Toward a model space and model independence metric

Gab Abramowitz[1,2] and Hoshin Gupta[3]

Received 29 November 2007; revised 25 January 2008; accepted 30 January 2008; published 4 March 2008.

[1] Understanding the relationship between computer-based models and the environment they simulate is becoming increasingly important as we try to predict how the earth's climate will change. As a surrogate for the representation of uncertainty in a prediction problem, it is common to use the range of behaviour from a set of models (an ensemble), and the ensemble mean as the 'best guess' prediction. We suggest a 'model space' metric, which, by providing one relevant definition of model independence, could allow us to begin to understand the relationship between model spread and prediction uncertainty. This in turn could allow the minimisation of bias from the inclusion of similar models in ensembles and quantification of how much independent information each model contributes to the prediction problem. **Citation:** Abramowitz, G., and H. Gupta (2008), Toward a model space and model independence metric, *Geophys. Res. Lett.*, *35*, L05705, doi:10.1029/2007GL032834.

1. Introduction

[2] Multi-model ensembles are widely used to increase predictive ability in climate and impacts modelling [e.g., *Intergovernmental Panel on Climate Change* (*IPCC*), 2001; *Gillett et al.*, 2002; *Palmer et al.*, 2005]. The rationale is that different climate modelling centres around the globe produce quite different models, and so provide independent estimates of what future climate may be like. The more independent estimates we have, the more errors tend to cancel, and the more confidence we can ascribe to predictions shared between the models. The impetus for model ensemble prediction has come from the recognition that for any measure or variable, several models may perform equally well [e.g., *Beven and Freer*, 2001; *IPCC*, 2001].

[3] Little work has been directed, however, to determining the extent to which different models give independent estimates [*Tebaldi and Knutti*, 2007]. There is a real chance that some models are similar and so do not contribute independent information, thus skewing the ensemble result. This may be because they share some parameterisations (conceptually or numerically) or are 'tuned' using the same (potentially biased) observational data, for example. To date, neither approaches that weight model members by performance on current data [e.g., *Krishnamurti et al.*, 2000] nor those that use equal weighting [e.g., *Räisänen and Palmer*, 2001] explicitly consider this question. This issue is particularly important for those frameworks that reduce the weight of projections that appear to be outliers when compared to the ensemble average [e.g., *Giorgi and Mearns*, 2002]. Attempting to formalise a measure of model independence may also be the only mechanism we have to assess the distribution of our models relative to the true climate system behaviour.

[4] A clear way to think of model dependence is in the context of models sharing biases. For simplicity, we choose to demonstrate the metric below using a single component of a global climate model (GCM) in an uncoupled environment. An uncoupled component model's response to smooth variations in its inputs is more likely to be smooth, allowing for a clearer characterisation of the conditions in which it shows strong biases. We suggest that a clear and intuitive way to consider bias for the purposes of identifying model dependence is in terms of 'conditional bias'. That is, a statistical characterisation of model bias under a particular subset (or small range) of its input and initial state conditions [e.g., *Tselioudis and Jakob*, 2002; *Williams and Tselioudis*, 2007]. The aggregation used to define bias in this way is 'conditional', rather than temporal.

[5] Even for a single component model there are many possible sources of bias. Figure 1 is a typical systems representation of a component model which shows these sources. While bias which stems from input, parameter (time-independent inputs), and state space is quantifiable, either through frequentist or Bayesian techniques, the space of all possible models for a given prediction problem is not a real number space and so considerably more difficult to conceptualise. A particular model is the result of a chain of model building steps, each of which may introduce bias into the final model. Important processes in the natural system may be unseen by us (not part of our perceptual model), not deemed important (not part of our conceptual model), inadequately represented (in our symbolic model) or suffer from poor numerical or scale implementation (issues associated with the numerical model).

[6] Understanding model independence is linked to understanding something of the nature of model space. Perhaps our only opportunity for characterisation of the model space is to examine 'projections' of it onto various performance measures. That is, how different models in the model space perform in different performance measures. Below we consider one such projection of the model space, onto what we might call the 'conditional bias space'. We also present a metric on the conditional bias space, which by allowing us to measure the distance between models, provides a proxy for model independence. While no single measure can serve as a conclusive indicator of performance [*Smith*, 2002; *Medlyn et al.*, 2005], we argue that the conditional bias approach is particularly useful for assessing

[1]Climate Change Research Centre, University of New South Wales, Sydney, Australia.
[2]Also at CSIRO Marine and Atmospheric Research, Melbourne, Australia.
[3]Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA.

0094-8276/08/2007GL032834

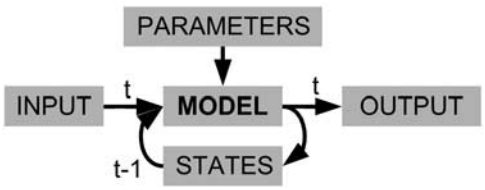

Figure 1. A typical systems representation of a model where the input, parameter, state, and output spaces are real number spaces.

independence, and provides valuable information that other performance measures do not.

2. Methodology

[7] We choose a single group of component models – land surface models, to illustrate. Three are compared: the Community Atmosphere Biosphere Land Exchange (CABLE) [*Kowalczyk et al.*, 2006]; Community Land Model (CLM) [*Oleson et al.*, 2004]; and the ORganising Carbon and Hydrology In Dynamic EcosystEms model (ORCHIDEE) [*Krinner et al.*, 2005]. All are executed using default parameter sets derived from coarse global grids (as would be used inside a GCM), for the sites simulated (corrected for vegetation and soil type). We examine two outputs from these models which affect the climate system at very different time scales, net ecosystem exchange of CO_2 (NEE) and latent heat flux (Qle). These models are driven by approximately 40 site-years of meteorological observations from 13 Fluxnet flux tower sites [*Baldocchi et al.*, 2001] across two continents, aggregated to the models' half hourly time step size. Site details are shown in Table 1.

[8] We want to construct a measure of model independence by examining how different models behave under similar sets of conditions. We consider models essentially as multi-dimensional functions, mapping from input and initial state space to output space (see Figure 1). We then define ''similar conditions'' by dividing the input/initial state space into several distinct regions, and then examine model behaviour in each of these regions. Here we use a Self Organising Map (SOM) [*Kohonen*, 1989] to define these regions (although there are other possible ways to do this), and choose only a subset of the total number of input and initial state variables to illustrate. We use the SOM to divide four meteorological inputs of the land surface models into nine behavioural regions or 'nodes'. Each node could be thought of as a mode of model behaviour (in terms of these inputs) and therefore collects together all time steps which are similar in this regard. Figure 2 shows the mean values of each of the four variables for time steps belonging to each SOM node. We now examine the output from the models associated with time steps belonging to each node.

[9] Figure 3 shows density functions for NEE outputs (for each node shown in Figure 2) from two models, CLM and CABLE. The intersection of these density functions is shaded, and represents a measure of similarity between the two models for the climatological regime represented by each node. We represent the area of this shaded region for the n^{th} node notationally as $\int \min(m_{1,n}, m_{2,n})dx$. Also shown in Figure 2 is the number of time steps belonging to each node, a_n. If the total number of time steps is A (here 726384), then we define the distance between model m_1 and model m_2 as:

$$d(m_1, m_2) = 1 - \sum_{n=1}^{N} \left[\frac{a_n}{A} \int \min\left(m_{1,n}, m_{2,n}\right) dx \right] \qquad (1)$$

where N is the number of SOM nodes. That is, we simply add all areas associated with the shaded regions, weighted by the number of time steps belonging to each node. Since these are density functions, this sum is between 0 and 1. Values closer to 0 suggest that two models are quite similar, and may be useful as a proxy for model dependence. As suggested above, this is a distance measure between models in a *projection* of the model space. The measure provides us with information similar to correlation, yet it is correlation in input/initial state conditions, rather than in time, is not sensitive to noise, and contains information about the mean.

[10] In this projected space, the measure appears a suitable candidate for a metric in the mathematical sense: $d \geq 0$ and it will only be zero when all density functions overlap perfectly, so $d(m_1, m_2) = 0$ if and only if $m_1 = m_2$ (at least if we consider a range of SOM resolutions). It should also be clear that $d(m_1, m_2) = d(m_2, m_1)$. We shall further see that at least for the models presented here, $d(m_1, m_2) + d(m_2, m_3) \geq d(m_1, m_3)$, the triangle inequality holds. If this is universally true, we may be able to use the metric space properties of this projected space to infer statistical characterisations of the model space.

3. Results and Discussion

[11] Figure 4 shows the distances between the three land surface model pairs for a variety of SOM resolutions, for

Table 1. Thirteen Flux Observation Sites

	Veg Type	Latitude	Longitude	Country	Annual Rain	Years/Length
Aberfeldy	forest	56°37′N	03°48′W	Scotland	1200	12/3/97 – 31/12/98
Bondville	cropland	40°0′N	88°18′W	USA-IL	760	1/1/97 – 31/12/99
Bordeaux	forest	44°42′N	00°46′W	France	950	12/7/96 – 31/12/98
Flakaliden	forest	64°07′N	19°27′E	Sweden	590	8/10/96 – 31/12/98
Hyytiala	forest	61°51′N	24°17′E	Finland	640	2/4/96 – 31/12/03
Little Washita	grassland	34°58′N	97°59′W	USA-OK	830	14/5/96 – 31/12/98
Loobos	forest	52°10′N	05°45′E	Netherlands	790	1/1/97 – 31/12/02
Metolius	forest	44°30′N	121°37′W	USA-OR	710	1/1/96 – 31/12/97
Norunda	forest	60°05′N	17°28′E	Sweden	530	1/1/96 – 31/12/98
Ponca City	grassland	36°46′N	97°08′W	USA-OK	800	1/1/97 – 31/12/97
Shidler	grassland	36°56′N	96°41′W	USA-OK	830	1/1/97 – 31/12/97
Tharandt	forest	50°58′N	13°38′E	Germany	820	1/1/96 – 31/12/00
Weiden Brunnen	forest	50°09′N	11°52′E	Germany	890	12/6/96 – 31/12/99

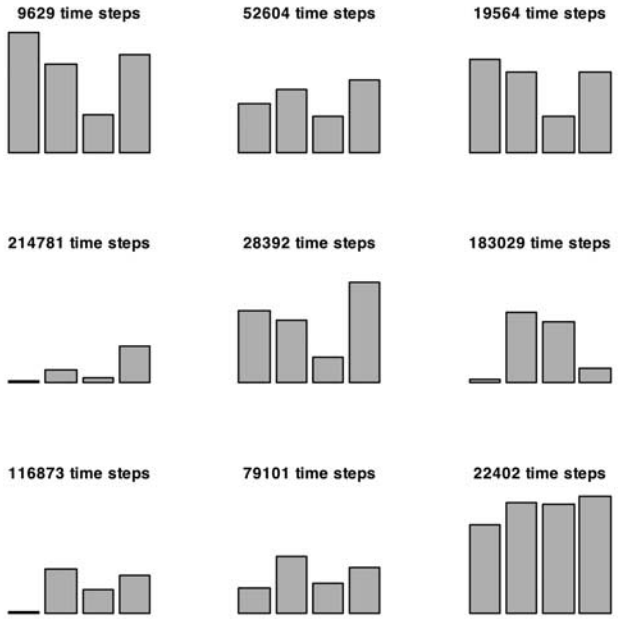

Figure 2. Mean values of 4 model input variables for time steps associated with each node in the SOM. The four variables, each represented by a column, from left to right are downward short-wave radiation, air temperature, specific humidity, and wind speed. Values are scaled to show relative differences between nodes.

both output variables, using the metric described above. The distance between CABLE and ORCHIDEE is shown as a solid line, CLM and ORCHIDEE dashed and CABLE and CLM dotted. While distance is dependent on the SOM size, as the size increases it appears that the distance may converge asymptotically. Note that the SOM size is limited by the number of data required at each node to construct a meaningful distribution. It is reassuring that within a small margin of error, the relative levels of dependence seem insensitive to the number of SOM nodes (the lines remain close to parallel). Similarly, for both model outputs, regardless of SOM resolution, this metric satisfies the triangle inequality discussed above.

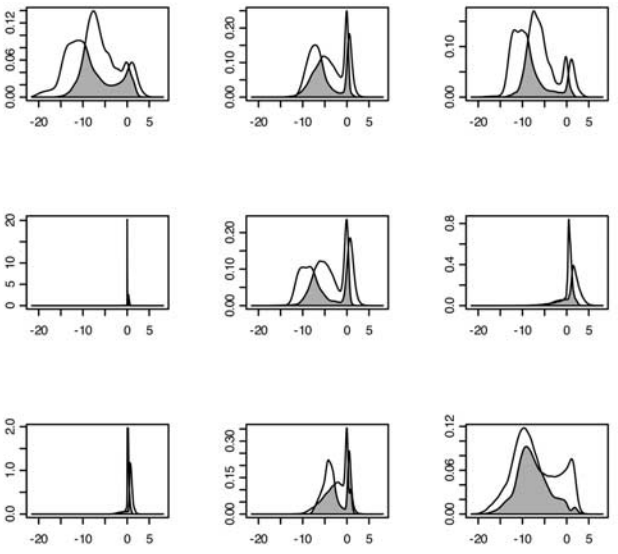

Figure 3. Density functions for NEE flux from CLM and CABLE for the nine nodes in Figure 2. Abscissa units are μmol/m^2/s. The shaded region represents the intersection of the two functions.

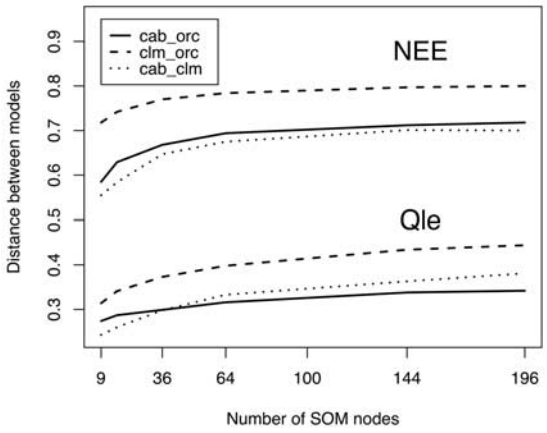

Figure 4. Distances between models as a function of the number of SOM nodes used to calculate it. Results are shown for all three model pairs for both carbon (NEE) and moisture (Qle) fluxes.

[12] There are good reasons for choosing this distance measure as a proxy for model independence over other projections of the model space (such as, mean or correlation). Firstly, as discussed above, it may well be a metric space, and so provide the possibility of statistical representation of the model space using metric space axioms. It also contains information about the mean, correlation and standard deviation in a single measure, and is not sensitive to noise. By including initial model states for each time step as part of the input ''conditions'' of simulation, the measure also implicitly contains information about model dynamics. It of course cannot tell us how dynamics will be manifest in longer term or coupled simulations, but it may tell us which models are likely to behave differently in these environments. We could also use this distance measure as a performance measure, by considering an observational data set as a model. Models close to the 'observed model' or 'true model' in the model space projection are better performing. We note however that models with significantly different state spaces, temporal or spatial scales of application cannot be compared with this measure.

[13] Although in Figure 4 the three models are further apart in the projected model space when predicting NEE than latent heat, it doesn't necessarily follow that an NEE prediction ensemble of these models would have any more predictive power than a latent heat ensemble. Increased independence alone does not equate to increased performance. For example, a model which produced Dirac δ-functions for each SOM node would clearly be independent of these models using this measure, yet perform poorly. Thus to choose the best model ensemble, we must consider both the independence and performance of potential ensemble members. To reinforce the duality of performance and

independence, we note that dependence as defined by this measure is not necessarily evident in conventional time-aggregated measures used for model evaluation. While we might infer from Figure 4 that CLM and ORCHIDEE provide the two most independent estimates of NEE flux, their RMSE score when compared with observations was closest (4.74 μmol/m^2/s and 4.60 μmol/m^2/s respectively versus CABLE's 4.28 μmol/m^2/s). A similar result applies for latent heat, with RMSE scores of 44.5, 45.4 and 50.9 W/m^2 and average fluxes of 26.2, 23.9 and 37.3 W/m^2 for CLM, ORCHIDEE and CABLE respectively.

[14] Choosing model weights for an ensemble is then a process of deciding on a performance measure (or aggregation of performance measures) and then using a weight description that values performance and independence in an appropriate ratio. This is clearly not a trivial exercise. We want to prescribe high value to models that are close together (dependent) if they are also very close to the observations, relative to other models. An example might be

$$w_{m_i} = \frac{p_{m_i} \cdot \sum_{j \neq i}^{n} d(m_i, m_j)}{w_{m_i} + \cdots + w_{m_n}} \quad (2)$$

where p_{m_i} is the performance measure for model i, $d(m_i, m_j)$ is the distance between model i and model j (the denominator is for normalisation). Such a weight construction shifts the focus from valuing similarity to observations alone to additionally valuing dissimilarity with other ensemble members. Using this weight definition and the inverse of RMSE as the performance measure, the latent heat weights for the three models above are 0.28, 0.35 and 0.37 for CABLE, ORCHIDEE and CLM respectively. For NEE, the weights are 0.34, 0.33 and 0.33 respectively. This is of a course a very crude choice of performance measure that is sensitive to the output variable units, but it gives a sense of how weighting might work.

4. Conclusion

[15] We have presented a proxy for model independence based on the notion of conditional bias. That is, models that consistently produce similar predictions under similar input and initial state conditions are deemed to show dependence. Furthermore, this dependence is quantified in a metric that allows us to measure the distance between models in a theoretical model space. Because this space also allows observational data sets to be considered as de facto models, we may be able to position models in an ensemble relative to the natural system behaviour and ascertain whether, by including dependent models, they produce unnecessarily biased mean predictions. The process of constructing model weights in this situation was briefly discussed. Currently, issues such as funding and computational resources decide which models contribute to international ensembles [*Tebaldi and Knutti*, 2007]. Arguably a more effective use of resources would be the development of models that contribute independent estimates to the climate change problem. A technique such as the one described here could help provide criteria to meet such a goal.

[16] **Acknowledgments.** Thanks to Kuo-lin Hsu for the SOM code, Andy Pitman for review and advice, and Fluxnet for the observed meteorological and flux data; this work exclusively used open source software.

References

Baldocchi, D. D., et al. (2001), FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor and energy flux densities, *Bull. Am. Meteorol. Soc.*, *82*, 2415–2434.

Beven, K., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the glue methodology, *J. Hydrol.*, *249*, 11–29.

Gillett, N. P., F. W. Zwiers, A. J. Weaver, G. C. Hegerl, M. R. Allen, and P. A. Stott (2002), Detecting anthropogenic influence with a multi-model ensemble, *Geophys. Res. Lett.*, *29*(20), 1970, doi:10.1029/2002GL015836.

Giorgi, F., and L. Mearns (2002), Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging" (REA) method, *J. Clim.*, *15*, 1141–1158.

Intergovernmental Panel on Climate Change (IPCC) (2001), *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, edited by J. T. Houghton et al., Cambridge Univ. Press, Cambridge, U. K.

Kohonen, T. (1989), *Self-Organization and Associative Memory*, Springer, New York.

Kowalczyk, E. A., Y. P. Wang, R. M. Law, H. L. Davies, J. L. McGregor, and G. Abramowitz (2006), *The CSIRO Atmosphere Biosphere Land Exchange (CABLE) Model for Use in Climate Models and as an Offline Model*, *CSIRO Mar. Atmos. Res. Pap. 013*, Commonw. Sci. and Ind. Res. Org., Aspendale, Vic., Australia. (Available at www.cmar.csiro.au/e-print/open/kowalczykea_2006a.pdf.)

Krinner, G., N. Viovy, N. de Noblet-Ducoudré, J. Ogée, J. Polcher, P. Friedlingstein, P. Ciais, S. Sitch, and I. C. Prentice (2005), A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, *Global Biogeochem. Cycles*, *19*, GB1015, doi:10.1029/2003GB002199.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. Larow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran (2000), Multimodel ensemble forecasts for weather and seasonal climate, *J. Clim.*, *13*, 4196–4216.

Medlyn, B. E., A. P. Robinson, R. Clement, and R. E. McMurtrie (2005), On the validation of models of forest CO_2 exchange using eddy covariance data: Some perils and pitfalls, *Tree Physiol.*, *25*, 839–857.

Oleson, K., et al. (2004), Technical description of the community land model (CLM), *Tech. Rep. TN-461+STR*, Natl. Cent. for Atmos. Res., Boulder, Colo.

Palmer, T. N., F. J. Doblas-Reyes, R. Hagedorn, and A. Weisheimer (2005), Probabilistic prediction of climate using multi-model ensembles: From basics to applications, *Philos. Trans. R. Soc., Ser. B*, *360*, 1991–1998, doi:10.1098/rstb.2005.1750.

Räisänen, J., and T. N. Palmer (2001), A probability and decision-model analysis of a multimodel ensemble of climate change simulations, *J. Clim.*, *14*, 3212–3226.

Smith, L. (2002), What might we learn from climate forecasts?, *Proc. Natl. Acad. Sci.*, *99*, 2487–2492.

Tebaldi, C., and R. Knutti (2007), The use of the multimodel ensemble in probabilistic climate projections, *Philos. Trans. R. Soc., Ser. A*, *365*, 2053–2075, doi:10.1098/rsta.2007.2076.

Tselioudis, G., and C. Jakob (2002), Evaluation of midlatitude cloud properties in a weather and a climate model: Dependence on dynamic regime and spatial resolution, *J. Geophys. Res.*, *107*(D24), 4781, doi:10.1029/2002JD002259.

Williams, K. D., and G. Tselioudis (2007), GCM intercomparison of cloud regimes: Present-day evaluation and climate change response, *Clim. Dyn.*, *28*, 231–235, doi:10.1007/s00382-007-0232-2.

G. Abramowitz, CSIRO Marine and Atmospheric Research, Private Bag 1, Aspendale, VIC 3195, Australia. (gabsun@gmail.com)

H. Gupta, Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721-0011, USA.