# 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.