The art and science of climate model tuning

FRÉDÉRIC HOURDIN*

Laboratoire de Météorologie Dynamique, IPSL, CNRS, UPMC, Paris, France

THORSTEN MAURITSEN

Max Planck Institute for Meteorology, Hamburg, Germany

ANDREW GETTELMAN

National Center for Atmospheric Research, Boulder, Colorado, USA

JEAN-CHRISTOPHE GOLAZ

NOAA Geophysical Fluid Dynamics laboratory, Princeton, NJ, USA Current affiliation: Lawrence Livermore National Laboratory, Livermore, CA, USA

VENKATRAMANI BALAJI

Cooperative Institute for Climate Science, Princeton University

QINGYUN DUAN

Beijing Normal University, Beijing, China

DORIS FOLINI

Eidgenössische Technische Hochschule, Zurich, Switzerland

DUOYING JI

Beijing Normal University, Beijing, China

DANIEL KLOCKE

Deutscher Wetterdienst, Offenbach, Germany

YUN QIAN

Pacific Northwest National Laboratory, Richland, USA

FLORIAN RAUSER

Max Planck Institute for Meteorology, Hamburg, Germany

CATHRINE RIO

LORENZO TOMASSINI

Max Planck Institute for Meteorology, Hamburg, Germany (Current affiliation: Met Office, Exeter, UK)

MASAHIRO WATANABE

University of Tokyo, Tokyo, Japan

DANIEL WILLIAMSON

University of Exeter, Exeter, UK

ABSTRACT

Abstract: The process of parameter estimation targeting a chosen set of observations is an essential aspect of numerical modeling. This process is usually named tuning in the climate modeling community. In climate models, the variety and complexity of physical processes involved, and their interplay through a wide range of spatial and temporal scales, must be summarized in a series of approximate sub-models. Most sub-models depend on uncertain parameters. Tuning consists of adjusting the values of these parameters to bring the solution as a whole into line with aspects of the observed climate. Tuning is an essential aspect of climate modeling with its own scientific issues, which is probably not advertised enough outside the community of model developers. Optimization of climate models raises important questions about whether tuning methods *a priori* constrain the model results in unintended ways that would affect our confidence in climate projections. Here we present the definition and rationale behind model tuning, review specific methodological aspects, and survey the diversity of tuning approaches used in current climate models. We also discuss the challenges and opportunities in applying so-called 'objective' methods in climate model tuning. We discuss how tuning methodologies may affect fundamental results of climate models, such as climate sensitivity. The article concludes with a series of recommendations to make the process of climate model tuning more transparent.

Capsule Summary: We survey the rationale and diversity of approaches for tuning, a fundamental aspect of climate modeling which should be more systematically documented and taken into account in multi-model analysis.

1. Introduction

As is often the case in sciences that address complex systems, numerical models have become central in climate science (Edwards 2001). General circulation models of the atmosphere were originally developed for numerical weather forecasting (e. g. Phillips 1956). The coupling of global atmospheric and oceanic models began with Manabe and Bryan (1969) and came of age in the 80s and 90s. Global climate models or Earth System Models (ESMs) are nowadays used extensively to study climate changes caused by anthropogenic and natural perturbations (Lynch 2008; Edwards 2010). The evaluation and improvement of these global models is the driver of many theoretical and observational research. Publications that analyze the simulations coordinated at an international level in the frame of the Coupled Model Intercomparison Project (CMIP) constitute a large part of the material synthesized in the IPCC Assessment Reports. Beyond their use for prediction and projection at meteorological to climatic timescales, global models play a key role in climate science. They are used to understand and assess the mechanisms at work while accounting for the complexity of the climate system and for the spatial and temporal scales involved (Dahan Dalmedico 2001; Held 2005).

The development of a climate model is a long-term project. When releasing a new model or new version of a model, a series of sub-models, sometimes developed or improved over years in separate teams, are combined and optimized together to produce a climate that matches some key aspects of the observed climate. While the fundamental physics of climate is generally well established, sub-models or parameterizations are approximate, either because of numerical cost issues (limitations in grid resolution, acceleration of radiative transfer computation) or more fundamentally because they try to summarize complex and multi-scale processes through an idealized and approximate representation. Each parameterization relies on a set of internal equations and often depends on parameters, the values of which are often poorly constrained by observations. The process of estimating these uncertain parameters in order to reduce the mismatch between specific observations and model results is usually referred to as tuning in the climate modeling community.

Climate model tuning is a complex process which presents analogy with reaching harmony in music. Producing a good symphony or rock concert requires first a good composition and good musicians who work individually on their score. Then, when playing together, instruments must be tuned, which is a well defined adjustment of wave frequencies which can be done with the help of electronic devices. But the orchestra harmony is reached also by adjusting to a common tempo as well as by subjective combinations of instruments, volume levels or musicians interpretations, which will depend on the intention of the conductor or musicians. When gathering the various pieces of a model to simulate the global climate, there are also many scientific and technical issues, and tuning itself can be defined as an objective process of parameter estimation to fit a predefined set of observations, accounting for their uncertainty, a process which can be engineered. However, because of the complexity of the climate system and of the choices and approximations made in each sub-model, and because of priorities defined in each climate center, there is also subjectivity in climate model tuning (Tebaldi and Knutti 2007) as well as substantial know-how from a limited number of people with vast experience with a particular model. One goal of this paper is to make this knowledge more explicit.

Choices and compromises made during the tuning exercise may significantly affect model results and influence evaluations that measure a statistical 'distance' between

*Corresponding author address: Frederic Hourdin, Laboratoire de Météorologie Dynamique, IPSL, 4UPMC, Tr 45-55, 3e et, B99, Jussieu, Paris, France.
E-mail: frederic.hourdin@lmd.jussieu.fr

the simulated and observed climate. In theory, tuning should be taken into account in any evaluation, intercomparison or interpretation of the model results. Although the need for parameter tuning was recognized in pioneering modeling work (e.g. Manabe and Wetherald 1975) and discussed as an important aspect in epistemological studies of climate modeling (Edwards 2001), the importance of tuning is probably not advertised as it should. It is often ignored when discussing the performances of climate models in multi-model analyses. In fact, the tuning strategy was not even part of the required documentation of the CMIP5 simulations. In the best cases, the description of the tuning strategy was available in the reference publications of the modeling groups (Mauritsen et al. 2012; Golaz et al. 2013; Hourdin et al. 2013a,b; Schmidt et al. 2014). Why such a lack of transparency ? Maybe because tuning is often seen as an unavoidable but dirty part of climate modeling; more engineering than science; an act of tinkering that does not merit recording in the scientific literature. There may also be some concern that explaining that models are tuned, may strengthen the arguments of those claiming to question the validity of climate change projections. Tuning may be seen indeed as an unspeakable way to compensate for model errors.

The purpose of this paper is to help making the process of model tuning more explicit and transparent. Tuning is an intrinsic and fundamental part of climate modeling that should be better documented and discussed as such in the scientific literature. Tuning can be described as an optimization step and follows a scientific approach. Tuning can provide important insights on climate mechanisms and model uncertainties. Some biases in climate models can be reduced or removed by tuning, while others remain stubbornly resistant. It is important to understand why if we want to improve models. Below, we present a definition of tuning, document current practices and methodologies, and address emerging issues. We conclude with recommendations on model tuning and its documentation.

2. Definition of climate model tuning

Model tuning or calibration is neither a new concept nor specific to climate modeling. In statistical sciences, Fisher introduced three steps in the process of modeling (Fisher 1922; Burnham and Anderson 2002): (i) model formulation, (ii) parameter estimation, and (iii) estimation of uncertainty. This categorization applies also to the wider context of numerical modeling. It is conceptually useful to discriminate between model formulation and parameter estimation, even if this distinction is by no means clear-cut in climate model tuning as explained below.

Climate model development is founded on well-understood physics combined with a number of heuristic process representations. The fluid motions in the atmosphere and ocean are resolved by the so-called 'dynamical core' down to a grid spacing of typically 25 to 300 km for global models, based on numerical formulations of the equations of motion from fluid mechanics. Subgrid scale turbulent and convective motions must be represented through approximate sub-gridscale parameterizations (Smagorinsky 1963; Arakawa and Schubert 1974; Edwards 2001). These sub-gridscale parameterizations include coupling with thermodynamics, radiation, continental hydrology, and optionally chemistry, aerosol microphysics, or biology.

Parameterizations are often based on a mixed physical, phenomenological and statistical view. For example, the cloud fraction needed to represent the mean effect of a field of clouds on radiation may be related to the resolved humidity and temperature through an empirical relationship. But the same cloud fraction can also be obtained from a more elaborate description of processes governing cloud formation and evolution. For instance, for an ensemble of cumulus clouds within an horizontal grid cell, clouds can be represented with a single mean plume of warm and moist air raising from the surface (Tiedtke 1989; Jam et al. 2013) or with an ensemble of such plumes (Arakawa and Schubert 1974). Similar parameterizations are needed for many components not amenable to first-principle approaches at the grid scale of a global model, including boundary layers, surface hydrology, ecosystem dynamics, and so on. Each parameterization, in turn, typically depends on one or more parameters whose numerical values are poorly constrained by first principles or observations at the grid scale of global models. Being approximate descriptions of unresolved processes, there exist different possibilities for the representation of many processes. The development of competing approaches to different processes is one of the most active areas of climate research. The diversity of possible approaches and parameter values is one of the main motivations for model intercomparison projects in which a strict protocol is shared by various modeling groups in order to better isolate the uncertainty in climate simulations that arises from the diversity of models (model uncertainty).

A model configuration is determined by two aspects, its complexity and resolution. For global climate models or ESMs, the configuration retained generally results from compromises between resolution, complexity and length and number of simulations. Different modeling groups may have different priorities in terms of scientific questions and applications, thus making different judgments on how to best balance finite resources. The choice of complexity and resolution itself can be considered as tuning in a wide sense, since it is often motivated by the ability of the model to reproduce with some realism key aspects of the climate system.

Here we focus on the classical definition of tuning, that corresponds to parameter estimation in Fisher's terminology. Once a model configuration is fixed, tuning consists in choosing parameter values in such a way that a certain measure of the deviation of the model output from selected observations or theory is minimized or reduced to an acceptable range. Defined this way, tuning is usually called calibration in other application areas of complex numerical models (Kennedy and O'Hagan 2001). Some climate modelers are reluctant to use this term however since they know that, by adjusting parameters, they also compensate, intentionally or not, for some (often unknown) deficiencies in the model formulation itself.

Parameter tuning itself occurs at various levels that correspond to stages of model development. An initial calibration may be performed during the development phase of a new parameterization, for instance using a single column version of the climate model. Although desirable in principle, this *parameterization tuning* is often difficult in practice because processes are strongly coupled to each other and to the large-scale dynamics. At the next stage, a number of parameterizations are tuned together when assembled into components: atmosphere, ocean, continental surface. This *component tuning* is performed by using standalone components with boundary conditions which would otherwise be provided by other components. For example, an ocean model with imposed surface wind stress, inputs of freshwater, precipitation and radiation might be tuned to get sea surface temperatures or meridional overturning circulation that match expectations. A *system tuning* is finally required to ensure consistency across the full climate system once components are coupled together.

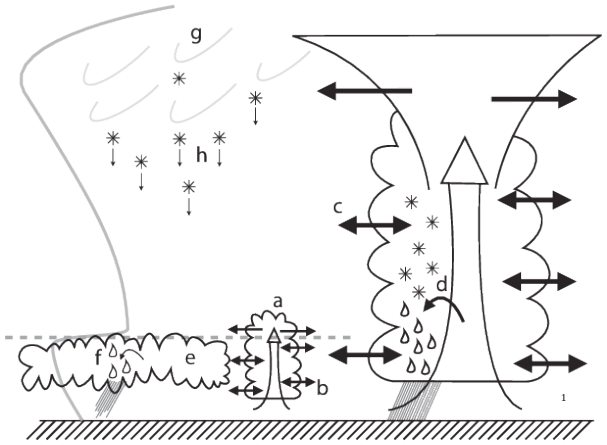

FIG. 1. Example of tuning approach for the ECHaM model (after Mauritsen et al. 2012). The figure illustrates the major uncertain climate-related cloud processes frequently used to tune the climate of the ECHAM model. Stratiform liquid and ice clouds, and shallow and deep convective clouds are represented. The grey curve to the left represents tropospheric temperatures and the dashed line is the top of the boundary layer. Parameters are a) convective cloud mass-flux above the level of non-buoyancy, b) shallow convective cloud lateral entrainment rate, c) deep convective cloud lateral entrainment rate, d) convective cloud water conversion rate to rain, e) liquid cloud homogeneity, f) liquid cloud water conversion rate to rain, g) ice cloud homogeneity, and h) ice particle fall velocity.

3. Common practices and targets

Tuning of coupled earth system models generally follows a common practice but with targets and priorities which may vary from group to group. This was confirmed by a poll conducted in August-September 2014 (See sidebar for results). Most of the major climate modeling groups (23 model centers) submitted answers to a questionnaire on why and how their models are tuned.

With the increasing diversity in the applications of climate models, the number of potential targets for tuning increases. There are a variety of goals for specific problems, and different models may be optimized to perform better on a particular metric, related to specific goals, expertise or cultural identity of a given modeling center. Groups more focused on the European climate may give more importance to the ocean heat transport in the North Atlantic whereas others may be more concerned with tropical climate and convection. Some groups may put more weight on metrics that measure the skill to reproduce the present-day mean climatology or observed modes of variability, while others may privilege process-oriented metrics targeting processes that are believed to dominate the climate change response to anthropogenic forcing.

There is, however, a dominant shared target for coupled climate models: the climate system should reach a mean equilibrium temperature close to observations when energy received from the sun is close to its real value ($\simeq 340$ W/m^2). This energy source will be balanced by the energy lost to space by reflected sunlight and thermal infrared radiation if the model conserves energy numerically (which can not always been imposed strictly). We know indeed that the system is nearly in balance but for the ocean heat uptake, believed to be of about 0.5 W/m^2 in our warming climate, a value much smaller than the model and observational uncertainties. This provides a strong large-scale constraint.[1]

A common practice to fulfill this constraint is to adjust the top-of-atmosphere or surface[2] energy balance in atmosphere-only simulations exposed to observed sea surface temperatures (component tuning) and check if the temperature obtained in coupled models is realistic. This energy balance tuning is crucial since a change by 1 W/m^2 of the global energy balance produces typically a change

[1] Even observations of the radiative fluxes are in fact adjusted using this constraint. The CERES-EBAF data stands for 'energy balance adjusted flux'.

[2] Top-of-atmosphere and surface energy balance should not differ if exact energy conservation in the atmosphere is ensured, which turns out not to be an easy task.

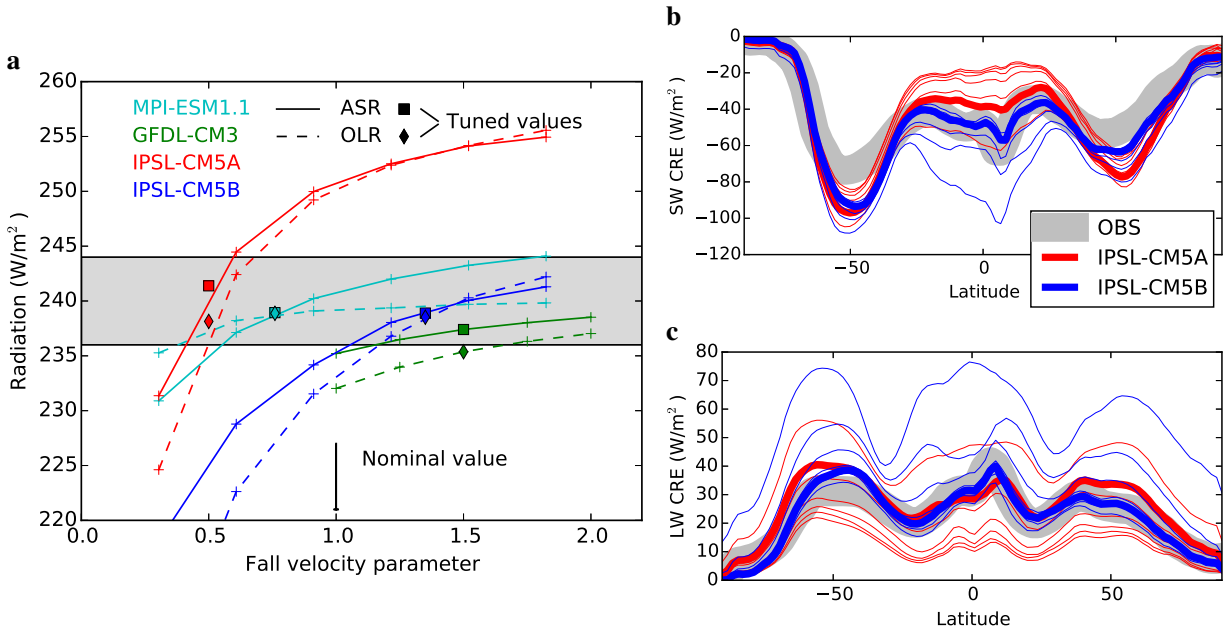

FIG. 2. Example of tuning of the global top-of-atmosphere energy balance with a cloud parameter for the GFDL-CM3, MPI-ESM1.1 and two versions A and B of the IPSL-CM5 model, that differ by the representation of the convective boundary layer, clouds and convection. **a)** Global absorbed short-wave radiation (ASR, full curve) and outgoing radiation (OLR, dashed) at top-of-atmosphere. The horizontal axis corresponds to the value of a scaling parameter in the ice crystal fall velocity equation, Eq. (5) of Heymsfield and Donner (1990) which is shared by the four models. The simulations are run over several years with imposed sea surface temperature. The difference between the dashed and full curves gives the global energy balance. The squares and diamonds correspond to default values retained after a tuning phase (for GFDL and IPSL-CM they correspond to the values retained for CMIP5 but, because the experiments were redone with recent versions of the same models, the balance is not completely satisfied with the selected values). For the IPSL models, we show how the tuning of the scaling parameter affects the latitudinal variation of cloud radiative effect computed as the difference of total and clear-sky radiation, for both **b)** short-wave and **c)** long-wave radiation. The thin curves correspond to the various values of the tuning parameter (the smaller the fall velocity the stronger the absolute cloud radiative effect both in the long-wave and short-wave radiation) and the thick curves to the values retained after tuning. The observations correspond to the CERES-EBAF L3b product for (Loeb et al. 2009). The height of the gray rectangle in **a)** and thickness of the gray curves in **b)** and **c)** correspond to an observation uncertainty of ± 4 W/m^2. Note however that true error bars are not available for these observations

of about 0.5 to 1.5 K in the global mean surface temperature in coupled simulations depending on the sensitivity of the given model.

In general, the parameters are given some a priori values, and ideally a range around this value. This information can come either from theory, from a back-of-the-envelope estimate, from numerical experiments (tuning an eddy-diffusion coefficient from explicit simulations of the turbulent process) or from observations (a mean effective cloud droplet for instance). Note that many internal parameters are not directly observable. Given this information, a common practice is to adjust the most uncertain parameters that significantly affect key climate metrics. Indeed, all parameters are not known with the same accuracy. There is fair consensus (see poll) that the most uncertain parameters that affect the atmospheric radiation are those entering in the parameterization of clouds and of the albedo of the earth's surface. Clouds exert a large net cooling effect (about -20 W/m^2), but this effect is uncertain to within several W/m^2 (Loeb et al. 2009). A 1 W/m^2 change in cloud radiative effects is only a 5% variation of the net cloud cooling effect, and 2% of the solar (or short-wave) effect, well below observational and model uncertainty (L'Ecuyer et al. 2015).

Most tuning parameters are specific to sub-model (parameterization) choices. Parameters controlling mixing of convective clouds with the environment will depend on the specific description of the convective vertical transport. Parameters controlling the size distribution of cloud droplets which will depend on the sophistication of the microphysics, et cetera. As an example, Fig 1 from Mauritsen et al. (2012) illustrates the various parameters which are used for tuning in one particular model.

Some parameterizations and associated tuning parameters are however shared by several models. We show in Fig 2 how a scaling factor on the ice crystal fall velocity (process **h** in Fig 1) is used to constrain both the global short-wave and long-wave radiation to match observed value of 240$\pm$4 W/m^2, in climate models that share the same formulation for the ice crystal fall velocity (Heymsfield and Donner 1990). A larger fall velocity systematically reduces the amount of ice clouds and thus increases both the absorbed short-wave radiation (reduced planetary albedo) and outgoing long-wave radiation (reduced

greenhouse effect). Beyond global values, tuning is sometimes applied to spatial variations of the radiative fluxes like latitudinal dependency that drives the general circulation or land-sea contrasts that drive monsoon circulations. Fig 2b,c illustrates for two models how the same factor on ice crystal fall velocity affects the latitudinal distribution of absorbed solar radiation and outgoing long-wave radiation.

After clouds, the most common tuning parameters are those entering in the parameterizations of snow and sea-ice albedo, ocean mixing and orographic drag. Soil and vegetation properties are also sometimes used for tuning.

Because of the uncertainties in observations and in the model formulation, the possible parameter choices are numerous and will differ from one modeling group to another. These choices should be more often considered in model inter-comparison studies. The diversity of tuning choices reflects the state of our current climate understanding, observation and modeling. It is vital that this diversity be maintained. It is however important that groups better communicate their tuning strategy. In particular, when comparing models on a given metric, either for model assessment or for understanding of climate mechanisms, it is essential to know whether some models used this metric as tuning target.

4. Applying objective methods

There exists a considerable literature on parametric tuning using objective approaches, developed in the statistics, engineering and computer science communities. By 'objective' methods, one means that a well founded mathematical or statistical framework is used to perform the model tuning, for instance by defining and minimizing a cost function or by introducing a Bayesian formulation of the calibration problem (Kennedy and O'Hagan 2001). The use of objective methods does not, however, in any way obviates the requirement for subjective judgment concerning the priorities and targets of the tuning process. An 'objective' algorithm merely identifies those parts of the procedure that require the subjective scientific expertise of the modeler. It requires that the modeler formulates this judgment in terms of numbers or mathematical formulas, which can be sometimes quite demanding but which also contributes to make the process of tuning more explicit and reproducible. Objective methods then provide an automatic tuning procedure based on those judgments.

Broadly speaking, objective methods fall into one of two categories. The first involves fast optimization of some cost function measuring the distance of model simulations to a small collection of observations. Applications of such methods in climate science include Bellprat et al. (2012); Yang et al. (2013); Zou et al. (2014); Zhang et al. (2015). The second class of methods represents a Bayesian approach and is now part of a class of methods under the banner of Uncertainty Quantification (UQ, Kennedy and O'Hagan 2001). UQ, for parameter tuning, aims to provide uncertainty for the parameters using a statistical model relating the climate model to observations that explicitly quantifies the key sources of uncertainty present in the problem: observational uncertainty, initial condition uncertainty (internal variability) and structural uncertainty (missing or incorrect physics). Applications of these methods to climate models include Rougier (2007); Jackson et al. (2008); Edwards et al. (2011); Williamson et al. (2013). UQ methods for example were used to provide the UK Climate Projections (Murphy et al. 2009; Sexton et al. 2012).

Both classes of objective methods (optimization and UQ) share advantages over more arbitrary trial and error approaches that focus on tuning only one or two parameters at a time. For example, by perturbing multiple parameters simultaneously and systematically, automatic methods can overcome concerns that a local optimum for one objective may not be a good solution for other objectives and may not even be a global optimum for the tuning metric (Qian et al. 2015; Williamson et al. 2015).

Both classes of methods also share some of the same challenges. The main challenge is computational cost of running the climate model with sufficient parameter choices to explore the parameter space. For high-resolution climate models (or even their components), available supercomputing power and the time available between tuning cycles – typically on the order of one to a few years between two model releases – limits even the best equipped institutions.

To overcome these computational issues, statistical emulators (also called meta-models) can be used. Developed by statisticians since the late 1980s (Sacks et al. 1989; Currin et al. 1991; Haylock and O'Hagan 1996), emulators use small training ensembles to train statistical models that can predict the climate model response very quickly (Neelin et al. 2010), reporting a measure of uncertainty (typically offering a full probability distribution for the climate model at any choice of the parameters). The emulator uncertainty must be included in Bayesian UQ methods for parameter tuning, though it is ignored in some applications of optimization methods with the emulator mean function used directly.

For high resolution models and models with long spin up time, running the model enough to build an emulator represents a huge challenge. Ensembles of shorter simulations to replace the traditional serial-in-time long-term climatology simulations have been proposed (Wan et al. 2014) and the UQ literature has long proposed and demonstrated the success of linked models of different resolution to build emulators. For example, Williamson et al. (2012) built an emulator for the CMIP5 model HadCM3 using only 16 integrations and a large ensemble of the

low-resolution version FAMOUS. This is an active area of research in UQ.

A principal challenge for automatic tuning methods is that tuning to a handful of metrics may risk achieving improved performance in those metrics at the expense of unphysical behavior in metrics or processes that were not used in tuning, i.e., we get some things 'right for the wrong reasons'. This problem, known as over-fitting or over-tuning, will arise as soon as a minimization or parameter selection is done that does not properly account for the observation and model structural uncertainties. It will also arise when tuning to partial observations (i.e. not tuning the whole state vector of the climate model), or over-fitting data that is partly simply natural variability (Notz 2015). Then tuning may be seen as an error compensation process rather than as model callibration. Over-tuning can also occur when tuning 'by hand', but blind trust in an automatic tool may be more risky in that it prevents from exercising the part of the expert judgment which can not easily be translated into objective functions, or expressed mathematically as uncertainties.

Over-tuning is a real concern and the raison d'être for Bayesian UQ methods. However, because the key sources of uncertainty in the tuning problem, observation uncertainty and structural error, are so poorly understood and difficult to quantify, automatic tuning has a long way to go before it is adopted routinely by the major modeling centers for CMIP integrations. A class of UQ methods that explicitly avoid over-tuning, called history matching, have recently been proposed for the climate model tuning community (Williamson et al. 2015). They avoid over-tuning by changing the problem from one of searching for a single best value of the parameters, to looking for unacceptable parameter values and ruling out the corresponding regions of the parameter space iteratively.

5. Tuning and model improvement

Although tuning is an efficient way to reduce the distance between model and selected observations, it can also risk masking fundamental problems and the need for model improvements.

There is evidence that a number of model errors are structural in nature and arise specifically from the approximations in key parameterizations as well as their interactions. For example, some models systematically underestimate rainfall over monsoon regions, whereas others will do the opposite. Other biases are systematic across models, like the presence of a persistent double Pacific Intertropical Convergence Zone (ITCZ), on both sides of the equator, or warm biases over the Eastern tropical oceans. Those model biases are indeed often resistant to model tuning. Tuning a model to improve its performance on a specific target also often degrades performance on other metrics. For example, tuning a model to improve the intraseasonal variability of precipitation in the tropics often comes at the cost of increased biases in the mean state (Kim et al. 2012).

Introduction of a new parameterization or improvement also often decreases the model skill on certain measures. The pre-existing version of a model is generally optimized by both tuning uncertain parameters and selecting model combinations giving acceptable results, probably inducing compensation errors (over-tuning). Improving one part of the model may then make the 'skill' relative to observations worse, even though it has a better formulation. The stronger the previous tuning, the more difficult it will be to demonstrate a positive impact from the model improvement and to obtain an acceptable retuning. In that sense, tuning (in case of over-tuning) may even slow down the process of model improvement by preventing the incorporation of new and original ideas. This difficulty has been known for decades in operational numerical weather prediction centers and could be overcome by not overweighting climate performance metrics (the ones which matter for the end users or for impact models) with respect to process-oriented ones. Process-oriented metrics are intended to help relate large-scale biases to the misrepresentation of specific sub-grid scale processes. Process oriented metrics include, for example: compositing cloud or precipitation characteristics by dynamical regimes (Bony et al. 2004), compositing relative humidity profiles based on precipitation percentiles to assess the sensitivity of convection schemes to relative humidity (Kim et al. 2014), or evaluating simulated cloud microphysical properties (and their co-variability) directly from satellite measurements (Suzuki et al. 2013).

On another hand, tuning may highlight where further model improvement is needed. If parameter values needed to satisfy a given metric are outside the acceptable range, or if different values are needed for different regions or climate regimes, developers may consider revisiting the formulation of the parameterization or develop new ones. Then, the tuning process can be pushed back to a deeper level inside the model while increasing the physical realism of the model.

For clouds and convection, parameterization development is often performed using single column versions of the global model compared to explicit high resolution simulations of the processes which are parameterized, following a strategy defined 20 years ago (see e. g. Ayotte et al. 1996; Liu et al. 2001). The explicit simulation gives access to variables hardly accessible by observation (like 3D fields of temperature and humidity or vertical velocities) but also to estimation of parameters which have no observational counterpart (like entrainment and detrainment rates between a mean bulk plume and its environment or a mean fall velocity for ice crystals at the model grid scale).

Such parameters can be derived by sampling and characterizing the equivalent of the parameterized structures in the explicit simulations, as done for example by Couvreux et al. (2010) to derive mixing rates between a mean bulk plume and its environment. The parameterization development process can thus help constrain some parameters but also propose physically-based sub-models for some others.

One way to make the reduction of model large scale biases and the parameterization development processes more "in tune" is by deriving an acceptable range of parameter values instead of a single value from the aforementioned process studies and use this range when tuning global simulations. To achieve this goal, UQ methods could be applied to the single-column model using explicit process simulations as a reference. It is important that the representation of turbulence, microphysics and radiation continue to be improved in explicit high resolution simulations, so that the parametrization can be evaluated not only in terms of subgrid-scale dynamics (as usually done so far) but also in terms of radiative effect of clouds.

Another emerging approach consists in using initialized or nudged simulations (Zhang et al. 2014) in the tuning process. In nudged simulations the model is forced to follow the observed trajectory by relaxing winds and also optionally temperature and humidity, toward meteorological analysis, with a time constant of typically a few hours. With initialized or nudged simulations, the simulated and observed meteorology follow the same trajectory and the comparison with observations can be done on a day-by-day basis. Wind-only nudging allows separation of parameterization tuning for a given meteorological situation (as is done in 1D mode) from that of the coupling of parameterization with large scale dynamics. Nudging with short enough time constants (typically of a few hours) removes the chaotic nature of the atmospheric large scale circulation, and slow feedbacks of that circulation on fast processes (such as clouds). Nudged or initialized simulations may also help accelerate tuning for high resolution climate models.

Whatever the approach, there is a need for relying more on observational studies at the process-scale to tune the radiative budget in a more physical way. Progress will be made by further incorporating model tuning as an uncertainty analysis into the parameterization development process.

6. Tuning to 20th century warming ?

The increase of about one Kelvin of the global mean temperature observed from the beginning of the industrial era, hereafter 20th century warming, is a *de facto* litmus test for climate models (Mauritsen et al. 2012). However, as a test of model quality, it is not without issues because the desired result is known to model developers and therefore becomes a potential target of the development.

The amplitude of the 20th century warming depends primarily on the magnitude of the radiative forcing, the climate sensitivity, as well as the efficiency of ocean heat uptake. By linearizing about a basic stationary climatic state, the global mean temperature change for a gradually increasing forcing can be approximated as:

$$\Delta T \approx \frac{F}{\kappa - \lambda} \tag{1}$$

where T denotes global mean surface temperature, F an imposed radiative forcing, κ the deep ocean heat uptake efficiency, and λ is the feedback parameter which is inversely proportional to equilibrium climate sensitivity ($ECS \approx -F/\lambda$). Climate models have values of λ that range from -0.6 to -1.8 Wm^{-2}/K and κ from approximately 0.5 to 1.2 Wm^{-2}/K. On average, in models the denominator $(\kappa - \lambda)$ is about 2 Wm^{-2}/K and in year 2003 forcing is around 1.7 Wm^{-2} (Forster et al. 2013).

The often deployed paradigm of climate change projection is that climate models are developed using theory and present-day observations, whereas ECS is an emergent property of the model and the matching of the 20th century warming constituting an *a posteriori* model evaluation. Some modeling groups claim not to tune their models against 20th century warming, however, even for model developers it is difficult to ensure that this is absolutely true in practice because of the complexity and historical dimension of model development.

The reality of this paradigm is questioned by findings of Kiehl (2007) who discovered the existence of an anti-correlation between total radiative forcing and climate sensitivity in CMIP3 models: High sensitivity models were found to have a smaller total forcing and low sensitivity models a larger forcing, yielding less cross-ensemble variation of historical warming than otherwise to be expected. Even if alternate explanations have been proposed and even if the results were not so straightforward for CMIP5 (cf. Forster et al. 2013), it could suggest that some models may have been inadvertently or intentionally tuned to the 20th century warming.

There is a broad spectrum of methods to improve model match to 20th century warming, ranging from simply choosing to no longer modify the value of a sensitive parameter when a match is already good for a given model (Mauritsen et al. 2012), or selecting physical parameterizations that improve the match, to explicitly tuning either forcing or feedback both of which are uncertain and depend critically on tunable parameters (Murphy et al. 2004; Golaz et al. 2013). Model selection could, for instance, consist of choosing to include or leave out new processes, such as aerosol-cloud interactions, to help the model better match the historical warming, or choosing to work on or

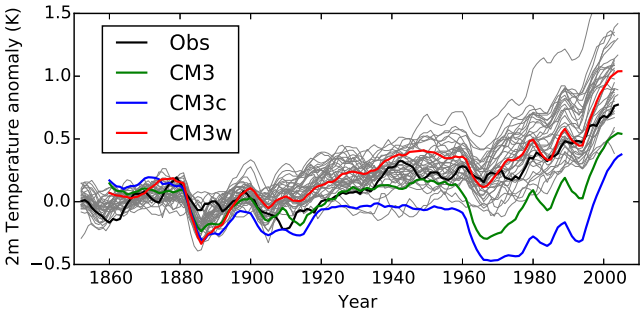

FIG. 3. Simulations of the 20th century temperature with the CMIP5 model Ensemble (grey curves). Each curve corresponds to a 5-year running mean of the anomaly of the global mean temperature at 2m above surface. The anomaly is computed using as a reference period years 1850-1899. The black curve corresponds to the version 4 of the HadCRUT observations. The colored curves correspond to 3 configurations of the GFDL-CM3 model. CM3 denotes the CMIP5 model, while CM3c and CM3w denote alternate configurations with larger, respectively smaller, cooling from cloud aerosol interactions.

replace a parameterization that is suspect of causing a perceived unrealistically low or high forcing or climate sensitivity.

An illustration of 20th century tuning with the GFDL-CM3 model is shown in Fig 3. The model (green) produces a relatively weak warming over the 20th century due to a strong cooling effect from aerosol-cloud interactions. Sensitivity tests, which were performed after the model was frozen, showed that it is possible to reduce this effect and thereby obtain a more realistic warming. However, this was achieved by lowering the threshold size for the conversion of cloud droplets to rain to values smaller than supported by observations (Golaz et al. 2013; Suzuki et al. 2013, and references therein).

Adjusting the 20th century warming would in principle require a series of multi-century simulations with the coupled ocean-atmosphere model, because of the long spin-up of the ocean state required before starting transient 20th century simulations. However, it has long been known that short atmospheric simulations can be used to estimate either adjusted forcing when forced with perturbed atmospheric composition (Hansen et al. 2005) or ECS when forced with perturbed sea surface temperature (Cess et al. 1989; Gettelman et al. 2012). Thereby it is possible to target specific values of F and λ thought to provide a good match to historical warming based on experience with previous model versions.

Any ECS tuning would need to take into account three main sources of uncertainties. First as usual, the uncertainty of the observation of the global mean surface temperature should not be forgotten even if it is believed today to be much smaller than the inter model dispersion. Then the radiative forcing F itself is uncertain. It is composed of a fairly well-known greenhouse gas forcing which is partly compensated by an uncertain aerosol forcing, and modified by a series of other less important forcing agents. Tuning of the 20th century could for instance be obtained with an overly large ECS balancing an overly strong aerosol radiative forcing. In such a case, and because the effect of greenhouse gases will dominate in the future, this would result in an overestimate of future global warming. The third important source of uncertainty comes from the internal climate variability which can cause variations among realizations with different initial conditions of typically ± 0.1 K to centennial warming; and since the observed only represents one such realization a model need not be closer than this to match the target. Trying to match the 20th century global warming without accounting for sources of uncertainty would inevitably lead to over-tuning.

The question whether the 20th century warming should be considered a target of model development or an emergent property is polarizing the climate modeling community, with 35 percent of modelers stating that 20th century warming was rated *very important* to *decisive*, whereas 30 percent would *not consider* it at all during development. Some view the temperature record as an independent evaluation data set not to be used, while others view it as a valuable observational constraint on the model development. Likewise, opinions diverge as to which measures, either forcing or ECS, are legitimate means for improving the model match to observed warming. The question of developing towards the 20th century warming therefore is an area of vigorous debate within the community.

However, the capability to control the modeled 20th century warming also offers new opportunities to explore the bounds of modeled climate sensitivity (Golaz et al. 2013): By combining altered ECS and aerosol forcing it is technically possible to construct outlier low- and high-sensitivity models that match the observed warming. Evaluating such models with other observed aspects, such as mid-century warming or modes of variability, and running them in pre-historic climates such as the last glacial maximum or the Pliocene, could potentially allow us to rule out extreme values of ECS and/or aerosol forcing.

The fact that some models are explicitly, or implicitly, tuned to better match the 20th century warming, while others may not be, clearly complicates the interpretation of the results of combined model ensembles such as CMIP. The diversity of approaches is unavoidable as individual modeling centers pursue their model development to seek their specific scientific goals. It is, however, essential that decisions affecting forcing or feedback made during model development be transparently documented.

7. Conclusions, implications and recommendations

There was a debate among authors on the idea of using the word ”art” in the title of the paper. Tuning is

seen by some modelers more as a pure engineering calibration exercise, which consists in applying objective or automatic tools, based on purely scientific considerations. Others see it as an experienced craftsmanship or as an art "a skill that is attained by study, practice, or observation" [3]. As in art, there is also some diversity and subjectivity in the tuning process because of the complexity of the climate system, and because of the choices made among the equally possible representations of the system. It is essential to maintain this diversity in model approaches and tuning because of the approximate nature of models, the lack of observational counterparts for many internal model parameters, and the importance of climate change predictions, for which no observation exist.

This subjectivity does not contradict the fundamental and two-fold scientific nature of climate tuning. On one side, the tuning process involves many scientific issues like the physical understanding of the phenomena to be modeled, algorithmic formulation of physical laws, mathematical basis of optimization, the statistics of internal variability. In turn, the understanding of climate mechanisms can be inspired by the act of tuning which is based intrinsically on a large exploration of possible climates through sensitivity experiments. It allows us to identify and understand the role of the various modeled processes and feedbacks involved. Tuning may also help identify model structural errors, for instance if the optimal value of a parameter falls outside the acceptable range, or if different values of the same parameter are optimal for different situations. In this sense, tuning is a form of uncertainty analysis.

Because tuning will affect the behavior of a climate model, and the confidence that can be given to a particular use of that model, it is important to document the tuning portion of the model development process. We recommend that for the next CMIP6 exercise, modeling groups provide a specific document on their tuning strategy and targets, that would be referenced to when accessing the dataset. We recommend distinguishing three levels in the tuning process: individual parameterization tuning, component tuning and climate system tuning. At the component level, emphasis should be put on the relative weight given to climate performance metrics versus process oriented ones, and on the possible conflicts with parameterization level tuning. For the climate system tuning, particular emphasis should be put on the way energy balance was obtained in the full system: was it done by tuning the various components independently, or was some final tuning needed? The degree to which the observed trend of the 20th century was used or not for tuning should also be described. Comparisons against observations, and adjustment of forcing or feedback processes should be noted. At each step, any occasion where a team had to struggle with a parameter value or push it to its limits to solve a particular model deficiency should be emphasized. This information may well be scientifically valuable as a record of the uncertainty of a model formulation.

It would also be valuable to produce and document two or more versions of the same model which would differ only by their tuning. One can imagine changing a parameter which is known to affect the sensitivity, keeping both this parameter and the ECS in the anticipated acceptable range, and retuning the model otherwise with the same strategy toward the same targets.

Finally, development of new methodologies is strongly encouraged. Some of the most promising ideas include (1) the systematic use of single column versus explicit simulations approach for parameterization tuning, (2) the use of process oriented metrics and (3) nudged simulations to fill the gap between parameterization and component tuning. The systematic use of objective methods at the process level in order to estimate the range of acceptable parameters values for tuning at the upper levels is probably one strategy which should be encouraged and may help make the process of model tuning more transparent and tractable.

There is a legitimate question on whether tuning should be performed preferentially at the process level, and the global radiative budget and other climate metrics used for a posteriori evaluation of the model performance. It could be a good way to evaluate our current degree of understanding of the climate system and to estimate the full uncertainty in the ECS. Restricting adjustment to the process level may also be a good way to avoid compensating model structural errors in the tuning procedure. However, because of the multi-application nature of climate models, because of consistency issues across the model and its components, because of the limitations of process studies metrics (sampling issues, lack of energy constraints), and also simply because the climate system itself is not observed with sufficient fidelity to fully constrain models, an a posteriori adjustment will probably remain necessary for a while. This is especially important for the global energy constraints that are a strong and fundamental aspect of global climate models. Adjustment will be done usually by tuning the most uncertain parameters involved in the representation of processes that most affect radiation such as cirrus clouds or low clouds within acceptable ranges. Tuning will probably induce some compensation of shortcomings or errors in the model parameterizations or configuration. However this error compensation is probably unavoidable and desirable for current models, due to the importance of the energetic tuning for a reasonable simulation of most aspects of the climate system. The level of accuracy required for the global energy tuning (of a few tenths of W/m^2) is for instance smaller than the error arising from not computing radiation at every time-step, as often done to save computational means (of the order of

[3] https://www.ahdictionary.com/word/search.html?q=art [www.ahdictionary.com]

several W/m^2, see e. g. Balaji et al. 2016). It is recommended however to ensure that the final global tuning is not obtained for a set of parameter values which would not be acceptable in terms of process studies and process oriented metrics.

The use of objective methods could also be promoted at all the stages of model tuning, in order to render the process more efficient. However, objective tuning approaches should be used with caution. Because of the approximate nature of models and because of observations uncertainties, it is impossible to retain one unique parameter set as an objective criteria. Formalizing the question of tuning addresses an important concern: it is essential to explore the uncertainty coming both from model structural errors, by favoring the existence of tens of models, and from parameter uncertainties by not over-tuning. Either reducing the number of models or over-tuning, especially if an explicit or implicit consensus emerges in the community on a particular combination of metrics, would artificially reduce the dispersion of climate simulations. It would not reduce the uncertainty, but only hide it.

We end by expressing the hope that this article will encourage both a systematic effort by the community to document this arcane aspect of model construction, and for more people to join a vigorous debate on model tuning and evaluation.

Acknowledgments. The authors would like to thank the World Climate Research Program and its Working Group on Coupled Modeling for initiating and helping organize the workshop on model tuning in October 2014 in Garmisch-Partenkirchen, Germany. Work at LLNL was performed under the auspices the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract No. DE-AC52-07NA27344. The National Center for Atmospheric Research is supported by the U.S. National Science Foundation. The contribution of Yun Qian was supported by the U.S. Department of Energy's Office of Science as part of the Earth System Modeling Program. The Pacific Northwest National Laboratory is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830.

APPENDIX

A1. SIDEBAR : How do modeling centers tune their models?

A survey was conducted in August-September 2014, polling 23 different modeling centers that develop coupled atmosphere and ocean models to find out how they tune models. Most centers had a number of people discuss the answers before submission (one answer per group). The full results can be found in the Supplementary Material. 22 of 23 groups reported adjusting model parameters to achieve desired properties such as radiation balance at the top of the atmosphere. Percentages are reported based on the fraction of respondents. 83% of centers use atmosphere & land only (fixed sea surface temperatures or a 'data-ocean') to adjust parameters, and 44% use single column models. 74% perform their adjustment with a pre-industrial (1850) coupled atmosphere-ocean configuration. 39% use coupled present day simulations. Many groups also adjust ocean (48%) and land (39%) model parameters using stand alone configurations. In addition, 21% use historical 20th century simulations, and 17% use slab ocean models.

The goals of tuning are fairly uniform. Groups were asked about 26 different metrics: a wide variety. About one third (8 of 26) of the metrics were rated as decisive or very important by at least one third (35%) of modeling centers. However, there was lots of agreement in the decisive (most important) metrics: global net top of atmosphere flux (69%) and then global mean surface temperature (26%). Based on these goals of tuning, there are a number of different parameterizations adjusted to achieve them. Since tuning is generally focused on the top of atmosphere and surface radiation balance, the most common properties adjusted are uncertain cloud properties, and then properties that affect surface albedo. 29% adjusted every parameterization asked about occasionally or frequently. The most common parameterizations frequently adjusted are clouds in the atmosphere, including cloud microphysics (65%), convection (52%) and cloud fraction (52%). The most common 'occasionally' adjusted parameters were snow (79%) and sea ice (57%) albedo, along with ocean mixing (57%), orographic drag (57%) and cloud optical properties (48%). Soil (43%) and vegetation (39%) properties were also adjusted. These adjustments are consistent with the feeling that atmospheric cloud physics and atmospheric convection were thought most likely to introduce biases in the model, with ocean physics and mixing third.

Finally, groups were asked whether different tuning practices were 'eligible' (justified) on a 5 point scale of disagree, somewhat disagree, neutral, somewhat agree, agree. All groups agreed or somewhat agreed that tuning was justified. 91% thought that tuning global mean temperature or the global radiation balance was justified (agreed or somewhat agreed). Given that these were groups attending a meeting on the subject, there is a self-selection bias. Using the same top 2 categories as registering agreement the following were considered acceptable for tuning by over half the respondents: atmospheric circulation (74%), sea ice volume or extent (70%), as well as cloud radiative effects by regime and tuning for variability (both 52%).

References

Arakawa, R. A., and W. H. Schubert, 1974: Interaction of a cumulus cloud ensemble with the large scale environment. part I. *J. Atmos. Sci.*, **31**, 674–701.

Ayotte, K. W., and Coauthors, 1996: An evaluation of neutral and convective planetary boundary-layer parameterizations relative to large eddy simulations. *Boundary-layer Meteorol.*, **79**, 131–175.

Balaji, V., R. Benson, B. Wyman, and I. I. Held, 2016: Bayesian calibration of computer models. *Geosc. Model Dev. Discussion*, 4464–4468, doi:doi:10.5194/gmd-2016-114.

Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär, 2012: Objective calibration of regional climate models. *J. Geophys. Res.*, **117**, doi:10.1029/2012JD018262.

Bony, S., J.-L. Dufresne, H. Le Treut, J.-J. Mo rcrette, and C. Senior, 2004: On dynamic and thermodynamic components of cloud change s. *Climate Dynamics*, **22**, 71–86.

Burnham, K. P., and D. R. Anderson, 2002: *Model selection and multimodel inference*. 2nd ed., Springer, 488 pp.

Cess, R. D., and Coauthors, 1989: Interpretation of Cloud-Climate Feedback as Produced by 14 Atmospheric General Circulation Models. *Science*, **245**, 513–516, doi:10.1126/science.245.4917.513.

Couvreux, F., F. Hourdin, and C. Rio, 2010: Resolved Versus Parametrized Boundary-Layer Plumes. Part I: A Parametrization-Oriented Conditional Sampling in Large-Eddy Simulations. *Boundary-layer Meteorol.*, **134**, 441–458, doi:10.1007/s10546-009-9456-5.

Currin, C., T. Mitchell, M. Morris, and D. Ylvisakerc, 1991: Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. **86**, 953–963, doi:10.1080/01621459.1991.10475138.

Dahan Dalmedico, A., 2001: History and epistemology of models: Meteorology (1946-1963) as a case study. *Archive for History of Exact Sciences*, **55 (5)**, 395–422, doi:10.1007/s004070000032, URL http://dx.doi.org/10.1007/s004070000032.

Edwards, N. R., D. Cameron, and J. Rougier, 2011: Precalibrating an intermediate complexity climate model. *Clim. Dyn.*, **37**, 1469–1482, doi:10.1007/s00382-010-0921-0.

Edwards, P. N., 2001: *Representing the Global Atmosphere: Computer Models, Data, and Knowledge about Climate Change*. Mit Press.

Edwards, P. N., 2010: *A vast machine: Computer models, climate data, and the politics of global warming*. Mit Press.

Fisher, R. A., 1922: On the mathematical foundations of theoretical statistics. Royal Society of London. *Philosophical Transactions (Series A)*, **222**, 309–368.

Forster, P. M., T. Andrews, P. Good, J. M. Gregory, L. S. Jackson, and M. Zelinka, 2013: Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models. *J. Geophys. Res.*, **118**, 1139–1150, doi:10.1002/jgrd.50174.

Gettelman, A., J. E. Kay, and K. M. Shell, 2012: The Evolution of Climate Sensitivity and Climate Feedbacks in the Community Atmosphere Model. *J. Climate*, **25**, 1453–1469, doi:10.1175/JCLI-D-11-00197.1.

Golaz, J.-C., J.-C. Golaz, and H. Levy, 2013: Cloud tuning in a coupled climate model: Impact on 20th century warming. *Geophys. Res. Lett.*, **40**, 2246–2251, doi:10.1002/grl.50232.

Hansen, J., and Coauthors, 2005: Efficacy of climate forcings. *J. Geophys. Res.*, **110**, D18104, doi:10.1029/2005JD005776.

Haylock, R., and A. O'Hagan, 1996: *On inference for outputs of computationally expensive algorithms with uncertainty on the inputs*, 629–637. Oxford University Press.

Held, I. M., 2005: The Gap between Simulation and Understanding in Climate Modeling. *Bull. Am. Meteorol. Soc.*, **86**, 1609–1614, doi:10.1175/BAMS-86-11-1609.

Heymsfield, A. J., and L. J. Donner, 1990: A Scheme for Parameterizing Ice-Cloud Water Content in General Circulation Models. *J. Atmos. Sci.*, **47**, 1865–1877, doi:10.1175/1520-0469(1990)047⟨1865:ASFPIC⟩2.0.CO;2.

Hourdin, F., and Coauthors, 2013a: Impact of the LMDZ atmospheric grid configuration on the climate and sensitivity of the IPSL-CM5A coupled model. *Clim. Dyn.*, **40**, 2167–2192, doi:10.1007/s00382-012-1411-3.

Hourdin, F., and Coauthors, 2013b: LMDZ5B: the atmospheric component of the IPSL climate model with revisited parameterizations for clouds and convection. *Clim. Dyn.*, **40**, 2193–2222, doi:10.1007/s00382-012-1343-y.

Jackson, C. S., M. K. Sen, G. Huerta, Y. Deng, and K. P. Bowman, 2008: Error Reduction and Convergence in Climate Prediction. **21**, 6698, doi:10.1175/2008JCLI2112.1.

Jam, A., F. Hourdin, C. Rio, and F. Couvreux, 2013: Resolved Versus Parametrized Boundary-Layer Plumes. Part III: Derivation of a Statistical Scheme for Cumulus Clouds. *Boundary-layer Meteorol.*, **147**, 421–441, doi:10.1007/s10546-012-9789-3.

Kennedy, M., and A. O'Hagan, 2001: Bayesian calibration of computer models. *Journal of the Royal Statistical Society (Series B)*, **68**, 425–464.

Kiehl, J. T., 2007: Twentieth century climate model response and climate sensitivity. *Geophys. Res. Lett.*, **34**, L22710, doi:10.1029/2007GL031383.

Kim, D., A. H. Sobel, A. D. Del Genio, Y. Chen, S. J. Camargo, M.-S. Yao, M. Kelley, and L. Nazarenko, 2012: The Tropical Subseasonal Variability Simulated in the NASA GISS General Circulation Model. *J. Climate*, **25**, 4641–4659, doi:10.1175/JCLI-D-11-00447.1.

Kim, D., and Coauthors, 2014: Process-Oriented MJO Simulation Diagnostic: Moisture Sensitivity of Simulated Convection. **27**, 5379–5395, doi:10.1175/JCLI-D-13-00497.1.

L'Ecuyer, T. S., and Coauthors, 2015: The Observed State of the Energy Budget in the Early Twenty-First Century. *J. Climate*, **28**, 8319–8346, doi:10.1175/JCLI-D-14-00556.1.

Liu, C., M. W. Moncrieff, and W. W. Grabowski, 2001: Hierarchical modelling of tropical convective systems using explicit and parametrized approaches. *Quart. J. Roy. Meteor. Soc.*, **127**, 493–515, doi:10.1002/qj.49712757213.

Loeb, N. G., B. A. Wielicki, D. R. Doelling, G. L. Smith, D. F. Keyes, S. Kato, N. Manalo-Smith, and T. Wong, 2009: Toward Optimal Closure of the Earth's Top-of-Atmosphere Radiation Budget. *J. Climate*, **22 (3)**, 748–766, doi:{10.1175/2008JCLI2637.1}.

Loeb, N. G., B. A. Wielicki, D. R. Doelling, G. L. Smith, D. F. Keyes, S. Kato, N. Manalo-Smith, and T. Wong, 2009: Toward optimal closure of the earth's top-of-atmosphere radiation budget. *Journal of Climate*, **22 (3)**, 748–766.

Lynch, P., 2008: The origins of computer weather prediction and climate modeling. *J. Computational Phys.*, **227**, 3431–3444, doi:10.1016/j.jcp.2007.02.034.

Manabe, S., and K. Bryan, 1969: Climate Calculations with a Combined Ocean-Atmosphere Model. *J. Atmos. Sci.*, **26**, 786–789, doi:10.1175/1520-0469(1969)026⟨0786:CCWACO⟩2.0.CO;2.

Manabe, S., and R. T. Wetherald, 1975: The Effects of Doubling the CO_2 Concentration on the climate of a General Circulation Model. *J. Atmos. Sci.*, **32**, 3–15, doi:10.1175/1520-0469(1975)032⟨0003:TEODTC⟩2.0.CO;2.

Mauritsen, T., and Coauthors, 2012: Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, **4 (3)**, doi:10.1029/2012MS000154, URL http://dx.doi.org/10.1029/2012MS000154.

Mauritsen, T., and Coauthors, 2012: Tuning the climate of a global model. 0 pp., doi:10.1029/2012MS000154.

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. **430**, 768–772, doi:10.1038/nature02771.

Murphy, J. M., and Coauthors, 2009: Uk climate projections science report: Climate change projections. Met Office Hadley Centre, Exeter, UK http://ukclimateprojections.defra.gov.uk/images/stories/projections_pdfs/UKCP09_projections_V2.pdf.

Neelin, J. D., A. Bracco, H. Luo, J. C. McWilliams, and J. E. Meyerson, 2010: Considerations for parameter optimization and sensitivity in climate models. *Proc. Natl. Acad. Sci. (USA)*, 21 349–21 354.

Notz, D., 2015: How well must climate models agree with observations? *Phil. Trans. R. Soc. A*, **(2052)**, doi:10.1098/rsta.2014.0164.

Phillips, N. A., 1956: The general circulation of the atmosphere: A numerical experiment. *Q. J. R. Meteorol. Soc.*, **82**, 123–164, doi:10.1002/qj.49708235202.

Qian, Y., and Coauthors, 2015: Parametric sensitivity analysis of precipitation at global and local scales in the Community Atmosphere Model CAM5. *J. Adv. Model. Earth Syst.*, **07**, doi:10.1002/2014MS000354.

Rougier, J. C., 2007: Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change*, **81**, 247–264.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn, 1989: Design and analysis of computer experiments. *Stat. Sci.*, **4**, 409–435.

Schmidt, G. A., and Coauthors, 2014: Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *J. of Adv. in Modeling Earth Systems*, **6**, 141–184, doi:10.1002/2013MS000265.

Sexton, D. M. H., J. M. Murphy, M. Collins, and M. J. Webb, 2012: Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. *Clim. Dyn.*, **38**, 2513–2542, doi:10.1007/s00382-011-1208-9.

Smagorinsky, J., 1963: General Circulation Experiments with the Primitive Equations. *Mon. Wea. Rev.*, **91**, 99, doi:10.1175/1520-0493(1963)091⟨0099:GCEWTP⟩2.3.CO;2.

Suzuki, K., J.-C. Golaz, and G. L. Stephens, 2013: Evaluating cloud tuning in a climate model with satellite observations. *Geophys. Res. Lett.*, **40**, 4464–4468, doi:10.1002/grl.50874.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projectionu. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 2053–2075.

Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Mon. Wea. Rev.*, **117**, 1179–1800.

Wan, H., P. J. Rasch, K. Zhang, Y. Qian, H. Yan, and C. Zhao, 2014: Short ensembles: An efficient method for discerning climate-relevant sensitivities in Atmospheric General Circulation Models. *Geosci. Model Dev.*, **7**, 1961–1977.

Williamson, D., A. T. Blaker, C. Hampton, and J. Salter, 2015: Identifying and removing structural biases in climate models with history matching. *Clim. Dyn.*, **45**, 1299–1324, doi:10.1007/s00382-014-2378-z.

Williamson, D., M. Goldstein, L. Allison, A. Blaker, P. Challenor, L. Jackson, and K. Yamazaki, 2013: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Clim. Dyn.*, **41**, 1703–1729, doi:10.1007/s00382-013-1896-4.

Williamson, D., M. Goldstein, and A. Blaker, 2012: Fast linked analyses for scenario based hierarchies. *J. R. Stat. Soc. Ser. C*, **61(5)**, 665–692.

Yang, B., and Coauthors, 2013: Uncertainty quantification and parameter tuning in the CAM5 Zhang-McFarlane convection scheme and impact of improved convection on the global circulation and climate. *J. Geophys. Res.*, **118**, 395–415, doi:10.1029/2012JD018213.

Zhang, K., and Coauthors, 2014: Technical Note: On the use of nudging for aerosol-climate model intercomparison studies. *Atmosph. Chemist. and Physics*, **14**, 8631–8645, doi:10.5194/acp-14-8631-2014.

Zhang, T., L. Li, Y. Lin, W. Xue, F. Xie, H. Xu, and X. Huang, 2015: An automatic and effective parameter optimization method for model tuning. *Geoscientific Model Development*, **8**, 3579–3591, doi:10.5194/gmd-8-3579-2015.

Zou, L., Y. Qian, T. Zhou, and B. Yang, 2014: Parameter Tuning and Calibration of RegCM3 with MIT-Emanuel Cumulus Parameterization Scheme over CORDEX East Asia Domain. *Journal of Climate*, **27**, 7687–7701, doi:10.1175/JCLI-D-14-00229.1.

Supplemental Material: to the BAMS paper on "The art and science of climate model tuning".

This supplementary Material presents the results of a survey organized in 2014, in preparation of the WCRP meeting on model tuning hold in Garmisch-Partenkirchen, Germany, in October 2014. 23 modeling groups involved in the Couple Model Intercomparison Project did contribute to the survey. The survey was organized on line using the "SurveyMonkey" web site. The results are given in figures and tables below. The questions are reproduced as they were presented on the survey web site. The number of groups is given inside ( ) together with the percentage relative to the groups that answered the given question.

For which purposes is your model being developed? Check all relevant boxes.

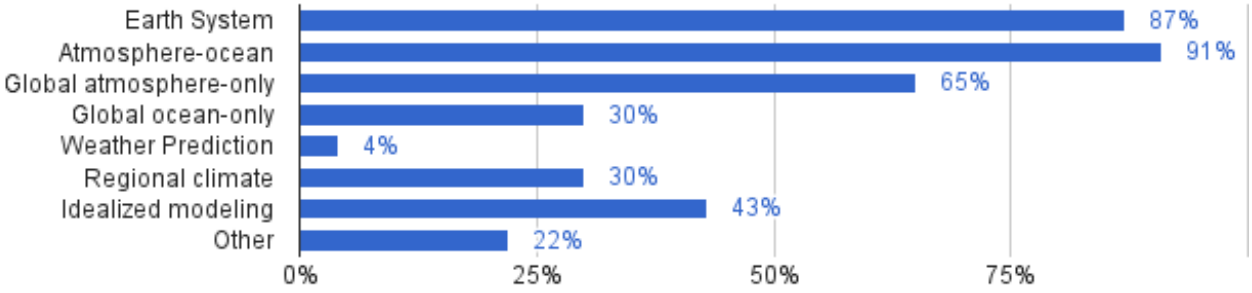

FIG. S1. For which purpose is your model being developed

Answer Choices	Responses
Earth System modeling (e.g. including carbon cycle)	87% (20)
Global coupled atmosphere-ocean climate modeling	91% (21)
Global atmosphere-only climate modeling	65% (15)
Global ocean-only climate modeling	30% (7)
Numerical weather prediction	4% (1)
Regional climate modeling	30% (7)
Idealized model studies	43% (10)
Other	22% (5)

TAB. S1. For which purpose is your model being developed, full results

Is your model being tuned by adjusting model parameters to obtain certain desired properties, e.g. radiation balance?

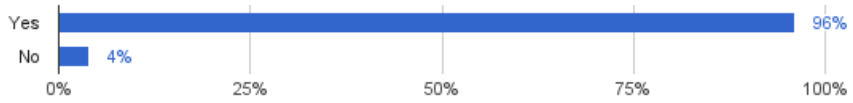

FIG. S2. Is your model being tuned ?

Answer Choices	Responses
Yes	96% (22)
No	4% (1)

TAB. S2. Is your model being tuned ? Full results

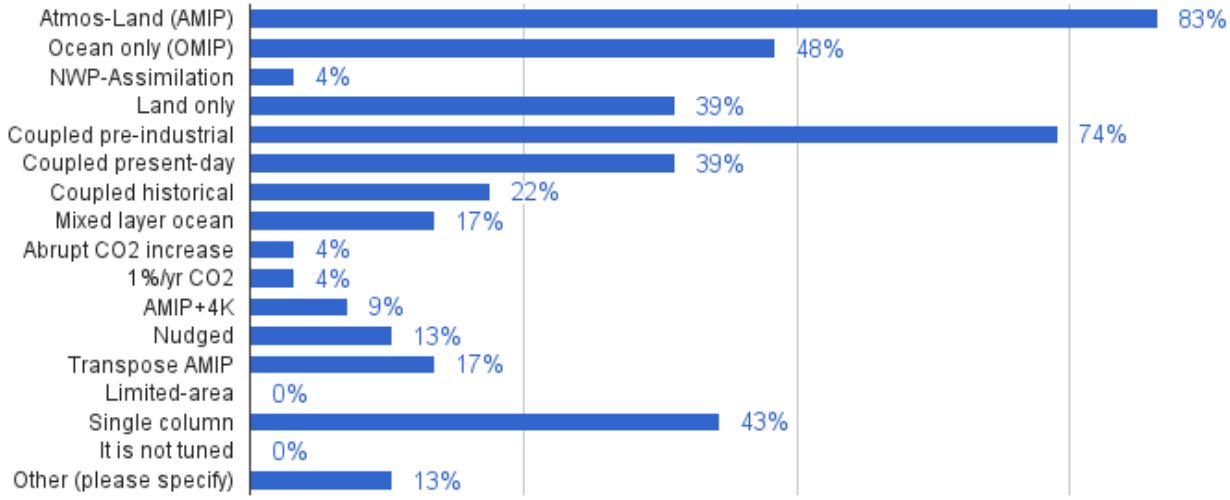

FIG. S3. In what modes is your model being tuned ?

Answer Choices	Responses
Atmosphere-Land (AMIP)	83% (19)
Ocean standalone (OMIP)	48% (11)
Weather forecasting assimilation cycles (NWP)	4% (1)
Land standalone	39% (9)
Coupled pre-industrial	74% (17)
Coupled present-day	39% (9)
Coupled 20th Century (historical)	22% (5)
Mixed layer/slab ocean	17% (4)
Abruptly increased CO2 (abrupt4xCO2)	4% (1)
1 percent per year CO2 (1pctCO2)	4% (1)
Cess-experiments, (AMIP+4K)	9% (2)
Nudged to observations	13% (3)
Transpose AMIP	17% (4)
Limited-area	0% (0)
Single column	43% (10)
It is not tuned	0% (0)
Other (please specify)	13% (3)

TAB. S3. In what modes is your model being tuned ? Full results

In which parameterizations do you apply changes when tuning your model, and how much?

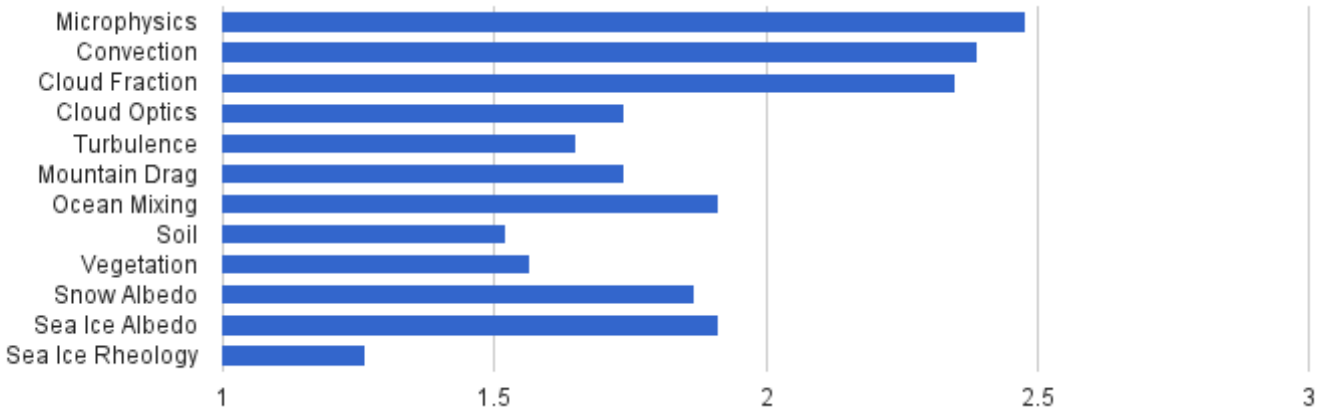

FIG. S4. In which parameterizations do you apply changes ?

	Not used 1	Occasionally 2	Frequently used 3	Total	Average Rating (/3)
Cloud micro- physical processes, e.g. droplet number concentration, conversion rates, fall velocities	17% (4)	17% (4)	65% (15)	23	2.48
Convection, e.g. entrainment rate, conversion to precipitation	13% (3)	35% (8)	52% (12)	23	2.39
Cloud fraction, e.g.critical RH threshold or assumptions about PDFs	17% (4)	30% (7)	52% (12)	23	2.35
Cloud optical properties, e.g. sub-grid inhomogeneity	39% (9)	48% (11)	13% (3)	23	1.74
Turbulent mixing, e.g. mixing length, explicit top entrainment	48% (11)	39% (9)	13% (3)	23	1.65
Orographics drag	35% (8)	57% (13)	9% (2)	23	1.74
Ocean physical properties, e.g. mixing, optics	26% (6)	57% (13)	17% (4)	23	1.91
Soil and run- off properties	52% (12)	43% (10)	4% (1)	23	1.52
Vegetation properties	52% (12)	39% (9)	9% (2)	23	1.57
Snow albedo	22% (5)	70% (16)	9% (2)	23	1.87
Sea ice albedo including meltponds	26% (6)	57% (13)	17% (4)	23	1.91
Sea ice rheology	74% (17)	26% (6)	0% (0)	23	1.26

TAB. S4. In which parameterizations do you apply changes ? Full results

What metrics of the state and variability are specifically used in the model tuning process, and how are they weighted in cases where compromises needs to be made?

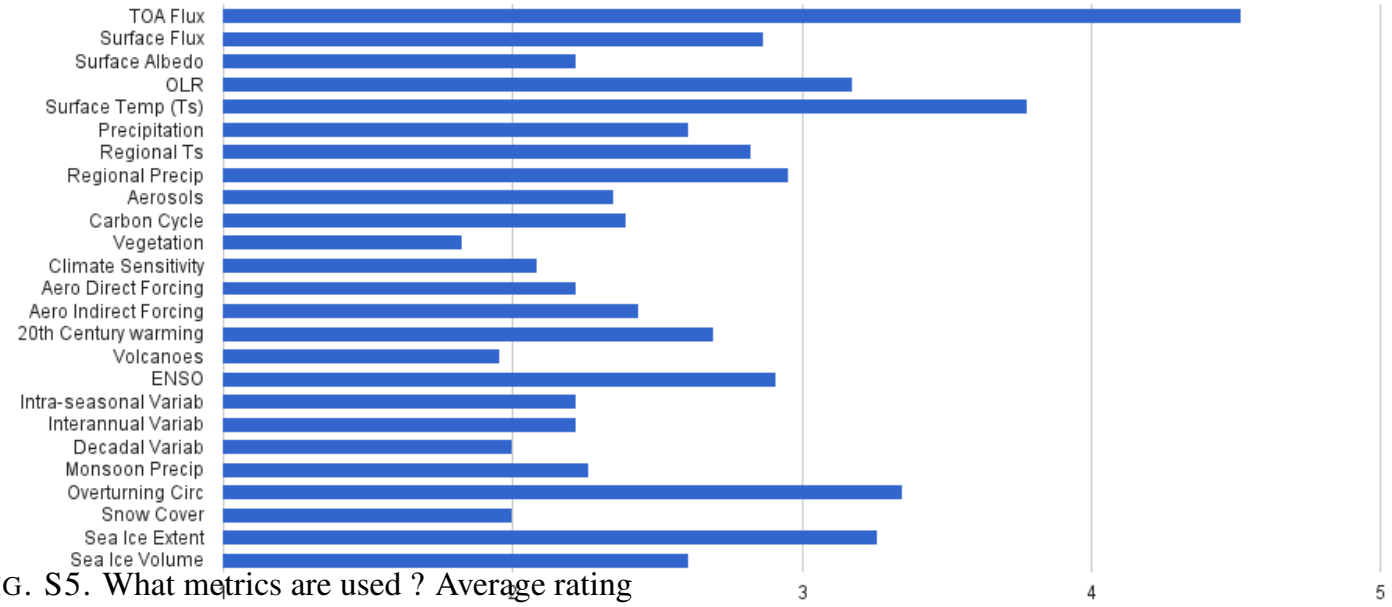

FIG. S5. What metrics are used ? Average rating

	Not considered 1	Less important 2	Important 3	Very important 4	Decisive 5	Total	Average Rating (/5)
Global mean TOA net flux	4% (1)	0% (0)	4% (1)	22% (5)	70% (16)	23	4.52
Global mean surface net flux	17% (4)	22% (5)	22% (5)	35% (8)	4% (1)	23	2.87
Global mean surface albedo	26% (6)	39% (9)	22% (5)	13% (3)	0% (0)	23	2.22
Global mean OLR	9% (2)	13% (3)	43% (10)	22% (5)	13% (3)	23	3.17
Global mean surface temperature	4% (1)	4% (1)	26% (6)	39% (9)	26% (6)	23	3.78
Global mean precipitation	13% (3)	39% (9)	22% (5)	26% (6)	0% (0)	23	2.61
Regional surf. temperature biases	9% (2)	22% (5)	48% (11)	22% (5)	0% (0)	23	2.83
Regional patterns of precip.	9% (2)	26% (6)	26% (6)	39% (9)	0% (0)	23	2.96
Aerosols, if applicable	30% (7)	13% (3)	48% (11)	9% (2)	0% (0)	23	2.35
Global carbon cycle, if applicable	35% (8)	22% (5)	17% (4)	22% (5)	4% (1)	23	2.39
Regional vegetation, if applicable	52% (12)	17% (4)	26% (6)	4% (1)	0% (0)	23	1.83
Climate sensitivity	43% (10)	17% (4)	26% (6)	13% (3)	0% (0)	23	2.09
Aerosol direct forcing	30% (7)	22% (5)	43% (10)	4% (1)	0% (0)	23	2.22
Aerosol indirect effects	35% (8)	13% (3)	26% (6)	26% (6)	0% (0)	23	2.43
20th Century warming	30% (7)	17% (4)	17% (4)	22% (5)	13% (3)	23	2.70
Response to volcanoes	39% (9)	35% (8)	17% (4)	9% (2)	0% (0)	23	1.96
ENSO variability	22% (5)	17% (4)	26% (6)	17% (4)	17% (4)	23	2.91
Intra- seasonal variability	30% (7)	30% (7)	30% (7)	4% (1)	4% (1)	23	2.22
Interannual variability	30% (7)	35% (8)	17% (4)	17% (4)	0% (0)	23	2.22
Decadal variability	35% (8)	43% (10)	9% (2)	13% (3)	0% (0)	23	2.00
Volcanic responses	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0	1.00
Monsoon rainfall	22% (5)	39% (9)	30% (7)	9% (2)	0% (0)	23	2.26
Ocean merid. overturn. circul.	13% (3)	13% (3)	13% (3)	48% (11)	13% (3)	23	3.35
Snow cover	39% (9)	26% (6)	30% (7)	4% (1)	0% (0)	23	2.00
Sea ice extent	13% (3)	13% (3)	17% (4)	48% (11)	9% (2)	23	3.26
Sea ice volume	13% (3)	35% (8)	26% (6)	26% (6)	0% (0)	23	2.65

TAB. S5. What metrics are used ? Full results

Which processes do you believe introduce the largest biases in your coupled model? Please rank from 1 to 5 with higher numbers indicating larger biases.

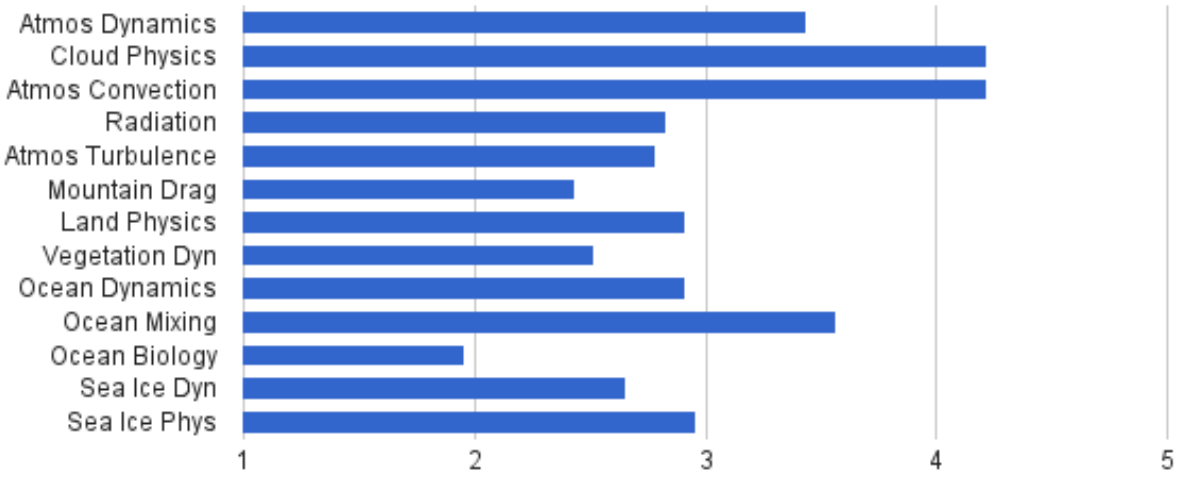

FIG. S6. Which processes do you believe introduce the largest biases ? Average Rating

	1	2	3	4	5	Total	Average Rating (/5)
Atmospheric dynamics, including model resolution	4% (1)	13% (3)	35% (8)	30% (7)	17% (4)	23	3.43
Atmospheric cloud physics	4% (1)	0% (0)	9% (2)	43% (10)	43% (10)	23	4.22
Atmospheric convection	0% (0)	0% (0)	22% (5)	35% (8)	43% (10)	23	4.22
Atmospheric radiation	13% (3)	26% (6)	30% (7)	26% (6)	4% (1)	23	2.83
Atmospheric turbulence	9% (2)	30% (7)	35% (8)	26% (6)	0% (0)	23	2.78
Orographic drag	17% (4)	26% (6)	52% (12)	4% (1)	0% (0)	23	2.43
Land physics	4% (1)	35% (8)	35% (8)	17% (4)	9% (2)	23	2.91
Vegetation physics	22% (5)	30% (7)	30% (7)	9% (2)	9% (2)	23	2.52
Ocean dynamics	4% (1)	26% (6)	52% (12)	9% (2)	9% (2)	23	2.91
Ocean physics/mixing	0% (0)	4% (1)	48% (11)	35% (8)	13% (3)	23	3.57
Ocean biology	39% (9)	35% (8)	22% (5)	0% (0)	4% (1)	23	1.96
Sea ice dynamics	4% (1)	43% (10)	35% (8)	17% (4)	0% (0)	23	2.65
Sea ice physics	0% (0)	35% (8)	39% (9)	22% (5)	4% (1)	23	2.96

TAB. S6. Which processes do you believe introduce the largest biases ? Full results

Please evaluate whether you agree that the following practices are eligible:

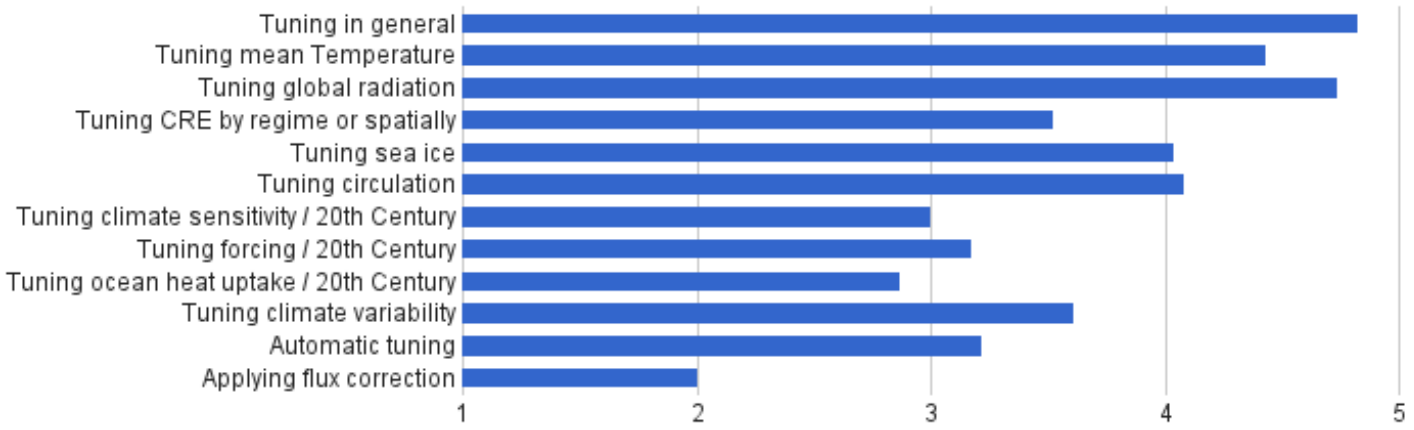

FIG. S7. Which practices you see as eligible ? Average Rating

	Disagree 1	Somewhat disagree 2	Neither Disagree Nor Agree 3	Somewhat agree 4	Agree 5	Total	Average Average Rating (/5)
Climate model tuning in general	0% (0)	0% (0)	0% (0)	17% (4)	83% (19)	23	4.83
Tuning global mean temperature	0% (0)	0% (0)	9% (2)	39% (9)	52% (12)	23	4.43
Tuning the global radiation balance	0% (0)	0% (0)	9% (2)	9% (2)	83% (19)	23	4.74
Tuning the cloud radiative effects by regime or spatially	9% (2)	13% (3)	26% (6)	22% (5)	30% (7)	23	3.52
Tuning sea ice volume and/or extent	4% (1)	0% (0)	26% (6)	26% (6)	43% (10)	23	4.04
Tuning the circulation, e.g. by gravity wave drag or turbulence	0% (0)	4% (1)	22% (5)	35% (8)	39% (9)	23	4.09
Tuning model sensitivity, e.g. to match 20th Century warming	9% (2)	35% (8)	22% (5)	17% (4)	17% (4)	23	3.00
Tuning model forcing, e.g. aerosol indirect effects, in order to match 20th Century warming	13% (3)	17% (4)	26% (6)	26% (6)	17% (4)	23	3.17
Tuning ocean heat uptake, e.g. to match 20th Century warming	13% (3)	26% (6)	35% (8)	13% (3)	13% (3)	23	2.87
Tuning variability, e.g. ENSO, MJO or decadal variability	0% (0)	17% (4)	30% (7)	26% (6)	26% (6)	23	3.61
Automatic tuning	4% (1)	13% (3)	57% (13)	9% (2)	17% (4)	23	3.22
Applying flux-corrections	52% (12)	17% (4)	17% (4)	4% (1)	9% (2)	23	2.00

TAB. S7. Which practices you see as eligible ? Full results