Collated Author's response

November 24, 2016

We'd like to thank all 3 referees for their constructive and considered reviews of our paper. The reviews we have received fall into two categories. The first two reviewers appear generally happy with the methodology and the application, but have comments, questions and concerns about our presentation and terminology relating to the statistical ideas in the paper. We will respond to each of their points in turn and add further clarifying remarks to the paper where appropriate. However, we would like to say, in general, that we have been very careful in our choices of language and terminology and have made these choices in order to resonate with and be familiar within the community of GCM/ESM developers and tuners at modelling centres. Experience of working with modellers from some of these centres for the last 5 years, and engaging with these important communities has led us to many of the choices that our statistical reviewers take objection to. Hence much of our response to these comments will be an attempt to explain and defend our choices. We certainly welcome the chance to do so in a public discussion and are grateful to both reviewers for raising the important questions that they do.

The response of the final reviewer, a model developer/tuner and a member of our target audience contains a number of points of clarification and we will endeavour to answer each of these and add suitable passages to the paper in response.

Referee 1

The authors propose a method for narrowing down the parameter space of a numerical model by ruling out parameter values which are not consistent with observations. The procedure is iterative. At each stage of the iteration additional observations are considered. Observational uncertainties are taken into account. The method avoids over-fitting since it does not aim at selecting a single optimal parameter combination which brings the model simulation closest to observations. The manuscript is well written and illustrates the methodology nicely using an ocean model. The main issue in my opinion is that the methodology is essentially a Bayesian parameter estima-

tion procedure and does not deserve a new name. As far as I can see, in a Bayesian formulation, at each step the authors basically rule out parameter values which have a posterior likelihood below a particular threshold.

In fact, the whole methodology would be much clearer and more transparent if the authors would acknowledge that what they do is a Bayesian procedure, and write down the prior distributions and the likelihood function. Also in the context of a carefully conducted Bayesian parameter estimation, the modeler would not just blindly select a single optimal value, but explore a range of parameter values which are broadly consistent with observations. In case the authors disagree, they should discuss the relation to Bayesian parameter estimation, highlight what is new in their approach, and indicate the advantage of their methodology. I am very doubtful that there is an aspect to "iterative refocussing" which is not naturally (and more transparently) part of an iterative Bayesian procedure.

History matching and iterative refocussing is not a (fully) Bayesian procedure, though this reviewer is not the first to claim it as such. The reason I use the term 'fully' above is that it could be argued that the procedure is a partial or second order Bayesian procedure, coming as it does from a Bayes linear methodological background. The essence of the Bayes linear procedure is to avoid altogether the specification of probability distributions and to work instead with partial moments and expectation primitive. The latter convention ensures that we do not even assume probability distributions must exist given the specification of a finite collection of moments. Without a probability distribution, either explicit or implicit, there is no justification for the claim that this is a Bayesian procedure. This is argued elsewhere, specifically in Craig et al. (1997), where the term "iterative refocussing" originates, Williamson et al. (2013) which discusses the use of expectation as primitive in history matching, and, effectively, in the discussion and rejoindre to Vernon et al. (2011).

Whilst accurate, a pragmatic argument is sometimes made that even though we never need probability distributions and do not assume them, you would get the same answer if you did, given a procedure based on removing near-zero likelihood regions of the parameter space. This is my interpretation of the reviewer's argument here. However, there is a subtlety that sets the methodology apart from Bayesian parameter estimation even if you were to interpret much of what we do probabilistically as the reviewer has done, rather than viewing expectation as primitive as we have. Bayesian parameter estimation, an approach first laid out for computer simulators fully by Kennedy and O'Hagan in 2001, requires a prior distribution $p(x^*)$ that is then updated by the ensemble, F, and the observations, z, through Bayes theorem to give $p(x^*|z, F)$.

Put technically, the parallel with our method would be if we were to

assume a uniform prior for x^*, then

$$p(x^*|z, F) \propto p(z, F|x^*)$$

and our procedure, roughly speaking, sets $p(z, F|x^*) = 0$ if $-log(p(z, F|x^*)) > a$ for some threshold a. But we would be doing this assuming that this implies $p(x^*|z, F) \approx 0$ due to its proportion to the likelihood in those regions of parameter space, which is not necessarily true. Within the probabilistic paradigm, $p(x^*|z, F)$ must integrate to 1, hence even when the model wildly differs from the observations and our discrepancy assessment, we will have a non-zero posterior. In fact, in this case, we will see a highly peaked posterior at the value of the model parameters where either uncertainty in the emulator is largest or at the value closest to reproducing the observations (even though it could be any number of standard deviations away under our model). The more inconsistent our model and specification with our data, the sharper the peak of this posterior. By taking the Bayesian route at all and having a prior, we assume x^* exists (with probabiliy 1), an assumption we can never coherently row back from.

What is happening here is even low near-zero likelihoods are modified by huge normalising constants to give considerable posterior probabilities to 'very far away yet closer than anywhere else' regions of parameter space. History matching has no problem here. We never assumed nor work with probability distributions and hence can coherently remove all of parameter space advising the modellers that the answer to the important question "can the model reproduce these observations/behaviours to within a given tolerance to error?" is no. How can the Bayesian proceed? The Bayesian has already ruled out the negative answer to the above in the prior. Bound (correctly) by coherence and in the face of the approximation $p(x^*|z, F) \approx 0$ for near zero likelihood breaking down, it seems that all is to be done is to revert to the posterior. Exactly what is to be reported or what inference can be drawn from here beyond pointing to the region of space where the posterior mass is concentrated as the most likely to contain good models is unclear.

Though we could imagine an iterative Bayesian analysis undertaken until convergence on a poor solution (if we have enough computing power to get there) could lead to the same conclusions (we simply get to a point where we stop and declare our statistical model invalid), we would question why should we be required to make infinitely more assumptions and judgements (in the form of full joint continuous probability distributions) in addition to new methodology required to establish and interpret convergence so that we can claim the Bayesian title for our procedure.

We say a lot more about the comparison to Bayesian parameter estimation in our response to reviewer 2 as they have raised a number of specific objections on this theme. To conclude the response to reviewer 1, we hope

this answer is sufficient to convince you that History Matching is fundamentally different to Bayesian parameter estimation. A full discussion of the respective merits of these two approaches falls without the scope of this paper, and we believe would be more at home in a statistics journal. Including such detail in this paper would, we believe, risk deterring our target audience of geoscientific model developers and tuners from engaging fully with the paper.

Reviewer 2

We are grateful for Dr Wilkinson's thorough and thoughtful review of our paper. His review largely focusses on our presentation of the statistical aspects of the methodology and its similarity to Bayesian calibration/parameter estimation, a different statistical methodology that is more popular in the statistical/UQ community, and arguably favoured by him (this methodology is also mentioned by the first reviewer).

Richard raises many important philosophical points and we are keen to debate the points he raises and the various merits of history matching vs calibration publicly. However, we are extremely conscious of the forum in which we engage in this debate and our target audience for this paper. Our target audience is primarily the scientists responsible for geoscientific model development and tuning, for example those involved with preparation of climate models for submission to CMIP(6) and the IPCC and our arguments and weighing of the literature is tailored for this audience.

It would be inappropriate to spend a considerable number of pages within this paper giving Bayesian calibration and its comparison to history matching/iterative refocussing the attention the subject properly deserves, because neither method is mainstream amongst the tuning community. A separate paper in a statistics oriented journal would be more appropriate for this discussion. In fact, our argument is that Bayesian methods have already been considered and dismissed by that community, as even though the method has been applied to simpler climate models many times since Rougier 2007 and has seen high profile application on simpler models (Sexton et al, 2011), the GMD/tuning community have not used it. It is our experience through engaging with scientists involved with model development and tuning at various modelling centres, even those which have used Bayesian calibration on climate models for other analyses, that those methods are not appropriate for their models and that that is a view shared by the modellers.

Instead, we give more time to the idea of optimisation of a cost function (usually the minimum of some distance from the model to data) versus our method and see little benefit to our audience to discuss every nuance of Bayesian calibration vs history matching when model tuning is currently done using neither. We will edit the paper where we feel the reviewer has a

point that deserves clarification in the text. We offer a lengthy response to Richard here and try to engage in the statistical debate perhaps more fully than is appropriate within this forum. However, we believe that a discussion paper in a statistical journal arguing the merits/similarities and differences of the two approaches would be a fantastic idea and the right forum for this important discussion.

Firstly, I agree with the other referee's comments that the additional terminology introduced by the paper, for concepts which already have established names, is unnecessary and potentially will further muddy the waters. The computer experiment literature already insists on using specialist language which confuses some other statisticians (emulators, calibration, etc), and I feel that "iterative refocussing" will only add to the problem, just at the stage where more people are becoming aware of history matching.

We do not believe we are at the stage where the model tuning community are becoming aware of history matching. In presenting to and working with this community, one of the biggest issues they have with the method is its name. It's not a descriptive name and its origin is in the oil industry.

Further it is clear that history matching is receiving more attention in the computer experiment literature and there have been some applications in climate (though not in tuning geoscientific models at most of the world leading centres). However, much of the history matching recently has been a one-wave analysis (Williamson et al. 2013, McNeall et al 2013, are examples), and we wanted to be very clear to demonstrate that the method is only really powerful through iteration. Its application to GCM tuning is only appropriate in our view as an iterative procedure over multiple waves as this mimics the way the model developers and tuners think about the problem and prioritise the various metrics and processes they are tuning to.

The term 'iterative refocussing' is not new. The phrase comes from the first papers on the method (e.g. Craig et al. 1997), as we stated in our paper. 'Refocussing' is a term that has been useful when interacting with the user community and they have responded well to it. We do not write this paper to further muddy any waters within the UQ literature. We write this for the model development and tuning community. We are principally concerned with reaching and engaging with this community, and we use the language that we have found gives us the best chance of doing this effectively.

I also dislike the use of "over-tuning". Overfitting a model occurs when we use a model which is too complex, so that we describe the noise not the signal. That may be happening here, but isn't the point that is being made. For example on p2, line 25-26, the authors are warning against the dangers of just using a single parameter value, rather than considering parametric uncertainty, a point I agree with, but which isn't related to overfitting or over-tuning (which as far as I can tell means the same thing).

We use 'over-tuning' as this is the language that the model development and tuning community have adopted to address the issue (e.g. Hourdin et al. 2016). Over-tuning occurs when we insist on matching our chosen metrics far closer than our uncertainties require and it is a very common problem in climate model tuning as the uncertainties are rarely known/given. We are not describing the noise by doing this so it is not the same as overfitting, though the idea is very similar. We are merely changing the parameters to get closer to the observations than we have a right to be, with the most obvious problem being that the time and effort required to do this could all be for nothing if it turns out the truth is closer to where our model is currently than the observed value. Worse still, every time we change the parameters, say to move the model closer to the observations than we have any right to do, we change all other model outputs that we are unable to compare to observations. Our point on line 25-26 is that by only selecting one value the temptation is to get this value as close as possible to the observations and that temptation is the risk we mention for over-tuning. Overfitting would involve us developing a more complex statistical model, yet here the model never changes and only different values of the parameters are tested.

The paper is also too dismissive of 'Bayesian calibration', which is what most Bayesians would just think of as inference. I like history matching (HM), and agree that is has some strengths that make it an attractive choice in situations where we want to avoid specifying a detailed statistical model.

We will address the individual arguments made by the reviewer on evidencing this point one by one. It is worth saying beforehand that I (the lead author) strongly identify as a Bayesian and, whilst I agree that 'most Bayesians' would think of Bayesian calibration as just Bayesian inference, *in the context of climate model tuning* I would view it as bad inference. That Bayesian inference is, in some sense, a gold standard when the likelihood is a good description of our uncertainty about the process and with carefully considered prior distributions is not really disputed by this author. Our view in this paper is that we are never in this position in climate model tuning (and rarely ever in large computer experiments, though this point would be better argued in the aforementioned statistical paper), and that not being in this position leads to serious and well understood problems with the inference. Hence we strongly argue for history matching and iterative refocussing in this paper for application to geoscientific model tuning and find Bayesian calibration as easy to dismiss as the other procedures used in model tuning that we discuss. However, we also note that we have not been overly dismissive of the method within the paper. We do mention it and cite papers and then discuss the benefits of HM/IF. The point of this paper is not to set up two solutions to tuning, calibration vs history matching and then dismiss the former in favour of the latter. Neither are used in the community and so only the method we argue for and, perhaps the most popular

methods involving optimising cost functions deserve special treatment.

the comment on page 3 line 7, that Bayesian calibration 'can also be described as forms of optimization and suffer from some of the drawbacks stated above' makes little sense to me.

What we mean by this is that calibration (and other methods that use MCMC to optimise cost functions, even though a distribution is computed/sampled from) still gives a probabilistic distribution for the true optimum. E.g. Bayesian calibration gives $p(x^*|z, F)$ which is a direct probabilistic statement about where the best input is and thus acts as a tool in the search for this optimum. The very nature of the Bayesian calibration solution to tuning is to say: "there is a best parameter setting and I want to find it, my prior for where this best model is is $p(x^*)$, I can now do Bayesian inference to update my beliefs about where the optimum is". Hence the method implicitly relies on the characterisation of the tuning problem as an optimisation problem and wraps the whole thing inside a Bayesian inference. Though you get a full distribution for this optimal value, the framing of the problem still assumes its existence.

We accept that we could have been more explicit on this point in the paper and have added a sentence at this point to give the essence of the above argument without labouring it too much (since our target audience is not the Bayesian calibration specialists).

Bayesian calibration can be made equivalent to history matching if we use a likelihood function which is an indicator function based on the implausibility metric, and flat priors (see for example Holden et al 2016 available as arXiv:1511.03475, where the similarity between ABC and HM is discussed).

We have read the paper referenced above to check and have concluded that we don't agree that the steps indicated above (and in the paper) make history matching and Bayesian calibration equivalent (even if we grant ignoring the ABC approximation that stating their equivalence here brushes over). We have been more detailed on this point in our response to the first reviewer. Part of our objection to the equivalence is that a posterior density, an object that by definition must integrate to 1, can be 0 everywhere (indicating that there are no matches with the given uncertainty specification). If, there is some subtlety we have not spotted in the exposition in the given paper that ensures this is not an issue (perhaps improper posteriors are allowed and claimed as indicative of this problem, though the paper doesn't make this clear unless we have misread it), then it is still our claim that the equivalence is an illusion. What has taken infinitely many judgements and assumptions (including that x^* exists with probability 1) to model probabilistically is being compared with an approach that assumes far less and offers a different interpretation. Numerically, in certain circumstances, the answers may coincide, but this is not enough for equivalence.

Even in the case of a non-empty NROY space after N waves of analysis, all the history matcher is able to say, before running the model further, is that she is unable to rule out any of the inputs in NROY space as being close enough to the observations with her given tolerance to error. The Bayesian inference gives a probability distribution over NROY. Unless the posterior is uniform over the NROY set (would it be following more than 1 wave and a Bayesian emulator?), we don't even have numerical equivalence, and if it is, the interpretation of the answer is still quite different: Namely, the best model is equally likely to be anywhere in NROY space.

The problem with Bayesian calibration occurs if we assuming likelihoods that are inappropriate. If we instead use conservative likelihoods as in HM, then we would also avoid the problems described.

History matching does not use likelihoods, it avoids probability and so avoids assuming that the NROY set is non-empty. We are also confused by the idea of assuming likelihoods and of conservative likelihoods in a Bayesian sense. Surely likelihood is a subjective description of beliefs? If so, then do we assume a set of beliefs, and how do we change our beliefs to ensure they are conservative (and what does that mean)?

In truth both approaches are simple and coherent. In the one, I specify my beliefs probabilistically, I use Bayes theorem, I make posterior inferences. There are a number of issues that not believing all of my assumptions here can unwittingly cause, we state some of them in the paper and in this discussion, though this is not the forum for a full treatment. In general, our preference for other approaches is due to not believing the assumptions for models of this scale. If we did believe them, we'd have no issues and have discussed following up history matching with Bayesian calibration on the NROY set in Williamson et al. 2013, 2015. The history matching approach is effectively a screening approach that takes every step possible to avoid assuming that a model satisfying our (second order) uncertainty specification exists until we really believe it does, and makes no inferences about models that cannot be removed.

I also dislike the description of history matching as a method that inherently is based upon the use of emulators. To me, history matching is defined by the use of implausibility metrics to find not ruled out yet (NROY) regions of parameter space. How we go about doing this is a matter of implementation, where we may (indeed most likely will) find that emulators are of benefit. The advantage of viewing HM as distinct from the algorithm used to implement it, is that we can then define the 'true' HM answer (which is a NROY set). This then allows us to independently answer the two questions

1. *How does the NROY set compare to answers from using other likelihoods (ie Bayesian calibration approaches) or implausibility measures?*

2. *How good is any given implementation/algorithm at finding the true NROY set?*

If we define HM to be iterative refocussing with GP emulators, then these two questions are conflated.

This is an interesting point and we could certainly be convinced that defining a true HM answer that can be subjected to the kinds of testing required in order to answer 1. and 2. above would be desirable as part of a paper developing, comparing or extending the methodology. However, our goal here was not to define history matching/iterative refocussing as a general method, though we know that it is. Our goal was to describe and demonstrate the use of the method for tuning geoscientific models such as climate models. This is a paper aimed at the geoscientific model development and tuning community and the building of emulators is essential in our view in order to tune such expensive models. It is also the main technical barrier to its implementation within the field, so we believe that our focus on this part of the method is justified.

If we separate these points, then I'm not sure that line 9 on page 9, that the order in which metrics are applied matters, makes sense. Unless something is happening I don't understand, the final true NROY set should be independent of the order in which the metrics are applied: if it isn't, then it says something worrying about HM as an approach. Note that I can see how the order matters in the implementation, and that some orderings will make finding the NROY set harder (something analogous to mixing in MCMC or SMC for example), but that relates to question 2 above (implementation) not question 1.

We have removed this line as we realised that we left the more detailed discussion of the point out of the paper. If we did define a true NROY set and left emulators out, then this would be true, but the ordering would still matter in implementation as you say.

The para around line 20 on page 3 claims several benefits for history matching, and the implication is, that these would not be available if one were to use for example, calibration, which simply isn't the case. They are benefits from the careful statistical analysis (particularly the emulation) that is done, and shouldn't just be claimed for HM (or iterative refocussing).

We do not imply that the benefits we describe are *only* available if history matching were used. We have simply advocated for our method. We have been specific in outlining some of the drawbacks of other methods that we believe HM/IF does not suffer earlier in the introduction. By explaining what we believe are positive features of our approach we are not implicitly bashing other methods. We have criticised some of them where we have. The rest of our paper is devoted to presenting the positive case for using

HM for climate model tuning and is not aimed at the statisticians/scientists using calibration (which is not the climate modelling community) who could be converted.

The debate as to whether all of the benefits of our approach are shared with other statistical approaches, such as calibration should happen in another forum.

Another case where HM is over-sold, is at the bottom of p12. In this case, the choice of a poor score (line 10) is to blame for the counter-intuitive result described. If the variance is not constant, then we should score using the log-likelihood say (or some other proper score), rather than the Mahalanobis distance, which is an improper score. If we added (or subtracted depending on your setup) log det σ to the Mahalanobis distance, then we would no longer find the optimum value occurs at an x where we are most uncertain.

This is an important, interesting and complicated point raised by the reviewer. To be clear though, we have not oversold HM here. We have outlined a flaw with optimisation approaches that use distance measures scaled by function uncertainty and to do so does not oversell HM at all. When evaluating climate models it is common to minimise RMSE, or to have cost functions that standardise by uncertainty. We think the point we raise in the paper is both valid, interesting and intuitive for our target audience.

As for the point about implausibility being an improper score, I (the lead author) have said in other communications with the reviewer that this is a subject worthy of investigation and possibly a paper. I am not sure implausibility has ever been considered as a score. It is a distance in the metric space induced by variance as a measure of uncertainty. The metric space wherein we are able to do Bayes linear analysis. Considering a scoring rule as a basis for ruling out parameter space is interesting but raises many unanswered questions. No such scoring rules have been proposed for UQ and been subject to scrutiny from the statistical community to date, and until they have been, we feel it would be premature to focus on these rather than the more common Mahalanobis type distance commonly used in HM and by the geophysical model development community.

In conclusion, I would like to see a more balanced discussion of HM in the paper. I think the approach makes a lot of sense for these models and can be argued for persuasively, without conflating the issues discussed above.

We have re-read our overall presentation of HM and we are not convinced that we have been overly imbalanced. From the point of view of a Bayesian practitioner that uses the standard Bayesian calibration ideas of Kennedy and O'Hagan (2001), we can see how our presentation may seem to be too easily dismissive of that method as we only really mention it in one paragraph during the introduction. That practitioner then may read some

of our criticisms of tuning as particular to calibration, when really we are making the case against optimisation based procedures that are far more popular within the climate literature (and we referenced these throughout the paper). However, we also feel our criticism of other methodologies is also relatively sparse as we have opted to focus on the positive benefits of HM for the geoscientific model development and tuning community. I (the lead author) would like to write a paper discussing why I believe the standard form of Bayesian inference is inappropriate for climate models. However, this is not that paper, nor is the readership of this journal the right audience. This paper is not intended to make the case for HM over Bayesian calibration specifically, though the reviewer may have read it as such.

We hope that by clarifying our purpose and our target audience, the reviewer will, upon a fresh reading of the paper, see that we have merely strongly advocated for HM/IR in geoscientific model tuning, without overly criticising calibration nor giving a detailed treatment of any of the other more popular methods within the literature on climate model tuning. We observe some of the flaws of general optimisation methods and highlight what we feel the benefits of our method are. That some other methods may share some benefits, is not our concern, as this paper is not a review of tuning methods in climate.

Finally, is there an error in equation 2 in the nugget term? Shouldn't the nugget only get added for the same model run, i.e.

$$\mathrm{Cov}(x_i, x_j) \propto I_{i=j}$$

rather than

$$\mathrm{Cov}(x_i, x_j) \propto I_{x_i=x_j}$$

if the nugget is there to represent internal variability?

Perhaps there is some confusion in notation as, by $x = x'$ in equation (2), we mean the same model run (we don't assume there are initial condition perturbations in the ensemble). We have added this to the text to clarify.

- *p1, line 21, 'and' −> 'which'?* Changed
- *p2, lines 15-19. Something about how this is structured confuses me. Perhaps consider rephrasing.* We have rephrased this.
- *p4, line 13 'depracated' −> 'deprecated'* Changed
- *p10, line 5, change 'it's' to 'its'* Changed
- *p10, line 15, an emulator isn't inherently Bayesian, unless you make it so.* We have re-written that sentence to make this clear.

- *p11, line 33, where can the code be downloaded from?* From the supplementary material in the cited paper. We have added a note to the paper to clarify.
- *p12, line 9, 'Mahalanobis' rather than 'malhalanobis' (spelling and capitalisation)* Changed
- *p12, line 18, I don't think you want the inverse variance (delete the '-1')?* Changed
- *p12, line 28, comma after x^** Changed
- *p17, line 7, has something gone wrong with this equation? It doesn't seem quite right.* We think this line is correct, though we think it is sufficiently difficult to interpret that we have removed it (the text version is far clearer).

Reviewer 3

We are grateful to reviewer 3 for the time they have taken to give this paper an extremely positive review. We feel it is very important that our target audience (geoscientific model developers and tuners) has been represented in the peer review of this paper, and we thank the reviewer for taking the time to engage with our ideas. We answer each of their minor points below.

p.3, ll.18-21. This is an important point. It is often the case that during development, the modeller is attempting to assess the model's ability to reproduce a physical phenomenon or feature of the real climate system (eg, Madden Julian Oscillation, the Quasi-Biennial Oscillation, ENSO, etc.). Is it difficult to determine whether the present formulation of its physical parametrizations allows such behaviour for "some" combination of values of its physical parameters or whether the representation of physical processes in the parametrizations are inadequate and require further development. The iterative refocussing method would seem to provide a powerful tool to help decide such issues.

We agree and we thank the reviewer for highlighting this.

p.6, ll. 1-2. "We also note that the real ocean has never been in equilibrium and hence a tuning procedure that works by comparison to observations may not require an equilibrated ocean." It may be true that the real ocean has never been in equilibrium but if a validation exercise against observations depended on its transient state, reproducing that transient state would seem to be a much more daunting task than what seems to be suggested in this passage of text. Perhaps I misunderstood the point that was trying to be made here.

Our point is that optimising by finding an equilibrated state that matches observations, even though the observations are not those of an equilibrated state themselves seems difficult to justify. That it would be harder to reproduce the transient state may be true, but is not a good reason for the optimisation approach based on equilibrium. We advocate our approach that is based on ruling out rather than optimising, which we introduced straight away at the start of the next section.

p.8, ll.1-2. I agree that uncertainty in the observations against which climate models are assessed is critical but do we even have this information from the observational community?

Rarely if ever. A point we devote a paragraph to in the discussion at the end of the paper. Our view is that routine reporting of uncertainty in observations and the gridded products on which they are built would be of enormous benefit, in particular to the modelling community. In some instances it may be relatively simple to provide, for example it may have been computed as part of the procedure for deriving a gridded product which incorporates many datasets, e.g. optimal interpolation based on kriging.

p.8, ll.28-29. "If errors can be "tuned out" with better choices of the free parameters, they are not structural at all, they are parametric." I agree that this would be very informative but it is a necessary rather than sufficient condition for this conclusion to be valid. A tuning exercise that gets some metric within observational error for some range of free parameters is suggestive but does not guarantee that such agreement is obtained for the "right" reasons. Further investigation would still be required to support such a conclusion.

We agree and have softened the text at this point in the paper to reflect this.

p.14, ll.12-13. "If the entire parameter space is ruled out using a certain metric, a structural error has been located." Again, this is a necessary but not sufficient condition. A potential issue/error in the estimate of observational uncertainty could also be the reason for such behaviour.

We have added a similar clarifier at this point in the text.

p.14, ll.19-22. I would also add that the final NROY space also nicely "defines/documents" all physical behaviour, for the set of metrics considered, of a particular model version (ie specific formulations of physics, the model resolution, numerics etc.). Currently, such behaviour is assessed from one set of model parameters and is used to drive decisions about further development of physical parametrizations. The current approach can be counterproductive if the issue is just parametric and not structural. The more complete description of potential model behaviour captured in the NROY space would

allow such decisions to be made in a more rational and effective manner.

We agree and have added a couple of sentences at this point in the paper to say that.

p.15, l.30. replace "at at" with "at".
Changed

p.20, Fig.4 It is suggested that there are lines with 7 different colours/patterns in this figure. I could only see 3 or 4. It might be better to show all 8 depths separately in addition to the continuous vertical profile. As it stands, it is not possible to see all of what is being described in this figure.

We have carefully considered this, as we agree with the reviewer that in a printed version the separation between lines is not clear. The figure currently follows the same format as the global mean salinity presented in figure 5, in which there are clear differences between the observation based datasets and the GO5 simulation. The fact that the lines in figure 4 are so close together demonstrates generally good agreement between the observational datasets at many depths. It is not helped by the fact that there is a large temperature gradient between the surface and the abyssal ocean. The image is produced as a high definition PDF so that in a digital version, the reader can zoom in to observe the behaviour of these different models/data sets. If the reviewer feels this figure needs revision, rather than plotting the 8 depths separately (we prefer to show the full structure of the profile) we could present the information in a similar manner to Fig. 6, right panel, with the x-axis scaled to reveal the detail near to EN3 (our target). We did not include this style of plot in the original submission because we wanted to capture the full range of solutions from both the first and second waves, and by zooming in many of these early wave simulations are off-scale for most of the depth range.

p.22, l.1 change "that fail 2 our more" to "that fail 2 or more".
Changed

p.36, ll.1-6. The history matching philosophy is one of identifying and then rejecting free parameter settings that are likely associated with unphysical model states or behaviour. As the authors correctly point out, following such a procedure, the NROY space is a residual of the exercise. As such it has passed necessary but not sufficient conditions in regard to the quality/plausibility of the underlying model. It is basically an efficient procedure to document what a specific model version is, and is not, capable of. From this perspective, it raises the question, is the parametric survey of model behaviour really a "tuning" exercise? The iterative refocussing approach discussed in this study would seem to be more of a tool to survey/discover an existing range of model behaviour associated with a specific set of frozen

physics and numerics. In this regard, "tuning" is not parametric (ie connected to the specific values of physical parameters), but rather structural (ie connected to the decisions related to how we choose to represent/model physical processes in our climate models). To me, this represents an important advance in our approach to the development and application of such large and complex models.

This is an interesting question, that speaks to the question of what is or is not 'tuning'. We agree entirely with the reviewer that to change the general approach to this activity and the more general activity of model development to include a full parametric survey of 'not implausible' model behaviour respecting key uncertainties would be an important advance in that area. In the final analysis, what we call 'tuning' maybe a semantic issue. For the foreseeable future at least, modelling centres will continue to develop their models and then adjust free parameters in order to provide submissions to CMIP-Next. Whether we call tuning the search for the best model for this submission, the locating of a representative set of models, or anything else, we believe our approach is at the very least an important part of this process and speaks to what is currently done in that field. We allude to this in section 7.

Ultimately, what tuning (the term currently in vogue amongst the climate model development community) involves and what the reported results look like will be determined by the community itself. We hope this paper can influence the direction of the discussion.