Supporting information for:

Considerations for parameter optimization and sensitivity in climate models.

J. David Neelin, Annalisa Bracco, Hao Luo, James C. McWilliams and Joyce E. Meyerson

1 On experimental design

The mixed-layer simulations use a 50-meter, fixed-depth mixed layer ocean, with atmospheric infrared transmissivity based on 280 ppm CO_2 for the preindustrial run, and 560 ppm for the doubled-CO_2 run. The imposed-SST simulations have an infrared transmissivity approximating present-day CO_2 values.

A number of strategies for high-dimensional optimization problems with computationally-expensive black-box functions (1; 2; 3; 5) may be suitable for climate modeling, so we briefly expand on choices made in the present work and related possibilities.

There is a large literature on variations of experimental design for approximating experimental results with fitted functions, termed response surfaces, for problems with relatively few design variables (6) and relatively low dimensional output quantities. Because climate models must deal, at least in initial exploration stages, with a larger number of parameters, the scaling of the experiment design with the number of parameters N is significant. A number of response surface strategies, designed to have properties such as protection against aliasing from higher order terms, have poor scaling properties (3). The minimum number of climate model runs required to fit the quadratic in Eq. **2** of the main text is $2N + N(N-1)/2$ in addition to the standard case which will commonly act like a center point in the experiment design. In our results, the standard error for simulation ensembles does not change strongly as a function of parameter (see, e.g., the spread of the ensemble in Fig. 1d of the main text). Thus a larger ensemble or longer run at the standard case can be used to establish the minimum ensemble size and simulation length to be used for a desired level of accuracy in the random error of estimation at each of the other points in parameter-space. For brevity, consider the case where for normalized parameter axes $\mu_{i\,\min}$ lies at -1, the standard case at 0 and $\mu_{i\,\max}$ at 1, although moderately uneven axes present no problem.

The most obvious way of carrying out the minimum set of runs, known as a minimum-point or saturated design, is a Koshal design for fitting a second-order model. Axial runs at three levels (-1,0,1), also known as star points, yield a_i, b_{ii} while b_{ij} require one point per plane of parameter combinations, for instance $(\mu_i, \mu_j) = (1, 1)$, with all other parameters 0. It is obviously desirable to have additional points for improved estimation, for testing fit, etc., referred to as an unsaturated design. A comparison in (7) finds desirable properties for a Koshal design augmented by 20-40% additional random points. Other, more complex, near-saturated designs exist (6). The results presented here with off-diagonal b_{ij} use RMS fits to values from all four corners of each pairwise parameter plane (where one is used in the Koshal design). For $N = 3, 4$ or 5 this corresponds to a Box-Behnken (8) design if the axial points are not used.

We use verification runs along lines in parameter space for ease of displaying the accuracy of the fit graphically, a nontrivial consideration at this exploration phase for climate models. However, it also corresponds to initial stages of a particular high-dimensional model representation (HDMR) algorithm known as cut-HDMR (3) which could prove useful for altering the quadratic metamodel if a particular subset of directions proved more highly nonlinear.

Because obtaining the off-diagonal quadratic coefficients is of order N^2 even for a minimum-point procedure, a screening phase is important. That is, computation of the off-diagonal b_{ij} should in practice be done for a subset N' of directions that a first screening step indicates to be significant in both sensitivity and nonlinearity. The form of the problem here differs from many response surface applications in which a scalar or low-dimensional vector is being fit, or from cases in which the objective function is being fit directly, in that the objective functions reconstructed from the high dimensional vector fields have quadratic terms arising from the linear terms of the metamodel. For many combinations of parameter direction and climate variable, the parameter dependence may be approximated as linear, and the concern is to guard against modifications of $a_i a_j$ terms by b_{ij} terms, as opposed to fundamentally depending on quadratic fit terms for the optimization. For the changes under global warming, the concern is whether quadratic terms alter the predicted change substantially in any particular region or variable that humans care about. In testing for nonlinearity, the criterion for conducting additional simulations should be based on the most non-linear variable to be assessed with the output.

The fact that the $b_{ij}\mu_i\mu_j$, $i \neq j$ term integrates to zero across a symmetric parameter domain has implications for the question of whether nonlinearity affects multimodel ensemble averages, and thus for aspects of the design of Sensitivity Model Intercomparison Project (SensMIP). Ideally, in averaging an ensemble distributed across the parameter domain D, one would like to assign a probability density function as a function of parameter, $p(\mu)$, so the estimated average response could be evaluated as $\int_D p(\mu)\Delta\tilde{\phi}d\mu$, where $\Delta\tilde{\phi}$ is the metamodel of the change under global warming, expanded as in Eq. **2** of the main

text. Assigning p is challenging; if one is forced, for lack of more specific information, to assume that each parameter value in a symmetric feasible range has an equal probability of being correct, i.e., uniform p, then the contributions associated with $b_{ij}\mu_i\mu_j,\ i \neq j$ integrate to zero. The same applies for any p symmetric in μ. It is thus the $b_{ii}\mu_i^2$ terms (provided by an order N procedure) that provide the first estimate of how nonlinearity will affect the multi-model ensemble mean.

Also worth considering is that the results of the metamodeling procedure may be to focus attention on possible revisions of a particular parameterization, so information about changing a single parameter can aid interpretation. It may often be the case that climate modeling groups already have runs that change a single parameter at a time. Thus while some classic screening strategies focus on runs at a fraction of the corners of a hypercube in parameter space, an appealing simple screening strategy is to use the standard case plus two axial points in each parameter direction. These $2N$ simulations, potentially using shorter simulations or smaller ensembles can thus be used to eliminate variables from the b_{ij} estimation if a_i, b_{ii} are small.

Carrying out optimization on the resulting metamodel, the fit with or without off-diagonal b_{ij} can be updated by least-squares estimates of the coefficients as subsequent ensembles of simulations are performed at points chosen based on the optima of the initial fit. Because these points are close to optima, they tend to improve the fit in regions of parameter space that are likely to be important. If the quadratic metamodel is insufficient to provide a good fit over the entire domain, limiting to a smaller trust region is an established strategy.

2 Computational cost and feasibility of a SensMIP

For estimation of the quadratic metamodel, there are two major factors in the computational cost. First, the $O(N^2)$ problem for the number of coefficients to be evaluated, corresponding to the minimum number of parameter points at which evaluations are required, where N is the number of parameters. This is vastly less costly at large N than the $O(s^N)$ cost that would be incurred by brute-force sampling at density s, and can be somewhat reduced by screening procedures, but is still substantial. Second, there is a factor associated with reducing estimation error associated with natural climate variability. For estimation of climate means at each of a given set of parameter points, the required ensemble size or number of years of simulation (for timescales longer than typical autocorrelation times of the variable of interest) scales as $O(\epsilon^{-2})$ where ϵ measures the desired standard error. The number of years required to achieve a given level of accuracy will depend on the climate variable of interest, and can be estimated from the standard case simulation.

Some back-of-the-envelope examples provide a sense of the size of the computational problem. For $N = 30$, 20, 10, 5 the minimum points in addition to the standard case are $2N + N(N-1)/2$=60+435, 40+190, 20+45, 10+10. For reference, a brute-force sampling with three levels, scaling as 3^N, would yield roughly 2×10^{14}, 3×10^9, 6×10^4 and 243 points, respectively. For the example of 10 parameters, the minimum of 65 points would be augmented by verification points by at least 40%, but even if doubled or quadrupled, the number of points remains of order 10^2, and can likely be reduced by screening. The substantial reduction relative to the order 10^4 points used in (4) potentially brings such computations from the realm of the extraordinary to one that can be done by multiple groups. The computational savings does depend on the conjecture that many parameter directions can be usefully represented by quadratic nonlinearity for climate variables of interest. If worse-than-quadratic nonlinearity is encountered in a subset of directions, computational resources can be focused on these; conversely, a subset of linear directions identified in screening can decrease the cost.

Consider a scenario that begins with screening 30 parameters, of which a plausible subset exhibits quadratic nonlinearity. If one can accept five times the standard error seen in Figs. 1 and 2 in the main text, 10 year runs might be used for screening. This would imply 600 simulated years for $2N$ screening runs in the 30 parameter case (plus a comparable length of spin-up toward equilibrium in coupled cases). Assuming 10 parameters prove important, an additional 450 years would be required for a first estimate of 45 off-diagonal b_{ij}. Prior to estimating these, a metamodel based on the $2N$ points can indicate which corners are likely to be closest to optima. If we further suppose that of the 10 key parameters, 5 directions have significant nonlinear interactions with each other, then in addition to 20 axial points for the 10 directions, there will be 10 points for b_{ij} for which longer runs are desired. If one accepts double the standard error seen here (assuming similar model properties and an interest in similar precipitation variables), an additional 50 years of run is required at each of these 30 points (spin-up does not have to be repeated), i.e., 1500 years of simulation, plus a certain percentage of verification points. The implied resources are substantial but feasible for a common class of climate model. Furthermore, the arguments in the previous section suggest that some estimates of nonlinearity useful for assessing multi-model averages can be obtained from just the axial points, which may assist prioritization of resources. Additional accuracy is nontrivial; for instance roughly 13,000 simulated years were used for 4 parameters for the imposed-SST runs here to assess the apparent smoothness. Fortunately, there is likely to be enough similarity among models that once a certain properties are established at high accuracy for some cases, less

costly estimates can be used routinely.

One can compare a design with relatively long simulations at a limited number of points to a space-filling design such as a Latin hypercube with many more points, but in which computational limitations imply much shorter runs. Designs with more points can help to guard against systematic error, or to test metamodel assumptions if random error is small. In the application here, computational limitations will tend to imply larger random error at each point as more points are used, making it difficult to evaluate whether the underlying surface is approximately smooth. Nonetheless, if the quadratic metamodel is a reasonably good representation of the parameter dependence, similar estimates should result from various designs. The key point is not a particular design, but practicality of estimating quadratic metamodels. The metamodel approach can be used to compare parameter space dependence, even for modeling centers that have used different approaches to the placement of simulations in parameter space.

Different climate models do not have identical parameters, although they each have parameters that affect similar physical processes, so the model intercomparison questions need to be phrased accordingly. While here we address only feasibility, rather than precise set up of a sensitivity model intercomparison, examples of targets include the following. In the main text, questions regarding regional climate change patterns were raised, focusing on the example of precipitation change. A systematic sensitivity study as a function of parameter within each model permits the questions of the degree of nonlinearity of the parameter dependence to be evaluated, as opposed to the ensemble-of-opportunity afforded by a set of different models whose parameter relationship is not known. An assessment of which climate variables exhibit strong nonlinearity in response to parameters affecting particular sets of physical processes, such as the representation of moist convection affecting precipitation distributions, would permit more accurate statements regarding multi-model ensemble averages. This, with an assessment of the degree to which the relative smoothness of the parameter dependence seen in the main text holds among climate models, would permit caveats on the use of quadratic metamodels to be established. Each model's feasible parameter domain establishes an uncertainty estimate on regional climate change response, and a comparison of these ranges among models is of interest. Assuming the metamodeling approach appears sufficiently accurate when applied to the parameter dependence of climatological variables, as well as to the parameter dependence of changes under global warming, the optimization procedures applied to the metamodel under normal climate might assist in constraining the parameter range (or probability density function) to be used in forming best estimate averages of climate change. In addition to regional climate change questions, the long-standing question of global-average climate sensitivity (4; 9) could be addressed from the same runs.

3 On the analytic solutions

In the main text Eq. **2** is written with a summation convention that corresponds to the matrix form

$$\tilde{\phi} = \phi_{\rm std} + \mu^T a + \mu^T \mathbf{b} \mu \qquad (1)$$

where $\mu = (\mu_1, ...\mu_N)^T$, $a = (a_1, ...a_N)^T$ are vectors over the parameter set and $\mathbf{b}$ is an N by N matrix with elements b_{ij}. Is worth reiterating that each element a_i, b_{ij} is itself a high dimensional vector over space and season. This notation gives the quadratic term in the main text Eq. **2** as $\sum_{i=1}^{N}\sum_{j=1}^{N} b_{ij}\mu_i\mu_j$ with $b_{ij} = b_{ji}$ rather than the more traditional sum $\sum_{i=1}^{N}\sum_{j=i}^{N} b'_{ij}\mu_i\mu_j$ which corresponds to the actual number of degrees of freedom, with $b'_{ij} = 2b_{ji},\ i \neq j$.

In this notation, the derivation of the gradient of a square-error objective function is just $\nabla_\mu f = \nabla_\mu f \langle(\tilde{\phi} - \phi_{\rm obs})^2\rangle = \langle 2(\tilde{\phi} - \phi_{\rm obs})(a + 2\mathbf{b}\mu)\rangle = \langle 2[\phi_{\rm err} + a^T\mu + \mathcal{O}(\mu^2)](a + 2\mathbf{b}\mu)\rangle$ which to first order in μ gives Eq. **3**-**5** of the main text. If the notation with b'_{ij} were used, terms in b'_{ii} would appear with differing factors of 2 than terms in $b'_{ij},\ i \neq j$.

When the Hessian has a negative eigenvalue, or the positive curvature is too weak, boundary solutions occur. When it is known which parameter boundary the solution lies on (determined numerically, from inspection of the gradient or from the eigenvector), the modification of the analytic solution comes from replacing $\mu_N = \mu_{N\,\max}$ in the derivation above—where for definiteness the relevant parameter has been defined to be μ_N and the boundary solution assumed to occur at the positive end of its range. Thus $\nabla_\mu f = \langle 2[(\phi_{\rm err} + a_N\mu_{N\,\max}) + a^T\mu + \mathcal{O}(\mu^2)][a + 2(b_{iN}\mu_{N\,\max}) + 2\mathbf{b}\mu)\rangle$ where μ, a and $\mathbf{b}$ have all been reduced from dimension N to $N-1$ by dropping elements associated with μ_N.

Thus the solution for interior points in the $N-1$ other dimensions is given by

$$\begin{aligned} \nabla_\mu f &= g + A\mu = 0, & (2)\\ g_i &= 2\langle a_i\phi_{\rm err} + a_i\phi'_{\rm err} + a'_i\phi_{\rm err}\rangle & (3)\\ A_{ij} &= A_{ji} = 2(\langle a_i a_j\rangle + 2\langle b_{ij}\phi_{\rm err}\rangle), & (4) \end{aligned}$$

where $\phi'_{\rm err} = a_N\mu_{N\,\max}$ is the correction to the error by linear changes in $\tilde{\phi}$ due to μ_N being at its boundary value and $a'_i = 2b_{iN}\mu_{N\,\max}$ comes from similar corrections to the gradient. Changes proportional to $\mu_{N\,\max}$ in Hessian terms are of order μ^2 in the gradient and thus are neglected, so the elements of the Hessian remain the same as in Eq. **5** of the main text. The Hessian no longer has a negative eigenvalue due to omission of the associated

parameter direction (with the caveat that the eigenvector must be closely enough aligned with one parameter direction that the boundary value is known). Extensions to more than one boundary value point should be obvious. The analytic solution is thus useful even in presence of a direction of negative curvature, such as occurs for the convective relative humidity parameter.

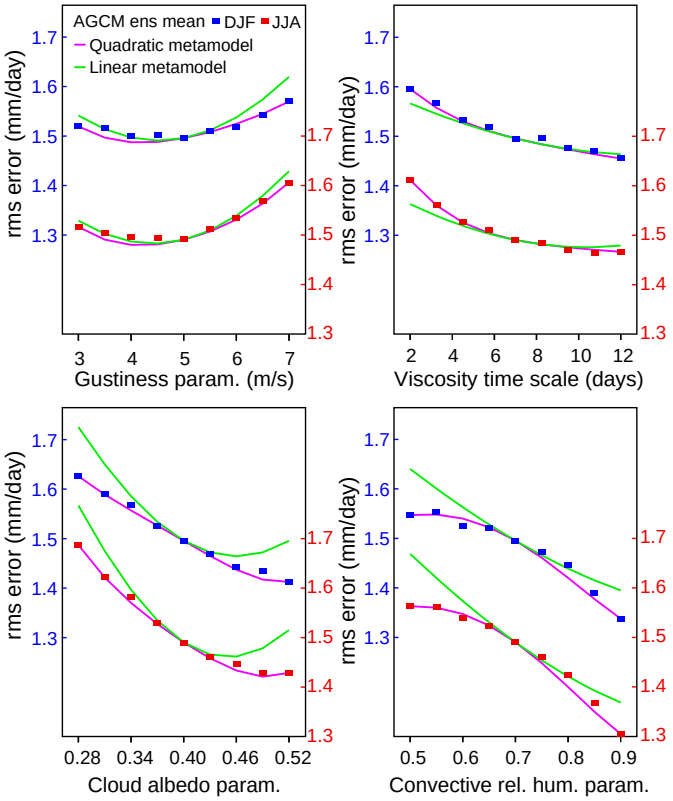

Figure 1: Root-mean-square (RMS) error relative to NCEP reanalysis of the ensemble mean AGCM precipitation similar to Fig. 1 of the main text, but for departures from annual average and comparing June-Aug. (JJA) to Dec.-Feb. (DJF). For clarity, the ordinate is shifted, with DJF values given by the left y-axis, JJA values by the right y-axis. The RMS error reconstructed from the quadratic metamodel and from its linear counterpart are shown for each. Parameters are given on the abscissa. The vertical size of the symbols gives the two standard error estimation range for the ensemble mean.

4 Seasonal dependence

All the quantities of interest are evaluated as a function of season. To emphasize this, a particular season (June-August) is used in the main text. Although there is no guarantee that parameter dependence should be similar among seasons, Fig. SI-1 shows that many of the important features tend to be reproduced. RMS error based on the departure of the precipitation field from its annual average is shown to make it clear that the similarity occurs within the parameter dependence of the seasonally varying signal. In other words, the similarity of the seasonal curves is not due simply to the contribution of the annual average, since this is removed. Overall, there is also a similarity in the curvature properties between the seasonal cycle shown here, and the total precipitation RMS error seen in Fig. 1 of the main text. For instance, the negative curvature for the convective relative humidity parameter is robust in these different measures. Indeed the magnitude of the quadratic terms and the sign of their contribution to the curvature is the same for both seasons in each parameter. When the seasonal dependence is qualitatively consistent as seen here, objective functions that sum over seasons (in addition to spatial averaging) are reasonable. Within a metamodeling, multi-objective framework the additional information that comes from seasonal objective functions has little cost.

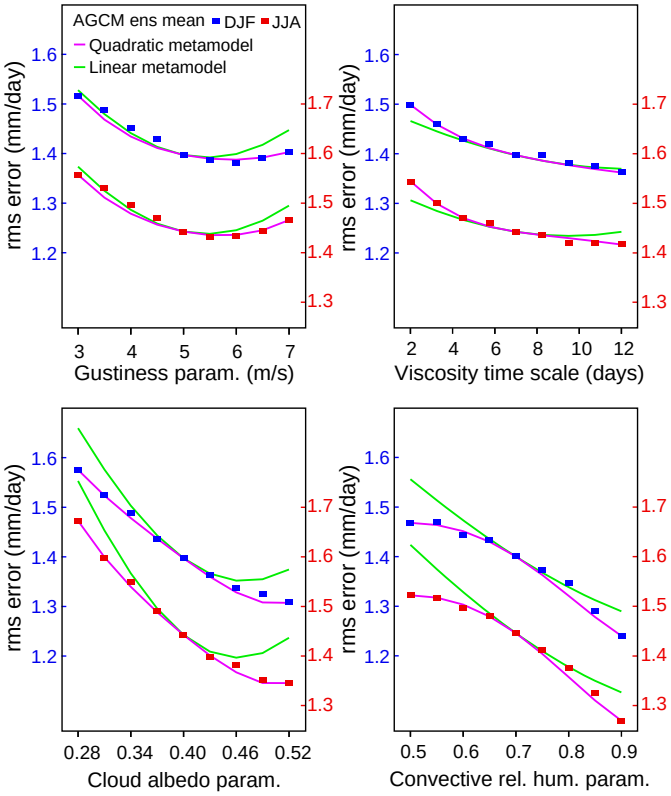

Figure 2: As in Fig. SI-1 but for RMS error relative to the CMAP precipitation data set (for departures from annual average for each season).

Fig. SI-2 evaluates the dependence on the data set used as the observed reference. In the main text, all verification quantities were taken from the NCEP reanalysis data set

for consistency among climate variables. Observational estimates can differ substantially, especially for precipitation, so we test also against the Climate Prediction Center Merged Analysis of Precipitation (CMAP) (10). Despite regional differences among the two data sets, the properties such as curvature of the RMS objective function, direction of slope and in most cases approximate location of the minima on parameter axes are similar in comparing Fig. SI-1 and Fig. SI-2. In the gustiness parameter, despite similarity of shape of the objective function, a shift in the minimum occurs that would be significant for quantitative optimization. Inclusion of objective functions evaluated with respect to different data sets is straightforward in a metamodeling framework. Overall properties including the relative smoothness in parameter space appear to be robust.

References

[1] Jones, D. R, Schonlau, M, & Welch, W. J. (1998) *J. Global Optimization* **13**, 452–492.

[2] Simpson, T. W, Peplinski, J. D, Koch, P. N, & Allen, J. K. (2001) *Engineering with Computers* **17**, 129–150.

[3] Shan, S & Wang, G. G. (2010) *Struct. Multidisc. Optim.* **41**, 219–241.

[4] Knight, C. G, Knight, S. H. E, Massey, N, & co-authors. (2007) *Proc. Nat. Acad. Sci.* **104**, 12259–12264.

[5] Wang, G. G & Shan, S. (2007) *J. Mechanical Design* **129**, 370–380.

[6] Myers, R. H, Montgomery, D. C, & Anderson-Cook, C. M. (2009) *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. (Wiley, New York), 3rd edition, p. 677.

[7] Carpenter, W. (1993) Effect of design selection on response surface performance, (NASA Langley Research Center, Hampton, VA), NASA Contract Report 4520. 174 pp.

[8] Box, G. E. P & Behnken, D. W. (1960) *Technometrics* **2**, 455475.

[9] Meehl, G., T. F. Stocker, W. Collins, and coauthors, Global climate projections, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon, D. Qin, and M. Manning, p. 996, Cambridge University Press, Cambridge and New York, 2007.

[10] Xie, P & Arkin, P. A. (1997) *Bull. Amer. Meteorol. Soc.* **78**, 2539–2558.