Practice and philosophy of climate model tuning across six U.S. modeling centers, by Schmidt et al.

Response to reviewers

We'd like to thank both reviewers and the executive editor for their constructive comments (in red). Author responses are in black.

Response to A. Kerkweg:

In my role as Executive editor of GMD, I would like to bring to your attention our Editorial [Policy] version 1.1: http://www.geosci-model-dev.net/8/3487/2015/gmd-8-3487-2015.html

In particular, please note that for your paper, the following requirements have not been met in the Discussions paper:

- "All papers must include a section, at the end of the paper, entitled 'Code availability'. Here, either instructions for obtaining the code, or the reasons why the code is not available should be clearly stated. It is preferred for the code to be uploaded as a supplement or to be made available at a data repository with an associated DOI (digital object identifier) for the exact model version described in the paper. Alternatively, for established models, there may be an existing means of accessing the code through a particular system. In this case, there must exist a means of permanently accessing the precise model version described in the paper. In some cases, authors may prefer to put models on their own website, or to act as a point of contact for obtaining the code. Given the impermanence of websites and email addresses, this is not encouraged, and authors should consider improving the availability with a more permanent arrangement. After the paper is accepted the model archive should be updated to include a link to the GMD paper."

I do not agree with your statement in the Code Availability Section of your article: "No data or code is presented in this paper". As you are presenting and discussing the tuning methods for six models and/or modeling centers, a statement how to access each of the discussed models has to be made here.

For completeness, we have added a table giving the URLs and details on the accessibility of the six centers' relevant codes. However, in justification of our initial text, we do not feel that this adds appreciably to the discussion since the processes discussed are not visible in a snapshot of final code.

Response to Referee #1 (S. Sherwood)

This review of model tuning practices is potentially a useful contribution to the literature. The complexity of modern models means tuning processes are complicated and can be opaque, but as pointed out by the authors, interpretation of model-data differences, and therefore model evaluation and improvement, depends crucially on how a model was tuned. I think this paper will be acceptable to the journal and a valuable contribution, once a few issues are addressed.

Thank you for your assessment.

1. The authors need to clarify how their contribution relates to the 2016 BAMS review article by Hourdin et al. They state that theirs can be viewed as a “followup” specific to US centres, but do not spell out what they are adding. I think what they are adding is more detail on the tuning practices at these six centres—but they need a clear statement. The issues discussion in Section 2 seems much too lengthy for a follow-up, unless there are important issues that were overlooked by Hourdin et al. It seems that many of these points were already made in the Hourdin et al. paper, or in another paper by Schmidt and Sherwood which is also frequently cited. I think the authors should shorten Section 2, summarise in Section 1 what they are adding to Hourdin et al., and indicate as appropriate within Sections 3-5 where they are repeating what was in Hourdin et al. vs. what is new.

We have added a clearer statement about what is being added here beyond what is in Hourdin et al (2016) (hereafter H16). While section 2 does contain a general discussion that covers similar ground to H16, it does differ in detail and in approach. We have condensed it slightly to avoid undue repetition. Note too that only three out of the six model groups here were surveyed by H16. We have added an explicit statement making that clearer.

2. I found the manuscript to be of uneven clarity in identifying whether the tuning practices are current and being used right now, or whether they only apply to existing, released model versions. Although some sections specified version numbers (e.g. GFDL), others (e.g. NCAR) did not, although at the end the NCAR section did discuss some issues that arose for CAM6—but with too little detail (e.g., “. . .with some tuning to those schemes, ENSO performance skill was en- hanced.”) Are any of the centres changing their practice? Is GFDL tuning climate sensitivity now that they know how to do so? Table 2 should note when the answers in the table hold, since in the future they may change. At least one of the authors (Golaz) has been outspoken in asking questions about tuning for climate sensitivity, but this manuscript remains strangely silent on what the US centres are planning (or currently doing), only arguing that this was not done in the past.

We have added version numbers where they were implicit in the original text. The discussion section now addresses how tuning efforts are changing (if at all) in the latest iterations. With reference to ENSO of course the complexity of a long-term coupled climate phenomena does not lend itself to obvious associations with atmospheric tuning parameters. However, published knowledge of basic state biases and their links to parameterization settings can be exploited. In the NCAR case modifying turbulence settings with an accepted range lead to improvements in basic state low-level zonal flow, which lead to improved ENSO amplified. The outcome is of course not guaranteed. If it were we would always get a good ENSO without fail!

3. The text mentions model selection in the introduction, but I did not see any further mentions of this. Have any of the modelling centres ever discarded a working model version because of its climate behaviour (e.g. climate sensitivity) or any other interesting reason? If not, a statement to this effect would be nice.

We are not aware of any of these centers discarding workable versions of their models for any ‘interesting reason’ (including climate sensitivity). That has been made clearer in the text.

4. I often hear grumbling about a hidden problem in GCMs being significant biases in their mean surface temperature, which are swept under the rug by using anomalies (itself a type of simple model calibration), and which some think should be a significant factor in evaluating models. On the other hand, Hourdin et al. claim that global mean surface temperature is the “dominant shared target” for tuning efforts around the world. If so, isn’t global-mean temperature a useless metric for evaluating models, since it only measures how hard the centres chose to tune for this particular target? Can you please say something about this, at least for the US centres? How hard do centres tune for this target, compared to other targets which may require compromising on global-mean T? What (if any) are these other, conflicting targets?

This does come up relatively often and there are two main things to say. Our experience is that global mean surface temperature in the coupled models is not ‘tunable’ in the sense that we can set it to a known value, but they are monitored during the development process as a diagnostic of processes that may be mis-configured in some way. However, we do not agree that this is not the “dominant shared target” for the model groups included here.

5. Related to (4), it would be nice for Section 4 (or, alternatively, Section 2) to give a better overview of the typical tuning sequence for a coupled model. For example, it seems that centres first tune the AGCM with observed SST to get the TOA flux (im)balance right (do they tune to get LW and SW separately correct?), typically by way of tuning things related to clouds, then tune the ocean (though much less is said about this and I am not sure what the target is), then probably retune the coupled model for global SST, ENSO / MJO, etc? Are the AMIP and CMIP versions of models used in CMIP5 tuned identically, or was there further tuning to the coupled model that is not retroactively put into the AGCM used in AMIP? Some of these details could be clarified also for the individual centres. It also looks like most centres that have aerosol indirect effects end up tuning those to be something they think is reasonable (which is a very important thing to know, probably the most important of all the information presented in this paper, since aerosol forcing in GCMs is a key source of information, even used by IPCC WGI Chapter 7 assessment of this, and many people may mistakenly believe this offers independent information!). Currently Section 4 summarises what is different between centres, but doesn’t give this typical set of steps in taken.

We agree that this should be clearer, but it does vary with the center. We have added some clarity on this in section 4.

Minor corrections

- 2:12-15. The fact that model behaviour depends on expert judgments about model design is no different to any other modelling exercise, and has been the situation in climate modelling since day dot.

True.

- Can you restate more precisely what new problem is brought on by recent developments? It seems like the new problem might be that modelling centres now have control over the climate behaviour of their model in ways that they did not before, and that the result could be that climate predictions begin to converge toward what modelling centres think is the most likely/plausible outcome even if it is wrong.

We don't really agree with this assessment. There is a clear convergence of functionality - processes shown to be important by one group get incorporated (often independently) by others. There is also a convergence in experimental design - which facilitates comparisons and across-model syntheses. Evidence for a convergence of results beyond what would be expected from the shared physical basis is lacking though. Indeed, we think it increasingly unlikely that this will happen since the independent complexificaiton of parameterisations is making it harder and more complex to tune for emergent behaviours. We anticipate that in the near future far more explicit and controlled PPEs will be generated that will make the structural uncertainty far more obvious and make the results even less prone to 'herding'.

- 3:21-23. We don't know the true aerosol indirect forcing, so this needs to be reworded—do you mean to say the model didn't warm enough globally compared to observations until the critical radius was changed? That artefacts in the geographic warming pattern were produced in the simulation that were judged to indicate too-strong indirect effects, and/or that were ameliorated by making the indicated change to the critical radius?

This has been reworded to refer to the (better known) trend in 20th Century temperature and to make clear that this was a finding that came after CM3 was frozen for CMIP5.

- 3:24-26. This sentence is too hard to understand.

This sentence has been deleted.

- 7:1-3. I assume you mean global, climatological (seasonal or annual mean) fields? Please specify

Yes. Clarified.

- 7:8. Please change “we” to “the DOE modelling group” or similar. “We” should refer to the authors, not the modellers at one centre.

Fixed.

- 9:18. “RFP” is introduced with no definition. This reviewer does not know what it means, which made the following text hard to review. I have no idea why the ratio of “RFP” to climate sensitivity is meaningful.

The definition was in the introduction, but we have added a sentence giving a rationale for its use.

- 16:1-3. Run-on sentence.

Fixed.

- 16:18-19. Rephrase; models are not tuned by models, but by model developers.

Indeed.

Response to Referee #2

General Comments

This paper describes the approach to model tuning of six U.S. modeling groups. It describes itself as a follow on from a paper by Hourdin et al (2016) paper (The art and science of climate model tuning) which was an outcome of a meeting of International modeling groups starting to discuss tuning practices and the implications thereof. I think the paper is potentially publishable, although I have some reservations over the balance of the content, notably what is new. I think the authors need to address the following issues;

1. Section 2 covers very much the same ground as Hourdin et al (2016). Whilst it is well written, this is not the new part of the paper. I think this needs to be reviewed and shortened. There are useful additional contributions such as using examples from the US models such as P4 second paragraph where parameter tuning vs structural uncertainty is discussed.

We have shortened this section slightly and highlighted the differences with H16 and how we are extending that paper.

2. Section 3 describes the specific practices of each modeling centre and hence is the new contribution. However this section is very uneven and there is no common format by which the reader can compare the six modeling centre approaches. At the very least this needs to be organised so that all groups describe e.g. first their use of component models (AMIP, forced ocean etc), then coupled (AOIL, PI control and/or PD control) models then ES models (if appropriate). Then they should describe how they use historic simulations and idealised futures (e.g. 4K SST or 4CO2) to look at climate sensitivity. In many cases there is some similar structure to this but the vagueness or lack of common use of specific terminology for

experiments or approaches makes it very hard to interpret.

We have endeavored to make these sections more uniform (and to summarise this in Table 2), but one of the key insights from the this exercise is that there is a great heterogeneity in how model centers do this. Key metrics and procedures in one center may not even be considered in another.

3. Also in Section 3, the language describing the methodology to tuning by each centre is often vague and not well quantified. The groups use terms like 'the magnitude of the aerosol indirect effect . . ...was adjusted if deemed to be inconsistent with. . .' or 'Configurations for which the ratio. . . fell substantially below. . . were rejected' or 'A key tuning target is matching the . . ..'. How was the model adjusted? What represents substantial? How was the model 'matched' to observations? I think if we are to describe the detail of the tuning process at this level we need to be completely clear about what we mean. I recognise of course that this might mean we have to say 'A subjective decision on the relative quality of the various configurations based on a set of X metrics was taken' but at least the reader then knows how a decision was made.

Fair point. We have tried to be more specific in the language, but these decisions often need to balance multiple tuning targets subjectively.

4. In a few places there is a description of what I would call 'traditional model development' which is here described as tuning (e.g. p 10 3rd paragraph). I think we must be really careful to separate improvement to convection e.g. by inclusion of cold pools which leads to better MJO variability from tuning of parameters to ensure e.g. balance of large-scale measures. Indeed I think it would be helpful to recognise that modeling centres often 'monitor' some of these tuning targets as models are developed (e.g. from bottom up) and this can avoid the need for a lot of final tuning in many cases.

It may be useful in theory to distinguish 'final (parameter) tuning' from model development, but harder in practice to find a clean dividing line between one and the other given that they are often occurring in tandem. Is an adjustment of the river runoff directions (as done by NCAR) a model development or a tuning? Nonetheless, we have added a brief note on this point and tried to be consistent across the paper.

Specific comments

P2, l10. I am not sure that model tuning has ever been transparent. It was simpler but still not documented in most cases.

We agree that that it has always been poorly documented, but the point is that there is now more to document, and so the gap between what is needed and what is available is now greater.

P3, l26. Not only possible but likely, I would suggest

Yes. But this line has been deleted.

P5, l17. I suspect the changes made affected the NH more widely than just the UK.

Changed.

P5,l24. I think the weak correlations between aerosol forcing and sensitivity do not provide strong evidence that there has been model tuning based on this trade-off but I don't think you can say that this means the CMIP3 models aerosols forcings were not tuned. I suspect it was done in some models but certainly not in all.

We are not aware of any group explicitly doing this.

P7,l4 'Cess climate sensitivity is evaluated using idealised SST +4K simulations' How is this then used? Are models thrown away if this is outside of some range (e.g. CMIP5)?

Climate sensitivity is being evaluated as part of the model development process for DOE ACME. To date, they have not encountered a situation where the estimated sensitivity was deemed to be unacceptable based on expert judgement. Should such a situation arise, the model would receive extra scrutiny to better understand what may have cause the climate sensitivity to change compared to previous developmental versions.

P7, l16. '...to monitor the combined impact of anthropogenic forcings and climate sensitivity' Again, what does 'monitoring' mean? Is action taken if it's deemed to be 'unacceptable'?

This is worth expanding on and we do so in section 2. Many diagnostics in GCMs are monitored during the development process to see if they remain within some a priori expected range. If they fall outside those limits, it is usually a sign of some inadvertent side effect of a related change, or conceivably due to a bug that was introduced into the code. Thus action is often taken to go into the details more closely and see what has happened - check recent changes, examine relevant budgets etc. Since these fields are often emergent and functions of many different parts of the physics, they are often not tunable in any simple sense.

P8, l29. Why were new model versions constrained to have a ratio of RFP to Cess sensitivity the same as the old model? Presumably so that the evolution of historic temperature will not differ substantially from that achieved by the old model – although it sounds like it didn't work very well. The target for this tuning needs to be said more explicitly.

This practice was adopted by GFDL in order to have a way of predicting how the coupled model would react using only the (cheaper) AGCM version.

P10, l6. What happens if the coupled model drifts are not ‘relatively small’? Do you go back to the start (i.e. component level tuning?)

In practice, longer integrations will help reduce drift, and the model state once stabilised can be assessed for suitability. Big drifts at the start of an integration can often be reduced by different tuning choices that either affect surface atmospheric fluxes or ocean mixing.

P11, l2 ‘The tuning suite includes present day climate simulations, . . .’ Does this mean AMIP or coupled PD?

Predominantly AMIP-style simulations. This has been added.

P11, l10 – same comment as above what do you mean by ‘present day climate’ here?

AMIP again

P11, l12 What does ‘matching’ mean. RMS?

It meant mean and variance of the difference from CERES. The text has been updated.

P11, l26 What does ‘bring together’ mean?

This was an inadvertent misstatement and the text has been changed to "bring to bear". The idea that GMAO use aerosol observations to further constrain the turbulence.

P12, second paragraph. Are the higher resolution simulations tuned independently from the lower resolution ones, even for parameters with no obvious resolution dependence? How does this fit with a seamless idea or is this an explicit recognition of the specific requirements of the different uses/customers?

Yes. Our experience is that resolution decisions almost always affect tunings (and development) and the goal that parameterized physics or models can be independent of resolution while a noble aim, is not yet a reality. Indeed, whether it will ever be possible is still an area of active research.

P13, l17. What is the basis for constraining the net aerosol forcing to be less than -1.5Wm-2?

NCAR used the guidance from IPCC AR5, that the range for total indirect+direct effects is likely to be weaker than -1.5 W/m2. This is considered alongside previous determinations that the equilibrium climate sensitivity is unlikely to be so high that it would require a very large net aerosol forcing in order to reproduce the 20th century surface temperature evolution. It is also worth noting that the correct simulation of low-cloud properties (water path, minimum drop number, fraction tend, radiative forcing) tends not to be associated with aerosol indirect effects in excess of -1.5 W/m2. Direct aerosol forcings are smaller and less sensitive to tuning choices.

P14, l13 Does ‘transient mode’ mean - historic simulation?

Yes. Clarified.

P14, l15 What tuning to the historic record does happen - no tuning or no fine-tuning?

That was explained in the previous line. There was no additional tuning after the specified changes.

P15, l4. Another example of model development. Increasing levels from 28 to 64 is not tuning.

Indeed. Going from 28 to 64 levels was not tuning, as such. The tuning was made to the convection parameterization to account for many more levels in the boundary layer. We have made this clearer in the text.

P16, l1-4 You talk about the value of evaluating fast physics in short range forecasts. It wasn’t clear that NASA GMAO used this capability for a seamless approach in their tuning approach?

This was described as part of the suite of approaches used by GMAO at the beginning of section 3.4.

P17, l14 I don’t understand what ‘using the decadal mean SST . . . and constant yr 2000 forcings’ means. What is the experimental design here?

It is a standard experiment run at multiple centers (including NCAR, GISS, GFDL and DOE) that has decadally-appropriate climatological SST but no interannual variability in ocean conditions. At GISS the same configuration is used for pre-industrial AGCM tests and tuning using the decade 1876-1885. Internal variability (mostly ENSO) dictates that much longer simulations would be required to gain a robust signal of pre-industrial minus present-day differences associated with changes in aerosols.

Practice and philosophy of climate model tuning across six U.S. modeling centers

Gavin A. Schmidt[1], David Bader[2], Leo J. Donner[3], Gregory S. Elsaesser[1,4], Jean-Christophe Golaz[2], Cecile Hannay[5], Andrea Molod[6], Rich Neale[5], and Suranjana Saha[7]

[1]NASA Goddard Institute for Space Studies, 2880 Broadway, New York
[2]DOE Lawrence Livermore National Laboratory, Livermore, California
[3]GFDL/NOAA, Princeton University Forrestal Campus, 201 Forrestal Rd., Princeton, NJ 08540
[4]Columbia University, New York, NY 10025
[5]National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA,
[6]Global Modeling and Assimilation Office, NASA GSFC, Greenbelt, MD 20771
[7]Environmental Modeling Center, NCEP/NWS/NOAA, NCWCP College Park, MD 20740

Correspondence to: gavin.a.schmidt@nasa.gov

Abstract. Model calibration (or "tuning") is a necessary part of developing and testing coupled ocean-atmosphere climate models regardless of their main scientific purpose. There is an increasing recognition that this process needs to become more transparent for both users of climate model output and other developers. Knowing how and why climate models are tuned and which targets are used is essential to avoiding possible misattributions of skillful predictions to data accommodation and vice versa. This paper describes the approach and practice of model tuning for the six major U.S. climate modeling centers. While details differ among groups in terms of scientific missions, tuning targets and tunable parameters, there is a core commonality of approaches. However, practices differ significantly on some key aspects, in particular, in the use of initialized forecast analyses as a tool, the explicit use of the historical transient record, and the use of the present day radiative imbalance vs. the implied balance in the pre-industrial as a target.

1 Introduction

Simulation has become an essential tool for understanding processes in the Earth system, interpreting observations and for making predictions over short (weather), medium (seasonal) and long (climate) terms. The complexity of this system is evident in the myriad processes involved (such as the microphysics of cloud nucleation, land surface heterogeneity, convective plumes, and ocean mesoscale eddies) and in the dynamic views provided by remote sensing. This complexity and wide range of scales that need to be incorporated imply that simulations will necessarily include approximations to well-understood physics and empirical formulations for unresolved effects. ~~Despite this, climate and weather simulations have demonstrated useful predictive skill in many emergent properties .~~

~~Computer simulations of this complexity occupy a middle ground between the two classic pillars of science; theory and experiment . They~~ The simulations are neither a straightforward encapsulation of some well-known theory, nor are they laboratory experiments probing the real world~~. Instead,~~ , though they have features of both ~~- there are many encapsulations of~~

üûúùøœöõôóòñðïîîë1llllllllllllllllllll0«/P(l)»1

~~theory within the models, but as described below, they also contain scale-dependent parameterizations and developer-dependent choices related to complexity and completeness. Yet the simulations should also be treated as numerical laboratories, with results that are always preliminary, subject to replication by other models and real world evaluation~~(Schmidt and Sherwood, 2014). Despite this, climate and weather simulations have demonstrated useful predictive skill across many emergent diagnostics (

?). Note that we distinguish fields or statistics in the model that arise from the interactions of multiple physical effects (“emergent properties”) from those that are closely related to single processes or parameterizations.

Since the pioneering work in climate modeling in the mid-20th Century (e.g. Phillips, 1956; Manabe and Bryan, 1969; Hansen et al., 1983), climate models have increased enormously in scope and complexity, going from relatively crude discretizations of atmospheric dynamics to now, far more detailed atmospheres, combined with ocean, sea ice, carbon cycles, and interactive composition in the atmosphere including chemistry and multiple aerosol species. As that complexity has grown, more processes are explicitly included and the parameterizations are pushed to a more detailed (and more fundamental) level, allowing for better constraints on unknown parameters. However, at the same time, the process of model development has become more convoluted and now involves many more components than it did originally. This has led somewhat predictably to an unfortunate reduction in transparency over time.

It is worth expanding on why this matters: First, model development ~~must involve~~ involves expert judgments which ~~, given a different set of experts, might have gone in a different direction. Had different choices been made, it's conceivable that this would impact on climate sensitivity and other~~ are inevitably subjective and with different choices there would be differences in emergent responses. ~~In~~ For instance, in the MPI model, Mauritsen et al. (2012) show that equally valid, but distinct tunings can impact model sensitivity. This kind of behavior should therefore be reported more widely to improve the assessment of the robustness of specific responses. Second, models used as part of international assessment projects (such as the Coupled Model Intercomparison Project, Phase 5 (CMIP5)) are increasingly being weighted or ~~subsetted~~ subset in order to refine predictions. If the skill measure that is used to filter or weight models has been tuned for in some cases rather than in others, the subset or weighted average will be biased towards models where that tuned over those that weren't, and that may not correspond to better physics nor better predictions (Knutti et al., 2010).

Thus it has become increasingly clear that a more transparent process is necessary. A survey of modeling groups ~~from~~ involved in CMIP5 ~~provided~~ (Hourdin et al., 2016, hereafter H16) provides a good background on tuning practices and makes a plea for better coordination of documentation of these issues. This paper ~~might therefore be seen as a followup~~ is a more detailed follow-up for a subset of climate models associated with laboratories in the US (3 of which were surveyed by H16, three of which weren't). The six modeling centers that are the focus of this paper have all developed and maintain Earth system models that (at minimum) have a dynamic atmosphere, coupled ocean components and are global in scope. Additional components (such as ice sheets, the carbon cycle, atmospheric chemistry and aerosols) are also common. While two of the models discussed (NCEP CFS and NASA GMAO's GEOS-5) are primarily used for short-term (daily to seasonal) predictions, there is sufficient overlap with the models focused on longer-term problems (decadal to multi-decadal periods) to warrant describing them all as ’climate models’ below.

2 Why is climate model tuning necessary?

Climate and weather models consist of three levels of representation of physical processes: fundamental physics (such as conservation of energy, mass and momentum), approximations to well-known physical theories (the discretization of the Navier-Stokes equations, broadband approximations to line-by-line radiative transfer codes, etc.) and empirical approximations (“parameterizations”) needed to match the phenomenology of unresolved or poorly understood sub-gridscale or excluded processes.

The degree of approximation and complexity in the empirical parameterizations vary greatly across models and processes, and the resolved scales. Many parameterizations employ an underlying paradigm that makes use of well known or well observed processes. Examples of this are land surface models that describe energy and mass exchanges in terms of a “big leaf” that occupies some fraction of a model grid box. Convective parameterizations describe a “big cloud” or a series of “big clouds”. In this way the fundamental dependence on the atmospheric state is ~~captured~~approximated, albeit only at a phenomenological level.

Parameters in climate models vary widely in their physical interpretation. Some parameters are well-determined physical values, such as the Coriolis parameter, the acceleration due to gravity, the Stefan-Boltzmann constant. Others, such as reaction rates for chemical or microphysical processes, may be inferred from laboratory or field measurements with some uncertainty.

Still others may emerge from the construction of parameterizations but not correspond directly to well-defined physical processes, e.g., “erosion rates” for clouds (Tiedtke, 1993). Others also emerge from the characterization of model sub-grid scale variations in the parameterizations, such as the “critical relative humidity” for cloud formation (Schmidt et al., 2006), or equivalent mixing rates for turbulent transport, and may be loosely approximated from either observations or higher resolution models, for example, Siebesma and Cuijpers (1995). Individual parameterizations for a specific phenomenon are generally calibrated to process-level data as much as possible using high resolution modeling and/or field campaigns to provide constraints. For instance, boundary layer parameterizations might be tuned to well-observed case studies such as in Larcfrom (Pithan et al., 2016) or DICE (http://appconv.metoffice.com/dice/dice.html). However, in some cases, even when parameter values are well-constrained physically or experimentally, simulations can often be improved by choosing values that violate these constraints. For example, Golaz et al. (2011) find that cooling by interactions between anthropogenic aerosols and clouds in GFDL’s AM3 model depends strongly on the volume-mean radius at which cloud droplets begin to precipitate~~; indirect aerosol forcing over the 20th Century is too large when the coupled~~. By altering CM3 ~~uses observationally constrained values of this radius but more realistic when values~~ after its configuration used for CMIP5 was established as described in Donner et al. (2011), Golaz et al. (2013) found that its 20th-Century temperature increase could be simulated more realistically (larger increase) using values for this threshold drop size smaller than observed ~~are used~~ (Pawlowska and Brenguier, 2003; Suzuki et al., 2013) (c.f. section 4.1 below). ~~While likely partly related to a rectified effect of unresolved variations following the same phenomenology~~

~~as the local effect but scaling differently, compensations among this behavior and other aspects of indirect aerosol forcing that have been modeled unrealistically are also possible.~~

~~Another example might concern the terms for horizontal and vertical diffusion, which reflect unresolved turbulence as well as the physical process of diffusion. Variations in the~~ Another example is the variation in the effective diffusion constants for momentum, moisture and temperature ~~have for instance~~ which has been used to decrease large root-mean-square errors in tropical winds in the NCEP model.

There remain a number of parameters that are not strongly constrained by process-level observations or theory but that nonetheless have large impacts on emergent properties of the simulation. It is these additional degrees of freedom that are used to "tune" or calibrate the emergent properties of the model against a selected set of target observations. The decisions on what to tune, and especially what targets to tune for, undoubtedly involve value judgments (Hourdin et al., 2016) (though see discussions in Winsberg (2012); Schmidt and Sherwood (2014); Intemann (2015) for more on the philosophical consequences of this). Notably, there isn't any obvious consensus in the modeling community ~~on~~ to what extent parameter choices should be guided by conforming to process-level knowledge as opposed to optimizing emergent behaviors in climate models. At many centers, the philosophy for the most part has been to tune parameters in ways that make physical sense, with the expectation that in the long run that should be the best strategy. Increasing skill in climate models over time does support this approach (Reichler and Kim, 2008).

Additionally, climate simulations depend not only parameter choices within an established model structure but also on the structural choices made in the parameterization itself. Examples include experimentation with alternate closures and triggers for the cumulus parameterization at GFDL during the development of GFDL AM3. Alternate closures and triggers yielded opposing effects on the realism of mean precipitation and the tropical wave spectrum (Benedict et al., 2013). In the development of the atmospheric component of the NCAR CESM2, the Atmospheric Model Working Group convened an expert panel to evaluate two candidate parameterizations for cloud macrophysics and convection (Bogenschutz et al., 2013; Park, 2014), with simulation characteristics central to the evaluation. Theoretically, all such structural choices could be coded to vary with a parameter and so there is no strong theoretical distinction between parameter and structural variations. In practice however perturbed physics ensembles (PPEs) do not span as wide a range of ~~results~~ structural variations as multi-model ensembles of opportunity (Yokohata et al., 2012).

~~The decisions on what to tune, and especially what targets to tune for, undoubtedly involve value judgments . This is more of a problem for complex simulations than it would be for a numerical calculation of the consequences of a well-specified theory, since there are far more degrees of freedom in building a climate model. This subjectivity has raised concerns that non-epistemic values (such as a modelers preference for inductive risk) might bias solutions , but this has not been demonstrated to be the case in practice~~ These examples suggest that it can be hard to distinguish model tuning from model development (writ large) in practice, since both happen concurrently. For our purposes, we define tuning as a change occurring within a fixed structural framework that does not involve adding new physics.

Targets for possible tuning fall into three classes. First there are targets that need to be satisfied in order for useful numerical experiments to be performed in the first place. The most important of these is a requirement of near energy balance at the top of the atmosphere and surface in an initial state of a coupled model. Without this, the coupled model will not be stable and will drift over time in order to compensate for the initial imbalance. Strictly speaking this is not tuning to an observed quantity, but rather is a tuning to a situation that was approximately inferred to hold in the “pre-industrial” (PI). Note that while the concept of a pre-industrial period is a little elusive (Hawkins et al., 2017), in this paper we refer to conditions around the mid-19th Century around 1850. To avoid dealing with the lack of sufficient observational data from the 19th Century, some modeling groups (see below) alternatively choose to tune to present-day (PD) conditions, including an energy imbalance at the top-of-the-atmosphere (TOA) as inferred from ocean observations today (Loeb et al., 2009). (Note that this imbalance is is ~~often~~ sometimes referred to as the radiative forcing perturbation (RFP) (or the effective radiative forcing) and is the change in net flux which occurs in a multi-year integration with specified, climatological present- day SSTs when emissions (primary aerosols and short-lived gases), long-lived greenhouse-gas concentrations (carbon dioxide, nitrous oxide, methane, and the halocarbons~~CFC-11, CFC-12, CFC-113, and HCFC-22~~) and solar irradiance are changed from ~~present-day to~~ pre-industrial to present day values). The consequences of these choices are discussed below. A second class of tuning targets are well-characterized climatological observations which might include annual means, average seasonal cycles or interannual variance. A third potential class are observations of transient events (at daily to centennial scales) or trends or even climate sensitivity itself.

~~It is important to note that some~~ Some observational targets have important (and sometimes unrecognized) structural uncertainties and therefore any tuning to those targets risks over-fitting the model to imperfect data, potentially reducing skill in “out-of-sample~~predictions~~” predictions (those for which the evaluation data either did not exist at the time of the prediction or was not used in model development or tuning). This is a particular problem for transient observations such as estimates of early 20th Century temperature changes (Thompson et al., 2008; Richardson et al., 2016), or pre-1979 sea ice extent (Meier et al., 2012; Walsh et al., 2016), of pre-1990 ocean heat content change (Levitus et al., 2000; Church et al., 2011), or water vapor trends (Dessler and Davis, 2010), which have all been corrected in recent years as non-climate artifacts in the raw observations have been found and adjusted for. In contrast, many climatologies over the satellite era are ~~far more~~ robust metrics whose estimates over any fixed period have not changed appreciably as understanding of the observations evolved.

Models equipped for data assimilation or that are used for operational forecasts have the additional possibility of tuning parameters to improve skill scores in those forecasts at multiple time scales - whether it’s 6 hours, daily, weekly, or even for many months for seasonal forecasts of, for instance, the state of the tropical Pacific.

We note here a distinction between fields that are closely monitored during the model development process (many examples are given below), and specific tuning targets. Changes in a monitored field, such as global mean temperature are kept track of, but unless the values stray beyond a nominal acceptable range no action to change the code would be taken. Monitored diagnostics tend to be complex emergent diagnostics that do not depend in any simple way to adjustable parameters, and thus are difficult (or impractical) to tune for. For example, note that the range of pre-industrial global temperatures in CMIP5 is [12.0,14.8]° C is noticeably wider than the uncertainties in that quantity (±0.5° C (Jones et al., 2012)).

The limitations of tuning are well-known (Mauritsen et al., 2012; Schmidt and Sherwood, 2014; Hourdin et al., 2016). First, it provides remarkably little leverage in improving overall model skill once a reasonable part of parameter space has been identified - for instance, tuning has been unable to resolve the persistent so-called “double ITCZ” problem (Lin, 2007; Oueslati and Bellon, 2015). Second, improvements in one field are often accompanied by degradation in others, thus the final choice of parameters involves (more) subjective judgments about the relative importance of different aspects of the simulations. For example, the Australian contribution to CMIP5 (ACCESS v.1) used a version of the UK Met Office atmosphere model with small modifications to mitigate problems in the tropics and Southern Hemisphere that affect Australian forecasts, at the ~~possible~~ expense of performance ~~in the UK~~ elsewhere (Bi et al., 2013). There are additionally many obvious biases in model simulations that persist across model generations, indicating that these aspects are robustly stubborn to development changes in the model (including the tuning) (Masson and Knutti, 2011).

Most discussions of tuning deal with explicit calibration of parameters to match a target observation. However, analysis of the CMIP3 ensemble (Kiehl, 2007; Knutti, 2008) suggested that there may have been some kind of implicit tuning related to aerosol forcing and climate sensitivity among a subset of models, with models with higher sensitivity having a tendency to have higher (more negative) aerosol forcing (this situation was less evident in CMIP5 (Forster et al., 2013)). Both of these correlations however seem rather low (CMIP3: 0.24; CMIP5: 0.19) and so do not provide evidence for a general tuning related to forcing and sensitivity. That models with accurate historical simulations must trade off forcing and sensitivity is not necessarily evidence they have been tuned to do so. Since the CMIP3 models’ aerosol forcings were not explicitly tuned to enforce the observed historical trend in temperature, the mechanisms that might explain this observation are unclear. With further data on the current top-of-atmosphere radiative imbalance (Allan et al., 2014; von Schuckmann et al., 2016), this issue will however need to be revisited for the latest generation of models.

Model selection can also act as an implicit form of tuning, even though this might be seen by others as simple model development. In deciding between two versions of a dynamical core or convection parameterizations, skill in El Niño/Southern Oscillation (ENSO) variability or reductions of ocean drifts may play an important role. Conceivably, a modeling center may decide not to release or use a particular version because it fails to meet certain criteria perceived to be essential ~~(see below for examples)~~though more generally this will simply spur further development. One candidate criteria would be a realistic simulation of the 20th Century, however the wide spread in 20th Century trends in the CMIP5 ensemble (Forster et al., 2013, Fig. 7) would indicate that this has not been ~~widely applied~~ generally applied (though see below for more detailed discussion of the use of historical changes).

Within climate models, there is always a choice as to whether to tune a specific component (such as the atmosphere, sea ice, land surface or ocean) with tightly constrained boundary conditions, or to tune the coupled model as a whole. In practice, both approaches are taken, though the relative importance and computation resources available vary across groups. Tuning components is generally fast and efficient, but does not necessarily prove robust when those components are coupled. However, coupled models take a very long time to equilibriate and their quasi-stable states may be too far from the observed climate to be useful. Assuming that models conserve energy appropriately, all control runs will eventually drift to a quasi-steady state with a ~~near zero~~ near-zero energy balance at the TOA and at the surface of the ocean. However, the realism of the final state is not guaranteed and indeed, given the long time constants in the ocean, might require many thousands of years of integration to get to the wrong answer. Thus a balance must be struck between approaches.

Table 1. Climate models discussed in the text.

Modeling group	Model	Reference
Department of Energy (DOE)	ACME 1.0	(in preparation)
NOAA Geophysical Fluid Dynamics Laboratory (GFDL)	CM3	Donner et al. (2011); Griffies et al. (2011)
NASA Goddard Institute for Space Studies (GISS)	GISS-E2/2.1	Schmidt et al. (2014)
NASA Global Modeling and Assimilation Office (GMAO)	GEOS5	Rienecker et al. (2008); Molod et al. (2015)
National Center for Atmospheric Research (NCAR)	~~CESM~~ CESM1	Gent et al. (2011); Hurrell et al. (2013)
NOAA National Center for Environmental Prediction (NCEP)	CFS v1 & v2	Saha et al. (2006, 2010, 2014)

3 Specific practices

Each of 6 US modeling centers described below have specific missions and foci that drive different aspects of their modeling. For instance, NASA GMAO and NCEP have operational data assimilation products for ~~short- term~~ short-term weather, longer seasonal forecasts and reanalyses, that form the core of their tasks. NCAR CESM, GFDL and NASA GISS have more long-term climate change issues at the forefront of their research, but each with different mandates - respectively, to be a community model, to advance NOAA's mission goal to understand and predict changes in climate, to help interpret and use NASA remote sensing products. The DOE ACME project has been tasked to a very specific role to serve DOE's energy planning and computational resource needs.

For each modeling group, we describe the principal targets and tuning strategies for their atmosphere-only GCM, their coupled ocean-atmosphere GCM, and additional ~~components (such as the carbon cycle or interactive atmospheric composition)~~earth system components as relevant. The specific models referred to are described in Table 1. We outline the commonalities of approaches and key differences in Section 4, and the discuss the implications and ways forward in Section 5.

3.1 DOE

The prototype version of DOE's Accelerated Climate Modeling for Energy (ACME v0) is closely related to the Community Earth System model (CESM). The initial version ACME v1 currently under development incorporates new ocean and sea-ice components (Model for Prediction Across Scales (MPAS)) (Ringler et al., 2013) as well as updated atmosphere and land components. ACME v1 is being developed at two horizontal resolutions. A low-resolution configuration which includes an atmosphere at approximately 1° and an ocean with varying resolution between 60 and 30 km. The high-resolution configuration is based on a 1/4° atmosphere and an eddy-permitting ocean resolution between 18 and 6 km.

Tuning is performed iteratively at the component levels and on the fully coupled system. Most of the component level tuning takes place in the atmosphere. The atmosphere is primarily tuned using short simulations (2 to 10 years) with climatological SSTs and sea ice boundary conditions, either for present-day (circa 2000) or pre-industrial conditions. The tuning targets a near zero TOA radiation balance for 1850 by adjusting cloud-related parameters. Overall simulation fidelity is another important aspect of the tuning process with the goal of minimizing errors in important climatological fields such as sea level pressure, short and long-wave cloud radiative effects, precipitation, near surface land temperature, surface wind stress, 300 hPa zonal wind, aerosol optical depth, zonal mean temperature and relative humidity~~, aerosol optical depth~~. The magnitude of the aerosol indirect effects is also evaluated and adjusted if deemed to be inconsistent with the observed historical warming. Cess climate sensitivity ~~is evaluated~~ (Cess et al., 1990) is monitored using idealized SST+4K simulations. The radiative imbalance in the 21st Century with observed SST must be positive with a target range of 0.5 to 1 W m^{-2} .

Most of the tuning is performed using the low-resolution atmosphere. However, cloud parameterizations need to be retuned separately for the high-resolution atmosphere. Because of the cost of the high-resolution atmosphere, ~~we have found it~~ it is more effective to use short hindcast simulations (Ma et al., 2015) to first evaluate the parameter space.

Tuning is also performed with the fully coupled system using perpetual pre-industrial or present-day forcing. Ocean and sea-ice initial conditions are either from rest (Locarnini et al., 2013; Zweng et al., 2013) or derived from separate CORE experiments (Griffies et al., 2009). Simulations vary in length from a decade to over a century. Priority metrics for the coupled pre-industrial simulations are top-of-atmosphere radiation, surface winds, sea ice extent and thickness (climatology and seasonal cycle), sea surface temperatures, stability of ocean heat content, meriodional heat transport, overturning circulations and the Nino3.4 index. Longer coupled simulations are often performed in pairs of perpetual present-day and pre-industrial forcing to monitor the combined impact of anthropogenic forcings and climate sensitivity and to maximize odds of successful historical simulations. To that end, parallel coupled simulations, one with perpetual 1850 forcings and one with perpetual 2000 forcings will be tested to ensure that the 2000 control simulation is indeed warmer that the 1850 control. Abrupt 4×CO2 experiments are also conducted to estimate the equilibrium climate sensitivity.

3.2 GFDL

In developing the GFDL atmospheric model AM3 and coupled model CM3, parameter choices and some structural choices as to how to deploy parameterizations were guided by multiple goals. In addition to choosing parameters within plausible ranges suggested by observations, experiments, theory, or higher-resolution modeling, these goals included simulating thermodynamic and dynamical fields, as well as TOA regional short-wave and long-wave fluxes, as realistically as possible. The global and annual mean net TOA radiative flux in integrations with specified, present-day (1981-2000) sea surface temperatures (SSTs) was tuned to a slight positive imbalance (0.8 W m^{-2}) within observational estimates (Loeb et al., 2009). Particular attention was also given to surface properties important for successfully coupling AM3 to models for sea ice (high-latitude surface energy balance) and ocean (wind stresses and implied ocean heat transports). Many of the changes in parameters from earlier

GFDL models or nominal values in literature describing the model parameterizations are summarized in Donner et al. (2011). For example, the momentum source in the Alexander and Dunkerton (1999) parameterization for gravity wave drag was chosen based on the stratospheric circulation it yielded. To facilitate optimizing input parameters to this parameterization, the orographic wave parameterization was limited in the vertical extent of its application. Additionally, the autoconversion threshold (volume-mean radius at which cloud droplets begin to precipitate), cloud erosion scales, and ice fall speeds in the Rotstayn (1997) and Tiedtke (1993) cloud microphysics and macrophysics parameterizations were tuned to improve regional patterns of TOA shortwave and long-wave fluxes, TOA shortwave and long-wave cloud radiative effects, the Earth's energy imbalance, precipitation, and implied ocean heat transports.

The choices of a closure based on convective available potential energy (CAPE) for the Donner (1993) deep cumulus parameterization and the relaxation time and CAPE threshold in that closure were primarily motivated by their effects on the precipitation simulation. Tuning vertical diffusion of horizontal momentum in the Donner (1993) deep and Bretherton et al. (2004) shallow cumulus parameterizations impacted tropical precipitation and surface wind stresses. Other tunings related to convection include changes in entrainment (partly to account for changes in vertical resolution), the moisture budget for mesoscale circulations associated with deep convection, and maximum heights for the mesoscale circulations. These tunings improved precipitation, shortwave cloud radiative effects, and implied ocean heat transports. Changes in lateral entrainment for shallow convection (Bretherton et al., 2004) also improved these fields, limiting excessive low cloudiness, in particular. The maximum heights of the mesoscale circulations also exerted a strong control on stratospheric water vapor. Between 100 and 10 hPa, zonally averaged water vapor mixing ratios are between 1.5 and 4 mg kg^{-1}, mostly within 0.5 mg kg^{-1} of HALOE (Halogen Occultation Experiment) and MLS (Microwave Limb Sounder) observations.

Aspects of AM3 related to variability, including stationary wave patterns, relationships between Niño-3 index and regional precipitation, relationships between the Northern Hemisphere Annular Mode and regional pressure and temperature patterns, tropical cyclones, and the tropical wave spectrum were monitored during AM3 development (Donner et al., 2011). Optimal tuning for mean state and variability in some cases conflicted. In AM3, this was particularly evident for the tropical wave spectrum, including the Madden-Julian Oscillation. Deep convective closures and triggers which produced a realistic mean simulation did so at the expense of of the tropical wave spectrum (Benedict et al., 2013).

AM3 includes prognostic aerosols based on emissions, transport, chemical processes, and dry and wet removal. An important aerosol tuning parameter is the strength of wet scavenging. In-cloud condensate fractions were prescribed to provide a reasonable simulation of the global mean and regional distribution of aerosol optical depth. These condensate fractions maintain relative solubilities among the various aerosols in AM3.

AM3 is the first GFDL model to include cloud-aerosol interactions. At the outset of this aspect of AM3 development, estimates of climate forcing by cloud-aerosol interactions ranged to -3 W m^{-2} (Lohmann and Feichter, 2005), and GFDL's AM2, modified to include cloud-aerosol interactions, yielded an associated climate forcing of -2.3 W m^{-2} (Ming et al., 2005). Since climate forcing by greenhouse gases is around 3 W m^{-2} (Stocker et al., 2013), the most extreme estimates of climate forcing by cloud-aerosol interactions would not be compatible with observed historical temperature increases. Given the approximate treatments of cloud-aerosol interactions in climate models, the possibility that some parameter combinations or formulations could lead to these extreme estimates could not be ruled out during model development. Indeed, Golaz et al. (2011) show that the magnitude of climate forcing by cloud-aerosol interactions depends strongly on the volume-mean drop radius at which cloud droplets begin to precipitate. Golaz et al. (2011) also find that assumptions regarding the sub-grid distribution of updraft speeds is an important control, though exerted through re-tuning for radiative balance as the distribution of updraft speeds is changed. The effective cloud-droplet radius and cloud droplet number concentration are both central to climate forcing by cloud-aerosol interactions and vary strongly with aerosol size distribution (Feingold, 2003; McFiggans et al., 2006). Ming et al. (2006), which is used to parameterize aerosol activation in AM3, supports a range of aerosol size distributions.

The TOA RFP was monitored during AM3 development, as was the Cess climate sensitivity. Configurations for which the ratio of the RFP to the Cess sensitivity fell substantially below its value for AM2 (The GFDL Global Atmospheric Model Development Team, 2004) were rejected. This imposes a bound on RFP which depends on AM3 sensitivity and the forcing/sensitivity ratio in AM2. Within the limitations of the Cess sensitivity, this ratio was used to compare without coupling the changes in global-mean surface temperature that might be expected from CM3 relative to CM2. The AM3 RFP is 0.99 W m^{-2} , with the aerosol contribution about -1.6 W m^{-2} (Golaz et al., 2013). The coupled model CM3 was not further constrained with respect to its simulation of 20th Century climate change. Although the ratio of RFP to Cess sensitivity for AM3 is only about 15% less than for AM2, 20th Century temperature increases in CM3 are less than observed, while CM2.1 temperature increases are greater than observed (Donner et al., 2011).

The CM3 coupled model was initialized from present-day ocean conditions and allowed to adjust to a pre-industrial, quasi-steady state with a small TOA energy imbalance (0.2–0.3 W m^{-2}). Tuning in CM3 was concentrated in the atmospheric component AM3. Outside of the atmospheric component, ~~in the~~ sea-ice~~model, dry-snow and ice albedowere set to values~~ , land, and vegetation albedo, along with snow masking, were tuned. These tunings improved the Atlantic Meridional Overturning Circulation in preliminary coupled configurations, prior to final tuning of the atmospheric component and subsequent initiation of the pre-industrial coupled control. The resulting albedos were generally more realistic than those ~~to which they had been tuned~~ used in CM2.1. The change was made possible by CM3's improved realism in regions with sea ice (Donner et al., 2011).

Note that the above description applies only to AM3/CM3. For CM4 (Zhao et al., 2016), development is ongoing, and the specific tuning practices will be documented in future papers.

3.3 NASA GISS

Tuning strategies in GISS ModelE2 are described in Schmidt et al. (2014). In the atmosphere-only simulation under 1850 pre-industrial conditions, the parameters in the cloud schemes that control the threshold relative humidity and the critical ice mass for condensate conversion are used to achieve global radiative balance and a global mean albedo of between 29 and 30%. Additionally, parameters in the gravity wave drag are chosen to optimize the simulation to the lower stratospheric seasonal zonal wind field and the minimum tropopause temperature. This also impacts high-latitude sea level pressure. In ocean-only simulations as described in the CORE protocol (Griffies et al., 2009), mixing parameters are chosen to minimise drift from observations in the basin-averaged temperature and salinity.

Upon coupling the ocean and atmosphere models, there is an initial drift to a quasi-stable equilibrium which is judged on overall terms for realism, including the overall skill in the climatological metrics for zonal mean temperature, surface temperatures, sea level pressure, short and long wave radiation fluxes, precipitation, lower stratospheric water vapor, and seasonal sea ice extent. For the configuration to be acceptable, drifts have to be relatively small and quasi-stable behavior of the North Atlantic meridional circulation and other ocean metrics, including the Antarctic Circumpolar Circulation, are required. While ENSO metrics are also monitored, they are not specifically tuned for.

Subsequent to CMIP5, further tuning exercises and development has occurred for the production of the E2.1 version of the model. One important tuning success ~~in developing the CMIP6 models~~ were the adjustments made to the convection scheme in order to allow for the simulation of the Madden-Julian Oscillation (MJO) (Kim et al., 2012). A combination of greater entrainment and the addition of a subgrid- scale explicit "cold pool" feature, greatly enhanced variability at MJO timescales and lead to greatly increased forecast skill in initialized 20-day simulations (Del Genio et al., 2015).

Further fine tuning in the coupled models, for instance for the exact global mean surface temperature, is effectively precluded by the long spin-up times and limited resources available. No tuning is done for climate sensitivity or for performance in a simulation with transient forcing or hindcasts. In transient simulations without an explicit ~~indirect aerosol effect, this~~ aerosol indirect effect, the AIE was preset to have a value of -1W m^{-2} in 2000 in the CMIP5 simulations (Miller et al., 2014), while configurations with aerosol microphysics have free latitude to produce whatever forcing is calculated.

~~For the CMIP6 submissions, the tuning will be done predominantly with pre-industrial and present-day fully interactive simulations (including chemistry and aerosols) and the non-interactive versions will use the composition derived from those simulations.~~ In simulations with interactive atmospheric composition, there are two specific tunings for ozone chemistry: the photolysis rate in the atmospheric window region for incoming solar radiation, and temperature threshold for the formation of polar stratospheric clouds (and hence the heterogeneous chemistry associated with them) (Shindell et al., 2013). The former is tuned so that N_2O and O_3 fields in the lower tropical stratosphere match observations, while the latter can be used to ensure that the polar ozone hole timing is correct despite potential biases in polar vortex temperatures. With respect to dust aerosols, emissions are tuned so that the model can match retrieved aerosol optical depths for the present-day (Miller et al., 2006), similarly tuning of the lightning parameterization (and associated source for NO_x) is done against modern observations of flash rate, and tropospheric ozone amounts.

For the E2.1 model and subsequent CMIP6 submissions, all tuning is being done with pre-industrial and present-day fully interactive simulations (including chemistry and aerosols and indirect effects) and the non-interactive versions will use the composition derived from those simulations and the same tuning.

3.4 NASA GMAO

The Goddard Earth Observing System (GEOS) model is currently in use at the NASA GMAO at a wide range of resolutions and for a wide range of applications. The range of resolutions and applications for the atmospheric model includes global mesoscale simulations/forecasts at approximately 7 km, atmospheric data assimilation and forecasts at 12 km (with ensemble members running at 50 km), seasonal coupled atmosphere-ocean forecasts at approximately 50 km, present day climate simulations at

100 km, and present day coupled chemistry climate simulations at resolutions from 12 km to 100 km. The tuning of the GEOS-5 AGCM physical parameterizations, therefore, is designed to allow the model to function across this range of uses and requires fidelity in many aspects of the simulation. The tuning also includes appropriate resolution dependence. Tuning targets differ among the many types of experiments that are conducted as part of the model validation suite. The tuning suite includes present day (AMIP-style) climate simulations, "replay" experiments at different resolutions (similar to nudging towards a reanalysis), coupled atmosphere ocean experiments, coupled atmosphere-chemistry simulations, short term forecasts and data assimilation experiments.

The tuning of the current version of the GEOS-5 AGCM is described in Molod et al. (2015) which shows the results of a series of sensitivity experiments demonstrating the impact of each change in tuning. The substantial majority of the tuning is focused on the behavior of the moist and turbulence parameterizations, and also includes a parameter change in the gravity wave drag scheme. For the lower resolution applications and uses, systematic comparisons of seasonal mean prognostic fields with different reanalysis estimates, and comparisons of cloud properties with satellite based estimates are used to identify errors in the mean present day climate. Iterative 30-year simulations at low resolution (100 km) and repeated comparisons ensures that a change in tuning to ameliorate one bias does not inadvertently exacerbate another.

A key tuning target is matching the mean and variance of the spatial distribution of CERES observations of all sky TOA long-wave and shortwave radiative fluxes, together with the daily TOA long- wave and short-wave distributions independently. The contribution of cloudy effects is approached by adjusting the parameters that describe the cloud radiative effect (cloud particle size and autoconversion rates). The clear sky portion of the TOA fluxes is matched by tuning the parameters that govern the mean atmospheric humidity and surface albedo over ice covered surfaces. The free atmosphere specific humidity is quite sensitive to the "critical relative humidity" specified in the cloud macro-physical scheme (Molod, 2012), and so although this parameter is largely dictated by observed subgrid scale moisture variations, the fine tuning and the details of the vertical profile are tuned to match a consensus of reanalysis estimates of specific and relative humidity and SSM/I total precipitable water.

The boreal winter mean circulation as compared to reanalyses (as seen by the 200 hPa eddy height or by the 300 hPa velocity potential) was found to be quite sensitive to the intensity of the hydrological cycle, largely dictated by the rates of re-evaporation or sublimation of rain and snow. These parameters are chosen so as to ensure agreement of the seasonal mean circulation with reanalysis, the seasonal mean precipitation with observations from GPCP and TRMM, and the agreement of the cloud radiative effects with CERES and with SRB at the surface. The behavior of the atmosphere-ocean coupled system is particularly sensitive to the geographical distribution of the surface shortwave cloud radiative forcing in the tropics.

At lower resolutions, coupled chemistry climate simulations are used to bring ~~together~~ to bear MERRA-2 (Gelaro, R. and co-authors, 2016) reanalysis estimates and satellite observations (MODIS) of seasonal mean aerosol content. These estimates are largely used to constrain the tuning of the GEOS-5 surface and atmospheric turbulence parameterizations. The choice of the turbulent length scale and the choice of parameters that govern the entrainment into buoyantly rising turbulent parcels of air are made so as to constrain the turbulent transport of aerosol. The extent of vertical mixing as well as the advective transport out of the source regions are governed by this choice of tuning parameters.

The GEOS-5 AGCM includes some resolution dependent parameters that govern the behavior of the moist processes. The two most important parameters that are specified to change with resolution in an ad hoc manner are chosen based on physical arguments and based on results from GEOS-5 global mesoscale simulations. The first of these is the critical relative humidity for condensation/evaporation, which accounts for subgrid scale variations of total water. Critical RH increases with resolution based on the expectation and evidence from global mesoscale model results that subgrid scale variations of total water decrease with increasing resolution (Molod, 2012). The second of the resolution dependent parameters is the so-called Tokioka limit in the convective parameterization. Again based on the expectation that the larger convective motions are resolved explicitly and on evidence from global mesoscale model results, the parameters that govern the stochastic Tokioka limit changes so as to restrict parameterized deep convection at higher resolutions.

At the higher resolutions (25 km and better) the tuning parameters are chosen based on short term forecasts and the behavior as part of the data assimilation system. Forecast skill scores, the fidelity of the spinup of tropical cyclones and the innovation vector for data assimilation (Observation-Forecast statistics) are critical relevant metrics for new tuning choices, and any new choices of tuning parameters are evaluated with an ensemble of forecasts. The analysis increments during both data assimilation and replay experiments provide the key guidance for choosing the parameters to tune. Under the general assumption that the mean analysis increments indicate systematic errors in the model physics (which is not always valid), correlations between the tendency term from any individual physical parameterization and the analysis increment reveals errors due to the behavior of that parameterization, and parameters of that scheme are adjusted so as to minimize the mean analysis increments.

High resolution forecasts are also evaluated and tuned based on comparisons with spatial and temporal variability of high resolution top of the atmosphere fluxes and radar-derived precipitation. As with the lower resolutions, the parameters which are adjusted to meet the tuning targets are the autoconversion/ice-fall rates and the cloud drop size. In addition to these parameters, high resolution tuning also includes adjustments of the Tokioka limit and the time scale of adjustment in the convective parameterization.

The ability to spin up tropical cyclones and match the correct track was found to be quite sensitive to the magnitude of low level drag. Based on theoretical considerations and the results of laboratory experiments, the model's function which relates surface stress to roughness height over the oceans (the "Charnock coefficient") was adjusted to decrease the drag at high wind speeds and resulted in substantial improvements in the simulation of tropical cyclones (Molod et al., 2013).

In addition to the tuning based on physical reasoning and diagnosis of errors using comparisons with observations, some tuning choices are based on trial-and-error experimentation. These include parameters that govern the magnitude of the different types of surface drag (more drag increases forecast skill score) and the adjustment time scale of mid-latitude parameterized convection (more mid-latitude convection increases forecast skill score).

The suite of different types of experiments with the GEOS-5 GCM at different resolutions are run iteratively as part of the overall tuning process, and the result is a model which meets the variety of tuning targets described here. The trade offs among the parameter choices to meet the different targets exist, and necessitate prioritization of the tuning targets, but in general this process results in a robust model that functions well in the various applications needed to fulfill the GMAO's goals and mission.

3.5 NCAR

The Community Earth System model (CESM, Hurrell et al. (2013)) is a joint NCAR and university-wide activity and governance takes place through a working group structure. Working groups are teams of scientists that contribute to the development of each individual component (atmosphere, land, ocean, sea-ice, land-ice, chemistry and bio-geochemistry) and relevant topics (such as climate variability or climate change).

Tuning begins as a generally separate activity for each components within the working groups. During this initial phase of tuning, periodic pre-industrial control coupled simulations are performed as a check on the impact of each components' developments to date on the whole coupled system, and to insure features of the simulation have not significantly degraded.

The atmosphere model tuning strategy initially performs 'stand-alone' experiments using the AMIP protocol with interactive land and atmosphere components and with prescribed observed Sea Surface Temperatures (SSTs) and sea-ice distributions. Initial development testing is performed using SSTs of the climatological period centered around the year 2000 for 5–10 year periods. This length of simulation is necessary due to the high-arctic variability. The first key measure of a simulation that will be appropriate to the fully coupled simulation is the TOA energy balance. Estimates of the observed present-day energy imbalance are of order 0.5–1.0 W m^{-2} (Loeb et al., 2009) and the aim is to achieve close to that through modification of cloud related fields that have an impact both on the short wave and long-wave components of the energy budget. The first quantitative assessment of simulation fidelity is given by summary RMSE and bias scores for a number of variables key to the fully coupled system including surface stresses, precipitation, temperature, cloud forcings and surface pressure. A secondary assessment involves 'pre-industrial minus present day' simulations to determine the aerosol indirect effects ~~we may expect to see~~ that would be expected in historical coupled simulations. This involves ensuring that the net aerosol forcing isn't greater in magnitude than about (negative) 1.5 W m^{-2} .

In parallel to the atmosphere component activities, the ocean and ice working groups perform equivalent 'stand alone experiments' with forcing provided by multiple cycles of the CORE forcing protocol (Griffies et al., 2009). The phenomena of key importance are the meridional overturning circulation, particularly in the North Atlantic, Gulf Stream separation, Drake passage flow, equatorial thermocline depth and SSTs in the Pacific. The land tuning approach uses land-only configurations forced by bias-corrected reanalysis-based meteorological forcing products. Metrics of performance are generally assessed for leaf area index, gross primary productivity, river discharge, latent heat flux and vegetation and soil carbon stocks. Other physical components of the coupled system, including land-ice and bio-geochemistry, will also be developed and tuned in parallel within their respective working groups.

Ideally, this would translate into a well-tuned atmosphere into a configuration with SSTs, sea-ice and land conditions relevant to the pre-industrial, and a simulation that should in ~~principal~~principle, would translate well to a coupled system close to energy balance i.e., with no net increase or decrease of energy into the whole coupled system. However, coupled system biases in the surface distribution of SSTs, sea-ice means that tuning also needs to be performed in the fully coupled system.

Coupled model tuning brings together the individual fully active 'tuned' components and their associated working groups to perform a series of pre-industrial climate experiments. The same performance metrics applied in atmosphere AMIP simulations apply to the coupled simulation; namely top of atmosphere zero energy imbalance. An equilibrium energy imbalance is the most challenging task in coupled CESM tuning. The difficulty lies in spin-up and drift of the system. Two ocean initialization approaches are used. The first is to use an observed Levitus temperature and salinity state with the ocean at rest. The second approach is to initialize from an ocean state of a previously run simulation. This has the advantages of a spun-up ocean state, and in particular the deep ocean and that is more 'familiar' with the overlying atmosphere component. However it is undesirable from the perspective of simulation provenance. A combination of the two are used. If the equilibrium energy imbalance is greater than 0.1–0.2 W m^{-2} then the system will need to be retuned, again most commonly through minor adjustments of cloud radiative impact parameters. If the energy imbalance and surface temperature drifts are observed to be small in short decadal runs, then longer 50–100 year simulations are performed to analyze longer to determine whether the performance of the ocean-ice only simulations translate to the fully coupled system.

For the coupled simulations to be considered successful, they have to satisfy many of the requirements outlined above in addition to the dominant ENSO mode of variability; also a very challenging task. For instance, the initial implementation of more advanced convection parameterizations in CAM6 gave rise to a degradation in ENSO performance, but with some tuning to those schemes, ENSO performance skill was enhanced. Another example of a coupled issues that arose in constructing the CMIP6 version of the code were a persistent cold bias and excessive sea ice in the Labrador Sea, which was mitigated by more accurate routing of local river runoff. In previous versions (such as CCSM4), there were evaluations of the coupled model in historical transient mode, specifically of the September Arctic sea ice trend from 1979 which was improved after adjustments to the sea ice albedo formulation to affect the PI ice thickness (Gent et al., 2011). A “reasonable” historical temperature trend remains the primary metric of success, but no attempts are made to fine-tune it.

3.6 NCEP

In recent history two fully coupled climate models have become operational at NCEP, the Climate Forecast System (CFS) version 1 (Saha et al., 2006) and CFS version 2 (Saha et al., 2010, 2014). For the most part, the CFS and its predecessors (since there have been global climate models at NCEP since 1995) have been developed in the same way as weather prediction models. Indeed, the atmospheric component of the CFS is taken from the Global Forecast System (GFS), which is the NCEP flagship that makes weather forecasts from day 1 to 15. Verification against independent future reality (the weather happening worldwide every day) shows the GFS and similar operational models elsewhere steadily improving their skill scores on independent data over the last 50 years.

The daily verification skill scores ~~obviously contain the seeds for~~ are the dominant source for tracking model improvement. This is a powerful target for tuning which confronts the model with real-time observations in evolving data assimilation (DA) systems and then verifying the forecasts, from the initial conditions provided by these DA systems, with independent observations.

A new CFS is built by taking a snap-shot of the latest state-of-the-art GFS as its atmospheric component, along with state-of-the-art ocean, sea-ice and land models which are available at that time. In developing CFSv1 in 2002, a 'large' (≈10) number of candidate coupled ocean-atmosphere models were constructed, which were then run on a limited number of test cases, with differing vertical and horizontal resolutions, as well as with different physics parameterizations, such as convection and radiation schemes. The results were then judged, along with the normal verification metrics, on whether the 9- month predictions produced skillful ENSO predictions. Our goal at that time was to be competitive with the statistical/empirical models that were predominantly being used for ENSO predictions. After initial testing, the model version that gave the best ENSO predictions was used to make retrospective forecasts over a 20+ year period (going back to 1982) in order to calibrate (remove the systematic bias in) the model forecasts and to make *a priori* skill assessments. These were then used in subsequent real-time operational forecasts made by the CFS. Since it is very expensive to make retrospective forecasts over long periods (20-30 years) for every imaginable model configuration, the preliminary test over a set of limited cases was extremely important. The dominant ~~change~~ changes that improved skill ~~(specifically in convection) was an increase in vertical layers~~ were associated with tunings in the convection as model vertical layers were increased from 28 to 64 levels.

Having achieved some success in the prediction of ENSO in seasonal forecasts out to 9 months in CFSv1, the goal for CFSv2 was to tackle sub-seasonal predictions, mainly of the MJO in the tropics. Prediction of the MJO from 5 days was successfully extended to nearly 21 days by improving model physics and having a high resolution state-of-the-art data assimilation system to assimilate direct satellite radiance data. Also, greenhouse gas (GHG) concentration changes were implemented in the NCEP forecast system. While the NCEP focus is short-term (seasonal) climate prediction, it has been recognized that even for these predictions, the forecast needs to be warmer than a ’normal’ that, by necessity, is based on past data. The increase in GHGs also played an important role in improving the data assimilation of satellite radiance data. Each satellite over the 1979-present history was calibrated using GHG concentrations observed at the time these satellite were operational. The result was a reasonable upward temperature trend over 1979-present period, much better than at the time of CFSv1, when the upward trend over land was brought about only by the warming in initial global ocean conditions. As described in Saha et al. (2014), the seasonal prediction model may not be exactly the same as the model used for weather forecasts. In the absence of data assimilation, coupled ocean- atmosphere models can drift and produce, for instance, a very cold Pacific ocean due to a boundary layer parameterization change in weather model that produced more marine stratus clouds, but which became excessive in the fully coupled runs. This change was thus reversed in the seasonal simulations.

Development is now underway for the next model, CFSv3. NCEP/EMC has a strategic plan to unify the global forecast systems and develop a Unified Global Coupled System (UGCS) for both weather and seasonal climate prediction. This system will have six fully coupled model components, namely the atmosphere, ocean, sea- ice, land, waves and aerosols. It will also have a strongly coupled data assimilation system in each of these six components.

4 Commonalities and differences

As might be expected the broad picture of tuning across the climate model groups is consistent. The key adjustable parameters are those associated with uncertain and poorly constrained processes such as clouds, convection, gravity wave drag, and ocean mixing parameters. Common too are the broad array of targets against which skill of the models are judged e.g. the TOA short-wave and long-wave radiation, 500 hPa geopotential height, surface temperatures, sea level pressure, precipitation, etc.

However it is also abundantly clear that the procedures at each group are quite distinct and can reasonably be surmised to reflect different scientific priorities and missions, and thus will produce different outcomes.

The model groups also differ in whether they focus on pre-industrial conditions or present day simulations. The former has the benefit of being closer to climate stability, while the latter has substantially more observational data. The groups focusing on the pre-industrial are judging (mostly correctly) that the errors in the control simulation (whether run for pre- industrial or present) are larger than the trends between those periods. A stark difference does exist between the models that have operational data assimilation products (NCEP and GMAO) and those who don't. The ability to assess improvements in fast physics based on short forecasts is an excellent resource that, even if the climate models were not run operationally in this way, ~~it~~ should become a more widely used ~~test~~methodology (e.g. Hurrell et al., 2006). Recent experience with this mode of testing in the GISS model has shown very positive results for representation of the MJO and tropical convection (Del Genio et al., 2015).

Groups also differ on ~~whether they judge metrics based on whether they~~ which metrics they monitor for whether the are within an acceptable range, ~~usually a spread that is wider than observational uncertainty,~~ or if a specific value is tuned for directly (for instance, as for the present day energy balance for some groups in Table 2).

There are also some clear commonalities in approaches. All groups focus on atmospheric models at first either in an AMIP-style mode (annually varying modern SST/sea ice), or using a climatological approach (decadal mean observed ocean conditions and forcings), or in weather forecast mode. Tunings for atmospheric composition and key atmospheric diagnostics use these experiments which have the advantage of fast equilibration times and reduced computational load. Tunings for ocean components can be done in stand-alone experiments, but often are done within the full coupled framework, with at least some model groups tuning sea ice and ocean mixing parameterizations to produce acceptable sea ice cover and ocean circulation metrics. ~~The former approach can produce a wider array of model outcomes, but the latter risks over-fitting and a potential loss of predictive skill.~~

4.1 Use of recent trends and present-day radiative imbalance

Because of the high importance and visibility of climate models' simulation of the historical period (PI to PD), model groups have to be particularly clear in how information that reflects the ongoing trends in temperature and ocean heat content have been used in the tuning process.

The descriptions above suggest increasing knowledge over time about the current radiative imbalance has clearly influenced model development. Developers prior to CMIP3 (circa 2004) had a general expectation that net radiative forcing over the 20th Century was positive, but they were not able to use a specific value for the present-day energy imbalance because oceanic analyses were not accurate enough: compare ~~to~~ Levitus et al. (2000) to Allan et al. (2014) for instance. Thus a posterior quantitative test of the model imbalance in coupled runs compared to (improving) observations was a valid test of skill (Hansen et al., 2005). This may not be true for a large fraction of simulations in CMIP6.

We ~~summarized~~ summarize the results in Table 2. None of the models described here use the temperature trend over the historical period directly as a tuning target, nor ~~do~~ are any of the models ~~here tune~~ tuned to set climate sensitivity to some pre-existing assumption. However, NCAR, GFDL and DOE do tune for a global radiative imbalance at near present-day

Table 2. Use of historical period trends and imbalances during the tuning process

Modeling Group	Historical Temp. Trend	Radiative Balance (PI)	Radiative Imbalance (PD)	Aerosol Forcing (as tunable parameter)	Aerosol Indirect Effect (AIE)
DOE	Yes[1]	Yes[A]	0.5–1.0 W m^{-2} [A]	Yes	Yes
GFDL	No	No	Yes, $<$1.0 W m^{-2} [A]	No[2]	Yes
GISS	No	Yes[A]	No	Yes/No[3]	Yes
GMAO	No	N/A	No	No	No (pending)
NCAR	Yes/No[4]	Yes[C]	0.5–1.0 W m^{-2} [A]	$<$1.5 W m^{-2}	Yes
NCEP	No	N/A	No	No	No

[A] Using atmosphere-only/AMIP simulations.

[C] Using coupled ocean-atmosphere simulations.

[1] PD has to be warmer than PI.

[2] However sensitivity and forcing were jointly constrained w.r.t. the previous model.

[3] Set in simulations with non-interactive composition only.

[4] It was a necessary criteria for CCSM4, but not specifically tuned for.

conditions. For instance, GFDL AM3 with observed SSTs was tuned to have a positive imbalance, with a magnitude less than about 1 W m^{-2} for 1981–2000.

As discussed above, the radiative imbalance can be affected in two ways, by adjusting internal parameters (mostly associated with clouds), and/or by using a different historical forcing. Four models adjust their historical aerosol forcing: GISS, though only in its non-interactive runs, aims for an indirect aerosol forcing of -1 W m^{-2} (Schmidt et al., 2014); NCAR CESM and DOE ACME tune for a substantive positive effective radiative forcing at near-present conditions (implying a limit of ~~about~~ -1.5 W m^{-2} for aerosols); GFDL AM3 constrained its ratio of Cess sensitivity to ~~total effective radiative forcing~~ RFP to be close to its value in its prior-generation coupled model, which implied an aerosol forcing around -1.6 W m^{-2} in AM3.

At least three of the model groups discussed here find a difference between the energy imbalance using year 2000 forcings together with observed SST and sea ice, and the transient coupled simulations for the same time-period and forcings. However the differences in how this calculation is done can be important and the implications for the coupled model simulations are unclear. For example in the GISS-E2 model, the decadal mean imbalance (1996–2005) in AMIP simulations, including all forcings and annually varying observed SST and sea ice, is 1.25 W m^{-2} (for 1981–2000 it is 0.6 W m^{-2}). However, using the decadal mean SST and sea ice for the same period and constant yr 2000 forcings, the imbalance ~~(RFP)~~ is much larger, 1.74 W m^{-2} . Furthermore, the decadal mean imbalance in coupled simulations with the same forcings is $\approx$1.0$\pm$0.1 W m^{-2} (Miller et al., 2014). Similarly, for the GFDL AM3 AMIP runs from 1981–2000, the radiative imbalance is 0.8 W m^{-2} , while the imbalance for the same period in the coupled historical runs has an ensemble mean of 0.4 W m^{-2} . The differences depend critically on the patterns of SST and sea ice~~–~~, related to both the rectification of interannual variability, and the offsets in the coupled model climatology compared to observations. The question that is raised by this is whether, given the increase in forcings over the historical period, and the sensitivity each model has, does tuning the present-day imbalance (however defined) determine (even to zeroth order) the coupled imbalance, the “committed warming” (at constant concentrations) for the model, or the historical trend? With a perfect coupled model, and perfect knowledge of the forcings, this might be the case, but the imperfections in both imply that tuning to the PD imbalance is ~~not a very strong constraint~~ less of a constraint than might be assumed.

5 Discussion and future approaches

As models are continually evaluated at the process-level against an increasing number of observations, analyses often show that existing parameterizations lack enough flexibility to represent the coupling between the sub-gridscale and the environment in all relevant climate regimes. The response is often to increase the complexity of a parameterization, which comes at the cost of an increased number of tunable parameterization parameters. With that increase, the challenges faced by the developers also rise and the potential for “local minima” to occur i.e. different parameter combinations have similarly good agreement according to standard GCM validation metrics (e.g. Taylor Diagrams, climate state mean biases, spatial correlations).

If these distinct/separate volumes of tuning parameter space ~~lend~~ lead to simulations that exhibit similarly good agreement with observations, there is no clear scientific reason to prefer one over another. ~~So the question arises as to whether~~ But will our decisions on parameter combinations today have noticeable impact on the simulated climate several centuries from now or to climate sensitivity more broadly? Specifically, does choosing different local minima in parameter phase space ’matter’?

With more combinations, is there room for improving regional biases in simulations while simultaneously making the tuning process more automated? These questions have motivated an effort, using the GISS model as a test-bed, for developing a more robust framework for assessing the true existence of local minima in a multi-dimensional space (see also Hourdin et al. (2016)). This is being explored by incorporating situational or regime-dependent errors in observations or regional biases in GCM fields in weighted cost functions that define model “goodness”. We hope that this endeavor will increase the objectivity for deciding on the most appropriate tuning parameters and either lead to improved metrics for diagnosing the fidelity of a particular model, or reveal the spread in simulated climate sensitivity arising from settling on very different, but seemingly optimal, combinations of tuning parameters.

More generally, the large variety of approaches demonstrated among just these 6 models indicates that the documentation of tuning procedures across a multi-model ensemble like CMIP6 will be quite challenging. What role should the degree of tuning matter when assessing the coupled model skill? Should simulations be up-weighted in the ensemble because of a closer climatology to observations, or down-weighted because this is partly due to accommodation? Should models that are tuned differently but have similar physics be treated as independent or not? (Annan and Hargreaves, 2016; Knutti et al., 2017). These questions play into more fundamental issues related to how one should think about an unstructured ~~multi-model~~ multi- model ensemble (i.e. Knutti et al. (2010, 2013)).

At minimum, we recommend that all future model description papers (or systematic documentation projects such as ES-DOC http://es-doc.org) include a list of tuned-for targets, ~~and describe~~ monitored diagnostics, and describe clearly (as in Table 2) their use of historical trends and imbalances in the development process.

Table 3. Code Availability at each Center

Model	Code URL and Collaboration Policies (if relevant)
DOE ACME	https://climatemodeling.science.energy.gov/projects/accelerated-climate-modeling-energy https://climatemodeling.science.energy.gov/sites/default/files/publications/ACME_collaboration_30Jun15.pdf
GFDL CM2.1	https://www.gfdl.noaa.gov/modeling-systems-group-public-releases/
GISS ModelE2	https://www.giss.nasa.gov/tools/modelE/
GMAO GEOS-5	http://geos5.org/wiki/index.php?title=GEOS-5_public_AGCM_Documentation_and_Access
NCAR CESM1	http://www.cesm.ucar.edu/models/cesm1.2 http://www.cesm.ucar.edu/models/cesm1.2/copyright.html
NCEP CFSv2	http://cfs.ncep.noaa.gov/cfsv2/downloads.html

While we have only discussed tuning in the context of historical and modern simulations, it is vital to ~~assessing~~ assess the credibility of models by examining ~~the performance of models~~ their performance in out-of-sample situations. This is easy for the models with an operational weather forecast mode (at least for some aspects of the climate system), and participation in paleo-climate model tests by NCAR and GISS are also invaluable. Medium term climate forecasts based on anticipated changes in forcings (such as the eruption of Mt. Pinatubo (1991) or the rise in greenhouse gases have been shown to have skill (Hansen et al., 1988, 1992; Hargreaves, 2010). The importance (or lack thereof) of tuning always needs to be seen within that context. This paper alone cannot hope to answer all of the above questions, but we hope that it can contribute to a more transparent and more widely usable discussion.

Data and Code availability

~~No data or code is presented~~ The availability of code for the models discussed in this paper is laid out in Table 3, though note that some decisions/tunings mentioned above are associated with current development which may only be accessible with a collaborative agreement in place.

Acknowledgements. Climate modeling at GISS and GMAO is supported by the NASA Modeling, Analysis and Prediction program and resources supporting this work were provided by the NASA High-End Computing (HEC) Program through the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center. 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. NCEP is a division of the National Weather Service in NOAA within the Dept. of Commerce. Discussions at the US Climate Modeling Summit convened by USGCRP in February 2016 were instrumental for putting this paper together. Manuscript reviews by Larry Horowitz and Levi Silvers (GFDL) are appreciated. We'd like to thank Steve Sherwood and an anonymous reviewer for constructive comments on the discussion version of the paper.

References

Alexander, M. and Dunkerton, T.: A spectral parameterization of mean-flow forcing due to breaking gravity waves, J. Atmos. Sci., 56, 4167–4182, 1999.

Allan, R. P., Liu, C., Loeb, N. G., Palmer, M. D., Roberts, M., Smith, D., and Vidale, P.-L.: Changes in global net radiative imbalance 1985-2012, Geophysical Research Letters, 41, 5588–5597, doi:10.1002/2014gl060962, 2014.

Annan, J. and Hargreaves, J.: On the meaning of independence in climate science, Earth System Dynamics Discussions, pp. 1–17, doi:10.5194/esd-2016-34, 2016.

Benedict, J., Maloney, E., Sobel, A., Frierson, D., and Donner, L.: Tropical intraseasonal variability in Version 3 of the GFDL atmosphere model, J. Climate, 26, 426–449, doi:10.1175/JCLI-D-12-00103.1, 2013.

Bi, D., Dix, M., Marsland, S., O'Farrell, S., Rashid, H., Uotila, P., Hirst, A., Kowalczyk, E., Golebiewski, M., Sullivan, A., Yan, H., Hannah, N., Franklin, C., Sun, Z., Vohralik, P., Watterson, I., Zhou, X., Fiedler, R., Collier, M., Ma, Y., Noonan, J., Stevens, L., Uhe, P., Zhu, H., Griffies, S., Hill, R., Harris, C., and Puri, K.: The ACCESS coupled model: description, control climate and evaluation, Aus. Met. Oceanogr. J., 63, 41–64, 2013.

Bogenschutz, P., Gettelman, A., Morrison, H., Larson, V., Craig, C., and Schanen, D.: Higher-order closure and its impact on climate simulations in the Community Atmosphere Model, J. Climate, pp. 9655–9676, doi:10.1175/JCLI-D-13-00075.1, 2013.

Bretherton, C., McCaa, J., and Grenier, H.: A new parameterization for shallow cumulus parameterization and its application to marine subtropical cloud-topped boundary layers, Mon. Wea. Rev., 132, 864–882, 2004.

Cess, R., Potter, G., Blanchet, J., Boer, G., Genio, A. D., D'equ'e, M., Dymnikov, V., Galin, V., and Co-Authors: Intercomparison and interpretation of climate feedback processes in 19 atmospheric general circulation models, J. Geophys. Res., 95, 16,601–16,615, doi:10.1029/JD095iD10p16601, 1990.

Church, J. A., White, N. J., Konikow, L. F., Domingues, C. M., Cogley, J. G., Rignot, E., Gregory, J. M., van den Broeke, M. R., Monaghan, A. J., and Velicogna, I.: Revisiting the Earth's sea-level and energy budgets from 1961 to 2008, Geophys. Res. Lett., 38, L18601, doi:10.1029/2011GL048794, 2011.

Del Genio, A. D., Wu, J., Wolf, A. B., Chen, Y., Yao, M.-S., and Kim, D.: Constraints on Cumulus Parameterization from Simulations of Observed MJO Events, Journal of Climate, 28, 6419–6442, doi:10.1175/jcli-d-14-00832.1, 2015.

Dessler, A. E. and Davis, S. M.: Trends in tropospheric humidity from reanalysis systems, J. Geophys. Res., 115, doi:10.1029/2010jd014192, 2010.

Donner, L.: A cumulus parameterization including mass fluxes, vertical momentum dynamics, and mesoscale effects, J. Atmos. Sci., 50, 889–906, 1993.

Donner, L. J., Wyman, B., Hemler, R., Horowitz, L., Ming, Y., Zhao, M., Golaz, J.-C., Ginoux, P., and Coauthors: The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component of the GFDL global coupled model CM3, J. Climate, 24, 3484–3519, doi:10.1175/2011JCLI3955.1, 2011.

Feingold, G.: Modeling of the first indirect effect: Analysis of measurement requirements, Geophys. Res. Lett., 30, doi:10.1029/2003GL017967, 2003.

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

Gelaro, R. and co-authors: The Modern-Era Retrospective Analysis for Research and Applications, Version-2 (MERRA-2), J. Climate, p. in press, 2016.

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., Lawrence, D. M., Neale, R. B., Rasch, P. J., Vertenstein, M., Worley, P. H., Yang, Z.-L., and Zhang, M.: The Community Climate System Model Version 4, Journal of Climate, 24, 4973–4991, doi:10.1175/2011jcli4083.1, 2011.

Golaz, J.-C., Salzmann, M., Donner, L., Horowitz, L., Ming, Y., and Zhao, M.: Sensitivity of the Aerosol Indirect Effect to Subgrid Variability in the Cloud Parameterization of the GFDL Atmosphere General Circulation Model AM3, J. Climate, 24, 3145–3160, doi:10.1175/2010JCLI3945.1, 2011.

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

Griffies, S., Winton, M., Donner, L., Horowitz, L., Downes, S., Farneti, R., Gnanadesikan, A., Hurlin, W. J., and Coauthors: GFDL's CM3 coupled climate model: Characteristics of the ocean and sea ice simulations, J. Climate, 24, 3520–3544,doi: 10.1175/2011JCLI3964.1., 2011.

Griffies, S. M., Biastoch, A., Böning, C., Bryan, F., Danabasoglu, G., Chassignet, E. P., England, M. H., Gerdes, R., Haak, H., Hallberg, R. W., Hazeleger, W., Jungclaus, J., Large, W. G., Madec, G., Pirani, A., Samuels, B. L., Scheinert, M., Gupta, A. S., Severijns, C. A., Simmons, H. L., Treguier, A. M., Winton, M., Yeager, S., and Yin, J.: Coordinated Ocean-ice Reference Experiments (COREs), Ocean Modelling, 26, 1–46, doi:10.1016/j.ocemod.2008.08.007, 2009.

Hansen, J., Fung, I., Lacis, A., Rind, D., Lebedeff, S., Ruedy, R., Russell, G., and Stone, P.: Global climate changes as forecast by Goddard Institute for Space Studies three-dimensional model, J. Geophys. Res., 93, 9341–9364, 1988.

Hansen, J., Lacis, A., Ruedy, R., and Sato, M.: Potential climate impact of Mount Pinatubo eruption, Geophys. Res. Lett., 19, 215–218, 1992.

Hansen, J., Nazarenko, L., Ruedy, R., Sato, M., Willis, J., Del Genio, A., Koch, D., Lacis, A., Lo, K., Menon, S., Novakov, T., Perlwitz, J., Russell, G., Schmidt, G. A., and Tausnev, N. L.: Earth's Energy Imbalance: Confirmation and Implications, Science, 308, 1431–1435, doi:10.1126/science.1110252, 2005.

Hansen, J. E., Russell, G. L., Rind, D., Stone, P., Lacis, A., Ruedy, R., and Travis, L.: Efficient three-dimensional models for climatic studies, Mon. Wea. Rev., 111, 609–662, 1983.

Hargreaves, J. C.: Skill and uncertainty in climate models, Wiley Interdisciplinary Reviews: Climate Change, 1, 556–564, 2010.

Hawkins, E., Ortega, P., Suckling, E., Schurer, A., Hegerl, G., Jones, P., Joshi, M., Osborn, T. J., Masson-Delmotte, V., Mignot, J., Thorne, P., and van Oldenborgh, G. J.: Estimating changes in global temperature since the pre-industrial period, Bulletin of the American Meteorological Society, doi:10.1175/bams-d-16-0007.1, 2017.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., Folini, D., Klocke, D. J. D., Qian, Y., Rauser, F., Rio, C., Tomassini, L., Watanabe, M., and Williamson, D.: The art and science of climate model tuning, Bull. Amer. Met. Soc., p. (in press), doi:10.1175/BAMS-D-15-00135.1, 2016.

Hurrell, J. W., Bader, D., Delworth, T., Kirtman, B., Meehl, J., Pan, H.-L., and Wielicki, B.: Seamless Prediction. White paper for the U.S. Inter-Agency Working Group on Climate Modeling, http://www.cgd.ucar.edu/staff/jhurrell/docs/SeamlessModellingDraft03302007.pdf, 2006.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J.-F., Large, W. G., Lawrence, D., Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins,

W. D., Hack, J. J., Kiehl, J., and Marshall, S.: The Community Earth System Model: A Framework for Collaborative Research, Bull. Amer. Meteor. Soc., 94, 1339–1360, doi:10.1175/bams-d-12-00121.1, 2013.

Intemann, K.: Distinguishing between legitimate and illegitimate values in climate modeling, European Journal for Philosophy of Science, 5, 217–232, doi:10.1007/s13194-014-0105-6, 2015.

Jones, P. D., Lister, D. H., Osborn, T. J., Harpham, C., Salmon, M., and Morice, C. P.: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010, Journal of Geophysical Research: Atmospheres, 117, D05127, doi:10.1029/2011jd017139, 2012.

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

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

Knutti, R.: Should we believe model predictions of future climate change?, Phil. Trans. R. Soc. A, 366, 4647–4664, doi:10.1098/rsta.2008.0169, 2008.

Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P. J., Hewitson, B., and Mearns, L.: Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections, Tech. rep., IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland, in: Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor and P.M. Midgley (eds.), 2010.

Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophys. Res. Lett., 40, doi:10.1002/grl.50256, 2013.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, Geophysical Research Letters, doi:10.1002/2016gl072012, 2017.

Levitus, S., Antonov, J. I., Boyer, T. P., and Stephens, C.: Warming of the world ocean, Science, 287, 2225–2228, doi:10.1126/science.287.5461.2225, 2000.

Lin, J.-L.: The Double-ITCZ Problem in IPCC AR4 Coupled GCMs: Ocean–Atmosphere Feedback Analysis, Journal of Climate, 20, 4497–4525, doi:10.1175/jcli4272.1, 2007.

Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P., Garcia, H. E., Baranova, O. K., Zweng, M. M., Paver, C. R., Reagan, J. R., Johnson, D. R., Hamilton, M., and Seidov, D.: World Ocean Atlas 2013, Volume 1: Temperature, NOAA Atlas NESDIS 73, U.S. Government Printing Office, Washington, D.C., s. Levitus (Ed.), A. Mishonov (Technical Ed.), 2013.

Loeb, N., Wielicki, B., Doelling, D., Smith, G., Keyes, D., Kato, S., Matalo-Smith, N., and Wong, T.: Toward optimal closure of the Earth's top-of-atmosphere radiation budget, J. Climate, 22, 748–766, doi:10.1175/2008JCLI2637.1, 2009.

Lohmann, U. and Feichter, J.: Global indirect aerosol effects: a review, Atmos. Chem. Phys., 5, 715–737, 2005.

Ma, H.-Y., Chuang, C. C., Klein, S. A., Lo, M.-H., Zhang, Y., Xie, S., Zheng, X., Ma, P.-L., Zhang, Y., and Phillips, T. J.: An improved hindcast approach for evaluation and diagnosis of physical processes in global climate models, J. Adv. Model. Earth Syst., 7, 1810–1827, doi:10.1002/2015ms000490, 2015.

Manabe, S. and Bryan, K.: 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, 1969.

Masson, D. and Knutti, R.: Climate model genealogy, Geophys. Res. Lett., 38, L08703, doi:10.1029/2011GL046864, 2011.

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., Jungclaus, J., Klocke, D., Matei, D., Mikolajewicz, U., Notz, D., Pincus, R., Schmidt, H., and Tomassini, L.: Tuning the climate of a global model, J. Adv. Model. Earth Syst., 4, M00A01, doi:10.1029/2012MS000154, 2012.

McFiggans, Artaxo, P., Baltensperger, U., Coe, H., Facchini, M., Feingold, G., Fuzzi, S., Gysel, M., and Co-Authors: The effect of physical and chemical aerosol properties on warm cloud droplet activation, Atmos. Chem. Phys., 6, 2593–2649, doi:10.5194/acp-6-2593-2006, 2006.

Meier, W. N., Stroeve, J., Barrett, A., and Fetterer, F.: A simple approach to providing a more consistent Arctic sea ice extent time series from the 1950s to present, The Cryosphere, 6, 1359–1368, doi:10.5194/tc-6-1359-2012, 2012.

Miller, R. L., Cakmur, R. V., Perlwitz, J., Geogdzhayev, I. V., Ginoux, P., Kohfeld, K. E., Koch, D., Prigent, C., Ruedy, R., Schmidt, G. A., and Tegen, I.: Mineral dust aerosols in the NASA Goddard Institute for Space Sciences ModelE atmospheric general circulation model, J. Geophys. Res., 111, D06208, doi:10.1029/2005JD005796, 2006.

Miller, R. L., Schmidt, G. A., Nazarenko, L. S., Tausnev, N., Ruedy, R., Kelley, M., Lo, K. K., Aleinov, I., Bauer, M., Bauer, S., Bleck, R., Canuto, V., Cheng, Y., Clune, T. L., Del Genio, A., Faluvegi, G., Hansen, J. E., Healy, R. J., Kiang, N. Y., Koch, D., Lacis, A. A., LeGrande, A. N., Lerner, J., Menon, S., Oinas, V., Perlwitz, J., Puma, M. J., Rind, D., Romanou, A., Russell, G. L., Sato, M., Shindell, D. T., Sun, S., Tsigaridis, K., Unger, N., Voulgarakis, A., Yao, M.-S., and Zhang, J.: CMIP5 Historical Simulations (1850-2012) With GISS ModelE2, J. Adv. Model. Earth Syst., 6, 441–477, doi:10.1002/2013MS000266, 2014.

Ming, Y., Ramaswamy, V., Ginoux, P. A., Horowitz, L. W., and Russell, L.: Geophysical Fluid Dynamics Laboratory general circulation model investigation of the indirect radiative effects of anthropogenic sulphate aerosol, J. Geophys. Res., 110, D22 206, 2005.

Ming, Y., Ramaswamy, V., Donner, L., and Phillips, V.: A new parameterization of cloud droplet activation applicable to general circulation models, J. Atmos. Sci., 63, 1348–1356, 2006.

Molod, A.: Constraints on the Profiles of Total Water PDF in AGCMs from AIRS and a High-Resolution Model, Journal of Climate, 25, 8341–8352, doi:10.1175/jcli-d-11-00412.1, 2012.

Molod, A., Suarez, M., and Partyka, G.: The impact of limiting ocean roughness on GEOS-5 AGCM tropical cyclone forecasts, Geophysical Research Letters, 40, 411–416, doi:10.1029/2012gl053979, 2013.

Molod, A., Takacs, L., Suarez, M., and Bacmeister, J.: Development of the GEOS-5 atmospheric general circulation model: Evolution from MERRA to MERRA2, Geoscientific Model Development, 8, 1339–1356, doi:10.5194/gmd-8-1339-2015, 2015.

Oueslati, B. and Bellon, G.: The double ITCZ bias in CMIP5 models: interaction between SST, large-scale circulation and precipitation, Climate Dynamics, 44, 585–607, doi:10.1007/s00382-015-2468-6, 2015.

Park, S.: A unfied convection scheme (UNICON). Part I: Formulation, J. Atmos. Sci., 71, 3902–3930, doi:10.1175/JAS-D-13-0233.1, 2014.

Pawlowska, H. and Brenguier, J.-L.: An observational study of drizzle formation in stratocumulus clouds for general circulation model (GCM) parameterizations, J. Geophys. Res., 108(D15), 8630, doi:10.1029/2002JD002679, 2003.

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

Pithan, F., Ackerman, A., Angevine, W. M., Hartung, K., Ickes, L., Kelley, M., Medeiros, B., Sandu, I., Steeneveld, G.-J., Sterk, H., Svensson, G., Vaillancourt, P. A., and Zadra, A.: Select strengths and biases of models in representing the Arctic winter boundary layer: The Larcform 1 single column model intercomparison, J. Adv. Model. Earth Syst., doi:10.1002/2016ms000630, 2016.

Reichler, T. and Kim, J.: How Well do Coupled Models Simulate Today's Climate?, Bull. Amer. Meteor. Soc., 89, 303–311, 2008.

Richardson, M., Cowtan, K., Hawkins, E., and Stolpe, M. B.: Reconciled climate response estimates from climate models and the energy budget of Earth, Nature Climate Change, 6, 931–935, doi:10.1038/nclimate3066, 2016.

Rienecker, M., Suarez, M., Todling, R., J. Bacmeister, L. T., Liu, H.-C., Gu, W., Sienkiewicz, M., Koster, R. D., Gelaro, R., Stajner, I., and Nielsen, J.: The GEOS-5 Data Assimilation System - Documentation of versions 5.0.1 and 5.1.0, and 5.2.0., Tech. rep., 2008.

Ringler, T., Petersen, M., Higdon, R. L., Jacobsen, D., Jones, P. W., and Maltrud, M.: A multi-resolution approach to global ocean modeling, Ocean Modelling, 69, 211–232, doi:10.1016/j.ocemod.2013.04.010, 2013.

Rotstayn, L.: A physically based scheme for the treatment of stratiform clouds and precipitation in large-scale models. I. Description and evaluation of microphysical processes, Quart. J. Roy. Meteor. Soc., 123, 1227–1282, 1997.

Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W., Zhang, Q., den Dool, H. M. V., Pan, H.-L., Moorthi, S., Behringer, D., Stokes, D., Peña, M., Lord, S., White, G., Ebisuzaki, W., Peng, P., and Xie, P.: The NCEP Climate Forecast System, Journal of Climate, 19, 3483–3517, doi:10.1175/jcli3812.1, 2006.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.-T., Chuang, H.-Y., Juang, H.-M. H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Delst, P. V., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Dool, H. V. D., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.-K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.-Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R. W., Rutledge, G., and Goldberg, M.: The NCEP Climate Forecast System Reanalysis, Bull. Amer. Meteor. Soc., 91, 1015–1057, doi:10.1175/2010bams3001.1, 2010.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., ya Chuang, H., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E.: The NCEP Climate Forecast System Version 2, Journal of Climate, 27, 2185–2208, doi:10.1175/jcli-d-12-00823.1, 2014.

Schmidt, G. A. and Sherwood, S.: A practical philosophy of complex climate modelling, European Journal for Philosophy of Science, 5, 149–169, doi:10.1007/s13194-014-0102-9, 2014.

Schmidt, G. A., Ruedy, R., Hansen, J. E., Aleinov, I., Bell, N., Bauer, M., Bauer, S., Cairns, B., Canuto, V., Cheng, Y., Del Genio, A., Faluvegi, G., Friend, A. D., Hall, T. M., Hu, Y., Kelley, M., Kiang, N. Y., Koch, D., Lacis, A. A., Lerner, J., Lo, K. K., Miller, R. L., Nazarenko, L., Oinas, V., Perlwitz, J., Perlwitz, J., Rind, D., Romanou, A., Russell, G. L., Sato, M., Shindell, D. T., Stone, P. H., Sun, S., Tausnev, N., Thresher, D., and Yao, M.-S.: Present day atmospheric simulations using GISS ModelE: Comparison to in-situ, satellite and reanalysis data, J. Clim., 19, 153–192, doi:10.1175/JCLI3612.1, 2006.

Schmidt, G. A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., Bauer, M., Bauer, S., Bhat, M. K., Bleck, R., Canuto, V., Chen, Y., Cheng, Y., Clune, T. L., Del Genio, A., de Fainchtein, R., Faluvegi, G., Hansen, J. E., Healy, R. J., Kiang, N. Y., Koch, D., Lacis, A. A., LeGrande, A. N., Lerner, J., Lo, K. K., Matthews, E. E., Menon, S., Miller, R. L., Oinas, V., Oloso, A. O., Perlwitz, J., Puma, M. J., Putman, W. M., Rind, D., Romanou, A., Sato, M., Shindell, D. T., Sun, S., Syed, R. A., Tausnev, N., Tsigaridis, K., Unger, N., Voulgarakis, A., Yao, M.-S., and Zhang, J.: Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive, J. Adv. Model. Earth Syst., 6, 141–184, doi:10.1002/2013MS000265, 2014.

Shindell, D. T., Pechony, O., Voulgarakis, A., Faluvegi, G., Nazarenko, L. S., Lamarque, J.-F., Bowman, K., Milly, G., Kovari, B., Ruedy, R., and Schmidt, G. A.: Interactive ozone and methane chemistry in GISS-E2 historical and future simulations, Atmos. Chem. Phys., pp. 2653–2689, doi:10.5194/acp-13-2653-2013, 2013.

Siebesma, A. P. and Cuijpers, J. W. M.: Evaluation of Parametric Assumptions for Shallow Cumulus Convection, Journal of the Atmospheric Sciences, 52, 650–666, doi:10.1175/1520-0469(1995)052<0650:eopafs>2.0.co;2, 1995.

Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nausels, A., Xia, Y., and Coauthors: Summary for policymakers, Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, pp. Cambridge University Press, Cambridge/New York, 2013.

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

The GFDL Global Atmospheric Model Development Team: The new GFDL global atmosphere and land model AM2-LM2: Evaluation with prescribed SST simulations, J. Climate, 17, 4641–4673, 2004.

Thompson, D. W. J., Kennedy, J. J., Wallace, J. M., and Jones, P. D.: A large discontinuity in the mid-twentieth century in observed global-mean surface temperature, Nature, 453, 646–649, doi:10.1038/nature06982, 2008.

Tiedtke, M.: Representation of clouds in large-scale models, Mon. Wea. Rev., 40, 3040–3061, 1993.

von Schuckmann, K., Palmer, M. D., Trenberth, K. E., Cazenave, A., Chambers, D., Champollion, N., Hansen, J., Josey, S. A., Loeb, N., Mathieu, P.-P., Meyssignac, B., and Wild, M.: An imperative to monitor Earth's energy imbalance, Nature Climate Change, 6, 138–144, doi:10.1038/nclimate2876, 2016.

Walsh, J. E., Fetterer, F., Stewart, J. S., and Chapman, W. L.: A database for depicting Arctic sea ice variations back to 1850, Geographical Review, doi:10.1111/j.1931-0846.2016.12195.x, 2016.

Winsberg, E.: Values and Uncertainties in the Predictions of Global Climate Models, Kennedy Institute of Ethics Journal, 22, 111–137, doi:10.1353/ken.2012.0008, 2012.

Yokohata, T., Annan, J. D., Collins, M., Jackson, C. S., Tobis, M., Webb, M. J., and Hargreaves, J. C.: Reliability of multi-model and structurally different single-model ensembles, Clim. Dynam., 39, 599–616, doi:10.1007/s00382-011-1203-1, 2012.

Zhao, M., Golaz, J.-C., Held, I. M., Ramaswamy, V., Lin, S.-J., Ming, Y., Ginoux, P., Wyman, B., Donner, L. J., Paynter, D., and Guo, H.: Uncertainty in Model Climate Sensitivity Traced to Representations of Cumulus Precipitation Microphysics, Journal of Climate, 29, 543–560, doi:10.1175/jcli-d-15-0191.1, https://doi.org/10.1175%2Fjcli-d-15-0191.1, 2016.

Zweng, M. M., Reagan, J. R., Antonov, J. I., Locarnini, R. A., Mishonov, A. V., Boyer, T. P., Garcia, H. E., Baranova, O. K., Johnson, D. R., Seidov, D., and Biddle, M. M.: World Ocean Atlas 2013, Volume 2: Salinity, NOAA Atlas NESDIS 74, U.S. Government Printing Office, Washington, D.C., s. Levitus (Ed.), A. Mishonov (Technical Ed.), 2013.