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†, and Myles R. Allen†

*Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester M1 7DN, United Kingdom; †Atmospheric, Oceanic, and Planetary Physics, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, United Kingdom; §Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, United Kingdom; ¶The Met Office, FitzRoy Road, Exeter, Devon EX1 3PB, United Kingdom; and ‖Science and Technology Facilities Council, Rutherford Appleton Laboratory, Didcot, Oxon OX11 0QX, United Kingdom

Edited by Stephen H. Schneider, Stanford University, Stanford, CA, and approved May 21, 2007 (received for review September 15, 2006)

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 >57,000 climate model runs in which 10 parameters, initial conditions, hardware, and software used to run the model all have 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 caused by a small subset of parameters. Notably, the entrainment coefficient in clouds is associated with 30% of the variation seen in climate sensitivity, although both low and high values can give high climate sensitivity. We demonstrate that 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.

classification and regression trees | climate change | distributed computing | general circulation models | sensitivity analysis

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 typically is 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 necessarily will be uncertain. A limiting factor in such simulations is the availability of computational resources. Thus, combinations of plausible parameter values rarely are tested, leaving the dependence of conclusions on 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 also may be little true replicate data available. For instance, there can be only one observational time series for global climate. Thus, if the same observations are used to fit parameter values, there is a severe risk of overfitting, gaining limited verisimilitude at the cost of the mechanistic insight and predictive ability for which the model originally was 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) soon may 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 refinement presents challenges: (*i*) to determine whether there is dependence of model behavior of interest on parameter variation within plausible bounds, (*ii*) to determine whether dependence applies to all uncertain parameters or only a more tractable subset, and (*iii*) to quantify the nature of parameter dependence. Because parameters interact in complex and unknown ways, meeting these challenges entails considering a very large parameter space.

In this article 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 57,000 model runs in which 10 parameters and initial conditions were systematically varied. Although large studies traditionally have been carried out on supercomputers, it currently only is possible to perform this many simulations via a distributed computing approach. Before this project, the largest published comparable ensemble was of 53 model runs (8, 9). We have achieved such a large data set via the climate*prediction*.net project (www.climateprediction.net) by 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 hardware/software may multiply to influence overall model behavior. Because the GCM is highly nonlinear, even small quantitative differences in model behavior of this sort in principle could produce qualitatively different results. We address this issue directly, treating hardware/software variation equivalently to parameter variation.

Considering plausible values of six parameters and a smaller number of model runs, Stainforth *et al.* (4) demonstrated that, although accepted predictions of 2–5 K global warming in response to a doubling in carbon dioxide (11) indeed were representative of model results, equally plausible parameter values gave global

Author contributions: C.G.K., S.H.E.K., N.M., T.A., C.C., D.J.F., J.A.K., A.M., S.P., D.A.S., and M.R.A. designed research; C.G.K. and N.M. performed research; C.G.K. contributed new reagents/analytic tools; C.G.K., N.M., and B.S. analyzed data; and C.G.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviations: CS, climate sensitivity to a doubling of carbon dioxide; GCM, general circulation model; CV, coefficient of variation (standard deviation as a percentage of the mean).

‡To whom correspondence should be addressed. E-mail: sknight@atm.ox.ac.uk.

This article contains supporting information online at www.pnas.org/cgi/content/full/0608144104/DC1.

GEOPHYSICS

Table 1. The explanatory variables used in this study

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	Nonspherical ice
midware	Client middleware
ice_size	Ice particle size
alpham	Albedo at melting point of ice
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

Further details are in SI Table 2.

warming of >8 K. 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 distributed unevenly among uncertain parameters. It thus is 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 to a doubling of carbon dioxide (CS) 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 hardware/software effects is very similar to that of initial conditions effects.

Results

Associations with CS. The data set comprised 57,067 climate model runs. These runs sample parameter space for 10 parameters [Table 1 and supporting information (SI) Table 2] with between two and four levels of each, covering 12,487 parameter combinations (24% of possible combinations) and a range of initial conditions. A total of 43,692 runs (77%) were allocated an equilibrium sensitivity of global mean temperature to a doubling of CO_2 CS (see *Methods*). Among the 13,375 failures, 295 had incomplete data; 1,897 failed to fit a temperature change curve; 11,441 did have fits, but they did not meet arbitrary criteria for acceptable error; and 530 failed to progress far enough toward equilibrium in the modeled timescale. Representative time series and fitted curves are shown in Fig. 1.

To determine relative contributions of explanatory variables (Table 1), we fit a regression tree for CS (Fig. 2). 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 1.7 K, which is 45% of the mean CS [3.7 K, i.e., the runs have a coefficient of variation (standard deviation as a percentage of the mean) (CV) of 45%]. The first split in the tree divides runs into subsets with high and low values of the entrainment coefficient, *entcoef*, respectively. The resulting subsets each have smaller CVs, 30% and 43%. These

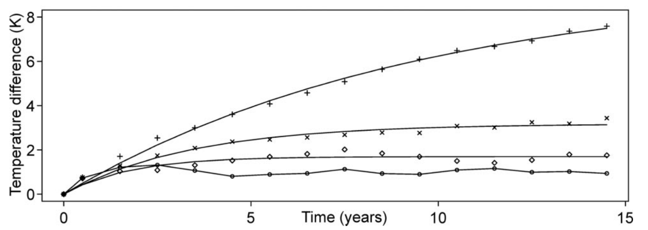

Fig. 1. Time series of temperature differences between control and doubled CO_2 phases for selected runs. Points are model outputs, and smooth lines are fitted curves. The runs shown illustrate the range of observed CS: ×, CS = 3.2 K (median); ◇, CS = 1.7 K (2.5% quantile); and +, CS = 9.5 K (97.5% quantile). Also shown is a run (○) not qualitatively different from the others, for which no curve could be fit.

subsets then can 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 (SI Table 3) contained 201 such splits based on parameter values and hardware and software explanatory variables. This tree explained a large majority of the variation in CS, 80% by cross-validation (see SI Fig. 7). Fig. 2 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 Fig. 3.

Summing over the optimal tree for each explanatory variable, the proportion of variation explained is shown in Fig. 4. Three-quarters of the total variation is explicable by just five 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, however, are all <1% of total variation and mostly <1% of the variation attributable to the more influential parameters. For example, the first split, described above, affects all of 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 1/4 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

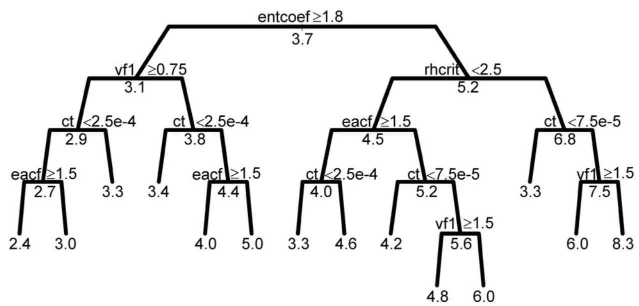

Fig. 2. Regression tree for equilibrium CS as a function of parameter, hardware, and software variation. The tree is read from top to bottom, starting with all model runs. At each split in the tree, the model runs are divided into two groups based on the statement given (either an inequality, for continuous variables, or an equality, for discrete variables). If the statement is true for any given model run, it passes to the left, if false, it passes 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 SI Table 3) and explains 64% of the variation in CS, whereas the optimal tree explains 80%. Hardware/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.

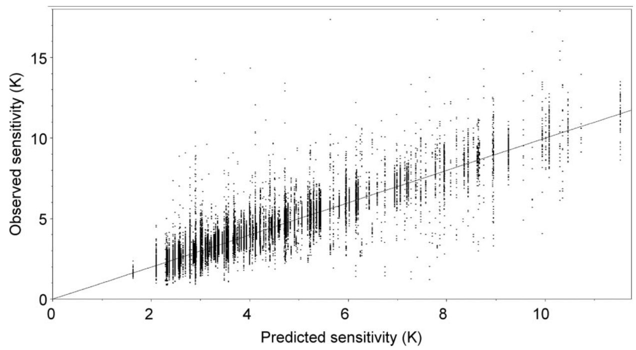

Fig. 3. Observed climate sensitivities for all 43,710 model runs where it was calculable plotted against those predicted by the optimal regression tree on the basis of their parameter, hardware, and software values (Fig. 2 and SI Table 3).

values behaving in unexpected ways or particular hardware/software giving spurious results. We used a similar tree-based approach fitting success or failure to the same explanatory variables as before (Table 1).

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 (SI Fig. 8 and SI Table 4). Although RAM size and the 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 (Fig. 4).

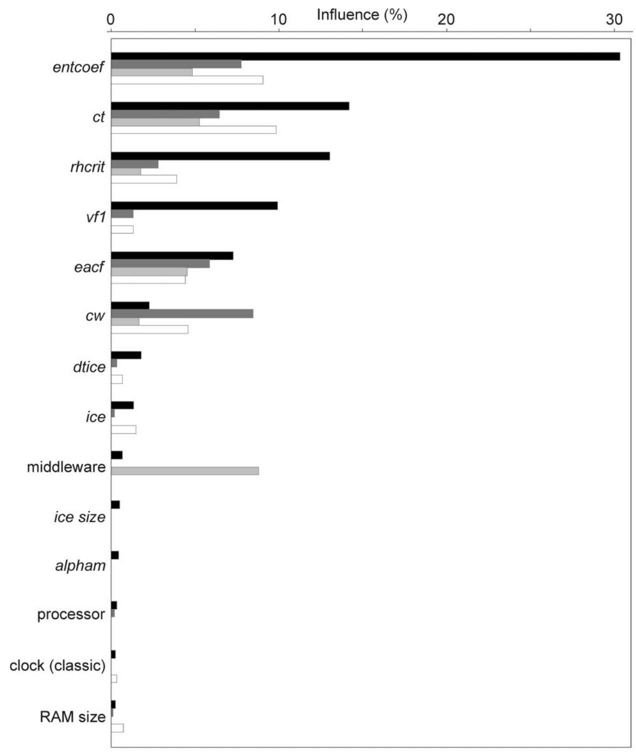

Fig. 4. Influence of variables in the trees. Each bar measures the percentage of the total variation explained by all splits based on that variable in one of the optimal trees. For each variable, there are four bars: 1, black, tree of the magnitude of CS (Fig. 2); 2, dark gray, tree of failure to fit an adequate CS (SI Fig. 8); 3, light gray, tree of variation attributable to hardware/software among otherwise identical runs (Fig. 6); and 4, white, tree of variation among runs with identical parameters but different initial conditions (SI Fig. 10). 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.

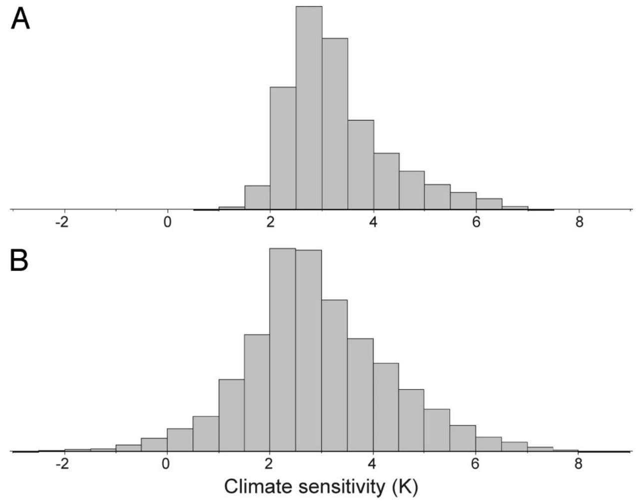

Fig. 5. Frequency distributions for CS as calculated by taking the difference of average global mean temperature for the latter half of the control and doubled CO_2 phases. (*A*) For the 43,677 model runs where a fitted CS as used for all other analyses was obtained. The relationship of these sensitivities to the fitted sensitivities is shown in SI Fig. 9. (*B*) For the 13,313 model runs where an adequate fitted sensitivity could not be obtained (26 outliers in *B* fall outside the range graphed).

Because there is a systematic element that depends on the parameter set in our failure to fit a CS, there also may 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 underestimates high CS, as compared with nonlinear fitting, but is a reasonable approximation at low values (SI Fig. 9). 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 (Fig. 5). The distributions are very similar in shape for sensitivities of $\approx$2.5 K and above with only a slight overrepresentation of high sensitivities in those not fitted. However, there is a tail of sensitivities $\leq$1 K that is missed almost entirely 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 (compare with the fact that 78% have fits in the rest of the data set). One example of a time series where no curve could be fit, but showing a 1-K sensitivity by the alternative measure, is shown in Fig. 1. 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^2 test), which also may 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. The number of combinations of parameters and initial conditions that had at least two and up to six runs giving a CS was 4,762 . Although many such "duplicate" sets (1,062 of them) gave identical results, most did not. For each parameter combination, we calculated the CV of the CS. We then fit a regression tree for this quantity in a similar way to earlier trees. An optimal tree (Fig. 6 and SI Table 5) explained

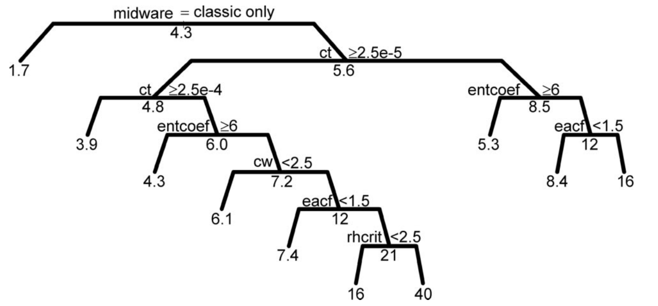

Fig. 6. Regression tree for percentage CV in CS among model runs with identical parameters and starting conditions. The tree is read from top to bottom in the same way as Fig. 2, starting with the 4,712 parameter and starting condition sets where a CV could be calculated. Full details are given in SI Table 5.

24% of variation by cross-validation; the relative effects of the explanatory variables are shown in Fig. 4.

The only hardware/software feature included in the duplicate divergence tree (Fig. 6) 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 initially used) and high levels with a mix of middlewares or BOINC middleware [developed for the SETI@home project (12) and subsequently used]. The five other explanatory variables in the tree are model parameters, precisely the same as the top five explanatory variables in the tree for failure to fit a sensitivity (Fig. 4 and SI Fig. 8).

We have characterized CS variation caused by differing hardware/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 hardware/software as equivalent to variation caused by different initial conditions. This approach would be valid only if the variation in CS from these sources is indistinguishable. The hypothesis that hardware/software affects CS indistinguishably from different initial conditions generates expectations:

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

Testing these expectations requires comparison of CS variation analyses for the two sources: hardware/software versus initial conditions. We already have an analysis of CS variation attributable to hardware/software (Fig. 6). Therefore, an equivalent analysis of CS variation attributable to initial conditions is required. We took the 8,196 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 Fig. 6 above. The optimal tree (SI Fig. 10 and SI Table 6) explains 32% of the total variation by cross-validation. To test the expectations, we compared predictions of this initial conditions tree to those of the hardware/software tree (Fig. 6) for the 4,709 models used to build both trees. Both trees are capable of predicting variation over a wide range, from CV $< 5\%$ to CV $\geq 40\%$. Expectation 1 predicts that CS variation in the initial conditions tree will be an upper bound for CS variation in the hardware/software tree. In 92% of models, the predicted CV is higher across initial conditions than across hardware/software differences, consistent with expectation 1. Expectation 2 predicts that 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, *entcoef*, higher values giving lower variation). Where the two trees first diverge, splitting on different parameters, in three of four 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 also is very similar between the two trees (Fig. 4, 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, $n = 4{,}709$, $P < 0.0001$). A more stringent similarity test asks whether 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, $n = 4{,}660$, $P < 0.0001$). From these tests, we conclude that both our expectations are met: hardware/software introduce 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 (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 on them is highly uneven. Over half of the variation associates with just three parameters (Fig. 4). 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 toward constraining the model by observations in general, or even constraining realistic values of parameters in general, but toward 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 split, the highest predicted CSs (>9 K) are all for low *entcoef* runs, associated with high *rhcrit* and *ct* and low *vf1* (Fig. 2 and SI Table 3), 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 (1.6 K in Fig. 3) are for low *entcoef* runs. The latter differ from other low *entcoef* runs in that they have high *rhcrit*, low *ct*, and high *eacf* and *cw* values (Fig. 2 and SI Table 3), i.e., the only consistent difference between the highest and lowest CS runs is *ct*, another cloud parameter. Thus, although 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. For example, if high *entcoef* best represents reality, this does not imply low real-world CS. In that case, focused studies would be required because, even in the current large ensemble, only 25 successful runs with high *entcoef* have the combinations of other parameters predicted to give CS >8 K.

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 by using a nonlinear fitting approach. This fitting assumes that 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 nonrandom subset with respect to CS. Specifically, a small tail of runs with low sensitivity cannot be assigned a CS (Fig. 5). In these cases, e.g., as shown in Fig. 1 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 "lost" low-sensitivity runs are not likely to affect estimates of real-world CS, both because they tend to agree poorly with observations (6, 7) and because we that find most display known nonphysical 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 pose risks here (14). Here, 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 or an AMD Athlon processor. However, there is no clear association, for example, that Intel chips give higher CS. Similarly, RAM size has an effect, but different model versions respond differently, in four of the six cases of splits based on RAM size, the smaller RAM size gives the higher sensitivity, but in two 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 data set, systematic hardware/software effects are reassuringly small relative to the effects of model parameters. Of the seven 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 ≈14,000 runs (33% of the total) each and accounting for >20 times as much variation as the processors (SI Table 4).

Important effects of hardware/software, however, may be less systematic. We identified a single software effect as important here. Runs with identical parameters and starting conditions average a 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 caused by different "controller" code that appeared more sensitive to small computational errors in the classic middleware. This sensitivity resulted in more crashes and thus failure to submit results for the classic middleware. BOINC was more likely to let the model run to completion despite computational problems. There also was 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 hardware/software and, unlike other hardware/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 GCMs 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 by using a conventional supercomputer (13). However, the challenge of uncertain or undefined parameters remains great. Even in this study, we only have been able to investigate 10 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 nonobvious 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 of 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 also will 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, for example, 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 also improved methods of exploring the complex parameter space. The latter requires a carefully designed experimental (15) and computing (16) strategy. It also may 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 also 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 hardware/software implementation of the model typically was small and indistinguishable from perturbations introduced by different initial conditions.

Methods

Model and Distributed Computing. The climate*prediction*.net project is the first multithousand member ensemble of climate simulations using a state-of-the-art 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 by using a heat-flux convergence field varying with position and season but not with 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 middle 15 years, the control phase, CO_2 is held constant at preindustrial levels (282 ppm). It is doubled for the last 15-year period.

Data Set. The first 57,067 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 Table 1 and SI Table 2.

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 CS, t is time, C is the effective heat capacity of the model, and F is the radiative forcing caused by a doubling of CO_2, taken to be 3.74 $W \cdot m^{-2}$. Fits that failed to converge after many iterations (1,000), gave a residual SE $>$0.2 K, 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 (Fig. 6 and SI Fig. 10), we created 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 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 SI Table 2).

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, in principle, can continue as long as there are multiple observations to be split and different levels of explanatory variables within subsets. However, although 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 cross-validation, i.e., splitting the data randomly into 100 equally sized subsets, with the 99th as a training set and the 100th as a test set. From the test set results, we calculated the error in prediction (cross-validation error) averaged over the 100 possible training and test sets. The optimal tree was chosen as the smallest where the cross-validation error lay within 1 SE of the minimum cross-validation error.

To identify unphysical cooling in the tropical East Pacific (4), surface temperature differences between the final year of the doubled CO_2 and calibration phases were taken for the 78.75 W, 2.5 N box and corrected for overall change by subtracting the figure for 48.75 W, 2.5 N in 13,983 BOINC runs. This quantity is distributed with the principal mode at 0 K and a secondary mode around -27.5 K; values less than -15 K 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 by using the rpart v.3.1–23 package.

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 the Natural Environment Research Council's Coupled Ocean Atmosphere Processes and European Climate, e-Science, and fellowship programs; Department of Trade and Industry; and Department for Environment, Food and Rural Affairs Grant PECD/7/12/37.

1. Karplus M, McCammon JA (2002) *Nat Struct Biol* 9:646–652.
2. Noble D (2002) *Science* 295:1678–1682.
3. Snoep JL (2005) *Curr Opin Biotechnol* 16:336–343.
4. Stainforth DA, Aina T, Christensen C, Collins M, Faull N, Frame DJ, Kettleborough JA, Knight S, Martin A, Murphy JM, *et al*. (2005) *Nature* 433:403–406.
5. Frame DJ, Booth BBB, Kettleborough JA, Stainforth DA, Gregory JM, Collins M, Allen MR (2005) *Geophys Res Lett* 32:L09702.
6. Knutti R, Meehl GA, Allen MR, Stainforth DA (2006) *J Climate* 19:4224–4233.
7. Piani C, Frame DJ, Stainforth DA, Allen MR (2005) *Geophys Res Lett* 32:L23825.
8. Murphy JM, Sexton DM, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) *Nature* 430:768–772.
9. Barnett DN, Brown SJ, Murphy JM, Sexton DMH, Webb MJ (2006) *Clim Dyn* 26:489–511.
10. He Y, Ding CHQ (2001) *J Supercomput* 18:259–277.
11. Houghton JT, ed (2001) *Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ Press, Cambridge, UK).
12. Anderson DP (2004) in *Fifth IEEE/ACM International Workshop on Grid Computing (GRID '04)* (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-Renaud C, Playez N (2003) in *Proceedings of the 17th Annual International Conference on Supercomputing* (ACM Press, New York), pp 226–233.
15. Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) *Statistical Sci* 4:409–435.
16. Casanova H, Zagorodnov D, Berman F, Legrand A (2000) in *Proceedings of the Ninth Heterogeneous Computing Workshop* (IEEE Computer Society, Washington, DC), pp 349–363.
17. Gramacy RB, Lee HKH, Macready WG (2004) in *Proceedings of the 21st International Conference on Machine Learning* (ACM Press, New York), Vol 69, pp 45–52.
18. Moles CG, Banga JR, Keller K (2004) *Appl Soft Comput* 5:35–44.
19. Pope VD, Gallani ML, Rowntree PR, Stratton RA (2000) *Clim Dyn* 16:123–146.
20. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) *Classification and Regression Trees* (Wadsworth International, Monterey, CA).
21. R Development Core Team (2004) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).
22. Sall J, Creighton L, Lehman A (2005) *JMP Start Statistics* (Thomson Learning, Belmont, CA).