Physical Sciences, Geophysics

Association of parameter, software and hardware variation with large scale behavior across 57,000 climate models

Christopher G. Knight[*]
Sylvia H. E. Knight[†¶]
Neil Massey[‡†]
Tolu Aina[†]
Carl Christensen[†]
Dave J. Frame[†]
Jamie A. Kettleborough[§‖]
Andrew Martin[‡]
Stephen Pascoe[‖]
Ben Sanderson[†]
David A. Stainforth[†]
Myles R. Allen[†]

[*] Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester M1 7ND
[†]Atmospheric, Oceanic and Planetary Physics, Clarendon Laboratory, Parks Road, Oxford OX1 3PU UK
[‡]Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD UK
[§] The Met Office, FitzRoy Road Exeter Devon EX1 3PB UK
[‖] CCLRC, Chilton, Didcot, Oxon, OX11 0QX UK
[¶]Corresponding Author : sknight@atm.ox.ac.uk

	Main	Supporting
Text	15pp	1
Figures	6	4
Tables	1	5

Abstract 208 words
Characters 38902

Abbreviations
CS Climate Sensitivity
CV Coefficient of Variation (standard deviation as a percentage of the mean)
GCM General Circulation Model
SST Sea Surface Temperature

In complex spatial models, as used to predict the climate response to greenhouse gas emissions, parameter variation within plausible bounds has major effects on model behavior of interest. Here we present an unprecedentedly large ensemble of over 57000 climate model runs in which 10 parameters, initial conditions and the hardware and software used to run the model have all been varied. We relate information about the model runs to large scale model behavior (equilibrium sensitivity of global mean temperature to a doubling of carbon dioxide). We demonstrate that effects of parameter, hardware and software variation are detectable, complex and interacting. However we find most of the effects of parameter variation are due to a small subset of parameters. Notably, the entrainment coefficient in clouds is associated with 30% of the variation in climate sensitivity (CS) seen, though both low and high values can give high CS. We demonstrate the effect of hardware and software is small relative to the effect of parameter variation and, over the wide range of systems tested, may be treated as equivalent to that caused by changes in initial conditions. We discuss the significance of these results in relation to the design and interpretation of climate modeling experiments and large scale modeling more generally.

Introduction

Simulation with complex mechanistic spatial models is central to science from the level of molecules (1) via biological systems (2, 3) to global climate (4). The objective is typically a mechanistically based prediction of system level behavior. However both through incomplete knowledge of the system simulated and the approximations required to make such models tractable, the 'true' or 'optimal' values of some model parameters will necessarily be uncertain. A limiting factor in such simulations is the availability of computational resources. Thus combinations of plausible parameter values are rarely tested, leaving the dependence of conclusions upon the particular parameters chosen unknown.

Observations of the modeled system are vital for model verification and analysis e.g. turning model output into probabilistic predictions of real-world system behavior (5-7). However typically, few observations are available relative to the complexity of the model. There may also be little true replicate data available. For instance, there can only be one observational time series for global climate. Thus if the same observations are used to fit parameter values, there is a severe risk of over-fitting– gaining limited verisimilitude at the cost of the mechanistic insight and predictive ability for which the model was originally designed.

To avoid fitting problems parameter estimates must be refined directly. In some biological systems, direct and simultaneous measurement of large numbers of system parameters (e.g. protein binding or catalytic constants) may soon be possible. In other systems such as climate models, this approach is not an option. Thus it is vital to focus efforts in parameter refinement. Deciding how to do this presents challenges: i) to determine whether there is dependence of model behavior of interest upon parameter variation within plausible bounds. ii) to determine whether dependence applies to all uncertain parameters or only a more tractable subset. iii) to quantify the nature of parameter dependence. Since parameters interact in complex and unknown ways, meeting these challenges entails considering a very large parameter space.

In this paper we address all three challenges for a state of the art General Circulation Model (GCM) of global climate. Without fitting to observations, we analyze an ensemble of over 57000 model runs in which 10 parameters and initial conditions were systematically varied. While large studies have traditionally been carried out on supercomputers, it is currently only possible to perform this many simulations via a distributed computing approach. Prior to this project the largest published comparable ensemble was of 53 model runs (8, 9). We have achieved such a large number data set via the climate*prediction*.net project using idle processing capacity on personal computers volunteered by members of the public. This approach entails variation in hardware and software used to run the model and serious concerns have been raised that results might depend *only* on this variation. Processes of rounding that vary between systems and lead to small differences in simple calculations are a well known issue highlighted in projects working with a similar distributed computing architecture (10). Given the enormous numbers of such calculations in a GCM, such miniscule effects of hard/software may multiply to influence overall model behavior. Since the GCM is highly nonlinear, even small quantitative differences in model behavior of this sort could in principle produce qualitatively different results. We address this issue directly, treating hard/software variation equivalently to parameter variation.

Considering plausible values of 6 parameters and a smaller number of model runs, Stainforth et al demonstrated that, while accepted predictions of 2-5K global warming in response to a doubling in carbon dioxide (11) were indeed representative of model results, equally plausible parameter values gave global warming of >8K (4). In that study the clustering of the climate change predicted by many model versions with perturbed parameters around the prediction from the unperturbed model hinted that influence on the results may be unevenly distributed among uncertain parameters. It is thus crucial to quantify the relative importance and interactions of different uncertain parameters in determining model behavior. Using a classification and regression tree approach we demonstrate that 80% of variation in climate sensitivity (CS) to carbon dioxide is associated with variation in a small subset of parameters mostly concerned with cloud dynamics. Initial conditions, hardware and software all have small but identifiable effects on model behavior. However the nature of hard/software effects is very similar to that of initial conditions effects.

Results

Associations with climate sensitivity (CS) The dataset comprised 57067 climate model runs. These sample parameter space for 10 parameters (Table1, supporting Table2) with between 2 and 4 levels of each, covering 12487 parameter combinations (24% of possible combinations) and a range of initial conditions. 43692 runs (77%) were allocated an equilibrium sensitivity of global mean temperature to a doubling of CO_2 (Climate sensitivity –CS–see methods). Among the 13375 failures, 295 had incomplete data; 1897 failed to fit a temperature change curve; 11441 did have fits but they did not meet arbitrary criteria for acceptable error; 530 failed to progress far enough towards equilibrium in the modeled timescale. Representative time-series and fitted curves are shown in Fig1.

To determine relative contributions of explanatory variables (Table1) we fit a regression tree for CS (Fig2). This is a recursive splitting technique in which the runs are split according to the value of one of the explanatory variables so as to make the variation about the means for the subsets of the data formed by the split as small as possible. E.g. the standard deviation of CS for all runs is 1.7K which is 45% of the mean CS (3.7K i.e. the runs have a Coefficient of Variation (CV) of 45%). The first split in the tree divides runs into subsets with high and low values of *entcoef* respectively. The resulting subsets each have smaller CVs, 30% and 43%. These subsets can then be split again based on any explanatory variable with multiple levels in the subset. These figures compare with an average CV for unforced runs (i.e. runs with the same parameter set but varying initial conditions) of 8.9%. An optimal tree (supporting Table3) contained 201 such splits based on parameter values, hardware and software explanatory variables. This tree explained a large majority of the variation in CS, 80% by crossvalidation (see supporting Fig7). Fig2 shows a subset of this optimal tree that is enough to explain most of the observed variation. A plot of CS observed in the runs against that predicted by the optimal tree is shown in Fig3.

Summing over the optimal tree for each explanatory variable, the proportion of variation explained is shown in Fig4. ¾ of the total variation is explicable by just 5 variables, all of them parameters to the original model. There is a measurable effect on CS of hardware (processor, RAM size and clock speed) and software (client middleware system). These effects are however all <1% of total variation and mostly <1% of the variation attributable to the more influential parameters. E.g. the first split, described above, affects all the runs and explains 29% of the variation whereas the most explanatory split based on RAM size explains only 0.09% of the variation and divides only a small subset of runs with a CV of 61% into smaller subsets with modestly reduced CVs of 61% and 52%.

Reasons for failure to fit CS Nearly ¼ of runs failed to give results from which an unambiguous CS could be calculated. Therefore we asked whether this failure was associated with particular parameter values behaving in unexpected ways, or particular hard/software giving spurious results. We used a similar tree based approach fitting success or failure to the same explanatory variables as before (Table1).

Unsurprisingly, the proportion of variation in fitting failure that can be explained is much less than the proportion of variation explicable for CS itself. Nonetheless 33% of total variation in the data can be explained by an optimal tree (supporting Fig8 and supporting

Table4). While RAM size and processor used to run the model do have an effect on failure to produce a fit, the most important factors, as for CS itself, are model parameters (Fig4).

Since there is a systematic element, dependent on the parameter set, in our failure to fit a CS, there may also be systematic loss of particular values of sensitivity. To test for systematic loss of particular CS values we considered an alternative estimate of CS, the average temperature difference between control and doubled CO_2 phases for the last 8 of the 15 years considered. This measure seriously under-estimates high CS, as compared with nonlinear fitting, but is a reasonable approximation at low values (supporting Fig9). We compared the frequency distribution of this measure for those runs where we obtained a CS by nonlinear fitting with those where we did not (Fig5). The distributions are very similar in shape for sensitivities of about 2.5K and above with only a slight over-representation of high sensitivities in those not fitted. However there is a tail of sensitivities ≤1K that is almost entirely missed by the fitting procedure. Overall, 985 runs (1.7%) show such a low sensitivity by the difference measure but only 6 of these (0.6%) have fitted CS (cf. 78% have fits in the rest of the dataset). One example of a time series where no curve could be fit, but showing a1K sensitivity by the alternative measure is shown in Fig1. These 'missing' low sensitivity runs show a larger than expected proportion (87% rather than 11%, $P<10^{-15}$, χ^2 test) of strong CO_2 phase cooling in the Eastern tropical Pacific characteristic of a known artifactual effect of mixed layer oceans (4). These 'missing' runs also drift more than expected in the control phase (85% rather than 46%, $P<10^{-15}$, χ^2test) which may also indicate unphysical behavior. Almost all low sensitivity runs that are missed by the fitting procedure (93.5% of them) have at least one of these issues, either drift or Eastern tropical Pacific cooling.

Role of hardware and software A subset of the runs analyzed above contained identical parameters and initial conditions. 4762 combinations of parameters and initial conditions had at least 2 and up to 6 runs giving a CS. While many such 'duplicate' sets (1062 of them) gave identical results, most did not. For each parameter combination we calculated the coefficient of variation (CV) of the CS. We then fit a regression tree for this quantity in a similar way to earlier trees. An optimal tree (Fig6 and supporting Table5) explained 24% of variation by crossvalidation; the relative effects of the explanatory variables are shown in Fig4.

The only hard/software feature included in the duplicate divergence tree (Fig6) is client middleware, the software used to implement the model on different computer systems. Low levels of duplicate divergence are associated with 'classic' middleware (the in-house software used initially) and high levels with a mix of middlewares or BOINC middleware (developed for the SETI@home project (12) and used subsequently). The 5 other explanatory variables in the tree are model parameters, precisely the same as the top 5 explanatory variables in the tree for failure to fit a sensitivity (Fig4, Supporting Fig8).

We have characterized CS variation caused by differing hard/software, but it remains unclear how it should be accounted for when analyzing model output. A simple solution would be to treat variation introduced by different hard/software as equivalent to variation due to different initial conditions. This approach would be valid only if the variation in CS from these sources is indistinguishable. The hypothesis that hard/software affects CS indistinguishably from different initial conditions generates expectations:

1) Deliberate sampling of initial conditions should cover CS responses better than incidental variation due to hard/software. Thus CS variation due to different initial conditions should be an upper bound on variation due to hard/software for any particular parameter set.
2) Variation in CS from the two sources– hard/software differences and initial conditions differences– should behave similarly.

Testing these expectations requires comparison of CS variation analyses for the two sources– hard/software versus initial conditions. We already have an analysis of CS variation due to hard/software (Fig6). Therefore an equivalent analysis of CS variation due to initial conditions is required: We took the 8196 parameter sets with at least two sets of initial conditions and calculated CS variation (CV) for each. We then fitted a regression tree in the same way as for Fig6 above. The optimal tree (supporting Fig10 and supporting Table6) explains 32% of the total variation by crossvalidation. To test the expectations we compared predictions of this initial conditions tree to those of the hard/software tree (Fig6) for the 4709 models used to build both trees. Both trees are capable of predicting variation over a wide range, from CV<5% to CV≥ 40%. Expectation 1) predicts CS variation in the initial conditions tree will be an upper bound for CS variation in the hard/software tree. In 92% of models the predicted CV is higher across initial conditions than across hard/software differences, consistent with expectation 1). Expectation 2) predicts the two trees will be similar. The earliest splits based on parameters are indeed similar in the two trees (on the accretion constant *ct*, higher values giving lower variation, then, for low *ct*, on the entrainment coefficient, higher values giving lower variation). Where the two trees first diverge, splitting on different parameters, in 3 out of 4 cases the top three alternative splits considered (competing with the chosen split) include the parameter actually chosen by the other tree (data not shown). The overall contribution of each parameter is also very similar between the two trees (Fig4 rank correlation coefficient between parameter contributions for each tree 0.90 N=8 P=0.0021). Another way that the trees might be similar is in their predictions. Predictions by the two trees are weakly but significantly correlated (rank correlation 0.10, N=4709 P<0.0001). A more stringent similarity test asks if real variation not captured by the trees is associated. This variation is found in the residuals for the tree predictions. There is a small but highly significant positive association between the residuals from the two trees (rank correlation 0.11, N=4660 P<0.0001). From these tests we conclude that both our expectations are met: hard/software introduces random differences that are similar in nature but typically smaller than deliberate initial condition perturbations. This finding is consistent with the hypothesis that hardware and software affect CS in ways indistinguishable from variation in initial conditions.

Discussion

Modeled sensitivity of equilibrium global mean temperature to a doubling of carbon dioxide (Climate Sensitivity, CS) shows strong dependence on model parameters whose values are uncertain (4). We demonstrate that of 10 parameters chosen for their relevance to this issue, the relative dependence of CS upon them is highly uneven. Over half the variation associates with just three parameters (Fig4). We cannot say how relevant constraining these parameters would be to other model behaviors of interest. However for questions directly related to CS, notably prediction of CO_2 mediated anthropogenic climate change at a global level, these results imply efforts would best be directed, not towards constraining the model by observations in general, or even constraining realistic values of parameters in general, but constraining the values of these few parameters in particular.

Such findings greatly simplify model refinement. However that the range of parameters involved is simple does not mean these parameters' effects are simple. The most influential parameter here is *entcoef*, defining the rate convective clouds mix with surrounding air. Strong *entcoef* effects on CS have been observed before (9). The first most explanatory split is into runs with high and low *entcoef*: high values typically give low CS and vice versa. Consistent with this, the highest predicted CSs (>9K) are all for low *entcoef* runs, associated with high *rhcrit* and *ct* and low *vf1* (Fig2, supporting Table3) a combination indicative of reduced cloud formation. However the association of *entcoef* and CS is not true absolutely– *entcoef* interacts in complex ways to give highly varied results. The assumption that parameter variation effects combine linearly, as required for extrapolation from smaller ensembles (8, 13) does not hold. Thus in contrast to the typical relation between CS and *entcoef*, the lowest predicted sensitivities (1.6K Fig3) are for low *entcoef* runs. The latter differ from other low *entcoef* runs in having high *rhcrit*, low *ct* and high *eacf* and *cw* values (Fig2, supporting Table3) i.e. the only consistent difference between the highest and lowest CS runs is *ct*, another cloud parameter. Thus while a better estimate of *entcoef* would best improve modeled predictions of CS, we cannot define any straightforward relationship between constrained *entcoef* values and the magnitude of predicted CS. E.g. if high *entcoef* best represents reality, this does not imply low real-world CS. In that case, focused studies would be required since, even in the current large ensemble, only 25 successful runs with high *entcoef* have the combinations of other parameters predicted to give CS >8K.

Simulation output is inevitably detailed and highly multivariate. To make it useful requires simplification and assumptions to derive humanly interpretable measures of interest. We have calculated CS as a quantity of interest using a nonlinear fitting approach. This fitting assumes there is an equilibrium difference in global mean temperature to be fit and it is approached via an arbitrarily good approximation to an assumed form of curve. For the large majority of runs these assumptions hold. However we find those runs where they do not hold are a non- random subset with respect to CS. Specifically, a small tail of runs with low sensitivity cannot be assigned a CS (Fig5). In these cases e.g. that shown in Fig1 where temperatures in the control and doubled carbon dioxide phases diverge very little, the signal to noise ratio is high, making adequate fits less likely (an effect that might be ameliorated by a longer run). If there is no divergence at all or a linear divergence, one of the two parameters in the fit is undefined so there will be no fit (an effect that would not be altered by longer runs). Here, by using more than one estimate of CS, we have demonstrated the effects of our assumptions. The low sensitivity runs 'lost' are not likely to affect estimates of real-world CS, both because they tend to agree poorly with observations (6, 7) and because we find most display known non-physical effects. However these findings highlight the care needed in parameter scanning modeling studies such as this to ensure important results from plausible parameter sets are not misinterpreted or excluded simply through their failure to fit prior assumptions.

Despite increases in supercomputing power, distributed and grid approaches are increasingly necessary to tackle ever more complex modeling studies. One result is a variety of hardware and even software being used to run the model. Such differences have systematic effects on calculations, a recognized issue (10) sometimes tackled as a subset of sabotage, that also poses risks here (14). Here for the first time we have quantified these effects on a model result of interest relative to the effects of parameter variation. Sometimes the CS predicted by the model did vary with whether the model was run on an Intel Pentium 4 processor or an AMD Athlon. However there is no clear association e.g. that Intel chips give higher CS. Similarly RAM size has an effect, but different model versions respond differently, in 4 of the 6 cases of splits based on RAM size, the smaller RAM size gives the higher sensitivity, but in 2 cases the reverse is true. It may be that RAM size is acting as a surrogate for other differing aspects of hardware. We have not covered all possible hardware and software variants, notably we have not used a 64-bit architecture. However in the large variety of permutations that are covered in this dataset, systematic hard/software effects are reassuringly small relative to the effects of model parameters. e.g. Of the 7 splits based on particular processors, at most 564 runs are affected (1.3% of the total) and together all 7 splits only account for 0.3% of the variation, whereas even the fifth most explanatory model parameter (*eacf*) gives 28 splits affecting up to ~14000 runs (33% of the total) each and accounting for over 20 times as much variation as the processors (Table4).

Important effects of hard/software may however be less systematic. We identified a single software effect as important here. Runs with identical parameters and starting conditions average a coefficient of variation (CV) in CS of only 1.6% when run exclusively under the original ("classic") climate*prediction*.net client middleware. However when run under a mixture of middlewares or the more widely used BOINC client middleware (http://boinc.berkeley.edu) that average can rise to 40% depending on parameter values. The causes of this difference are unclear. We speculate that it may be due to different 'controller' code that in the classic middleware, appeared more sensitive to small computational errors. This sensitivity resulted in more crashes and thus failure to submit results for the classic middleware; BOINC being more likely to let the model run to completion despite computational problems. There was also a change of compiler for the underlying code between the two middlewares that could have had an effect. Whatever the cause, it is clear client middleware is much more important than other hard/software and, unlike other hard/software, can be controlled by the experimenter.

The computing power of distributed systems offers an approach to explore large tracts of plausible parameter space for a complex model. Alternative and potentially complementary approaches for climate models have focused on speeding up models by simplifications (reduced temporal or spatial resolution, dimensionality, physics or dynamics) relative to state-of-the-art general circulation models such as that used here. For instance FAMOUS is based on a similar model (HadCM3) but with reduced temporal and spatial resolution so it runs ~10 times quicker. This speed made it possible to tune parameter values using a conventional supercomputer (13). However the challenge of uncertain or undefined parameters remains great. Even in this study we have only been able to investigate ten model parameters. Expert choice decided the parameters to investigate and the range of levels they should take. With models as complex as these, such reliance on human skill may miss parameters that affect the results through non-obvious mechanisms. Even for parameters we considered, the number of levels used may be insufficient to define adequately their complex influences (e.g. the variety of high and low climate sensitivities associated with particular levels of *entcoef* discussed above). Both these observations suggest that investigating more parameters in more detail would be desirable and perhaps necessary to tune the model adequately. However the vast numbers of model runs involved in comprehensively scanning combinations of parameters would exceed the resources of any distributed computing setup or speeded up model on a conventional system. To add to the challenge recent work suggests it is unreasonable to hope for a generally optimized climate model, the model parameters need tuning to the specific question being asked (5). Ultimately such questions undoubtedly extend beyond the timescale of decades used here to computationally extensive questions involving paleoclimate. It will also be important to compare different models. Findings for one model may not be transferable to another, or even to different versions of the same model e.g. with altered resolution. If these modeling challenges are to be met computationally, it will require not only improvements in model speed and access to computing power, but improved methods of exploring the complex parameter space. This latter requires a carefully designed experimental (15) and computing strategy (16). It may also entail adaptive techniques, adjusting the model versions run in response to results received, which poses particular challenges in a distributed computing context where it is uncertain when or whether any particular run will be returned. Adaptive techniques include evolutionary computation and refined combinations of approaches including recursive splitting of parameter space as used in this study (17). Similar methods have been applied successfully in many fields including identification of optimal model versions given uncertain parameters in computationally simpler but nonetheless nonlinear and complex, economic climate models (18).

In conclusion, by considering an unprecedentedly large ensemble of climate model runs, we have a series of findings relevant not only to the implementation, interpretation and improvement of models predicting climate change, but to studies using large and complex models more generally. Our findings reinforce the fact that variation of parameters within plausible bounds may have a substantial systematic effect on large scale model behavior. However we find only a small subset of parameters to be associated with most of the variation in a specific behavior (CS). Those associations are complex and interacting but the small number of parameters involved provides a focus for future model refinement. In addition we have identified how the very process of making model results interpretable affects the findings. The effect of the precise hard/software implementation of the model was typically small and indistinguishable from perturbations introduced by different initial conditions.

Methods

Model and Distributed computing The climate*prediction*.net project is the first multi-thousand member ensemble of climate simulations using a state-of-the-art General Circulation Model (GCM). Members of the public worldwide download an executable version of the Met Office Unified Model. This model comprises the HadAM3 atmosphere (19) at standard resolution (3.75° longitude by 2.5° latitude, 19 vertical levels) with increased numerical stability, coupled to a mixed-layer ocean with heat transport prescribed using a heat flux convergence field varying with position and season but not year.

Participants are allocated a particular set of parameter perturbations and initial conditions enabling them to run one 45 year simulation. For each simulation the heat-flux convergence field is calculated in the first 15 years simulated, where Sea Surface Temperatures (SSTs) are fixed. In the subsequent 30 years simulated the SSTs vary according to the atmosphere-ocean heat flux. In the first 15 years of this, the control phase, CO_2 is held constant at preindustrial levels (282ppm). It is doubled for the last 15 year period.

Dataset The first 57067 simulations returned to climate*prediction*.net servers were considered. Each simulation was classified according to parameter set, initial conditions, hardware and software used to run the model. These 18 explanatory variables are listed in Table1 and Supporting Table2.

Analysis Simulated CS is taken as the predicted equilibrium difference between global mean temperature in the doubled CO_2 and control phases. This quantity was calculated via a self-starting nonlinear regression fit, using a Gauss-Newton algorithm, to the difference in the annual global mean temperatures between the doubled CO_2 and control phases. The curve fit had the form $\Delta T = S(1-\exp(-Ft/SC))$ derived from an energy balance model where ΔT is difference in global mean temperature, S is climate sensitivity (CS) t is time, C is the effective heat capacity of the model and F is the radiative forcing due to a doubling of CO_2, taken to be 3.74 Wm^{-2}. Fits that failed to converge after many iterations (1000) or gave a residual standard error >0.2K or failed to reach half their predicted equilibrium temperature in the period of the fit were rejected. Runs with a full set of data were deemed to have failed only on the basis of our failure to produce an adequate fit by these criteria, not on the bases of either temperature drift in the control phase or the relationship to observations, constraints that have been used in previous studies using similar data (4, 5).

For analyses of CS variation (Fig6, supporting Fig10) we created new explanatory variables capturing variation in the hardware and software for each set of duplicate runs: for continuous variables– RAM size and clock measures, we used coefficient of variation (CV) for discrete variables– processor, operating system and middleware, we created a discrete variable detailing whether the duplicate runs had a particular level or a mix of levels. We used these quantities as explanatory variables in these trees alongside the parameters used (see supporting Table2).

To determine the association of explanatory variables with model response we used classification and regression trees (20). These techniques recursively split data to minimize variation (measured as deviance for continuous variables and entropy for categorical variables) for the two resulting subsets of data. Splitting can in principle continue as long as there are multiple observations to be split and different levels of explanatory variables within subsets. However while the fit of the resulting tree to the data used to create it will only improve by further splitting, the ability of the tree to predict data not used to create it will not. We used a standard approach of creating large trees (considering splits down to those reducing the lack of fit by a factor 1×10^{-4}) and then pruning them to an optimal size. This size was determined by 100-fold crossvalidation i.e. splitting the data randomly into 100 equally sized subsets, using 99 as a training set and the 100^{th} as a test set. From the test set results was calculated the error in prediction (crossvalidation error) averaged over the 100 possible training and test sets. The optimal tree was chosen as the smallest where the crossvalidation error lay within one standard error of the minimum crossvalidation error.

To identify unphysical cooling in the tropical East Pacific (4) surface temperature between the final year of the control and calibration phases was taken for the 78.75W, 2.5N box and corrected for overall change by subtracting the figure for 48.75W 2.5N in 13983 BOINC runs. This quantity is distributed with the principal mode at 0K and a secondary mode around -27.5K; <-15K was deemed to show strong evidence for this cooling.

Software Statistical analyses used R 2.0.1 (21) and JMPIN 5.1 (22). Within R, classification and regression trees were fitted using the rpart v.3.1-23 package.

Acknowledgements

We thank all climate*prediction*.net participants, Tessella Support Services plc, Research Systems Inc., Numerical Algorithms Group Ltd, the CMIP II modeling groups, Joshua Knowles, Ed Tredger and Philippa Knight. This work was supported by NERC's COAPEC, e-Science and fellowship programs, the DTI and DEFRA (PECD/7/12/37).

References

1. Karplus, M. & McCammon, J. A. (2002) *Nat Struct Biol* **9,** 646-52.
2. Noble, D. (2002) *Science* **295,** 1678-82.
3. Snoep, J. L. (2005) *Curr Opin Biotechnol* **16,** 336-43.
4. Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D. J., Kettleborough, J. A., Knight, S., Martin, A., Murphy, J. M., et al. (2005) *Nature* **433,** 403-6.
5. Frame, D. J., Booth, B. B. B., Kettleborough, J. A., Stainforth, D. A., Gregory, J. M., Collins, M. & Allen, M. R. (2005) *Geophys Res Lett* **32,** L09702.
6. Knutti, R., Meehl, G. A., Allen, M. R. & Stainforth, D. A. (2006) *J Climate* **19,** 4224-4233.
7. Piani, C., Frame, D. J., Stainforth, D. A. & Allen, M. R. (2005) *Geophys Res Lett* **32,** L23825.
8. Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M. & Stainforth, D. A. (2004) *Nature* **430,** 768-72.
9. Barnett, D. N., Brown, S. J., Murphy, J. M., Sexton, D. M. H. & Webb, M. J. (2006) *Clim Dyn* **26,** 489-511.
10. de Dinechin, F., Lauter, C. & J., M. (in press) *Theoret Informatics Appl*.
11. Houghton, J. T. & Intergovernmental Panel on Climate Change. Working Group I. (2001) (Cambridge University Press, Cambridge).
12. Anderson, D. P. (2004) in *International Workshop on grid computing* (IEEE, Pittsburgh), pp. 4-10.
13. Jones, C., Gregory, J., Thorpe, R., Cox, P., Murphy, J., Sexton, D. & Valdes, P. (2005) *Clim Dyn* **25,** 189-204.
14. Germain-Renau, C. & Playez, N. (2003) in *International Conference on Supercomputing* (ACM, San Francisco), pp. 226-233.
15. Sacks, J., Welch, W. J., Mitchell, T. J. & Wynn, H. P. (1989) *Statistical Sci* **4,** 409-435.
16. Casanova, H., Legrand, A., Zagorodnov, D. & Berman, F. (2000) in *Heterogeneous computing workshop* (IEEE, Cancun), pp. 349-363.
17. Gramacy, R. B., Lee, H. K. H. & Macready, W. G. (2004) in *International Conference on Machine Learning* (ACM, Banff), pp. 45-52.
18. Moles, C. G., Banga, J. R. & Keller, K. (2004) *Appl Soft Comput* **5,** 35-44.
19. Pope, V. D., Gallani, M. L., Rowntree, P. R. & Stratton, R. A. (2000) *Clim Dyn* **16,** 123-146.
20. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984) *Classification and regression trees* (Wadsworth International, Monterey).
21. R development core team (2004) (R foundation for statistical computing, Vienna).
22. Sall, J., Creighton, L. & Lehman, A. (2005) *JMP Start Statistics* (Thomson Learning, Belmont, CA).

Figure Legends

Fig1 Time series of temperature differences between control and doubled CO_2 phases for selected runs. Points are model outputs and smooth lines are fitted curves. The runs shown illustrate the range of observed CS: crosses CS=3.2K (median); diamonds CS=1.7K (2.5% quantile); pluses CS=9.5K (97.5% quantile). Also shown is a run (connected circles) not qualitatively different from the others, for which no curve could be fit.

Fig2 Regression tree for equilibrium climate sensitivity to a doubling of carbon dioxide (CS) as a function of parameter, hardware and software variation. The tree is read from top to bottom, starting with all model runs. At each split in the tree the model runs are divided into two groups based on the statement given (either an inequality, for continuous variables, or an equality, for discrete variables). If the statement is true for any given model run it passes to the left, if false to the right. The average CS for the subset of the model runs reaching that point is given below the split or tip. This tree is a subset of the optimal tree (fully defined in supporting Table3) and explains 64% of the variation in CS, where the optimal tree explains 80%. Hard/software explanatory variables were included in the creation of this tree and do appear in the optimal tree. However this subset only contains splits based on model parameters.

Fig3 Observed climate sensitivities for all 43710 model runs where it was calculable plotted against those predicted by the optimal regression tree on the basis of their parameter, hardware and software values (Fig2, Supporting Table3).

Fig4 Influence of variables in the trees. Each bar measures the percentage of the total variation explained by all splits based on that variable in one of the optimal trees. For each variable there are 4 bars: 1. black, tree of the magnitude of CS (Fig2); 2. dark grey, tree of failure to fit an adequate CS (Supporting Fig8); 3. light grey, tree of variation due to hard/software among otherwise identical runs (Fig6); 4. white, tree of variation among runs with identical parameters but different initial conditions (Supporting Fig10). Residual variation (unexplained by any of the parameters) is not shown but, estimated by cross-validation, is 18%, 67%, 73% and 66% respectively. Only parameters with at least 0.1% influence in at least one tree are shown.

Fig5 Frequency distributions for CS as calculated by taking the difference of average global mean temperature for the latter half of the control and doubled CO_2 phases. (*A*) For the 43677 model runs where a fitted CS as used for all other analyses was obtained. The relationship of these sensitivities to the fitted sensitivities is shown in supporting Fig9. (*B*) For the 13313 model runs where an adequate fitted sensitivity could not be obtained (26 outliers in *(B)* fall outside the range graphed).

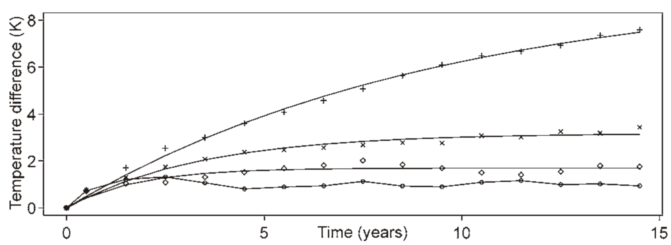
Fig6 Regression tree for percentage coefficient of variation (CV) among model runs with identical parameters and starting conditions. The tree is read from top to bottom in the same way as Fig2 starting with the 4712 parameter and starting condition sets where a CV could be calculated. Full details in supporting Table5.

Tables

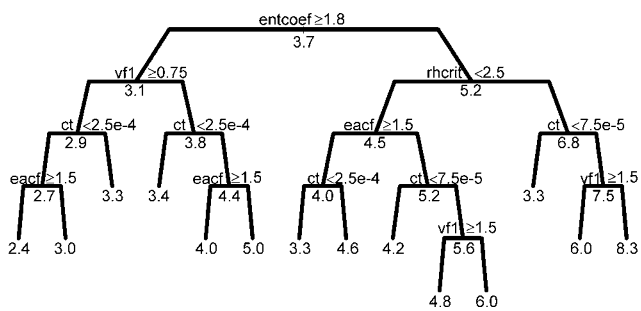
Table1 The explanatory variables used. Further details in supporting Table 2

Explanatory variable	**Meaning**
entcoef	entrainment coefficient
ct	Accretion constant
rhcrit	critical relative humidity
vf1	Ice fall speed through clouds
eacf	Empirically adjusted cloud fraction
cw	Threshold for precipitation
dtice	temperature range of ice albedo variation
ice	Non-spherical ice
midware	Client middleware
ice_size	Ice particle size
alpham	Albedo at melting point of ice

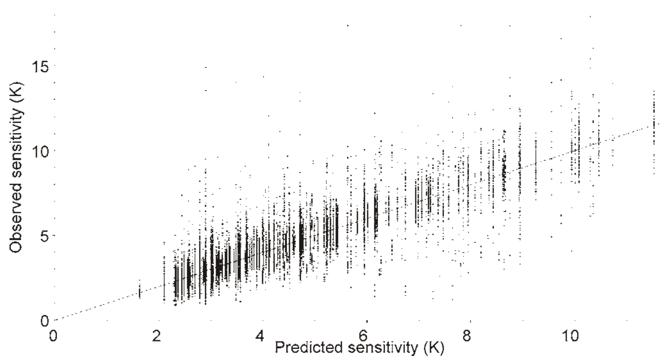
Explanatory variable	Meaning
processor_name	CPU Classification
clock_classic	Processor clock speed recorded under classic middleware
ram_size	Hardware RAM
clock_boinc_i	Integer processor clock speed recorded under BOINC middleware
clock_boinc_f	Floating point processor clock speed recorded under BOINC middleware
os_name	Operating system
dtheta	Perturbations to initial conditions on a given level



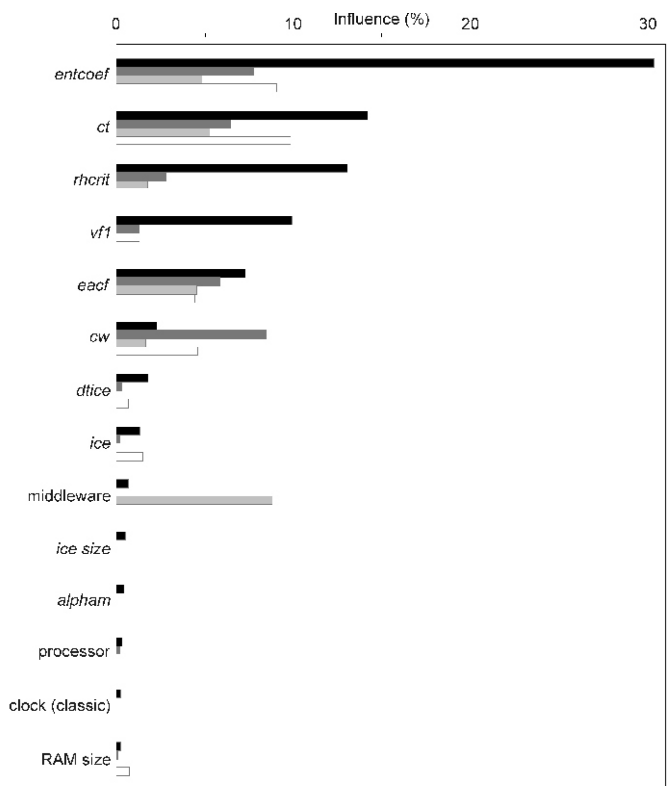
Temperature difference (K)
0
2
4
6
8
0
5
10
15
Time (years)



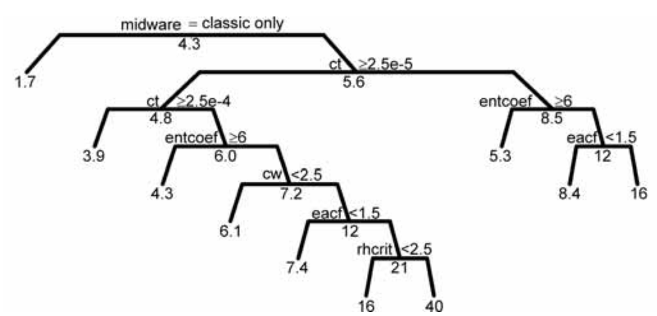
entcoef ≥1.8
3.7
vf1 ≥0.75
3.1
rhcrit <2.5
5.2
ct <2.5e-4
2.9
ct <2.5e-4
3.8
eacf ≥1.5
4.5
ct <7.5e-5
6.8
eacf ≥1.5
2.7
3.3
3.4
eacf ≥1.5
4.4
ct <2.5e-4
4.0
ct <7.5e-5
5.2
3.3
vf1 ≥1.5
7.5
2.4
3.0
4.0
5.0
3.3
4.6
4.2
vf1 ≥1.5
5.6
6.0
8.3
4.8
6.0

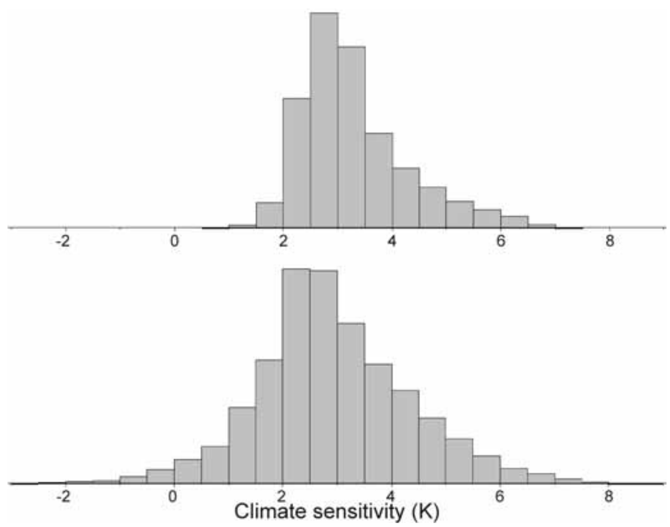


Observed sensitivity (K)
Predicted sensitivity (K)



Influence (%)
0
10
20
30
entcoef
ct
rhcrit
vf1
eacf
cw
dtice
ice
middleware
ice size
alpham
processor
clock (classic)
RAM size


midware = classic only
4.3
1.7
ct ≥2.5e-5
5.6
ct ≥2.5e-4
4.8
3.9
entcoef ≥6
6.0
4.3
cw <2.5
7.2
6.1
eacf <1.5
12
7.4
rhcrit <2.5
21
16
40
entcoef ≥6
8.5
5.3
eacf <1.5
12
8.4
16



-2
0
2
4
6
8
-2
0
2
4
6
8
Climate sensitivity (K)