS climate model, has indicated that although the standard version of FAMOUS appears to have an acceptable representation of the modern AMOC, its sensitivity to freshwater forcing is unrealistically low. It therefore does not generate a realistic Younger Dryas in response to an appropriate amount and duration of freshwater forcing. However, within a perturbed physics ensemble of FAMOUS versions, the standard version turns out to be somewhat unusual: several other versions of the same model show greater and more realistic sensitivity to freshwater forcing. Some have speculated that widely used climate models may commonly underestimate this sensitivity, because of an inherent preference by modellers for a 'more stable' control climate.

4. Concluding remarks

Gavin Schmidt (NASA- GISS) had the last word, which resonated strongly with participants. He opined that the use of evaluation metrics *should* be an integral part of development process of Earth System models—but it is not, because of the lack of a concerted approach by the research community. He noted that over 600 papers have been published to date describing various diagnostics for different aspects of Earth System models. No single group could possibly ever perform all of the diagnostics that have been proposed. This is a key issue for the Earth System modelling community.

There is a huge gain in efficiency when the set- up of a diagnostic (assembly of data sets, specification of the modelling protocol, definition of metrics for data- model comparison) is done *once*, by one group, and made freely available to others. A more open and interactive process is urgently required to make this the norm rather than the exception.

Acknowledgement

This final QUEST Scientific Liaison Group meeting was held as a session at Earth System Science 2010, the IGBP-AIMES Open Science Conference, sponsored and organised by QUEST. See earthsystemscience2010.org. Contributors to the discussion included: E. Blyth (CEH); B. Booth (Met Office); E. Buitenhuis (UEA); P. Friedlingstein (QUEST/LSCE); G. Schmidt (NASA); M. Scholze (University of Bristol); E. Wolff (BAS); and other participants.