Multiobjective constraints for climate model parameter choices: Pragmatic Pareto fronts in CESM1

B. Langenbrunner[1]* , **J. D. Neelin**[1]

[1]Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, Los Angeles, USA

Key Points:

- Global climate model parameter tuning suffers from tradeoffs where one error metric or objective function is improved while others degrade
- An evolutionary algorithm is used to find Pareto fronts, or surfaces in objective function space on which model performance tradeoffs occur
- Pareto fronts allow the modeler to quickly visualize these tradeoffs, identify physics at play, and choose Pareto-optimal parameter updates

*Current address, Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, 7127 Math Sciences Bldg., 405 Hilgard Ave., Los Angeles, CA 90095.

Corresponding author: B. Langenbrunner, baird@atmos.ucla.edu

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1002/2017MS000942

Abstract

Global climate models (GCMs) are examples of high-dimensional input-output systems, where model output is a function of many variables, and an update in model physics commonly improves performance in one objective function (i.e., measure of model performance) at the expense of degrading another. Here, concepts from multiobjective optimization in the engineering literature are used to investigate parameter sensitivity and optimization in the face of such tradeoffs. A metamodeling technique called cut high-dimensional model representation (cut-HDMR) is leveraged in the context of multiobjective optimization to improve GCM simulation of the tropical Pacific climate, focusing on seasonal precipitation, column water vapor, and skin temperature. An evolutionary algorithm is used to solve for Pareto fronts, which are surfaces in objective function space along which tradeoffs in GCM performance occur. This approach allows the modeler to visualize tradeoffs quickly and identify the physics at play. In some cases, Pareto fronts are small, implying that tradeoffs are minimal, optimal parameter value choices are more straightforward, and the GCM is well-functioning. In all cases considered here, the control run was found not to be Pareto-optimal (i.e., not on the front), highlighting an opportunity for model improvement through objectively informed parameter selection. Taylor diagrams illustrate that these improvements occur primarily in field magnitude, not spatial correlation, and they show that specific parameter updates can improve fields fundamental to tropical moist processes—namely precipitation and skin temperature—without significantly impacting others. These results provide an example of how basic elements of multiobjective optimization can facilitate pragmatic GCM tuning processes.

1 Introduction

Uncertainties noted in present-day global climate model (GCM) simulations are complex, region-dependent, and occur across a broad range of time scales. GCMs must correctly simulate coupling among the land, ocean, and atmosphere as well as the interplay between large- and small-scale dynamics, which themselves rely heavily on sub-grid scale physics and their parameterizations. This study focuses on the tropical Pacific climate at the seasonal time scale, where uncertainty is largely due to under-constrained moist processes. An inexhaustive list of GCM issues in this region includes excessive precipitation in the Southern Hemisphere and the double inter-tropical convergence zone [ITCZ; *Dai*, 2006; *Lin*, 2007], issues with dynamics related to the El Niño-Southern Oscillation [ENSO; *Latif et al.*, 2001] and the South Pacific Convergence Zone [SPCZ; *Brown et al.*, 2010; *Lintner et al.*, 2016], sea surface temper-

ature biases leading to the excessive equatorial cold tongue [*Li and Xie*, 2014], issues in simulating the three-dimensional structure of moisture and temperature in the atmosphere [*Tian et al.*, 2013], persistent errors representing clouds and microphysics [*Bony and Dufresne*, 2005], and uncertainty related to land-sea contrasts and representation of topography, particularly over the Amazon [*Yin et al.*, 2013].

These uncertainties are each present to an extent in the National Corporation for Atmospheric Research (NCAR) Community Earth System Model version 1 [CESM1; *Kay et al.*, 2012; *Gettelman et al.*, 2012a,b; *Neale et al.*, 2013]. Additional issues have been noted in the ability of CESM to represent tropical Pacific dynamics at interannual time scales, specifically the frequency and seasonal timing of ENSO events and other modes of variability [*Deser et al.*, 2012; *Capotondi*, 2013], as well as its ability to simulate tropical wave dynamics associated with the Madden-Julian Oscillation [MJO; *Boyle et al.*, 2015]. Although some aspects of the physics at fast time scales relevant to convective processes are reasonably well simulated—including the pickup of deep convective precipitation and how this depends on column-averaged temperature and column water vapor [*Sahany et al.*, 2012, 2014; *Kuo and Neelin*, 2016]—disparities in convective measures between CESM and observations have also been noted [e.g., *Zhang and Chen*, 2015]. Poorly represented microphysics is also a persistent issue affecting CESM and leads to errors in radiation statistics, cloud cover, and feedbacks [*Gettelman et al.*, 2012b; *Trenberth et al.*, 2015; *Zheng et al.*, 2016].

In the process of improving GCMs through parameter tuning and calibration, a common phenomenon can occur where updating a parameter value can cause one metric of model performance to improve while another gets worse. Perturbed physics ensembles, in which a single model is integrated across a range of parameter values, allow the modeler to isolate these parameter uncertainties in a single climate model [*Allen and Stainforth*, 2002; *Murphy et al.*, 2004; *Stainforth et al.*, 2005; *Collins et al.*, 2006]. Recent studies have used perturbed physics ensembles to identify sensitive parameters in climate models [*Guo et al.*, 2014, 2015; *Boyle et al.*, 2015; *Qian et al.*, 2015], determine optimal parameter ranges [*Jackson et al.*, 2004, 2008; *Annan et al.*, 2005; *Severijns and Hazeleger*, 2005], and explore nonlinearity in parameter sensitivity using polynomial-based metamodels [*Neelin et al.*, 2010; *Bellprat et al.*, 2012; *Bracco et al.*, 2013] or more involved techniques like artificial neural networks [*Sanderson et al.*, 2008] or Bayesian inference strategies [*Sacks et al.*, 1989; *Rougier*, 2007; *Rougier et al.*, 2009; *Lee et al.*, 2011, 2012]. The majority of these studies is done in the context of model calibration or parameter optimization, and multiobjective methods offer a powerful approach for constrain-

ing high-dimensional parameter space [e.g., *Price et al.*, 2009]. The utility of multiobjective techniques, however, has not been exploited in the context of fully-coupled atmosphere-ocean GCMs, nor have they been combined with metamodels or model emulators.

In this paper, we use a perturbed physics ensemble to train reduced-complexity models that reconstruct the parameter space and sensitivity of the deep convection scheme within a fully-coupled GCM. This technique is termed metamodeling, model emulation, or surrogate modeling, and the primary method employed here is cut high-dimensional model representation (cut-HDMR), a technique adapted from the engineering literature. The output from cut-HDMR is then used within multiobjective optimization methods that quantify parameter-based tradeoffs in model performance and facilitate selections for parameter value updates. Central to our methodology is the Pareto front or frontier, which is a surface in objective function space along which model performance in one metric or objective function cannot improve without degrading another, and it is used to characterize the tradeoffs encountered in a GCM. By considering multiple objective functions simultaneously, one can better understand the tradeoffs involved in GCM parameter updates as well as calibrate or optimize parameter choices in an objective way.

In the remainder of this manuscript, we leverage a perturbed physics ensemble described in *Bernstein and Neelin* [2016] to build simple metamodels and use their output for multiobjective parameter optimization. Following steps that a modeler might take in the tuning phase of GCM development, we focus on the tropical Pacific climatology of precipitation, column water vapor, and skin temperature. Sections 2 and 3 contain a description of the data, methods, and concepts used in this paper. Section 4 demonstrates the use of metamodels to reconstruct the parameter space of the the GCM used here. We then pivot to the primary goal of this paper in Section 5, where we use concepts from multiobjective optimization to visualize tradeoffs in GCM performance and identify parameter values that optimize a set of observed metrics. Section 6 examines these tradeoffs for precipitation over the Amazon and over the tropics. These results are shown on Taylor diagrams in section 7, and parameter updates are explored across additional fields in section 8. Summary and conclusions are contained in section 9.

2 Data

Unless otherwise noted, climatological fields for models, observations, and reanalysis products are analyzed on a domain that encompasses the tropical Pacific, including all land and ocean grid points between 40°S–40°N, 120°E–300°E.

2.1 Perturbed physics ensemble setup

Many parameter space sampling strategies exist, though some are more naturally associated with particular metamodeling approaches. We choose a perturbed physics ensemble that aims at estimating nonlinearity first, for one parameter at a time, and then for pairwise parameter combinations. The first-pass estimate uses on-axis runs—where "axis" refers to a single axis in parameter space—because it allows the modeler to build intuition about the magnitude and nonlinearity of the parameter sensitivity to each parameter acting alone, and the computational costs associated with this step are order-N (where N is the number of parameters sampled). Parameter interactions are then estimated with pairwise parameter perturbation runs that are informed by the first pass. The organization of this approach leverages the construction of cut-HDMR, which orders approximations by their degree of parameter interaction (see section 3.1 and the Supporting Information [SI]). This allows leading aspects of high-order nonlinearity to be estimated at order-N in the number of computations.

Our approach assumes that the modeler knows only qualitative (if any) information about the nonlinearity of the parameter space that is being sampled, and so an ensemble with along-axis perturbations gives direct and quantitative insight into this nonlinearity in directions of parameter space that are more straightforward to interpret. Another important aspect considered in our ensemble is that precipitation exhibits substantial internal variability, and so model runs are several decades long to better sample the GCM climatology. Other parameter space sampling strategies are discussed more comprehensively in the SI (and in references in the introduction) and would also be compatible with the multiobjective optimization techniques that are the main thrust of this paper.

2.1.1 On-axis runs

Bernstein and Neelin [2016] have created a branch run perturbed physics ensemble for the fully coupled CESM1 (subversion 1.0.5) by perturbing four parameters in the deep convection scheme of the Community Atmosphere Model version 4 (CAM4). Note that both CAM4

and CAM5, as well as the upcoming CAM6, all use the Zhang-MacFarlane scheme for deep convection [*Zhang and McFarlane*, 1995] with modifications incorporated from *Richter and Rasch* [2008] and *Raymond and Blyth* [1986, 1992].

These integrations have been performed in a way that mimics experiments in the Climate Model Intercomparison Project phase 5 (CMIP5) ensemble [*Taylor et al.*, 2012]. To build the perturbed physics ensemble, the GCM was first integrated using transient climate forcing during the 1850–1975 period. From there, branch integrations were performed for each parameter value during an additional 30 years, producing a total of 20 integrations (including a control run) that each spans 1975–2005. The name, units, perturbations, and a short description for each parameter in the ensemble are listed in Table 1. We use monthly fields and calculate December-January-February (DJF), June-July-August (JJA), and annual climatologies for each run. The first 10 years are discarded to allow for model equilibration, so climatologies represent a 1985–2005 average. *Bernstein and Neelin* [2016] show that the hydrological cycle tends to adjust quickly to parameter changes (on timescales less than 10 years). Small remaining imbalances in top-of-atmosphere (TOA) radiation associated with adjustment of the deep ocean could be relevant to some climate quantities but are not a strong effect for those examined here.

2.1.2 Off-axis runs

In addition to the single-parameter perturbation runs, we leverage off-axis experiments (in this case, two parameters varied simultaneously) to more accurately interpolate into the four-dimensional parameter space. Table 2 lists the additional off-axis integrations that are used as interaction terms in the metamodel calculations shown later. A grand total of 51 integrations have been produced: the initial ensemble of 20 mentioned above, the six in Table 2 for off-axis terms, and 25 additional runs that were created for validation and exploration purposes. These 25 additional runs consist of 20 integrations where two parameter were varied at once, and five integrations where three parameters were varied. The parameter information for these runs is not listed explicitly, though they are shown in some figures and will be discussed later.

2.2 Observations and reanalyses

The primary data sets used to constrain the GCM in this study are precipitation from the Global Precipitation Climatology Project [*Adler et al.*, 2003; *Huffman et al.*, 2009] and column

Table 1. The four parameters modified in the perturbed physics ensemble. The first column lists the parameter shorthand as used in this chapter, with the full parameter name in the second column. The third column shows the parameter values used in the CESM1 integrations, with the control run values (the parameter default values) shown with an asterisk. Note in the the text that the first two dmpdz values are discussed as a dangerous range, so quadratic metamodel fits in this chapter exclude the model runs for the bracketed values of dmpdz (third column). Units and a short description are listed in the fourth and fifth columns. More information on this ensemble can be found in *Bernstein* [2014] and *Bernstein and Neelin* [2016]. For more information on CESM1 or the deep convection scheme, see the Community Atmosphere Model version 4 (CAM4) documentation, *Neale et al.* [2010].

Parameter	**Name (and units)**	**Values**	**Description**
dmpdz	deep convective entrainment parameter ($\times 10^{-3}$ m^{-1})	[0, 0.08,] 0.16, 0.25, 0.5, **1***, 1.5, 2	Turbulent entrainment of environmental air into deep convective plume
τ	deep convective timescale (minutes)	30, **60***, 120, 180, 240	Time scale for consumption rate deep of Convective Available Potential Energy (CAPE) by cumulus convection; necessary for closure of deep convection scheme
α	downdraft fraction (unitless, out of 1.0)	0, **0.1***, 0.25, 0.5, 0.75	Fraction or proportionality factor that determines the mass flux of an ensemble downdraft, taking into account precipitation and evaporation
k_e	evaporation efficiency ($\times 10^{-6}$ kg [m^{-2}s^{-1}]$^{-1/2}$ s^{-1})	0.1, 0.5, **1***, 5, 10	Evaporation efficiency of precipitation

bold* (with asterisk) indicates standard or control value

Table 2. Off-axis runs (i.e., integrations in which two parameters are changed at a time) used to fit nonlinear interaction terms in the quadratic and cut-HDMR metamodels. Control values for each parameter are marked with bold font; see Table 1 for units and description of parameters.

dmpdz	τ	α	k_e
1.0	**60**	0.5	5.0
1.0	120	0.5	**1.0**
1.0	180	**0.1**	5.0
1.5	**60**	0.5	**1.0**
1.5	**60**	**0.1**	5.0
1.5	180	**0.1**	**1.0**

water vapor and skin temperature from the ERA-Interim reanalysis [*Dee et al.*, 2011]. Several other data sets evaluated here are described in Table 3; monthly fields were downloaded for each variable during 1985–2005, unless otherwise noted.

3 Methods

3.1 Metamodels

The GCM in this study is an example of a high-dimensional input-output problem, in which a physical system has a large number of input parameters that can be varied independently and will produce a complex response in the output (a GCM simulation). If N parameters are sampled K times each, the number of model integrations required for brute-force sampling is K^N. Sampling the full parameter space of the 20-member ensemble used in this study requires 1000 model integrations at the combinations of values in Table 1. Such a task is computationally unfeasible, so metamodels are borrowed from the engineering literature to accomplish this task.

The first type of metamodel used here is the polynomial-based metamodel (here, linear and quadratic) described in *Neelin et al.* [2010]. This technique fits the parameter dependence of a given field at each grid point and time step or climatological average to a quadratic function, allowing for linear (single-parameter) and nonlinear (parameter interaction) effects. This metamodel can be trained by using on-axis information only or can be further approximated using nonlinear interaction terms, requiring at least one off-axis integration in each pairwise

Table 3. The observational and reanalysis data sets used as model constraints in this study. The full field name is listed in the first column, with the CESM nomenclature given in parentheses. The observational data sets are listed in the second column, the time period used is listed in the third column, and references are given in the fourth column. Note that the CERES and ERS data are more recent satellite products and are only available for the listed time frames.

Field name (with CESM shorthand)	**Data set**	**Period analyzed**	**Citation**
precipitation (PRECT)	Global Precipitation Climatology Project version 2.2 (GPCP)	1985–2005	*Adler et al.* [2003]
column water vapor (TMQ) skin temperature (TS) sea-level pressure (PSL) 300 hPa zonal winds (U300)	ERA-Interim Reanalysis	1985–2005	*Dee et al.* [2011]
2-meter air temperature or reference height temperature (TREFHT)	Willmott and Matsuura version 1.02	1985–2005	*Willmott and Matsuura* [2001]
Zonal wind stress (TAUX)	European Remote Sensing satellites 1 and 2 (ERS-1 and ERS-2)	1991–2001	*PODAAC* [1994]
Longwave cloud forcing (LWCF) Shortwave cloud forcing (SWCF)	Clouds and the Earth's Radiant Energy System - Energy Balanced and Filled (CERES-EBAF) edition 2.8	2000–2016	*Wielicki et al.* [1996]

parameter plane, or $N(N-1)/2$ additional integrations. An even simpler linear metamodel can be calculated by neglecting second-order terms.

The second type of metamodel we use is a modification of cut-HDMR [e.g., *Rabitz and Alis*, 1999; *Rabitz et al.*, 1999; *Li et al.*, 2001; *Wang and Shan*, 2007]. This technique is a simple yet flexible expansion that orders terms by degree of interaction. We use empirical orthogonal functions (EOFs) across each parameter axis separately to approximate the on-axis nonlinearity in the cut-HDMR expansion, and we incorporate a nonlinear interaction in a manner similar to the quadratic metamodel to fit the residuals.

For a detailed description and comparison of these specific techniques, a discussion of their performance, and their context in the greater literature, we refer the reader to the SI. For the majority of the results presented here, we have used the cut-HDMR metamodel with interaction terms included. As discussed in the SI, the quadratic and cut-HDMR metamodels give similar results for response surfaces, and we find them to be reasonable choices for reconstructing the full climatological fields. Cut-HDMR is more adept at capturing model fields directly along the axes, though both metamodels show degradation at points in parameter space far from parameter axes and the control parameter set (i.e., in regions near the edge of the feasible parameter space).

These metamodels are used to reconstruct total fields of seasonal precipitation, column water vapor, and skin temperature. Visualization techniques are then applied to these reconstructions with the goal of finding the parameter combinations in the full parameter space that minimize model error relative to observations and reanalyses. These measures of model error, referred to in this paper as objective functions, include use latitude-weighted root-mean-square error (rmse) and mean-square error (mse) for December-January-February (DJF), June-July-August (JJA), and annual climatologies in a domain that includes the tropical Pacific Ocean. We note that this study does not advocate for any metamodeling technique in particular, as the ideal candidate will change based on the perturbed physics ensemble and goals of the modeler. Here, we use these techniques as a means to an end: to visualize and quantify tradeoffs in model performance.

3.2 Pareto-optimal sets

When multiple objective functions are considered at once, a GCM's performance can be viewed in objective function space. A schematic is shown in Fig. 1a, where two objective

functions ‘a’ and ‘b’ are the axes, and the goal of the modeler is to minimize both simultaneously. Each dot represents a different point in parameter space for the high-dimensional model, and we will use this type of plot to examine the performance of CESM1. The black dots represent the Pareto-optimal set, and the curve connecting them denotes a Pareto frontier, which gets its name from Vilfredo Pareto, an Italian engineer-turned-economist who first described these concepts in the context of optimal resource allocation at the turn of the 20th century. In this schematic, optimum GCM performance is confined to move along the front, where performance in one measure cannot improve without degradation in another. The extreme ends highlight where either a or b are optimized individually, though if they are equally important, the modeler might opt for a happier medium. These Pareto fronts not only help identify tradeoffs that occur but also the GCM physics that warrant revision. Also note that while we visualize the Pareto front as a curve in two dimensions, it can be scaled up to a multi-dimensional surface based on the number of objective functions considered. Figure 1b shows a similar Pareto front but with different objective functions ‘c’ and ‘d.’ In this case, the front is shorter in length and has fewer Pareto-optimal points on it, implying that the tradeoffs are less extensive and the decision for the modeler more straightforward.

Figure 1. A schematic Pareto front is shown for a general high-dimensional model like a GCM. Dots show points in parameter space, plotted as a function of two objective functions, a and b. The black dots show Pareto-optimal solutions, and the line portrays the Pareto front.

In the multiobjective optimization literature, tradeoffs in a system are largely governed by the details of the system, itself. As an example, a common tradeoff experienced in car manufacturing might be luxury versus affordability of a new car. True constraints that govern manufacturing costs imply that a cheaper car will be made with less expensive materials, and because of this, fewer resources will be dedicated to the comfort and features (the luxury) of its design. The manufacturer would weigh these tradeoffs when designing a car for the consumer market, and the tradeoff frontier would ideally be broad—a situation akin to Fig. 1a. GCMs are somewhat different in that observations serve as a truth, and if a GCM (and observations) were functioning perfectly, the optimal parameter set would be a single point, not a curve or surface. In reality, GCMs experience Pareto fronts that vary between the likes of Figs. 1a and 1b, and in cases where a small Pareto front *is* encountered, this is a positive sign indicating the model is functioning well in the aspects that impact these objective functions.

There is an extensive literature on how to calculate a Pareto-optimal set, and this can be a thorny and computationally expensive problem when there are many dimensions over which a modeler wishes to optimize. Iterative or evolutionary algorithms are popular, and the general approach of these methods—inspired by concepts in biological evolution and natural selection—is to select a 'population' of individuals (here, points in parameter space), test the performance of that population using a fitness function, and iterate or evolve over successive populations until an optimal set of points is found. In this paper, the optimality condition or fitness function will be to minimize the mse of multiple fields relative to observations or reanalyses simultaneously. The Pareto-optimal sets discussed in this study have been calculated using a Python package obtained through the Github repository of *Woodruff and Herman* [2014]. This code implements a non-dominated evolutionary sorting algorithm (NGSA) originally introduced by *Laumanns et al.* [2002]. For a more thorough discussion of NGSA and related methods, useful starting points are *Deb et al.* [2002] and *Deb et al.* [2005].

Multiobjective methods have had limited application so far in climate modeling literature. *Price et al.* [2009] built a metamodel for the response surface of an intermediate-complexity global energy and moisture balance model, Grid ENabled Integrated Earth system (GENIE), using a parameter set that represented physical processes in ocean, atmosphere, and sea ice dynamics. They used a kriging method to model the response surface and then employed a version of NGSA [*Deb et al.*, 2002] to find Pareto-optimal solutions. The metamodels we employ here are chosen for their simplicity and reasonable performance, as discussed in the SI.

Figure 2. Rmse values as a function of parameter for the tropical Pacific domain. The left, center, and right columns show precipitation, column water vapor, and skin temperature. Dots show values of original model integrations along each parameter axis, and lines show metamodel reconstructions. Dark solid lines represent the cut-HDMR metamodel reconstruction using the leading three modes, and dark dashed lines represent this for the leading two modes. Light solid lines show quadratic metamodel reconstructions, and light dashed lines show linear metamodel reconstructions. Blue, red, and green correspond to DJF, JJA, and annual analyses.

In the sections that follow, we inspect the parameter sensitivity of tropical Pacific precipitation, column water vapor, and skin temperature across single parameter axes and then use metamodels to do this in multi-dimensional parameter and objective function space. Response surface methodology (discussed in the SI and in more detail below) is used to guide choices for interaction terms and metamodel adjustment, and the results are used to construct

Pareto fronts and explore optimal points in parameter space. A parameter update is proposed in section 6 that improves the tropical Pacific simulation of precipitation, column water vapor, and skin temperature climatologies. Taylor diagrams are used to give an alternative view of the Pareto fronts as well as compare a larger set of model fields to observations before and after the update. As the update demonstrates, improvement in some fields can lead to degradation in others, and these parameter changes are therefore suggested alongside caveats.

4 Parameter sensitivity across multiple fields

Figure 2 shows model rmse as a function of parameter value for precipitation, column water vapor, and skin temperature fields in the tropical Pacific domain. A notable theme here is that all parameters show some degree of nonlinearity as a function of parameter value for a given field—even when the parameter dependence appears fairly linear in others—and these results parallel those of *Bernstein and Neelin* [2016]. For example, the parameter dependence across α appears linear for column water vapor but more notably nonlinear at low α values for both precipitation and skin temperature. Because of this behavior, it is not possible to assume a consistent functional form of parameter dependence across all variables, and the most useful metamodeling techniques are those that can account for this.

Tradeoffs in model performance can also be found in the objective functions of Fig. 2. Examining precipitation rmse as a function of τ (Fig. 2d), values of τ near or just above the control (60 minutes) lower the error relative to GPCP and are therefore candidates for model improvement. In contrast, much higher values of τ (in the range of 200 minutes) are desired in order to better constrain column water vapor or skin temperature against the ERA-Interim reanalysis. Multiobjective tradeoffs become quickly apparent: by changing τ from its default value, one cannot improve model simulation of precipitation without degrading that of column water vapor or skin temperature. Tradeoffs like these are encountered frequently in such a high-dimensional optimization problem, and improving the tropical Pacific climate in the face of such constraints is the focus of this study.

Finally, we note the skill of the cut-HDMR metamodel at capturing on-axis sensitivity (see the SI for a lengthier discussion). The dark solid lines pass through all of the integrated points of Fig. 2, lending confidence to this method for cases when parameters are varied close to the axes. Comparing this with the linear or quadratic metamodel—especially for parameters like dmpdz or k_e—cut-HDMR does significantly better along the axes.

Figure 3. Rmse values as a function of two parameters, α and dmpdz, with τ and k_e held at their control values. Calculations were done over the tropical Pacific domain. Surfaces show the metamodel reconstruction of parameter dependence. Small filled circles show on-axis rmse values (as seen in Fig. 1), and large filled circles show off-axis values extrapolated by the metamodel. The numbers to the right of each large circle show the true rmse value (top) and the percent error of the metamodel surface relative to this point (positive values imply the metamodel overestimates the rmse at that parameter combination). Rows show off-axis reconstructions for (a–d) precipitation, (e–h) column water vapor, and (i–l) skin temperature. The left two columns show cut-HDMR results using on-axis information for DJF and JJA, and the right two columns show these reconstructions with an added interaction term outlined in red.

5 Objective functions in parameter space: Response surfaces

Figure 3 displays results from using cut-HDMR to interpolate off parameter axes from information in Fig. 2 in the α-dmpdz plane, shown for precipitation (top row), skin temperature (middle row), and column water vapor (bottom row). These figures are called response surfaces and fall into the discipline of response surface methodology that arose in the statistics literature [*Box and Wilson*, 1951]. This methodology discusses the relationship between a nonlinear response function (here, the rmse values of a climate field relative to observations) and explanatory variables (parameters). Response surfaces are often thought of as univariate in one objective function, but note here that we use this terminology in a multivariate sense—i.e., considering multiple response surfaces of different seasons or fields simultaneously.

The first and second columns show the results using on-axis information in the metamodel reconstruction, while the third and fourth columns show results when refining cut-HDMR to incorporate the off-axis (interaction) term, outlined in red at (α,dmpdz) = (0.5,1.5). General qualities are similar between cut-HDMR with and without the interaction terms, and they commonly overestimate the curvature of the response surface, particularly at the edges of the parameter ranges (e.g., low values of dmpdz). This is visually apparent when comparing the value of the rmse values from integrations (filled circles) to that of the cut-HDMR results (underlying response surface). For example, the integration at (α,dmpdz) = (0.5,0.16) shows a large discrepancy between the true model rmse value and the underlying surface. This sensitivity is not ideal for the metamodel, though it doesn't devalue it, since the modeler will typically avoid parameter values at the edge of the feasible range and is likely to consider parameter updates that are in the vicinity of the control axes.

We note that a similar analysis for Fig. 3, but in this case using the (α,dmpdz) = (0.25,1.5) point as the interaction term, leads to a much more substantial curvature effect, so we choose to use (α,dmpdz) = (0.5,1.5) for the nonlinear interaction fit in this plane. Our process in selecting this was iterative, informed by integrating the GCM at several well-chosen off-axis points (larger circles in Fig. 3) and comparing their true rmse values to those predicted by the metamodel. This method helps bring the box contained within $\alpha \in [0.25, 0.5]$ and dmpdz $\in [1.0, 2.0]$ into focus as a likely region or window for a parameter update. By starting with on-axis information and using the response surface to guide where further integrations should be placed, the modeler can narrow to a region of interest in parameter space that can help improve the accuracy of the metamodel and improve decisions about GCM parameter updates. Note that if the curvature effect is severe and the metamodel is deemed untrustworthy, one could incorporate more off-axis terms into the metamodel calculation or even scale up to higher-order terms in HDMR for that particular plane.

6 Parameters in objective function space: Pareto fronts

6.1 Pareto front visualization

Figure 3 allows the modeler to visualize initial tradeoffs when two parameters are varied at once, though the goal here is to achieve full exploration of the ensemble at hand. The cut-HDMR metamodel is used to interpolate into the four-dimensional space, and Fig. 4 shows this for an approximately global domain (60°S to 60°N, top row) and for the tropical Pacific domain shown in other figures (bottom row). The vertical and horizontal axes show metamodel-estimated mse values for precipitation and column water vapor. The familiar Pareto front, as schematized in Fig. 1, is approximated by plotting the first three successive Pareto-optimal sets—as calculated by the evolutionary algorithm described in section 3—as black squares.

For a well-tuned model, the control run (yellow star) would ideally lie on the Pareto front in Fig. 4, though in this case one can see notable uncertainty in column water vapor that displaces the control run along the horizontal axis. Note that for measures like these, annual cases (not shown) tend to perform better, likely because the model itself has been historically tuned to annual averages. Each square in Fig. 4 represents one of the 1000 possible parameter combinations that have been reconstructed using the cut-HDMR metamodel. The color of each square represents the Euclidean distance in parameter space between a point's parameter values and the control values; the distances are normalized to have a maximum of one (light yellow) for

Figure 4. Cut-HDMR reconstructions of the full parameter space, showing tradeoffs in objective functions of precipitation and column water vapor mse. (a,b) show seasonal results for all longitudes between 60°S and 60°N, and (c,d) show results for the tropical Pacific domain. The control run is shown in each figure as a yellow star, and squares represent the 1000 possible parameter combinations based on the sampling discussed in the methods. Color and shading of squares denotes the Euclidean distance of a given parameter set to the control, with each parameter axis normalized by its range to contribute equally in computing the Euclidean distance. Darker red squares represent parameter combinations closer to the control run, and yellow squares imply parameter combinations further from the control; the minimum and maximum distances are represented by the shaded red and yellow squares in the legend, respectively. Black squares mark points that approximate the Pareto front, selected through an iterative procedure that collects the first three successive sets of Pareto-optimal solutions using the evolutionary algorithm.

the combination of all parameters at their farthest endpoints, and zero (dark red) for all parameters at the control. That the darkest red squares tend to occur nearest the control run—and the yellow squares farthest—serves as a check for the smoothness of the response surface.

The 1000 parameter combinations have been reconstructed based on the original on-axis parameter sampling in the perturbed physics ensemble. While it would be possible to interpolate between these values and sample at a density greater than 1000, we have chosen to stop here with the knowledge that parameter sensitivity is smoothly varying (see Figs. 2 and 3), and for the size of improvements we are getting, this resolution is adequate. A notable point from Fig. 4 is that the shape of the Pareto front varies seasonally and regionally. For the global domain in JJA (Fig. 4b), it is curved more smoothly, with a continuous set of tradeoffs. The global domain during DJF (Fig. 4a) is sharper, by contrast. In addition, a significant portion of the Pareto front in the tropical Pacific domain during both DJF and JJA is nearly parallel to the horizontal axis, highlighting a region in parameter space where precipitation error will not change measurably, but as much as a 50% improvement can be made to column water vapor relative to the control run.

6.2 Two-dimensional Pareto fronts in detail

Figure 5 shows a zoomed version of the Pareto front for different combinations of objective function planes. The control is now plotted as a gray star, and the shaded squares from Fig. 4 are in grayscale. All on- and off-axis validation (i.e., true GCM) runs are also included

Figure 5. Zoomed in version of cut-HDMR reconstructions for the full parameter space, calculated over the tropical Pacific domain and shown for different combinations of precipitation, column water vapor, and skin temperature mse. The control run (star) and the shading of squares as in Fig. 3 now appear in grayscale, with dark gray squares representing parameter combinations close to the control, and light gray far from the control. The colored shapes shown in the legends indicate mse values from full model integrations. The first four colored shapes after the control run represent model integrations varying parameters on-axis separately. Shapes plotted with darker hues relative to the legend imply parameter values that are less than the control, and shapes plotted with lighter hues relative to the legend imply parameter values greater than the control. Colored squares represent integrations where parameter values were varied two at a time, according to Table 2. Thick light gray lines represent a fit Pareto-optimal set using univariate spline interpolation. In (c,d), three points are circled in yellow, red, and blue corresponding to the off-axis integrations at (τ,α) = (120,0.5), (α,dmpdz) = (0.5,1.5), and (k_e,α) = (5.0,0.5), respectively, conducted at key points suggested by the metamodel.

as distinct shapes, and these represent all of the integrations performed with CESM1 as part of the iterative search process in narrowing to an optimal parameter window. Pareto-optimal sets were calculated using the evolutionary algorithm, and the resulting Pareto fronts are drawn as thick, light gray curves. Spline interpolation has been used to smooth each of these for display purposes, and the width of the gray curves is chosen to schematically convey the width that arises from an iteration procedure as described in Fig. 4.

Inspecting Fig. 5 more closely, it is clear that the control run is at times displaced from the Pareto front. Along-axis runs (large shapes) are plotted for the four parameters that have been perturbed in this study, as well as off-axis integrations (colored squares) where two parameters were perturbed at once. In Fig. 5c, three off-axis integrations have been circled in orange, red, and blue to represent different locations along the Pareto front in this plane (see Table 4 for specific parameter values of these runs). The selection process was based in particular on DJF, though the same points are also labeled for JJA. We emphasize that wherever the metamodel produces an optimal region on the Pareto front with points that are not close to an existing model run, additional CESM integrations should be conducted (and potentially incorporated into the metamodel) if that region is being considered for a parameter update.

These figures also give a comprehensive sense of the tradeoffs to be expected in two dimensions. For example, tradeoffs are minor for precipitation versus column water vapor (Figs.

5a,b), indicated by the small Pareto front. For precipitation versus skin temperature, the Pareto front is much larger for DJF, implying substantial tradeoffs in model performance (Fig. 5c), though the opposite is true for JJA (Fig. 5d). A similar story exists for column water vapor versus skin temperature, with DJF exhibiting considerable tradeoffs relative to JJA (Figs. 5e,f). As discussed for Fig. 1, in cases where the Pareto front is small and the tradeoffs slight, the GCM is well-optimized for the objective functions being considered, making the modeler's decision somewhat easier. But when the Pareto front is large, tradeoffs can lead to a larger range of equivalently performing (i.e., Pareto-optimal) parameter updates, depending on what is important to the modeler. For example, Fig. 5c shows that parameter adjustments can decrease skin temperature mse as much as 50%, though that comes at the expense of an increase in precipitation mse of about 30%.

6.3 Tests for identifying optima for higher-dimensional Pareto fronts

Extending the Pareto fronts above to include more than two dimensions—i.e., incorporating three or more objective functions at once—is a useful next step in the optimization process. Different Pareto-optimal solutions will exist for different combinations of seasons, fields, and regions (all of which embody unique objective functions). Because of this, we consider these tradeoffs simultaneously when searching for Pareto-optimal solutions. We use DJF, JJA, and annual climatologies for three separate fields—precipitation, skin temperature, and column water vapor—over the tropical Pacific region. This gives $3 \times 3 = 9$ separate objective functions over which the evolutionary sorting algorithm optimizes simultaneously. The goal is to find the parameter combinations that lie on the Pareto front when all nine dimensions are considered together.

To establish which points in parameter space perform best, we have developed several tests for determining what points lie on or closest to the nine-dimensional front using the evolutionary algorithm. The first test is the most stringent and is only passed when a given (dmpdz,τ,α,k_e) combination is on the Pareto front for both the quadratic and cut-HDMR metamodels (once each with the interaction terms included, and once each without). In other words, points that pass this test are Pareto-optimal solutions for all versions of metamodel used here. For the 1000 parameter combinations approximated by the metamodel, as well as all 51 true integrations in the perturbed physics ensemble, only six passed this test and are marked in Table 4 with one asterisk. Three additional parameter combinations in table 3—denoted with two asterisks—passed a test discussed in the next paragraph, and a final tenth combination performs well but

Table 4. Rows above the line detail the six parameter combinations that are contained in the Pareto-optimal set across all four metamodels; rows below the line show three additional parameter combinations that perform well across other tests (see text for details). Units for each variable are as follows: dmpdz $\times 10^{-3}$ m^{-1}, τ minutes, α as a fraction out of 1.0, and $k_e \times 10^{-6}$ kg [m^{-2}s^{-1}]$^{-1/2}$ s^{-1}. Control values for each parameter are marked with bold font. Asterisks give information about the tests that each case passes and are discussed in the text; colors correspond to parameter combinations in Figs. 5c,d and 6.

dmpdz	τ	α	k_e	
1.0	120	0.5	0.1	*
1.0	120	0.25	**1.0**	*
1.0	120	0.5	0.5	*
1.5	**60**	0.25	**1.0**	*
2.0	**60**	0.25	0.5	*
2.0	120	**0.1**	**1.0**	*
1.0	120	0.5	**1.0**	**
1.5	**60**	0.5	**1.0**	**
1.5	120	0.25	**1.0**	**
1.0	**60**	0.5	5.0	

passes no tests, and it is included for presentation purposes later. Note that the control run is *not* part of the optimal parameter sets in Table 4, and none of the cases suggest changing just one parameter or all parameters at once.

The test described above requires agreement among the different forms of metamodel. An alternative and complementary screening process for Pareto-optimal solutions might relax the requirement that the points fall directly *on* the Pareto front and instead fall somewhere *near* it. The final three cases marked with two asterisks in Table 4 meet criteria for three additional screenings designed to accommodate this. The first was to search for optimal parameter combinations for each pairwise combination of season and field (as done in Fig. 5), then evaluate which cases occur most frequently on these fronts. The second test was to calculate the leading points on the Pareto front by iterating over successive Pareto-optimal layers until at least 100 cases were collected (i.e., the top 10% of points). The third and final test was to calculate the Euclidean mse distance from the origin in each pairwise plane and extract the lowest 100 values (i.e., the 10% of cases closest to the origin). This final option is the most relaxed in that points are not selected based on whether they lie on a Pareto front, but instead are located closest to the origin in each pairwise mse plane. Such a test overlaps with information gained from solving for points on the Pareto front, since many of these cases will coincide. Three parameter combinations that score well on these alternative tests are listed in Table 4, denoted with two asterisks. Though more points fit the necessary criteria, we choose to show three that represent physically plausible parameter values that are reasonably distinct from one another. For this process to be truly useful, this element of human judgment and decision-making is crucial.

Figure 6. Zonally averaged anomalies of precipitation and skin temperature relative to observations and reanalyses, calculated for the tropical Pacific region between 120°E and 300°E during (a,b) DJF and (c,d) JJA. Yellow, red, and blue lines correspond to the points circled along the Pareto front in Fig. 5c, the dark solid lines show the control run, and the dashed line represents the GPCP and ERA-Interim data sets, about which the integrations are centered.

6.4 Comparing Pareto-optimal cases

Figure 6 shows zonal averages in the tropical Pacific region for precipitation and skin temperature, calculated as anomalies relative to the GPCP and ERA-Interim data sets, respec-

tively. These averages are shown for DJF (top row) and JJA (bottom row) and lend insight into the kind of tradeoffs encountered along the Pareto front as well as the improvements that can be achieved with multiobjective optimization. The blue lines, corresponding to the parameter update (α,k_e) = (0.5,5.0) shown in Fig. 5 and Table 4, show a slight decrease in the maximum error of precipitation in both hemispheres (i.e., improvement as much as 20%) relative to the control run, but skin temperature simulation is worse, shifting the model from initially overestimating zonal temperatures to fully underestimating them. The yellow lines, corresponding to (τ,α) = (120,0.5), achieve the opposite effect: precipitation quality gets worse relative to the control, and skin temperature simulation improves slightly (by about 10%). A happy medium can be found in the red lines, which correspond to (α,dmpdz) = (0.25,1.5) and improve both precipitation (by about 10%) and skin temperature (by about 5%) relative to the control. This improvement happens in both the DJF and JJA seasons, even though the integrations themselves were based on the DJF Pareto front in Fig. 5c. Such an outcome—an update that improves fields of interest across multiple seasons—is unusual but certainly advantageous to a modeler.

Figure 7. Zoomed in cut-HDMR reconstruction of the full parameter space, showing tradeoffs between precipitation over the tropical Pacific (vertical axis) and over the Amazon (horizontal axis) during DJF and JJA. Shading, shapes, and Pareto fronts are plotted as in Fig. 5.

Armed with the information gained from tests leading to Table 4, as well as the details of Figs. 5 and 6, we now evaluate the model simulated with the update corresponding to the red point, which represents increasing α and dmpdz to 0.25 and 1.5, respectively—leaving τ and k_e the same. We do this with informed confidence that the seasonal climatology for both precipitation and skin temperature will improve in the tropical Pacific region. One caveat discussed later is that other fields will also be affected in a coupled climate system, and so the parameter value that improves precipitation and skin temperature most may not do so for other fields of interest to the modeler. Note that this particular update was not the combination used to fit the interaction term of the metamodel in the α-dmpdz plane of Fig. 3, which was (α,dmpdz) = (0.5,1.5). The point of the interaction term is therefore not to serve as an *optimal* point in an objective function plane, but rather to help adjust the metamodel in its vicinity so that potential nearby optima are brought into sharper focus and can then be integrated in the full GCM for subsequent analysis.

6.5 An application to precipitation tradeoffs between the tropical Pacific and the Amazon

Another relevant tradeoff in CESM1 occurs between simulating climatological rainfall over the Amazon and over the broader tropical Pacific. This can be seen in Fig. 7, which shows precipitation mse over the tropical Pacific (vertical axis) versus that over the Amazon (horizontal axis, defined here as 30°S–20°N, 270°E–330°E). Making parameter changes to improve one of these aspects of model performance comes at the expense of degrading the other. The control run sits fairly close to the Pareto front for the JJA season, but it is much further during DJF, especially for the Amazon. Some suggestions for parameter updates during DJF involve modifications to α, either by increasing this parameter along its axis (light upside-down triangles) or higher values of α coupled with either changes to dmpdz (red squares) or changes to k_e (blue squares). For JJA, higher dmpdz values appear to be favored (light red triangles), though red squares (α,dmpdz) and blue squares (k_e,α) are also close to the Pareto front. The circles from Figs. 5c,d have been included on the corresponding points here, highlighting once again that the shape and composition of the Pareto front can vary across region and that optimal parameter combinations are likely to change across season, domain, or objective function.

Figure 8. cut-HDMR metamodel reconstruction of the full parameter space plotted as a Taylor diagram. Calculations were done for (a) precipitation, (b) column water vapor, and (c) skin temperature over the tropical Pacific domain during the DJF season. Black dots show the full parameter space reconstruction from cut-HDMR, gray dots represent all validation runs for the GCM, and the red stars highlight the location of (α,dmpdz) = (0.25,1.5) model integrations. Angular values are latitude-weighted spatial correlations relative to observations and reanalyses, and radial values are normalized spatial standard deviations for each field (where observations and reanalyses correspond to the "REF" or reference value of 1.0).

7 Pareto fronts visualized on Taylor diagrams

Taylor diagrams [*Taylor*, 2001] are a common way to visualize and compare multiple aspects of GCM performance, and Fig. 8 shows these results from the full ensemble integration for the red point in Table 4 for DJF. The angular direction in these plots is the latitude-weighted spatial correlation for each field relative to the control or reanalysis data sets, and the radial value is the field's latitude-weighted spatial standard deviation divided by that of the

observations. Black points represent the parameter space reconstruction of the cut-HDMR metamodel with interaction terms from Table 2; gray dots show all on- and off-axis runs available in the ensemble. The control run is shown as a circled yellow star, and the suggested parameter update discussed previously, (α,dmpdz) = (0.25,1.5), is shown as a circled red square. In each plot, a zoomed inset is included to see the edge of points more clearly. Note that the 'tails' of these clouds are due to incorporating interaction terms and the curvature effects caused by them. The modeler must make a compromise when including the interaction terms: the metamodel is improved in the vicinity of these points and closer to the axes, though it can be degraded at the edges of the parameter ranges, where (as stated previously) parameter values are physically less reasonable. This is not a large point of concern, however, as the behavior of the metamodel is less important far from the Pareto front.

Across DJF, JJA, and annual climatologies, the control run falls close to the edge of the cloud of points at higher correlation values, meaning the spatial correlation for the control parameter set is nearly maximized in the parameter space explored here. The magnitude of spatial variability in this domain, however, is overestimated for precipitation and column water vapor and underestimated for skin temperature.

For precipitation, column water vapor, and skin temperature, the parameter update helps nudge model performance closer to the observed magnitude of spatial variability. As discussed previously, the relatively short or narrow Pareto front seen in Fig. 8c implies that there are no significant tradeoffs between pattern and amplitude in skin temperature. Precipitation in Fig. 8a is an example where further improvement could happen, though the current update still does some good. In contrast, the update for column water vapor (Fig. 8b) does show a clear tradeoff. A schematic Pareto front has been drawn as a light gray line in the inset of Fig. 8b. The parameter update here causes the model magnitude to be underestimated, though the spatial correlation increases slightly. The slope of this line is important and depicts a true Pareto front along which the GCM performance can improve its spatial correlation only by degrading its magnitude.

These Taylor diagrams separate spatial correlation from magnitude and therefore give the modeler information not captured in measures of rmse or mse in previous figures. The relatively flat shape of the cloud of points in each case indicates that parameter optimization is mainly changing the magnitude of the fields rather than their spatial correlation with the observations and reanalyses. This outcome implies that the GCM uncertainty explored here is

likely a deeper issue within the model's structural physics and dynamics themselves and can only be alleviated to an extent by parameter optimization. This result is important, since many changes expected as a result of anthropogenic global warming alter the magnitude of precipitation, and this is a large source of uncertainty in diagnosing end-of-century changes in multi-model ensembles [e.g., *Xie et al.*, 2015]. Placing constraints on GCMs that can improve the representation of this magnitude of variability is a necessary step in understanding GCM uncertainty in end-of-century changes.

Figure 9. Taylor diagrams showing model performance for the (α,dmpdz) = (0.25,1.5) update (colored markers) relative to the control run (black markers). These plots are modeled after those in the CESM Atmospheric Model Working Group (AMWG) diagnostics package, distributed by AMWG at http://www2.cesm.ucar.edu/working-groups/amwg. Correlation and amplitude values were calculated over the tropical Pacific region discussed in the text.

8 Standard model diagnostics

While the model improvements gained from the α and dmpdz updates are small for most fields, they are not inconsequential. GCMs are complex systems that change from one generation to the next, and model improvement is a stepwise process. Figure 9 shows how GCM performance in multiple fields changes as the model is updated from its control parameter set to the modified (α,dmpdz) = (0.25,1.5). The information presents a collection of different fields typically analyzed in the CESM Atmospheric Model Working Group (AMWG) diagnostics package (https://www2.cesm.ucar.edu/working-groups/amwg/amwg-diagnostics-package). Details of the observational or reanalysis data sets are listed in Table 3.

Similar to the discussion above, the improvement in model fields tends to occur primarily in the magnitude of spatial variability and not correlation. Sea level pressure and precipitation improve across DJF, JJA, and annual climatologies. The simulation of other fields improves in certain seasons but degrades in others (e.g., compare zonal wind stress during DJF and JJA), though shortwave and longwave cloud forcing get consistently worse across all climatologies. These changes in model performance occur while having only a minimal effect on fields that already perform well, however (e.g., skin temperature or 2-meter air temperature), and this is a nontrivial result, given that precipitation can exhibit significant internal variability, and objectively improving its quality without significantly altering other aspects of the

climate system is historically a thorny problem. In the case of the update (α,dmpdz) = (0.25,1.5), if the goal is to improve precipitation or sea level pressure at seasonal timescales, then such an update might be worth the tradeoff of slightly degraded quality in other fields. If the tradeoffs for other fields are severe, then it would be straightforward to repeat analyses with objective functions corresponding to these fields, as well. And finally, it is important to keep in mind that the measures plotted in these diagrams are distinct from rmse or mse values, so an improvement in the objective functions evaluated earlier will not translate directly into improvement in correlation or spatial variability, though there will be significant overlap.

9 Summary and conclusions

In this paper, we showcase concepts from multiobjective optimization to answer questions about parameter optimization in a perturbed physics ensemble, which samples four parameters from the deep convection scheme of CESM1.

Parameter sensitivity is visualized in one dimension, along each parameter axis, and in multiple dimensions, as a response surface that describes model rmse as a function of two or more parameters. This method gives quantitative information on how objective functions vary in parameter space, and by incorporating interaction terms into the response surface, we demonstrate how one can adapt the metamodel to focus on regions or windows of interest where likely candidates for parameter updates exist. We then employ metamodeling techniques to evaluate combinations of parameters in objective function space, yielding information about which parameter combinations can improve multiple metrics simultaneously. Our optimization approach is iterative and is done in several steps. First, a metamodel is used to generate hypotheses for where in parameter space to place more model runs, and the GCM is integrated at these well-chosen points. Next, this information is used to further refine the metamodel and produce a more accurate Pareto front in the vicinity of these selected points. Different choices of metamodel or emulator can be made, depending on how the parameter space is initially sampled and what the overall goals of the modeler are. In our case, understanding along-axis sensitivity was a first-order concern. Cut-HDMR is a useful choice for this, though we do not recommend or promote any metamodeling technique in particular.

Tradeoffs in model performance are visualized for GCM precipitation, column water vapor, and skin temperature climatologies, and an evolutionary algorithm is used to find Pareto-optimal sets in objective function space. These results are used to estimate the Pareto front,

which is a surface in objective function space along which the optimal configuration of GCM parameters exist, and where improvements in one dimension cause degradation in another.

Both quadratic and cut-HDMR metamodels are used in a series of tests to pinpoint the most likely candidates for a GCM parameter update. Our results show that the control parameter set is not located on the Pareto front for any of the cases considered (and in some examples is notably displaced). These outcomes depend entirely on the objective functions of interest to the modeler and the observational constraints used, so these results will likely vary for other fields, processes, and domains of interest. One parameter update in particular, which passed the testing criteria and represented an increase in both α (the evaporation efficiency) and dmpdz (the fractional entrainment rate), improved zonal mean skin temperature and precipitation magnitude relative to the control parameter set. We note that the details of the Pareto front will depend on the parameter space sampling strategy, the metamodeling approach, and the observations used as objective constraints. Furthermore, while the metamodel is adept at reconstructing climatological fields near parameter axes and in the vicinity of parameter space where off-axis runs have been used for fitting, its performance quality degrades away from these regions. We therefore emphasize that we do not trust the metamodel for final conclusions about optimal parameter updates but rather use it to suggest locations for new GCM runs and to provide context for interpreting them.

These results are also displayed using Taylor diagrams. Incorporating multiple fields shows that this parameter update modestly improves CESM1 precipitation during DJF without significantly affecting column water vapor or temperature. These diagrams highlight that the improvement from the proposed (α,dmpdz) updated arises primarily in the magnitude measure of the field, and not the correlation. That the parameter optimization has such a modest effect on the correlation score implies that model error may be rooted in larger issues underlying GCM dynamics or physics, or in other parameters not sampled in this ensemble.

The performance of CESM1 with the control parameter set and the proposed update is then compared to six additional observational and reanalysis data sets commonly used in the NCAR Atmospheric Model Working Group (AMWG) diagnostics package. Improvements can be seen in precipitation and sea level pressure across DJF, JJA, and annual climatologies. Improvement in other fields only happens in certain seasons, and simulation of both longwave and shortwave cloud forcing gets slightly and consistently worse. In considering such parameter updates identified by this multiobjective approach, one would ideally also incorporate out-

side information available to modelers—including reasonable evidence from process-based model studies or observations—and make an informed decision about whether the sub-grid scale processes involved in these updates are well-represented by the parameter values and GCM code. In addition, this approach serves as a way of identifying processes in need of additional scrutiny for observational constraints in the coupled climate system.

Acknowledgments

This work was supported in part by National Science Foundation (NSF) grant AGS-1540518 and National Oceanic and Atmospheric Administration (NOAA) grants NA14OAR4310274 and NA15OAR4310097. We also acknowledge high-performance computing support from Yellowstone (ark:/85065/d7wd3xhc) provided by NCAR's Computational and Information Systems Laboratory and sponsored by NSF. CESM data used in this study are available from the authors following guidelines in the CESM data plan (www.cesm.ucar.edu/management/docs/data.mgt.plan.2011.pdf). We would like to thank Charles Jackson and two anonymous reviewers for their helpful comments and feedback, as well as Diana Bernstein for her help in producing and maintaining the CESM data used here. The scripts used in this manuscript can be accessed at http://research.atmos.ucla.edu/csi/ or by emailing the corresponding author at baird@atmos.ucla.edu.

References

Adler, R. F., G. J. Huffman, A. Chang, R. Ferraro, P.-P. Xie, J. Janowiak, B. Rudolf, U. Schneider, S. Curtis, D. Bolvin, A. Gruber, J. Susskind, P. Arkin, and E. Nelkin (2003), The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present), *Journal of Hydrometeorology*, *4*, 1147–1167.

Allen, M. R., and D. A. Stainforth (2002), Towards objective probabalistic climate forecasting, *Nature*, *419*(6903), 228–228.

Annan, J. D., J. C. Hargreaves, N. R. Edwards, and R. Marsh (2005), Parameter estimation in an intermediate complexity earth system model using an ensemble kalman filter, *Ocean Modelling*, *8*(1–2), 135–154.

Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär (2012), Objective calibration of regional climate models, *Journal of Geophysical Research: Atmospheres*, *117*(D23).

Bernstein, D. N. (2014), Evaluation of regional sensitivities in climate modeling, Ph.D. thesis, Hebrew Univeresity of Jerusalem.

Bernstein, D. N., and J. D. Neelin (2016), Identifying sensitive ranges in global warming precipitation change dependence on convective parameters, *Geophysical Research Letters*, p. 2016GL069022.

Bony, S., and J.-L. Dufresne (2005), Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models, *Geophysical Research Letters*, *32*(20).

Box, G. E. P., and K. B. Wilson (1951), On the experimental attainment of optimum conditions, *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(1), 1–45.

Boyle, J. S., S. A. Klein, D. D. Lucas, H.-Y. Ma, J. Tannahill, and S.-P. Xie (2015), The parametric sensitivity of CAM5's MJO, *Journal of Geophysical Research: Atmospheres*, *120*, 1424–1444.

Bracco, A., J. D. Neelin, H. Luo, J. C. McWilliams, and J. E. Meyerson (2013), High dimensional decision dilemmas in climate models, *Geoscientific Model Development*, *6*(2), 2731–2767.

Brown, J. R., S. B. Power, F. P. Delage, R. A. Colman, A. F. Moise, and B. F. Murphy (2010), Evaluation of the south pacific convergence zone in ipcc ar4 climate model simulations of the twentieth century, *Journal of Climate*, *24*(6), 1565–1582.

Capotondi, A. (2013), Enso diversity in the ncar ccsm4 climate model, *Journal of Geophysical Research: Oceans*, *118*(10), 4755–4770.

Collins, M., B. B. B. Booth, G. R. Harris, J. M. Murphy, D. M. H. Sexton, and M. J. Webb (2006), Towards quantifying uncertainty in transient climate change, *Climate Dynamics*, *27*(2-3), 127–147.

Dai, A. (2006), Precipitation characteristics in eighteen coupled climate models, *Journal of Climate*, *19*(18), 4605–4630.

Deb, K., A. Pratap, S. Agarwal, and T. Meyarivan (2002), A fast and elitist multiobjective genetic algorithm: NGSA-ii, *IEEE Transactions on Evoluationary Computation*, *6*(2), 182–197.

Deb, K., M. Manikanth, and S. Mishra (2005), Evaluating the ϵ-domination based multiobjective evolutionary algorithm for a quick computation of Pareto-optimal solutions, *Evolutionary Computation*, *13*(4), 501–525.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haim-

berger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J. J. Morcrette, B. K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J. N. Thépaut, and F. Vitart (2011), The era-interim reanalysis: Configuration and performance of the data assimilation system, *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597.

Deser, C., A. S. Phillips, R. A. Tomas, Y. M. Okumura, M. A. Alexander, A. Capotondi, J. D. Scott, Y.-O. Kwon, and M. Ohba (2012), Enso and pacific decadal variability in the community climate system model version 4, *Journal of Climate*, *25*(8), 2622–2651.

Gettelman, A., J. E. Kay, and K. M. Shell (2012a), The evolution of climate sensitivity and climate feedbacks in the Community Atmosphere Model, *Journal of Climate*, *25*, 1453–1469.

Gettelman, A., J. E. Kay, and J. T. Fasullo (2012b), Spatial decomposition of climate feedbacks in the community earth system model, *Journal of Climate*, *26*(11), 3544–3561.

Guo, Z., M. Wang, Y. Qian, V. E. Larson, S. Ghan, M. Ovchinnikov, P. A. Bogenschutz, C. Zhao, G. Lin, and T. Zhou (2014), A sensitivity analysis of cloud properties to CLUBB parameters in the Single-column Community Atmosphere Model (SCAM5), *Journal of Advances in Modeling Earth Systems*, *6*, 829–858.

Guo, Z., M. Wang, Y. Qian, V. E. Larson, S. Ghan, M. Ovchinnikov, P. A. Bogenschutz, A. Gettelman, and T. Zhou (2015), Parametric behaviors of CLUBB in simulations of low clouds in the Community Atmosphere Model (CAM), *Journal of Advances in Modeling Earth Systems*, *7*, 1005–1025.

Huffman, G. J., R. F. Adler, D. T. Bolvin, and G. Gu (2009), Improving the global precipitation record: GPCP version 2.1, *Geophysical Research Letters*, *36*(17).

Jackson, C., M. K. Sen, and P. L. Stoffa (2004), An efficient stochastic bayesian approach to optimal parameter and uncertainty estimation for climate model predictions, *Journal of Climate*, *17*(14), 2828–2841.

Jackson, C. S., M. K. Sen, G. Huerta, Y. Deng, and K. P. Bowman (2008), Error reduction and convergence in climate prediction, *Journal of Climate*, *21*(24), 6698–6709.

Kay, J. E., B. R. Hillman, S. A. Klein, Y. Zhang, B. Medeiros, R. Pincus, A. Gettelman, B. Eaton, J. Boyle, R. Marchand, and T. P. Ackerman (2012), Exposing global cloud biases in the community atmosphere model (cam) using satellite observations and their corresponding instrument simulators, *Journal of Climate*, *25*(15), 5190–5207.

Kuo, Y.-H., and J. D. Neelin (2016), Tropical convective transition statistics and causality in the water vapor-precipitation relation, *Journal of Climate*, *revised*.

Latif, M., K. Sperber, J. Arblaster, P. Braconnot, D. Chen, A. Colman, U. Cubasch, C. Cooper, P. Delecluse, D. Dewitt, L. Fairhead, G. Flato, T. Hogan, M. Ji, M. Kimoto, A. Kitoh, T. Knutson, H. Le Treut, T. Li, S. Manabe, O. Marti, C. Mechoso, G. Meehl, S. Power, E. Roeckner, J. Sirven, L. Terray, A. Vintzileos, R. Voß, B. Wang, W. Washington, I. Yoshikawa, J. Yu, and S. Zebiak (2001), Ensip: the el niño simulation intercomparison project, *Climate Dynamics*, *18*(3-4), 255–276.

Laumanns, M., T. Lothar, K. Deb, and E. Zitzler (2002), Combining convergence and diversity in evoluationary multiobjective optimization, *Evolutionary Computation*, *10*(3), 263–282.

Lee, L., K. Carslaw, K. Pringle, G. Mann, and D. Spracklen (2011), Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters, *Atmospheric Chemistry and Physics*, *11*(23), 12,253–12,273.

Lee, L., K. Carslaw, K. Pringle, and G. Mann (2012), Mapping the uncertainty in global ccn using emulation, *Atmospheric Chemistry and Physics*, *12*(20), 9739–9751.

Li, G., and S.-P. Xie (2014), Tropical biases in CMIP5 multimodel ensemble: The excessive equatorial pacific cold tongue and double itcz problems, *Journal of Climate*, *27*(4), 1765–1780.

Li, G., C. Rosenthal, and H. Rabitz (2001), High dimensional model representations, *The Journal of Physical Chemistry A*, *105*(33), 7765–7777.

Lin, J.-L. (2007), The double-ITCZ problem in IPCC AR4 coupled GCMs: Ocean–atmosphere feedback analysis, *Journal of Climate*, *20*(18), 4497–4525.

Lintner, B. R., B. Langenbrunner, J. D. Neelin, B. T. Anderson, M. J. Niznik, G. Li, and S.-P. Xie (2016), Characterizing cmip5 model spread in simulated rainfall in the pacific intertropical convergence and south pacific convergence zones, *Journal of Geophysical Research: Atmospheres*, *121*(19).

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth (2004), Quantification of modelling uncertainties in a large ensemble of climate change simulations, *Nature*, *430*(768–772).

Neale, R. B., J. H. Richter, A. J. Conley, S. Park, P. H. Lauritzen, A. Gettelman, D. L. Williamson, P. J. Rasch, S. J. Vavrus, M. A. Taylor, W. D. Collins, M. Zhang, and S.-J. Lin (2010), Description of the NCAR Community Atmosphere Model (CAM 4.0),

NCAR Technical Note TN-485, National Center for Atmospheric Research.

Neale, R. B., J. Richter, S. Park, P. H. Lauritzen, S. J. Vavrus, P. J. Rasch, and M. Zhang (2013), The mean climate of the community atmosphere model (cam4) in forced sst and fully coupled experiments, *Journal of Climate*, *26*(14), 5150–5168.

Neelin, J. D., A. Bracco, H. Luo, J. C. McWilliams, and J. E. Meyerson (2010), Considerations for parameter optimization and sensitivity in climate models, *Proceedings of the National Academy of Sciences*, *107*(50), 21,349–21,354.

PODAAC (1994), ESA ERS-1 and ERS-2 scatterometer ocean surface winds, Physical Oceanography Distributed Active Archive Center. Produced by Jet Propulsion Laboratory, NASA. Available at the NCAR Research Data Archive, http://rda.ucar.edu/datasets/ds744.0/.

Price, A. R., R. J. Myerscough, I. I. Voutchkov, R. Marsh, and S. J. Cox (2009), Multi-objective optimization of genie earth system models, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1898), 2623.

Qian, Y., H. Yan, Z. Hou, G. Johannesson, S. Klein, D. Lucas, R. Neale, P. Rasch, L. Swiler, J. Tannahill, H. Wang, M. Wang, and C. Zhao (2015), Parametric sensitivity analysis of precipitation at global and local scales in the Community Atmosphere Model CAM5, *Journal of Advances in Modeling Earth Systems*, *7*(2), 382–411.

Rabitz, H., and O. F. Alis (1999), General foundations of high-dimensional model representations, *Journal of Mathematical Chemistry*, *25*(2–3), 197–233.

Rabitz, H., O. F. Alis, J. Shorter, and K. Shim (1999), Efficient input–output model representations, *Computer Physics Communications*, *117*(1–2), 11–20.

Raymond, D. J., and A. M. Blyth (1986), A stochastic mixing model for nonprecipitating cumulus clouds, *Journal of the Atmospheric Sciences*, *43*(22), 2708–2718.

Raymond, D. J., and A. M. Blyth (1992), Extension of the stochastic mixing model to cumulonimbus clouds, *Journal of the Atmospheric Sciences*, *49*(21), 1968–1983.

Richter, J. H., and P. J. Rasch (2008), Effects of convective momentum transport on the atmospheric circulation in the community atmosphere model, version 3, *Journal of Climate*, *21*(7), 1487–1499.

Rougier, J. (2007), Probabilistic inference for future climate using an ensemble of climate model evaluations, *Climatic Change*, *81*(3-4), 247–264.

Rougier, J., D. M. Sexton, J. M. Murphy, and D. Stainforth (2009), Analyzing the climate sensitivity of the HadSM3 climate model using ensembles from different but related

experiments, *Journal of Climate*, *22*(13), 3540–3557.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989), Design and analysis of computer experiments, *Statistical science*, pp. 409–423.

Sahany, S., J. D. Neelin, K. Hales, and R. B. Neale (2012), Temperature–moisture dependence of the deep convective transition as a constraint on entrainment in climate models, *Journal of the Atmospheric Sciences*, *69*(4), 1340–1358.

Sahany, S., J. D. Neelin, K. Hales, and R. B. Neale (2014), Deep convective transition characteristics in the community climate system model and changes under global warming, *Journal of Climate*, *27*(24), 9214–9232.

Sanderson, B., C. Piani, W. J. Ingram, D. A. Stone, and M. R. Allen (2008), Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics gcm simulations, *Climate Dynamics*, *30*(2-3), 175–190.

Severijns, C. A., and W. Hazeleger (2005), Optimizing parameters in an atmospheric general circulation model, *Journal of Climate*, *18*(17), 3527–3535.

Stainforth, D. A., T. Alna, C. Christensen, M. Collins, N. Faull, D. J. Frame, J. A. Kettleborough, S. Knight, A. Martin, J. M. Murphy, C. Planl, D. M. H. Sexton, L. A. Smith, R. A. Spicer, A. J. Thorpe, and M. R. Allen (2005), Uncertainty in predictions of the climat response to rising levels of greenhouse gases, *Nature*, *433*, 403–406.

Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183–7192.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experiment design, *Bulletin of the American Meteorological Society*, *93*(4), 485–498.

Tian, B., E. J. Fetzer, B. H. Kahn, J. Teixeira, E. Manning, and T. Hearty (2013), Evaluating CMIP5 models using airs tropospheric air temperature and specific humidity climatology, *Journal of Geophysical Research: Atmospheres*, *118*(1), 114–134.

Trenberth, K. E., Y. Zhang, and J. T. Fasullo (2015), Relationships among top-of-atmosphere radiation and atmospheric state variables in observations and cesm, *Journal of Geophysical Research: Atmospheres*, *120*(19), 10,074–10,090.

Wang, G. G., and S. Shan (2007), Review of metamodeling techniques in support of engineering design optimization, *Journal of Mechanical Design*, *129*(4), 370–380.

Wielicki, B. A., B. R. Barkstrom, E. F. Harrison, R. B. Lee, G. Louis Smith, and J. E. Cooper (1996), Clouds and the earth's radiant energy system (CERES): An earth observing system experiment, *Bulletin of the American Meteorological Society*, *77*(5),

853–868.

Willmott, C. J., and K. Matsuura (2001), Terrestrial air temperature and precipitation: monthly and annual time series (1950–1999).

Woodruff, M. J., and J. D. Herman (2014), Pareto.py: Nondominated sorting for multi-objective problems, last acessed August, 2016 from https://github.com/matthewjwoodruff/pareto.py.

Xie, S.-P., C. Deser, G. A. Vecchi, M. Collins, T. L. Delworth, A. Hall, E. Hawkins, N. C. Johnson, C. Cassou, A. Giannini, and M. Watanabe (2015), Towards predictive understanding of regional climate change, *Nature Climate Change*, *5*(10), 921–930.

Yin, L., R. Fu, E. Shevliakova, and R. Dickinson (2013), How well can CMIP5 simulate precipitation and its controlling processes over tropical south america?, *Climate Dynamics*, *41*(11-12), 3127–3142.

Zhang, G. J., and N. A. McFarlane (1995), Sensitivity of climate simulations to the parameterization of cumulus convection in the canadian climate centre general circulation model, *Atmosphere-Ocean*, *33*(3), 407–446.

Zhang, Y., and H. Chen (2015), Comparing CAM5 and Superparameterized CAM5 simulations of summer precipitation characteristics over continental East Asia: mean state, frequency–intensity relationship, diurnal cycle, and influencing factors, *Journal of Climate*, *29*(3), 1067–1089.

Zheng, X., S. A. Klein, H. Y. Ma, P. Bogenschutz, A. Gettelman, and V. E. Larson (2016), Assessment of marine boundary layer cloud simulations in the CAM with CLUBB and updated microphysics scheme based on ARM observations from the Azores, *Journal of Geophysical Research: Atmospheres*, *121*(14), 8472–8492.

(b)