Climate Dynamics manuscript No.
(will be inserted by the editor)

Identifying and removing structural biases in climate models with history matching

Daniel Williamson · Adam Blaker · Charlotte Hampton · James Salter

Received: date / Accepted: date

Abstract We describe the method of history matching, a method currently used to help quantify parametric uncertainty in climate models, and argue for its use in identifying and removing structural biases in climate models at the model development stage. We illustrate the method using an investigation of the potential to improve upon known ocean circulation biases in a coupled non-flux-adjusted climate model (the third Hadley Centre Climate Model; HadCM3). In particular, we use history matching to investigate whether or not the behaviour of the Antarctic Circumpolar Current (ACC), which is known to be too strong in HadCM3, represents a structural bias that could be corrected using the model parameters. We find that it is possible to improve the ACC strength using the parameters and observe that doing this leads to more realistic representations of the sub-polar and sub-tropical gyres, sea surface salinities (both globally and in the North Atlantic), sea surface temperatures in the sinking regions in the North Atlantic and in the Southern Ocean, North Atlantic Deep Water flows, global precipitation, wind fields and sea level pressure. We then use history matching to locate a region of parameter space predicted not to contain structural biases for ACC and SSTs that is around 1% of the original parameter space. We explore qualitative features of this space and show that certain key ocean and atmosphere parameters must be tuned carefully together in order to locate climates that satisfy our chosen metrics. Our study shows that attempts to tune climate model parameters that vary only a handful of parameters relevant to a given process at a time will not be as successful or as efficient as history matching.

Keywords Tuning · Ensembles · Emulators · HadCM3 · Climate Model

D. Williamson · J. Salter
College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK
E-mail: d.williamson@exeter.ac.uk

A. T. Blaker
National Oceanography Centre, Southampton, UK, SO14 3ZH

1 Introduction

One of the principal challenges facing the climate modelling community is the removal of systematic or structural errors in generalised circulation models (GCMs) (Randall et al, 2007). So called "known biases" in a GCM drive the development and improvement of these models. For example, a motivation for the development of HadGEM2 was the improvement of the model ENSO compared with its predecessor HadGEM1 (Martin et al., 2010). The Hadley Centre models are not alone in this regard, with each new version of a group's GCM containing biases that the modellers speculate can be improved or removed with better parameterization schemes or finer resolution (for example, Watanabe et al., 2010; Gent et al., 2011).

As the desire to remove these structural errors drives increases in model resolution and development of new code and parameterization schemes, it is important to know that the errors in question really do represent structural deficiencies of the model and are not merely an artefact of poor tuning of the current parameterization schemes. GCMs necessarily consist of many parameterized schemes designed to approximate the physics in the real world on the grid scale of the model. Each scheme contains a number of parameters whose value must be fixed in order to run the climate model. A major part of the development of a new climate model represents the tuning of these parameters and schemes in order to ensure the resulting model climate is consistent with observations over a number of chosen metrics.

A climate model bias represents a structural error if that bias cannot be removed by changing the parameters without introducing more serious biases to the model. Hence, to state that a climate model bias represents a structural error is to assume that the model has been optimally tuned and yet fails to adequately represent the metric in question. In this paper we argue that any GCM is highly unlikely to be optimally tuned due to the way the parameters are usually selected by modellers.

Climate models are usually tuned on a process by process basis and by trial and error (Severijns and Hazeleger, 2005). An individual process or module is selected and one or two parameters thought to drive that process are changed. If the change moves the process closer to reality, the change is accepted. For example, Acreman and Jeffery (2007) change two parameters in the Kraus and Turner (1967) mixed layer scheme. The study was used to fix these parameters in studies using the UK Met Office's 3rd Hadley centre model HadCM3 (Gordon et al., 2000; Pope et al., 2000; Collins et al., 2007). Martin et al. (2011) reduce background tracer diffusivity in the ocean by an order of magnitude to improve SST profiles for HadGEM2. There are very few guidelines for tuning the parameters of a climate model. The 4th IPCC report (Solomon et al., 2007, sect. 8.1.3) gives 2 guidelines for tuning. The first is that observation-based constraints on parameter ranges should not be exceeded. The second is that climate model performance is only judged with respect to observational constraints not used in tuning. Whilst the second of these seems sensible, the first is questionable (why is it necessarily true that a numerically integrated solution to climate model equations over a relatively coarse spatial grid be most informative for the true climate when theoretically observable parameters are within their real-world ranges?). Taken together, these guidelines offer little actual instruction for tuning.

Mauritsen et al. (2012) offer a tuning protocol, which was used to develop the latest version of the MPI-ESM model. Their protocol is based on identifying

model biases and targeting those in particular by iteration through steps that first involve short runs with prescribed SSTs to find promising parameter choices. These choices are then subjected to longer simulations and compared to observed climate. If they are still promising, they are changed in the coupled model and the resulting model climate is evaluated.

Though the use of short runs in preliminary tuning steps seems promising, the focus on tuning only parameters influencing specific processes using an uncoupled version of the model is problematic. Experiments such as these represent "one factor at a time" (OFAT) designs. The hope is that after changing the parameter choices individually in order to improve each process, the coupled climate model will also have improved. This type of experimental design is well known in the statistics literature for being both inefficient and dangerous (Fisher, 1926; Friedman and Savage, 1947; Daniel, 1973). In particular, if parameters controlling different processes interact, that is, if by changing them simultaneously in some way the effect is different to changing them separately, OFAT type designs cannot find these interactions. Partly because of this and partly due to inefficiency, this type of design is prone to missing optimal settings of the parameters.

More formal procedures for climate model tuning in the literature do exist. For example, Severijns and Hazeleger (2005) treat tuning as a global optimization problem which they solve using the downhill simplex method (a numerical minimisation algorithm). A class of data assimilation methods approach tuning with respect to the key uncertainties: observation error and structural error. These methods, for example, based on the ensemble Kalman filter, combine the parameters with the climate model state vector in order to fine tune a model, and have been applied to intermediate complexity climate models (Annan et al., 2005c; Hargreaves et al., 2004) and to the atmosphere-only component of a GCM (Annan et al., 2005a,b). However, as yet, they have not been applied to successfully tune a coupled atmosphere-ocean GCM and, Rougier (2013) states that there is reason to think that this type of tuning method is intractable in the full parameter space.

Another problem with data assimilation approaches to parameter tuning is that the parameters, and hence the model physics, are allowed to vary in time. This means that the final parameter choice, that which corresponds to the value of the augmented state vector following assimilation of the most recent observations, need not represent a model that would reproduce an acceptable solution to the underlying model equations over the full assimilation period. In fact the model is constantly tuned so that the solution is not dynamically consistent. We might think of these solutions as representing worlds with "transient physics", which can be seen as undermining the key assumptions made in using a climate model for long term projections, i.e., that the model represents the physics well enough to trust the projections as long as the initial conditions are captured well. Ideally, we would like to find a setting of the model parameters at which, when the model is run without assimilation, the output most closely approximates the physical behaviour and evolution of the climate system. At such a parameter choice, we may then assimilate key data assuming that once assimilation is complete, the free running model will "drift" back to its attractor slower than otherwise (as its attractor is closer to the data than at any other parameter setting). In theory then, with better parameter choices, short to medium term forecasts based on data assimilation will be more accurate.

In this paper we present a statistical approach to climate model tuning using an existing technique called history matching (Craig et al., 1996). It was initially presented as a methodology for finding parameter settings of computationally expensive oil well models that led to output consistent with the observed history of a well. It is currently used as a tool for quantifying parametric uncertainty in computer models and has been applied to ice sheet models (McNeall et al., 2013), intermediate complexity climate models Edwards et al. (2011) and to GCMs (Williamson et al., 2013). The idea is for all parameters to be varied simultaneously in the generation of a perturbed physics ensemble (PPE). The PPE is then used to train emulators (fast statistical approximations to the climate model that give a prediction of the climate model output for any setting of the parameters with an associated uncertainty on the prediction) that are then used, in tandem with observations, to cut out regions of parameter space that lead to models deemed "too far" from the observations according to a robust geometric measure (we note briefly here that if the model is extremely fast and uses no computational resource, that emulators will not be required to map and reduce parameter space (Gladstone et al., 2012). However, we present history matching as a method requiring emulation to assist climate model tuning as these models have the opposite property). Emulators have also been used in tuning exercises by Bellprat et al. (2012).

We argue that history matching is an effective and intuitive tool for tuning and that it can be used to determine whether a perceived structural error actually exists or if it can be corrected by changing the model parameters. History matching can be applied on the most computationally expensive climate models and with small ensembles, and represents a far simpler undertaking than a data assimilation based approach. We illustrate the method by investigating a number of known ocean circulation biases in HadCM3.

In particular we investigate the perceived structural bias in the Antarctic Circumpolar Current (ACC) strength. This current is known to be too strong in the Hadley Centre climate models (Russell et al., 2006; Meijers et al., 2012); however, we show that this may not represent a structural error at all. We show that by jointly varying both ocean and atmosphere parameters together, it is possible to find models that cannot be ruled out as having physical global surface air temperature and precipitation using the metrics defined by Williamson et al. (2013), that also have no ACC bias. We explore the properties of the ocean and atmosphere circulations in one of these models and compare them to the standard HadCM3 and to observations. We use history matching to identify a region of parameter space containing not implausible ACC strengths and further refine this region using a constraint on North Atlantic sea surface temperatures (SSTs). We investigate qualitative features of the parameter space not ruled out by the observations of these metrics and illustrate why ocean and atmosphere parameters must be varied jointly in the coupled model when tuning.

In section 2 we briefly describe emulation and history matching and discuss its implementation for tuning expensive climate models. In section 3 we use history matching to search a subset of the HadCM3 parameter space with not implausible SAT and precipitation profiles found by Williamson et al. (2013) for models with not implausible ACC strength. We identify a subset of this space predicted to contain not implausible ACC strengths and find some models therein. We investigate properties of the ocean circulation for a run without the usual ACC bias and compare them to the standard HadCM3. In section 4 we include SSTs in the

sub tropical gyre into our history match and discuss features of the parameter space that has not been cut out. Section 5 contains discussion and the appendices present details of the emulation, the parameters varied in the ensemble and present further pictures.

2 Emulation and history matching

We write the climate model as the vector valued function $f(x)$ where x corresponds to a vector of climate model parameters. History matching requires an emulator for $f(x)$ to be fitted so that, for any setting of the parameters x, an expectation and variance for those elements of $f(x)$ we intend to compare with observations ($\mathrm{E}[f(x)]$ and $\mathrm{Var}[f(x)]$) may be computed from the emulator. There is a vast and growing literature on using ensembles to fit statistical emulators, so we don't go into mathematical details here. We refer the reader to Craig et al. (2001); Rougier (2008); Haylock and O'Hagan (1996); Sacks et al. (1989) and the book by Santner et al. (2003), for general information on building emulators; and to Rougier et al. (2009); Challenor et al. (2009); Sexton et al. (2011); Williamson et al. (2012); Schmittner et al. (2011); Lee et al. (2011); Williamson et al. (2013) and Williamson and Blaker (2014) for application of emulators to climate models.

Once an emulator is fitted so that we can compute $\mathrm{E}[f(x)]$ and $\mathrm{Var}[f(x)]$ for any x, history matching proceeds by ruling out choices of x as being inconsistent with chosen observational constraints, z, using an implausibility function $\mathcal{I}(x)$. A common choice is $\mathcal{I}(x) = \max_i\{\mathcal{I}_i(x)\}$ and

$$\mathcal{I}_i(x) = \frac{|z_i - \mathrm{E}[f_i(x)]|}{\sqrt{\mathrm{Var}[z_i - \mathrm{E}[f_i(x)]]}}, \tag{1}$$

but others do exist (Craig et al., 1996; Vernon et al., 2010). Large values of $\mathcal{I}(x_0)$ at any x_0 imply that, relative to our uncertainty, the predicted output of the climate model at x_0 is very far from where we would expect it to be if $f(x_0)$ were consistent with z. A threshold a is chosen so that any value of $\mathcal{I}(x_0) > a$ is deemed implausible. The remaining parameter space, $\{x \in \mathcal{X} : \mathcal{I}(x) \leq a\}$ is termed Not Ruled Out Yet (NROY). The value of a is often taken to be 3 following the 3 sigma rule (Pukelsheim, 1994), which states that for any unimodal continuous probability distribution, at least 95% of the probability mass is within 3 standard deviations of the mean.

The form of $\mathrm{Var}[z_i - \mathrm{E}[f_i(x)]]$ will depend on any statistical model used to establish a relationship between observations of climate and output of the climate model. The most popular model, termed the 'best input approach' (Kennedy and O'Hagan, 2001) expresses the observations via

$$z = y + e$$

where y represents the underlying aspects of climate being observed and e represents uncorrelated error on these observations (perhaps comprising instrument error and any error in deriving the data products making up z). The best input approach then assumes that there exists a 'best input' x^* so that

$$y = f(x^*) + \eta$$

where η is the model discrepancy (or structural error) and is assumed independent from x^* and from $f(x)$ at any x. Model discrepancy, being independent from any evaluation of the climate model, represents the extent to which the climate model fails to represent actual climate owing to missing or poorly understood physics, parameterisation schemes and the resolution of numerical solvers.

The best input approach has been used in studies with climate models by Murphy et al. (2009) and Sexton et al. (2011) and is described by Rougier (2007). The statistical model leads to

$$\mathrm{Var}\left[z_i - \mathrm{E}\left[f_i(x)\right]\right] = \mathrm{Var}\left[e\right] + \mathrm{Var}\left[\eta\right] + \mathrm{Var}\left[f_i(x)\right]$$

where $\mathrm{Var}\left[e\right]$ is the variance of the observation error, $\mathrm{Var}\left[\eta\right]$ is model discrepancy variance and $\mathrm{Var}\left[f_i(x)\right]$ is a component of the emulator for $f(x)$.

2.1 A tuning procedure: history matching in waves

History matching represents a formal statistical procedure for tuning climate models by iteratively ruling out implausible regions of parameter space. We advocate tuning a climate model through a series of "waves" of history matching, where a "wave" involves running a new PPE in the current NROY space, building new emulators for each of the currently considered metrics and for a series of new metrics to be introduced for this wave, and using these emulators to further cut down NROY space. Structural errors are identified when a particular metric, once introduced, can rule out the whole space, indicating that, given all of the other metrics are NROY, the chosen metric cannot be reproduced to within the model discrepancy.

This approach has been demonstrated to be successful in other fields. For example, Vernon et al. (2010) demonstrate this procedure through five waves on a computer model simulating the evolution of galaxies after the big bang. After 5 waves (with each ensemble containing 1000 different parameter settings) they found hundreds of computer model runs that were consistent with their observations when, prior to the study it was thought that no such parameter settings existed. Given this success in other fields, we believe that it is highly likely that at least some of the perceived "structural errors" in modern GCMs will be eliminated by history matching without compromising model performance with respect to other physically important metrics.

The crucial decision to be made by the modellers when using history matching in this way, is what metrics should be used to tune the model and in what order should they be applied. There are aspects of real world physics that we know, a priori, that the model does not capture. For example, sub grid scale processes such as eddies in HadCM3, or convective plumes in a $1/4^o$ model. These are definitely not part of the model, and so if we were to history match to them, we would rule out the whole parameter space. Hence the choice of metric is important.

Further, the order in which they are introduced through the different waves is also important. Note that within a given wave, all metrics have the same level of importance. If a parameter choice is ruled out because of one metric, it is ruled out no matter if it is NROY with respect to others in the same wave or not. However, the wave at which each metric is introduced should reflect the order of importance of any particular metric when it comes to trusting the output of a

climate model. For example, it may be that the model must have a reasonable global SAT profile and that this is more important than its AMOC strength. In this case, by first matching to SAT in wave 1, then searching the space of models with NROY SAT profiles for reasonable AMOC strengths in wave 2, we only search a sub-space of models for good AMOC strengths. If we fail to find any, we would declare that AMOC was a structural error in the model. However, it might be that certain parameter choices that lead to poor SAT profiles do have not implausible AMOCs. By choosing the order in which metrics are introduced over successive waves, we effectively define what it means for the model to have structural error. Models are currently tuned by comparing various metrics to observations by the modellers. Hence there is already an implicit sense of what metrics are important and which are more important than others.

An important issue if adopting our approach to tuning through history matching is that of stopping rules. How many wave of refocussing and history matching are required before one or more of the NROY models can be adopted as a tuned run (or collection of them)? The answer to this will be very problem dependent, however we can offer guidelines. If the set of chosen metrics is fixed then once the emulator variance is a great deal smaller than the denominator in the implausibility calculation, then it is unlikely that further waves will change the implausibility very much, and further ensembles for history matching purposes may be considered to be an inefficient use of resources. In these situations, parameter choices will generally have low implausibility because we are relatively sure that the model when run at those choices is genuinely close to the observations (with respect to the uncertainty and model discrepancy). How many waves are required in order to reach this situation will depend on a number of factors. These include, the available ensemble size at each wave, the complexity of the metric and its behaviour in parameter space (is it easy to emulate with few runs, or is a lot of data required to capture the parameter dependencies?), the size of the sub-volume of parameter space that, if we did run the model there, would be close enough to the observations (which depends on their uncertainty on on the model discrepancy) and the wave number at which new metrics are introduced.

Before we move on, we address the issue of specifying model discrepancy. A reader might object that we are advocating a methodology for locating structural errors that requires us to know already what the structural errors are by providing a model discrepancy variance. Certainly, if model discrepancy variance for any metric can be specified or estimated by experts, then history matching can proceed straightforwardly. However, when tuning we do not expect this to be the case. If it is not, we can treat the discrepancy variance as our tolerance to structural error. This enables us to explore parameter space and discover whether or not regions containing parameter settings that are not inconsistent with the observations we would like to match to exist with respect to different tolerances to this error.

The notion of specifying a tolerance to error should not be unfamiliar to model developers tuning their climate models, where the goal is often to tune components of the climate model so that they are "close to" observations. How close is acceptable will be known to the modellers who are often varying one or a handful of parameters thought relevant to that process at any one time until an "acceptable" setting of the model parameters is found (or it is thought that a structural error exists). Hence, part of the definition of a structural error, is what the tolerance to model error is. For example, we might be more tolerant to errors in the location

of the Gulf Stream in a 2^o model than we would be to errors in its global mean temperature.

2.2 Computationally expensive models

One potential objection to adopting a rigorous statistical approach to climate model tuning that uses PPEs is that the latest climate models are too expensive to run, so that PPEs large enough to build emulators with cannot be obtained. This is not a problem for history matching.

History matching only requires an emulator for a climate model. Though one effective way to emulate a climate model is to use a large PPE, it is not the only way. In fact an emulator can be built for a model for which you have no data at all (Goldstein and Rougier, 2009; Williamson and Goldstein, 2013)! The most practically effective way to build an emulator for a slow, expensive climate model is to use a large PPE on a coarse resolution version of it. For example, as mentioned earlier, the ocean component of HadGEM3 is the 0.25^o resolution version of the NEMO ocean model. This model can also be run much more quickly at 2^o and 1^o resolution.

The idea is to use a large ensemble of coarse resolution models and to write down an emulator for the expensive model as a function of the emulator for the coarse version. Note that this is an emulator and that we need no runs of the expensive model to construct it. Though this emulator is likely to have large uncertainties on the predictions it makes, particularly when changes in resolution lead to changes of parameterization schemes, it can then be efficiently tuned using very small ensembles from the expensive model in order to reduce these uncertainties.

To be clear, we are not suggesting that a 1^o model run at some parameter choice x_1 would be informative for the 0.25^o output at x_1. For example, some parameters in x_1 may not be needed in the 0.25 degree model, and other new parameters may be required to run it. Our claim is that because the models are all simulating climate, then unless we claim that one of the versions has no skill even in reproducing large scale features of the climate (such as global mean temperatures or circulations), the different parameter spaces must be related. Large ensembles of the coarse model and small ensembles of the high resolution model can be used to establish this relationship statistically through an emulator for the high resolution version. New coarse experiments can reduce some of the uncertainty in this emulator as can new high resolution experiments, and the emulator can be used in history matching as with any other emulator.

For example, Williamson et al. (2012) emulate 200 year time series of the Atlantic Meridional Overturning Circulation (AMOC) in the coupled HadCM3 using a PPE with just 16 members and a large ensemble of a coarse version called FAMOUS. Given the prior emulator for the expensive model, one can use its uncertainty specification to aid experimental design decisions so that a finite budget of model evaluations can be spent on removing as much parameter space as possible. This will be more efficient than the "one factor at a time' type of approach that is currently used. For more information on emulating expensive models using coarse resolution versions see Cumming and Goldstein (2009); Kennedy and O'Hagan (2000); Le Gratiet (2014) and Williamson (2010).

Time and budget constraints prevent us from obtaining further ensembles of HadCM3 for this study, so that we cannot demonstrate iterative tuning for this model. However, in the rest of the paper we demonstrate the potential effectiveness of history matching for tuning climate models by using further constraints on the NROY space found in Williamson et al. (2013). These constraints are designed both to remove regions of parameter space with poor ocean circulations (including the standard HadCM3) and regions with the observed SST biases. Following this second history match, we can plot 1 and 2D projections of NROY parameter space and find which parameters drive the majority of the reduction of parameter space.

3 The Antarctic Circumpolar Current in HadCM3

At a recent workshop on ocean model discrepancy, a group of oceanographers and statisticians discussed key processes in the ocean that drive the AMOC and that would have to be modelled correctly in order for them to have confidence in the modelled transient response of the AMOC to CO_2 forcing. Of the processes mentioned, some of those deemed more important included location and strength of the sub-polar and sub-tropical gyres, temperature and salinity in the sinking regions in the North Atlantic and the strength of currents in the Southern Ocean. These discussions also led to a number of ocean processes in HadCM3 that were thought to impact upon AMOC strength being identified as having "known structural biases". We are motivated in this illustration of history matching as a tool for tuning climate models by investigating the nature of the biases that our experts deemed influential on the AMOC. We begin with the ACC strength.

ACC strength (Sv, 1 Sverdrup $= 1 \times 10^6 m^3 s^{-1}$), measured across Drake Passage, is an ocean transport that has proved difficult to capture accurately in AOGCMs. In the multi-model ensemble used to support the Intergovernmental Panel on Climate Change's fourth assessment report (IPCC-AR4 Solomon et al., 2007) known as CMIP3 (Coupled Model Intercomparison Project phase 3 Meehl et al., 2007), the range of ACC transports given by the then state of the art climate models (including HadCM3) was huge compared to the observations and their associated uncertainty (134 ± 15 to 27 Sv though this error is misquoted as being 11.2 Sv Cunningham et al., 2003). The CMIP3 models ranged from -6 to 336 Sv, but, perhaps more surprisingly, only two of the models returned an ACC strength consistent with the observations (see Russell et al., 2006, for details). The CMIP5 models (Meijers et al., 2012) fare a little better with a range of $90 - 245$ Sv but still only 2 models consistent with the observations. The Hadley centre models are all too strong in CMIP5.

If an overly strong ACC strength represents a structural error in HadCM3, this would imply that it is not possible for HadCM3 to simulate a realistic climate with an ACC strength close to observations. We investigate this possibility using a large PPE of HadCM3 runs described below.

3.1 The ensemble

We designed a large PPE on the coupled, non-flux-adjusted, climate model HadCM3 (Gordon et al., 2000; Pope et al., 2000). This ensemble varied 27 parameters con-

trolling both the model physics in the atmosphere and ocean of HadCM3 and was generated using Climate Prediction Dot Net (CPDN, `http://climateprediction.net`). CPDN is a distributed computing project through which different climate models are distributed to run on personal computers volunteered by members of the public. A copy of the model, along with a specific prescribed setting of the model parameters, is downloaded by the "client" computer, where it runs in the background using any spare computing resources available. Data is returned to CPDN where it is stored and made available for access by the general public.

The ensemble consists of a 10,000 member design in the chosen parameters submitted in April 2011. At the time of writing there are over 3500 unique ensemble members that have completed 120 years of integration with preindustrial boundary conditions. Information on the design of the ensemble can be found in Williamson et al. (2013) and Yamazaki et al. (2012). A comprehensive list of the parameters varied appears in appendix B.

Though we were able to use CPDN to generate a very large ensemble, our method does not rely on ensembles that are so large that only a program similar to CPDN would render it practical. Loeppky et al. (2009) suggest that a good rule of thumb for ensemble size, in the absence of information from previous experiments, is 10 times the number of parameters. Sexton et al. (2011) built GCM emulators for UKCP09 using slightly larger ensembles than this using the UK Met Office supercomputer. As discussed in section 2.2, emulators, and thus history matching, can be achieved using extremely small ensembles of very expensive models, if large ensembles (e.g. large enough to satisfy the rule of thumb suggested by Loeppky et al. (2009)) on related coarser resolution models are available.

3.2 NROY space

Williamson et al. (2013) perform a history match on HadCM3 using 4 observational metrics to cut out over half of the original parameter space. The NROY space for HadCM3 derived in Williamson et al. (2013) consists of all those parameter settings that couldn't be ruled out using global mean surface air temperature (SAT), global mean precipitation (PRECIP), the global mean surface air temperature gradient (SGRAD) and the global mean seasonal cycle in surface air temperature (SCYC). Hence any parameter setting in NROY space already has a not implausible global mean surface air temperature profile and global mean precipitation with respect to the chosen constraints.

3.3 NROY ACC

In order to further constrain NROY space using the ACC strength by history matching, we require an emulator for the ACC strength as well as an observational error variance and a discrepancy variance. We describe the emulator for ACC strength in appendix A and we interpret the Cunningham et al. (2003) range ($134\pm$ 15 Sv) as 3 standard deviations using the 3 sigma rule (Pukelsheim, 1994). This gives $\mathrm{Var}\,[e] = 25$. We note here that there are many ways one might interpret the error quoted in data range statements such as this. One is that the range represents hard boundaries on the value of the true process. Under this interpretation our

representation of the range as 3 standard deviations leads to a larger error variance than necessary and so less parameter space ruled out through history matching. The interval might be viewed as a confidence interval for the true value of the data, and our interpretation is consistent with the quoted range as a 95% confidence interval under the assumption that the underlying distribution of the observations is unimodal (Pukelsheim, 1994). A third way might be to interpret the quoted range as 1 standard deviation, however to use this interpretation here would imply that the search for models with an ACC strength any closer than 45 Sv (3 standard deviations) from the observations would be overfitting (even with a perfect model and perfect emulator). We know from conversations with NEMO developers and from the discussion of the performance of the CMIP5 models that the data are treated as being more accurate than this and that the field looks for models that are within the quoted data range. Treating the quoted range as 3 standard deviations is consistent with the desire to search for model runs that meet this constraint.

We specify zero tolerance to climate model error via a model discrepancy variance of 0, so that we demand that the model output lies within the range of the observation uncertainty. This assumption is not one we would make if our goal were to tune HadCM3, as we do have tolerance to model error. This study aims to explore the capabilities of HadCM3, hence, in specifying zero tolerance to model error, we are testing to see if the model is capable of replicating the observations whilst having a reasonable global temperature and precipitation profile. If we were actually looking to tune the climate model these tolerances would not be zero and may well be correlated across constraints as a modeller may tolerate more error in one type of constraint (e.g. AMOC strength) in favour of less error in another (e.g. SST). One argument, for example, for including a non-zero model discrepancy variance here, is that our simulations are preindustrial and hence we might want to account for climate change in the observations. However, we might also view this uncertainty as part of the observation error.

From Williamson et al. (2013) we know that 56% of the defined parameter space is removed using the first 4 constraints. Demanding not implausible ACC strength reduces the remaining space by 90.4% leaving just 4.3% of the parameter space not ruled out yet. We explore the properties of the parameters in this NROY space in section 4, however, we note that we have ensemble members that satisfy each of our 5 constraints and focus the rest of this section on exploring the behaviour of one of these models in particular. Figure 1 plots the mean ACC strength for the final decade of every member of our ensemble against the mean AMOC strength. Dashed lines represent the lower and upper bounds on the observations. Points outside of this box have ACC strength, AMOC, or both outside of the range given by the observations and may be thought of as having unphysical ocean circulations. We colour points ruled out by our wave 1 history match in Williamson et al. (2013) in grey and add the NROY members from this analysis in cyan. We colour those points NROY to the additional constraint of ACC strength in dark blue. Standard HadCM3 is plotted as the red triangle.

From this plot we can see that we have a number of not ruled out yet ensemble members with a not implausible ACC strength. We also see that the standard HadCM3 (plotted as the pink triangle) has an overly strong ACC. Note that many (more than the 5% expected) of the now NROY ensemble members (the blue points) still have ACC strengths outside of the observation range. This is a feature of history matching with emulators. We only rule parameter choices out

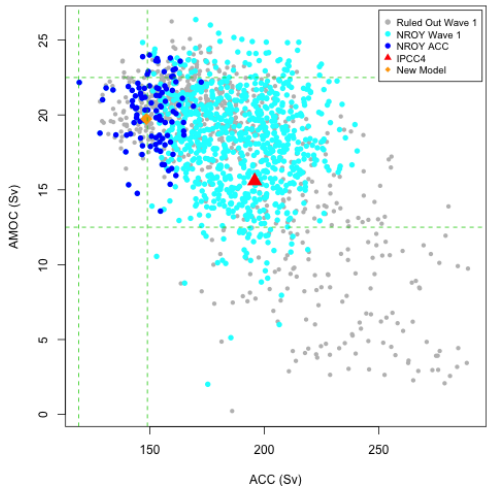

Fig. 1 The ACC (Sv) through Drake Passage plotted against the AMOC (Sv) at 26^{o}N in our HadCM3 ensemble. The dashed lines represent upper and lower bounds on observations of ACC and AMOC. We colour those ensemble members ruled out by the initial history matching in Williamson et al. (2013) grey, with NROY members from this wave in cyan. Members in NROY space when adding ACC strength as a constraint appear in blue. The standard HadCM3 is highlighted on this plot as the red triangle. The new run we examine in further detail in this section is shown as the orange diamond.

if we are sure they lead to unphysical circulations, and part of our uncertainty comes from the quality of our emulator (and the predictability of the model). Our particular choice of emulator (see appendix A) does not interpolate the ensemble members and give zero variance at those parameter locations because we have accounted for internal variability in our modelling. So we expect to see parameter choices, and therefore existing ensemble members, that are predicted to be (or have been observed to be) outside the target data range as NROY because our uncertainty (driven by internal variability in this case) is such that we can't be sure that they really do lie outside the data range for any setting of initial conditions.

3.4 Comparison with the standard HadCM3

The NROY members identified with the ACC constraint all pass the large scale constraints imposed by the initial history match (SAT, PRECIP, SGRAD, SCYC). We now examine the state of the climate of one of these members (the orange diamond in figure 1) in more detail. The ensemble member selected for more detailed analysis is typical of the NROY members identified in blue on figure 1, and whilst for any individual metric it is possible to select other equally good or better example (we present one such example for the barotropic streamfunction in appendix C), this member represents a good compromise of all the metrics we examine.

Although there are no direct observations of the barotropic streamfunction, comparison with other models and reanalyses, as well as observations of sea surface

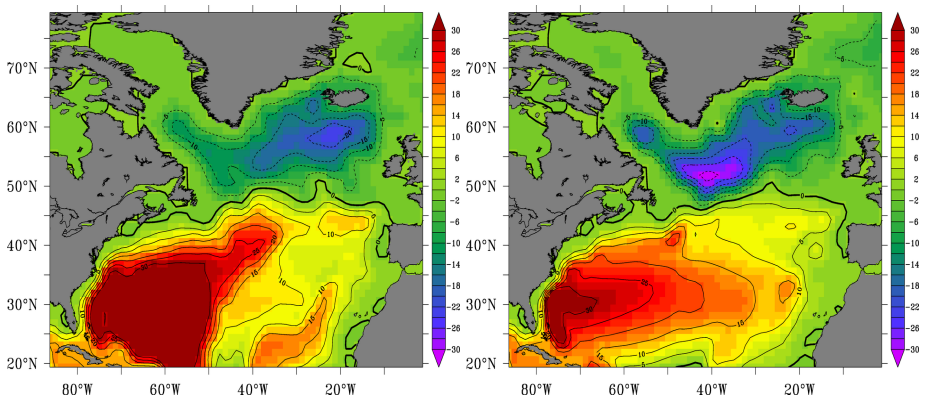

Fig. 2 The barotropic streamfunction (BSF) for the standard HadCM3 (left) and an ensemble member with realistic ACC strength (right). Units are Sv. The domain is smoothed spatially to remove grid point noise.

height from altimetry can be made. These comparisons indicate that the subpolar gyre is too far east in the standard HadCM3 (see figure 2), with its centre located around 25^{o}W, 55^{o}N (south of Iceland), and the subtropical gyre is too broad and diffuse, with most of the southward return flow occurring in a narrow band between 50-55^{o}W. In comparison, the alternative model with the more realistic ACC has a subtropical gyre which is much more tightly constrained to the western boundary and has more uniform southward return distributed across the rest of the basin. It also has a westward shifted subpolar gyre with a maximum transport located around 45^{o}W, 55^{o}N (south of Cape Farewell), though there is still a strong gyre near Iceland. Some of the ensemble members do exhibit weaker subpolar gyre circulation south of Iceland (see appendix C), although we note that modifications to the bathymetry in HadCM3 (excavation of the sills and a submerged Iceland) may force the model to produce an unrealistic eastward extension of the SPG circulation.

We continue our examination of the chosen NROY model with an assessment of the SST, SSS and circulation represented in the North Atlantic and Nordic Seas, a region considered very important because of the formation of North Atlantic Deep Water (NADW), which forms the lower limb of the AMOC. Figure 3 plots sea surface salinity (SSS) (left panels) and sea surface temperature (SST, right panels) anomalies with respect to the observations (Ingleby and Huddleston, 2007) for the standard HadCM3 (top panels) and the chosen NROY model in the bottom panels. A number of supposed "structural errors" in HadCM3 can be identified on the upper panels and have been improved by the alternative parameter choice. For example, the standard HadCM3 has a fresh bias of -0.5 PSU to -1.5 PSU in the subpolar gyre and along the Greenland coast. Extending out into the Norwegian sea the fresh bias can exceed -3. The fresh bias in the subpolar gyre is not present in the alternative model and the one along Greenland is halved. We also improve the salty bias which extends all the way down the eastern boundary and across the Atlantic at 20-25^{o}N. These improvements are at the expense of a slight freshening

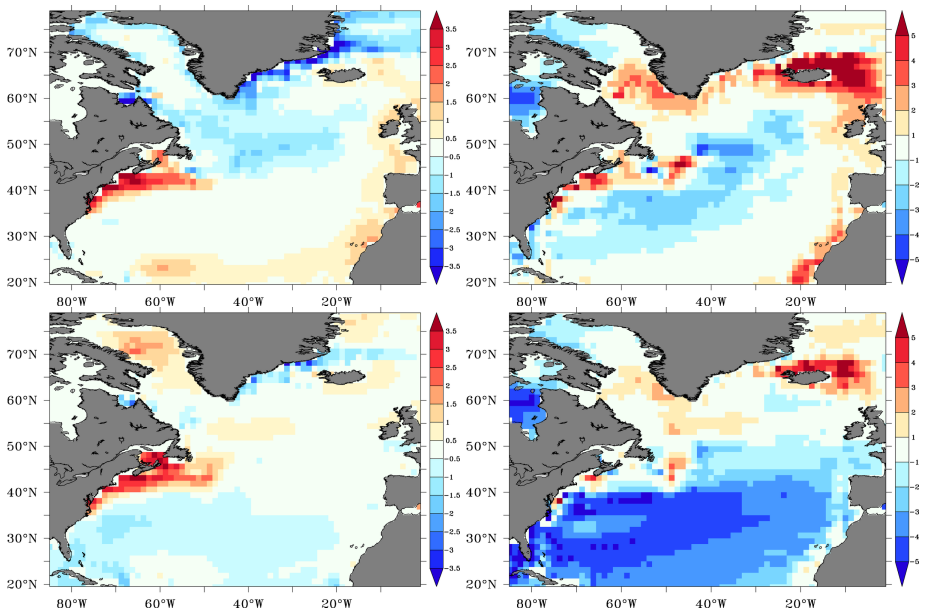

Fig. 3 Sea surface salinity (SSS) anomalies (PSU, left panels) and sea surface temperature (SST) anomalies (^{o}C, right panels) for the standard HadCM3 (top 2 panels) and the ensemble member with realistic ACC strength (bottom two panels). Both sets of anomalies are calculated as the difference of the mean of the last ten years in our ensemble and EN3 (Ingleby and Huddleston, 2007).

of the subtropical gyre. However, it is arguably more important that the salinity is correct in the AMOC sinking regions.

Note that both models exhibit the same positive salinity anomaly at the region of the Gulf Stream separation, indicative of a structural error which arises because of the model resolution. We believe that this anomaly is not possible to correct in HadCM3 by tuning parameters. The SST is also closer to observations in the North Atlantic sinking regions. Most notably the warm bias around Iceland, and extending west round Greenland and into the Labrador sea is reduced. However, these improvements are accompanied by the development of a larger and stronger cold bias in the sub tropical gyre. This cold bias is undesirable, and we would try to address this in the next wave. Many, though not all, of our NROY members have a similar cold bias in the sub tropical gyre. We believe that this bias (and the others introduced, see below) can be tuned out as its strength and sign vary across our ensemble, unlike, say, the salinity bias in the gulf stream, which was present in every ensemble member.

We can examine the circulation of the subtropical gyre in more detail. Figure 4 shows a cross section of the meridional velocity at 26^{o}N for the standard HadCM3 (top) and the alternative model with improved ACC (bottom). There is no available climatology to compare with for this field, but we refer the reader to observations presented in Johns et al. (2008) (their Fig. 7), and to the high resolution model climatology validated and used by Meinen and Garzoli (2014) (their Fig. 6). In the standard model the deep western boundary current is too

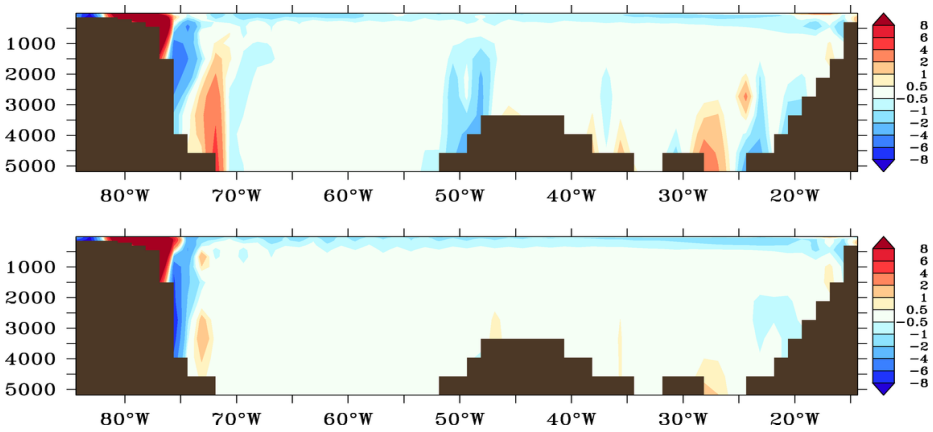

Fig. 4 Cross section of the meridional velocity (ms^{-1}) at 26^{o}N for the standard HadCM3 (top) and the ensemble member with realistic ACC strength (bottom). Red indicates northward flow, and blue indicates southward flow. Units are cm/s.

broad and shallow at the western boundary and there is a substantial northward transport below 2500 m between 72-74^{o}W. The intense southward transport indicated by the tight contours in figure 2 (50-55^{o}W) can be identified as a strong return flow at the western flank of the mid-Atlantic ridge and there are additional spurious transports on the eastern boundary. The alternative model with improved ACC transport exhibits a more physical deep western boundary current, which is tightly constrained to the western boundary. It does not have the large return flow on the western flank of the mid-Atlantic ridge and transports in the eastern basin are more realistic. By simply finding a NROY model with a more physical ACC transport through Drake Passage, we have found a model with an improved representation of the ocean circulation in the North Atlantic. However, we must also verify that these improvements have not arisen at the expense of the model developing serious problems elsewhere in the global climate.

Figure 5 compares the global SSS anomaly (left panels) and SST anomaly (right panels) fields for the standard HadCM3 (top panels) and the improved ACC member (bottom panels). SSS is improved almost everywhere outside of the Arctic Ocean. We note that HadGEM1, the successor of HadCM3, showed similar improvements to SSS globally, also replacing the Arctic SSS dipole anomaly with a pan Arctic positive anomaly (Johns et al., 2006). The SSS anomalies were also improved in CHIME (Megann et al., 2010), a coupled model closely related to HadCM3 where the ocean component was replaced by HYCOM and interestingly CHIME also exhibits the same pan Arctic positive SSS anomaly.

SST anomalies still present a problem. The alternative model shows the same gradient in SST anomalies, with the northern hemisphere exhibiting a cold bias and the southern hemisphere exhibiting a warm bias. The North Pacific cold bias is much stronger in the NROY simulation, exceeding 5^{o}C, and a cold bias associated with excessive equatorial upwelling can be seen in the Pacific. South of 20^{o}S the alternative model shows substantial improvement, with the warm biases being reduced both in area and amplitude. Interestingly, the North Pacific cold bias is

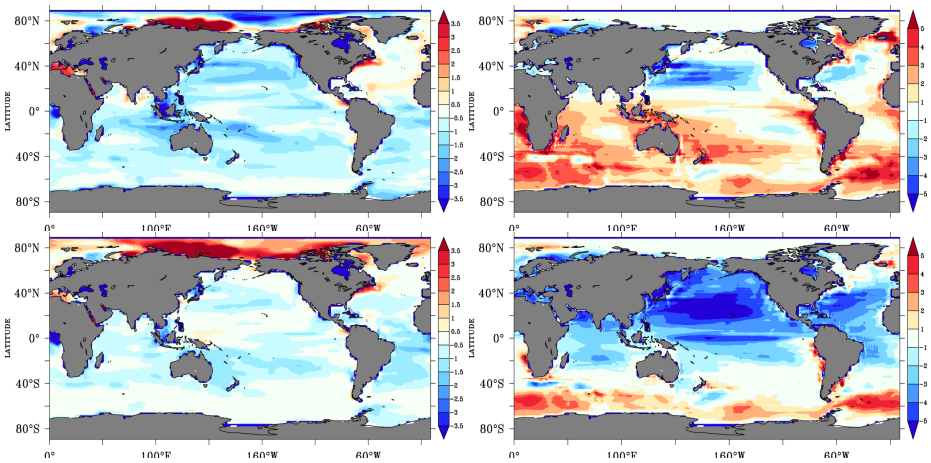

Fig. 5 Left panels are global SSS anomalies (PSU) and right panels are global SST anomalies (^{o}C), both with respect to EN3 (Ingleby and Huddleston, 2007) climatology. The top panels are for the standard HadCM3 and the bottom panels are from an ensemble member with realistic ACC strength.

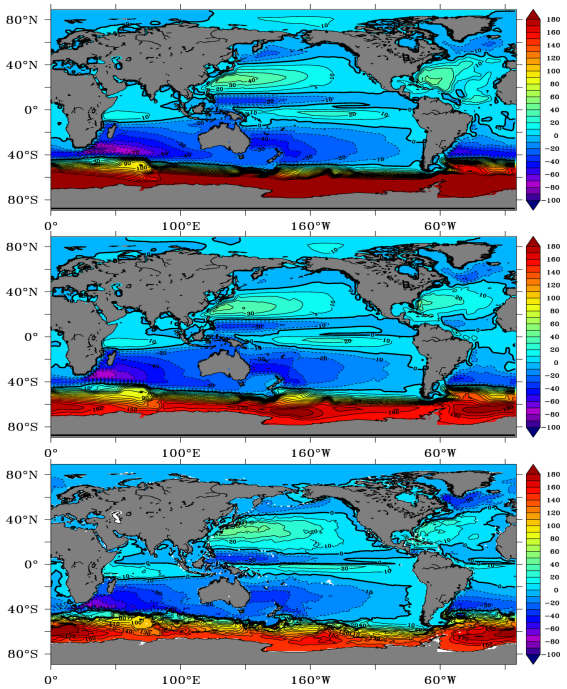

Fig. 6 Global barotropic streamfunction for the standard HadCM3 (top), the improved ACC run (middle), and for a NEMO ORCA025 (1/4^o) ocean only simulation

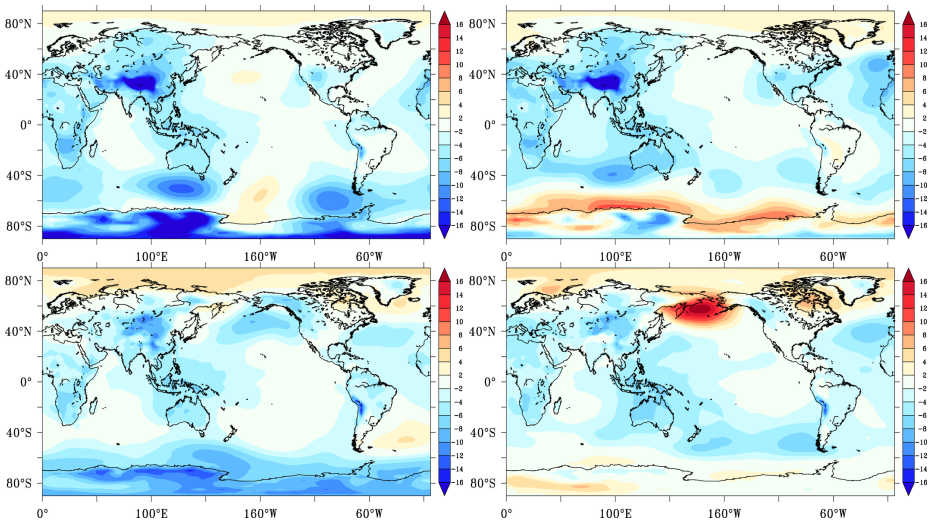

Fig. 7 Sea level pressure (SLP) anomalies (*mbar*) with respect to ERA-40 (Uppala et al., 2005) 1960-1990 climatology for summer (left) and winter (right). Anomalies for the standard HadCM3 are shown in the top panels and the bottom panels are anomalies from the improved ACC strength run.

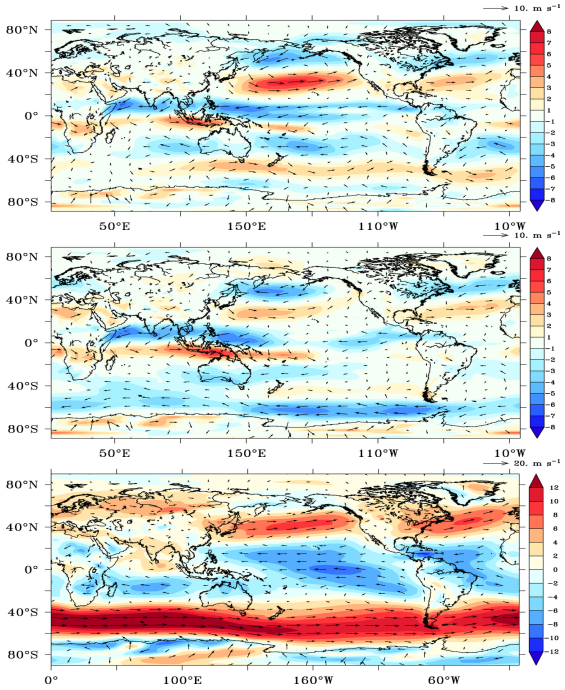

Fig. 8 Global wind field anomalies from the ERA-40 reanalysis at 850hpa (ms^{-1}) for the standard HadCM3 (top) and the improved ACC run (middle). Shading represents the zonal wind anomaly. The ERA-40 reanalysis climatological mean is also presented for reference (bottom)

present in both HadGEM1 (Johns et al., 2006) and HiGEM (Shaffrey et al., 2009) is much less evident in CHIME (Megann et al., 2010), but the North Pacific is still recognisably biased cold compared with the global mean. The gradient in SST anomalies (generally cold in the northern hemisphere with warmer biases in the southern hemisphere) may be indicative of different biases in the air-land and air-sea interactions, suggesting that the northern and southern hemisphere biases could be controlled by different parameters and that there is therefore scope to reduce the slope and improve the bias overall.

HadCM3 is a rigid lid model, so there isn't a sea surface height field to compare with observations, and the rigid lid pressure is not saved as a time mean output. We do, however, have the time mean barotropic streamfunction, which we can compare with state-of-the-art high resolution ocean only simulations forced with observed surface fluxes. Comparing the global barotropic streamfunction in the standard and improved ACC simulations to a simulation of NEMO ORCA025 ($1/4^o$) referred to as N206 (figure 6; see Blaker et al. (2014) for a detailed description of N206), substantial improvements exist over both the Atlantic and Southern Oceans. N206 has a mean ACC transport of 139 Sv, and an AMOC transport of 14.9 Sv. The barotropic streamfunctions of the two HadCM3 simulations are similar over the Indian and Pacific oceans, with the standard HadCM3 producing a slightly stronger Pacific subtropical gyre and Kuroshio, as seen in N206.

There are improvements in the SLP (Figure 7), particularly over the Southern Ocean and Antarctic continent, but also over land across much of the globe, most notably over the Himalayas. The improvements in SLP are closely linked to the wind field (figure 8). Comparing the 850 hpa wind anomalies for the two simulations (top two panels), differenced from the ERA-40 (Uppala et al., 2005) reanalysis 1960-1990 mean (bottom panel) there are improvements in the mean wind field over much of the globe, with the improved ACC model showing more realistic wind strength over northern extent of the Southern Ocean, and better easterly winds over the tropical Pacific and Atlantic. The zonal wind over the southern extent of the Southern Ocean is anomalously weak in the improved ACC run. Both simulations exhibit a too-zonal storm track over the North Atlantic, whilst in the Pacific the storm track is too far south from around 150^oE to 180^oE. The standard HadCM3 storm tracks are stronger than the climatology, whilst the storm tracks simulated by the alternative model with improved ACC are slightly weaker. The model with improved ACC representation displays a localised, strong positive SLP anomaly in the North Pacific between Kamchatka and Alaska, related to the weaker and more southerly Pacific storm track. This may be considered undesirable in a model used for UK weather and climate prediction if it affects the characteristics of the northern hemisphere storm track, however the limited data available from the CPDN ensemble means we cannot investigate this more closely.

Comparing precipitation anomalies between standard HadCM3 and the alternative model we see reductions in the error almost everywhere, particularly over the maritime continent and along the ITCZ, but also in the subtropical regions (figure 9). However, we note that there are large uncertainties in precipitation climatologies so these improvements should be regarded with caution.

In figure 10 we plot the Meridional Heat Transport (MHT) and the AMOC for every member of our ensemble, and highlight the standard HadCM3 (red) and the improved ACC (blue) members. The grey lines are ruled out (RO) ensemble members and the other colours correspond to the value of entcoef, the parameter

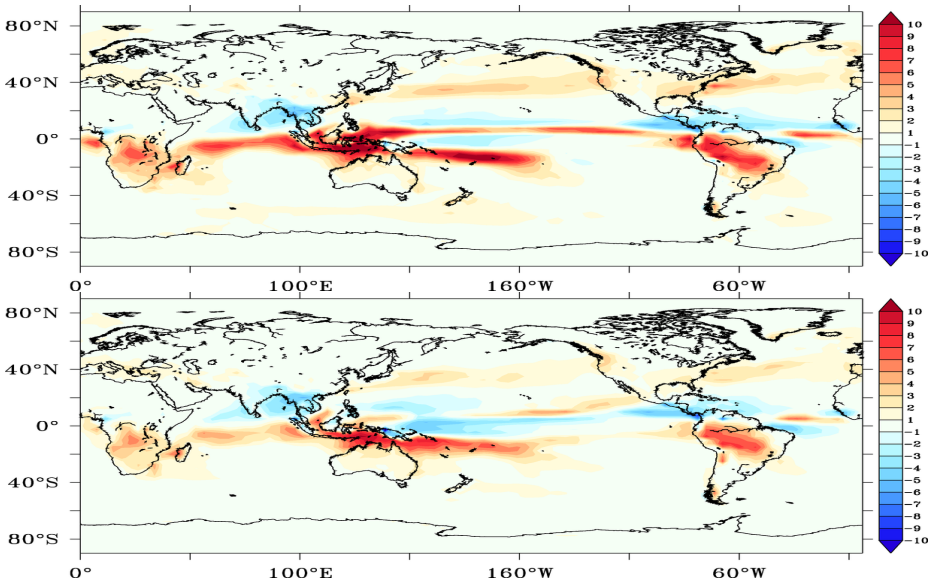

Fig. 9 Precipitation anomalies (mm/day) from ERA-40 (Uppala et al., 2005) 1960-1990 climatology for the standard HadCM3 (top) and the improved ACC model (bottom).

which governs the convective cloud entrainment rate. We see that both the MHT and AMOC are stronger and closer to observations. We also note that the variability in the AMOC is larger for the improved ACC run, an observation that was true of each of the improved ACC runs we looked at.

4 Refining the search

Through jointly varying atmosphere and ocean parameters and using history matching we have found a region of parameter space with a predicted not implausible global mean temperature profile, global mean precipitation and ACC transport through Drake Passage. This region of parameter space represents only 4.3% of the original space. Within this space we have found a version of HadCM3 that outperforms the standard version in many aspects of its ocean and atmosphere, and is arguably a better model. However, some might argue that the improvements to certain aspects of the ocean and atmospheric circulation have come at the too high price of exacerbating the cold bias in the north Pacific. To head off such objections, we are not claiming to have found the "best" HadCM3 in any sense, nor that if the goal were to tune HadCM3, one should choose ACC transport through Drake Passage as a primary metric to cut down parameter space. It might be that, if a certain combination of metrics were used and ACC were left out, then something close to standard HadCM3 would be found (though our results make us doubt this). Instead, we claim that this method will find models that will exhibit improvements in key metrics chosen by the modeller if all of those metrics can be improved by changing the model parameters.

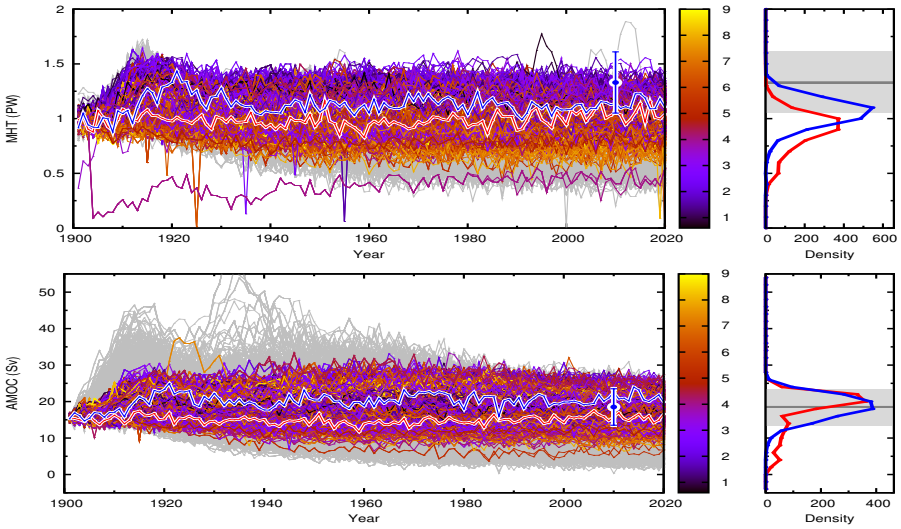

Fig. 10 The Meridional Heat Transport (top panel) and AMOC time series for each member of our ensemble. Grey runs have been ruled out by history matching. NROY runs are coloured by their value of entcoef. The highlighted red line in each image is the time series for the standard HadCM3 and the highlighted blue line is the time series for our alternative with realistic ACC transport. The curves on the right shows the unscaled density of the final year data for the NROY part of the ensemble (blue) and the ruled out part of the ensemble (red). The blue point represents the observations and includes error bars (also reproduced as grey shading in the unscaled density plots on the right hand side).

Further, the first model that is found within the NROY space that satisfies current metrics will not be the optimal version of the climate model unless we have been extremely fortunate to have hit the exact optimal parameter setting within our 27 dimensional space of continuous parameters. Instead, this model provides insight into which metrics have been improved and which must be used in further history matching (as long as this is physically appropriate). In our application to HadCM3, we would want to refine our model search to only include models that correct the ACC bias whilst simultaneously improving the representation of global SST.

In order to refine our search for HadCM3's with fewer, less serious structural biases using history matching, the process is to first select metrics on which to match, to emulate these within the current NROY space, and to use the results to design a new ensemble within the latest NROY space, to refocus our statistical models, then repeat in order to converge either on a set of models that reproduces all specified metrics to within the chosen error tolerances, or upon a number of metrics that cannot be simultaneously reproduced and can thus be correctly identified as structural errors. Time and budget constraints have left us unable to run any further ensembles of HadCM3 with CPDN, however, we can perform the first part of this task, in order to gain insight into features of the NROY space of HadCM3 when we require that SAT, Precip, ACC and SST profiles are not unphysical. A future project might then seek to populate this space with models in

order to investigate their properties further, to further cut down NROY space, and to look at transient simulations of the most physical looking models if appropriate.

To do this, we include the SST anomaly in the sub-tropical gyre as a chosen metric. We could have included features of the Pacific SST or even certain spatial patterns as metrics for history matching, however, it was felt that the most crucial region to get right in order to have confidence in the AMOC, was the North Atlantic, and our model exhibits a large SST bias there too. It was also felt that correcting this bias, if possible, would simultaneously improve the temperature of the North Pacific. We define our metric to be the mean SST in a box from 70^oW to 30^oW and from 26^oN to 36^oN. The "improved" HadCM3 we found has anomalies up to 5^oC in this region. We specified a tolerance to error of half of that, so that our model discrepancy has a 3 standard deviation range of 2.5^oC. Our discrepancy variance is therefore 0.69. The region of the Atlantic we are assessing here is very well observed compared with global SST, so the observation error variance is likely to be low (calculations based on Ingleby and Huddleston (2007) estimated the observation error variance on global mean SST observations over a 30 year period as around 0.003). We therefore ignore observation uncertainty for this constraint, taking the view that it is negligible relative to model discrepancy variance.

4.1 Results

In this section we explore the NROY space left when history matching to our 4 prior constraints as well as the ACC strength and the SST in the sub-tropical gyre. To do this we evaluate implausibilities, via equation (1), for millions of untried points in parameter space. We first estimate the volume of NROY space relative to the original parameter space. This is done by Monte Carlo simulation where a large number of points are uniformly drawn from parameter space and the proportion within NROY space recorded. After matching to the ACC strength in section 3 we had ruled out over 95% of the parameter space of HadCM3. Our current NROY space is just 0.7% of the original space now that we include the North Atlantic SST constraint, an estimate based on 10^6 Monte Carlo samples.

We can investigate features of the shape of the NROY parameter space by sampling implausibilities. By looking at 1 and 2 dimensional representations of NROY space, we can assess how different parameters combine to improve the model. Figure 11 shows marginal density plots for 9 of the more interesting atmosphere and ocean parameters. The first panel, showing the convective cloud entrainment rate coefficient, entcoef, indicates that low values of entcoef are implausible, as shown by Joshi et al. (2010) for HadSM3, and that there are more NROY models at the upper end of its range than in the range between 2-4 that is often determined to contain the best models (Sexton et al., 2011; Rowlands et al., 2012). Isopycnal diffusivity in the ocean (ahi1_si) is also very active, with values towards the top end of its range favoured. The cloud droplet to rain conversion rate, ct, and the relative humidity threshold for cloud formation, rhcrit, have similar profiles as do the cloud droplet to rain threshold over land, cwland, and the boundary layer cloud fraction at saturation, eacf. There is also a similarity in the marginal form for the ocean mixed layer parameters lamda, the wind mixing energy scaling factor and delta_si, the wind mixing energy decay depth, with lamda in particular quite active.

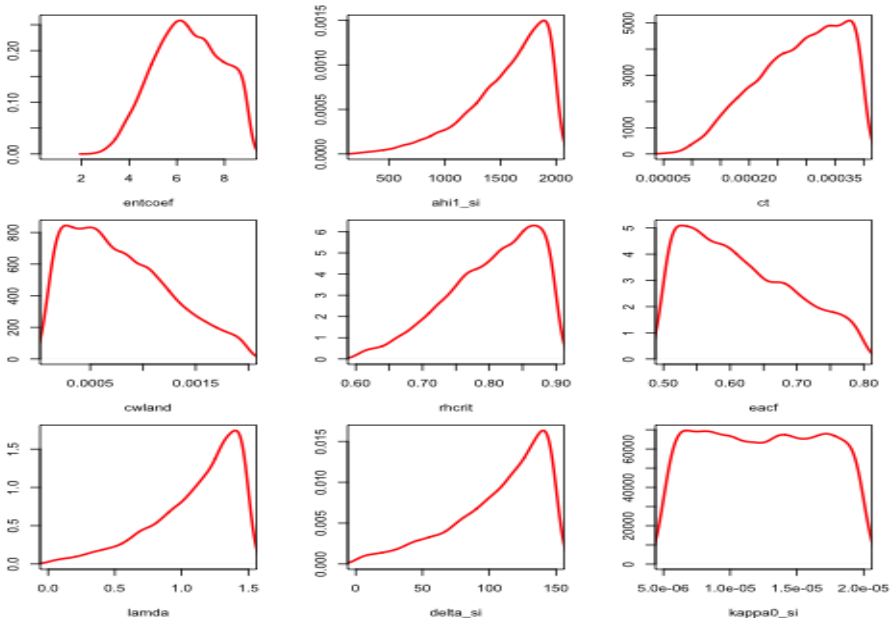

Fig. 11 Marginal NROY density plots 9 for the parameters.

Note that a collection of 1D marginals can be difficult to interpret. For example, the parameter setting fixed on each of the marginal modes is ruled out by SAT. More informative are the NROY density and minimum implausibility plots for 2D projections of the parameters, shown in figure 12. Each panel on the upper triangle shows the proportion of parameter settings behind each pixel that are NROY. The map is drawn by fixing the two parameters labelled for each panel at the value of a pixel, and exploring a 1000 point Latin Hypercube in the other 25 dimensions of HadCM3, plotting the proportion of samples in NROY space. Hence each upper triangle image can be viewed as a 2D projection of the density of NROY space. Grey regions are completely ruled out, meaning that, for any grey pixel, we were unable to find any NROY parameter setting in the other 25 dimensions for the given value of the other 2.

The standard version of HadCM3, which is ruled out by our history match to ACC strength, is plotted as the solid triangle in each panel. The version of HadCM3 we explored in section 3 is the circular point. From this picture we notice that though, perhaps, entcoef was a little low for the standard setting, the combination of low rhcrit and ct rules out the standard parameter setting anyway.

The NROY density plots reveal a great deal of non-linear structure to NROY space. We see that ocean parameters, such as the isopycnal diffusivity (ahi_si) must be varied jointly with cloud parameters such as ct, cwland and eacf in order to find NROY models. This result is much stronger than saying that tuning procedures that only vary one parameter at a time will not be successful for HadCM3, it says that one must tune the atmosphere and the ocean together. Parameters that appear to be reasonable at tuning a particular process or even the atmosphere

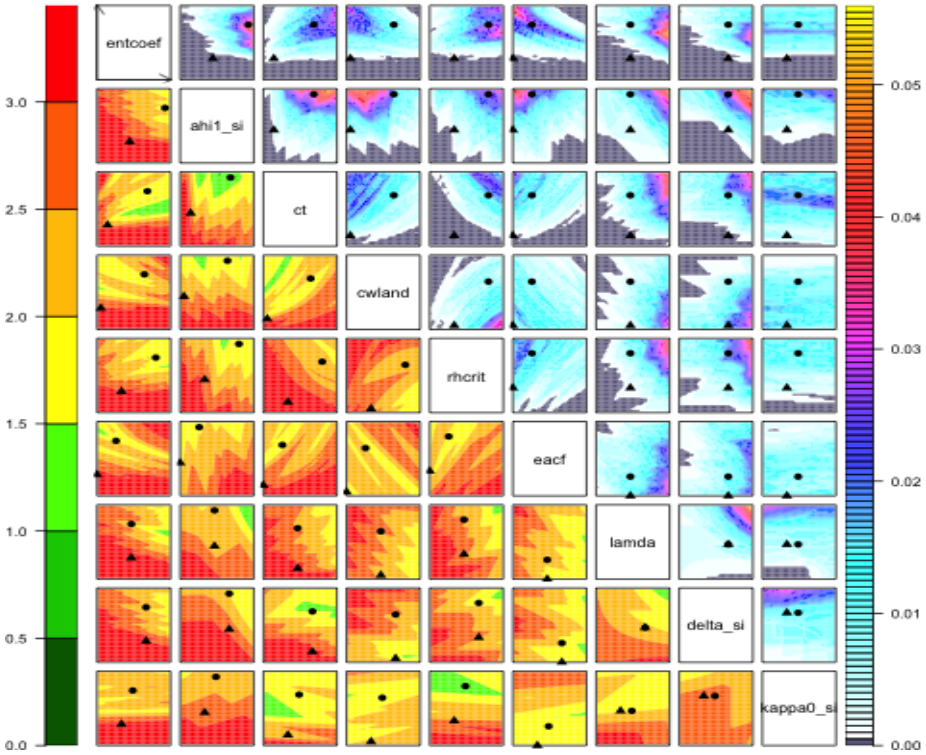

Fig. 12 NROY density plots (upper triangle) and minimum implausibility plots (lower triangle) for 2D projections of NROY space. Each panel plots either NROY density or minimum implausibility for a pair of HadCM3 parameters. NROY densities, for each pixel on any panel in the upper triangle, represent the proportion of points behind that pixel in the remaining 25 dimensions of HadCM3's parameter space that are NROY and are indicated by the colour whose scale is indicated on the right. Minimum implausibilities, for each pixel on any panel on the lower triangle of the picture, represent the smallest implausibility found by fixing the two parameters at the plotted location and searching the other 25 dimensions of the HadCM3 parameter space. These plots are orientated the same way as those on the upper triangle, for ease of visual comparison. Standard HadCM3 is depicted on each panel as the triangular point. The parameter setting we explored in section 3 is the circular point.

only model, may not be close to optimal for different, and better configurations of the model ocean.

The lower triangle shows minimum implausibility plots. Similar to those on the upper triangle, for each pixel, representing a fixed value of a pair of parameters, a Latin Hypercube in the other 25 dimensions of HadCM3 is searched for NROY parameter values. We plot the value of the smallest implausibility found in each LHC. The plots have the same orientation as those on the upper triangle so that comparisons are easier to make. If our emulators were extremely accurate, green and yellow areas of this plot would indicate the location of potentially "good" settings of the model parameters. Though history matching restricts itself to ruling out the bad parameter settings, if that has been done and the emulators were extremely accurate (with little uncertainty in the posterior), we would look to

further explore the green and yellow areas of these plots as containing points that are actually consistent with the data.

Certain panels in the lower triangle reinforce the case for varying all of the parameters of the model simultaneously. Take the plot of ct against rhcrit for example (and remember the orientation mimics the upper triangle). Suppose we keep one parameter fixed and vary the other, starting from low values of ct and rhcrit (as in standard HadCM3, though perhaps with a lower value of rhcrit if we are being strict). By doing this, according to the figure, we would never escape the red zone, meaning that we would never find a model that satisfies all of our metrics to within our tolerance to error. We might then come to the conclusion that we had identified a structural error. But note that this figure actually implies that if just one of the parameters were held fixed, and all 26 other parameters varied, we would still not come across any NROY models. We would believe that we had found structural errors, and yet, by varying all of the parameters, we find better models (as we have) and can point to even better ones at untried parameter settings.

The upper triangle shows that, with the exception of entcoef, whose values cannot be low, no matter what the setting of the model parameters, values can be found for each of the other parameters that would lead to a NROY model, though most do not. For example, in general, the lower the isopycnal diffusivity in the ocean is set, the larger the mixed layer parameter delta_si (which governs the decay of wind mixing energy with depth) must be in order to avoid ruling out that parameter choice, independent of any of the other parameters. These restrictions on any given parameter range would become greater as either more metrics were included or as further ensembles allowed more accurate emulators to be built within parameter space (reducing the denominator in (1) and thus increasing the number of models ruled out).

As noted previously (see figure 1), our model described in section 3 is NROY, whilst the standard setting is ruled out. However, it is clear that though our model may have reasonable settings for some of the parameters, there are regions of parameter space with a higher density of NROY points that we would like to explore, and that our model is not one of the lowest implausibility models found during sampling the emulator. If we had the resource, the next step in this type of analysis would be to design an ensemble within NROY space, run it, re-emulate and perform another history match to further refocus the search.

We note that there are lots of NROY models in some of the corners of parameter space. This raises questions about whether or not the pre-defined ranges of NROY space were wide enough. These are valid questions which underline the importance of exploring the widest physically plausible parameter ranges right from the start (see discussion of this topic in Williamson et al., 2013). However, during the first 1 or 2 waves of a history match, the more likely explanation for seeing a lot of NROY models in corners of parameter space is due to a feature of the statistical modelling.

We design our emulators so that the uncertainty outside of the convex hull of points in the ensemble (the smallest convex region containing all ensemble members) increases asymptotically. This is a feature of any emulator with a mean function containing non-constant functions of the parameters (Draper and Smith, 1998). We design our emulators in this way in order to avoid extrapolation issues whereby the emulator reverts back to the prior mean outside the convex hull of

explored parameter settings, but with low uncertainty so that we might incorrectly rule out points on the edges and in the corners. HadCM3 has 2^{27} corners and our ensemble is far smaller than this, so we have many unexplored corners in parameter space. This is one of the reasons we are so careful with our language. Those high density regions in the corners are Not Ruled Out Yet, but it is likely that they will be once our emulators can be better tuned there. If there were still high density regions in corners or on edges after multiple waves, we might suspect that there really were good models on the edges, and that would give us cause to consider parameter values beyond the current boundary.

Note that it may seem natural to ask about the behaviour of models (or even the model) with the lowest values of implausibility. At this stage of the analysis, that would not be appropriate. The careful language and philosophy behind history matching supports only the notion that models with large implausibilities can be ruled out and says nothing about those with small implausibilities (hence "not ruled out yet" as opposed to "acceptable" or even "likely"). The reason is the emulator. During these early waves of analysis, much of the denominator in the implausibility calculation is emulator variance, so low implausibility can mean that we are simply uncertain regarding what will turn out to be a poorly matching model. We advocate history matching as a method for tuning only if used in multiple waves as discussed in section 2.1. The waves allow emulator uncertainty to be reduced to the point where NROY models are close to the data (or there are no NROY models). They also allow more complex joint emulators to be built more easily facilitating approaches to finding the "best" models through either multivariate implausibility or otherwise (e.g. calibration as in (Edwards et al., 2011)). We provide a large sample of NROY parameter values and their implausibilities in the supplementary material for readers interested in further exploration of this NROY space. We also include the experiment IDs and parameter values for our NROY ensemble members in the supplementary material.

5 Discussion

Tuning a climate model with a high dimensional parameter space and a long run time is a difficult task. Currently this task is undertaken without taking advantage of the latest statistical technologies for managing uncertainty in complex models. These methods allow for a targeted and comprehensive search of parameter space for models satisfying numerous criteria. We have argued that many perceived structural errors or "known biases" in climate models may be present due to an inefficient search of the existing model parameter space during model development.

We have presented history matching, a technique already used to quantify parametric uncertainty with climate models, as a method for climate model tuning based on sound principles of statistical design. The method seeks to tune all parameters simultaneously by using PPEs and emulators to rule out regions of parameter space that lead to models that do not satisfy observational constraints imposed by the model developers. We describe how the procedure should be undertaken iteratively, with new constraints and new PPEs used to refine the search for models without perceived structural biases. We also discuss how to use coarse resolution versions of an expensive model so that history matching can be used to

assist in tuning the expensive version without the requirement for large ensembles or long runs.

We have illustrated the power of this technique in investigating perceived structural biases in the HadCM3 ocean circulation. We found that the perceived structural bias in the ACC strength could be corrected by jointly varying both atmosphere and ocean model parameters and showed that these changes also improved other important physical properties of the ocean circulation, without compromising the surface air temperature profile.

We showed that the location of the sub-polar gyre was more realistic in the model we found than in the standard HadCM3 and that the western boundary current intensification in the sub-tropical gyre was greatly improved. We showed that the depth profile of meridional velocities in the North Atlantic deep, unrealistic in the standard HadCM3, compares favourably with the current physical understanding of these flows. We showed that the global sea surface salinity was closer to observations, but that the models found in this wave of history matching had a larger cold bias in the northern hemisphere SST, though the Southern Ocean warm bias in the standard HadCM3 was improved. We showed that the pressure and wind fields, particularly in the Southern Ocean were far more realistic in the model without the ACC bias, and showed that precipitation anomalies, particularly around the ITCZ were also improved. The AMOC and MHT are increased in the improved ACC model, though the values are still consistent with observations.

We then illustrated the method of iterative history matching by imposing a further constraint on parameter space designed to look for models without the cold bias in the northern hemisphere SST. We ruled out over 99% of the model parameter space as possibly containing models that satisfied our constraints and showed the joint structure of the remaining space using 1 and 2 dimensional projections of it. We found that jointly tuning atmosphere and ocean parameters, instead of tuning them one or even a few at a time was important for finding these regions of parameter space. We found complex interactions between cloud, ocean and convection parameters that would likely confound any one at a time approach to tuning and lead us to different results than would be obtained, for example, by not tuning the parameters together. For example, we found that the marginal distribution for the convection parameter entcoef had much of it's lower range removed, yet had more density at the higher values of its range, contrary to the findings of previous studies that have not simultaneously varied the ocean parameters along with the atmosphere. Though this result may depend heavily on our choice to tune to ACC strength, we have established through this, the need for joint exploration of parameter space.

Though we have no further access to ensembles of HadCM3 through CPDN as part of this work, further work could run an ensemble within the 1% of parameter space that is NROY in order to search for even better models by history matching in multiple waves. History matching is most effective with multiple waves of PPEs, as the emulators improve in the region of parameter space potentially containing good climates due to a higher density of model runs there. The improved emulators have lower variances, which serves to increase implausibilities and rule out more space.

Our work suggests that an overly strong ACC strength in HadCM3 is not a structural error, but a calibration error. However, it may be the case that more realistic ACC strengths are only possible at the expense of introducing new biases

in processes deemed more important than the ACC by model developers, for example, it may turn out that the SST cold bias in the northern hemisphere in the alternative model we studied cannot be improved without compromising the ACC strength, though the results we presented in section 4.1 suggest that this would not be the case. However, the best way we know of to find out for sure is to use history matching with all of the important constraints included. If this is done at the model development stage, structural errors in a process can be identified by an attempt to history match using that process ruling out all of the parameter space. If the constraints are introduced iteratively, in order of importance, the modellers can determine where the structural errors are and use this information to focus their research into improving the model in order to reduce or remove these errors.

Given the cost of developing GCMs and their importance for decision making and global policy strategy, it is important that every opportunity to improve the accuracy of these models is taken. History matching offers a robust and rigorous statistical methodology that is easy to implement and can be used to help to efficiently tune the parameters of GCMs.

Acknowledgements

This research was funded by the NERC RAPID-RAPIT project (NE/G015368/1). Daniel Williamson was also funded by an EPSRC fellowship, grant number EP/K019112/1. We would like to thank the CPDN team for their work on submitting our ensemble to CPDN users. We'd also like to thank the Institute of Advanced Study at Durham University for funding and hosting our workshop on ocean model discrepancy which formed the motivation for these investigations. In addition, we thank the oceanographers who participated in this workshop. We'd like to thank the CPDN users around the world who contributed their spare computing resource as part of the generation of our ensemble. NCEP and CMAP Precipitation data were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their website at `http://www.esrl.noaa.gov/psd/`

References

Acreman, D. M. and Jeffery, C. D. (2007), "The use of Argo for validation and tuning of mixed layer models," *Ocean. Model.*, 19, 53–69.

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

Annan, J. D., Lunt, D. J., Hargreaves, J. C., Valdes, P. J. (2005), "Parameter estimation in an atmospheric GCM using the ensemble Kalman filter", *Nonlinear Processes in Geophysics*, 12, 363–371.

Annan, J. D., Hargreaves, J. C., Ohgaito, R., Abe-Ouchi, A., Emori, S. (2005) "Efficiently constraining climate sensitivity with ensembles of paleoclimate simulations", *SOLA*, 1, 181-184, doi:10.2151/sola.2005-047.

Bellprat, O., Kotlarski, S., Luthi, D., Schar, C. (2012). "Objective calibration of regional climate models", *Journal of Geophysical Research*, 117, D23115.

Blaker, A. T.; Hirschi, J. J-M.; McCarthy, G.; Sinha, B.; Taws, S.; Marsh, R.; de Cuevas, B. A.; Alderson, S. G.; Coward, A. C. (2014), "Historical analogues of the recent extreme minima observed in the Atlantic meridional overturning circulation at 26N". *Climate Dynamics*, 10.1007/s00382-014-2274-6

Challenor, P., McNeall, D., and Gattiker, J. (2009), "Assessing the probability of rare climate events," in *The handbook of applied Bayesian analysis*, eds. O'Hagan, A. and West, M., Oxford University Press, chap. 10.

Collins, M., Brierley, C. M., MacVean, M., Booth, B. B. B. and Harris, G. R. (2007) "The Sensitivity of the Rate of Transient Climate Change to Ocean Physics Perturbations," *J. Clim.*, 20, 23315–2320.

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1996), "Bayes Linear Strategies for Matching Hydrocarbon Reservoir History," in *Bayesian Statistics 5*, eds. Bernado, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford University Press, pp. 69–95.

Craig, P. S., Goldstein, M., Rougier J. C., and Seheult, A. H. (2001), "Bayesian Forecasting for Complex Systems using Computer Simulators," *J. Am. Stat. Assoc.*, 96, 717–729.

Cumming, J. A. and Goldstein, M. (2009), "Small sample designs for complex high-dimensional models based on fast approximations," *Technometrics*, 51, 377–388.

Cunningham, S. A., Alderson, S. G., King, B. A. (2003) "Transport and variability of the Antarctic Circumpolar Current in Drake Passage", *Journal of Geophysical Research*, 108, No. C5, 8084, doi:10.1029/2001JC001147.

Daniel, C. (1973) "One at a time plans", *Journal of the American Statistical Association*, 68, 353–360.

Draper, N. R., Smith, H. (1998), "Applied Regression Analysis," 3rd Edition, John Wiley and Sons, New York.

Edwards, N. R., Cameron, D., Rougier, J. C. (2011), "Precalibrating an intermediate complexity climate model", *Clim. Dyn.*, 37, 1469–1482.

Fisher, R. (1926) "The arrangement of field experiments", *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513.

Friedman, M., Savage, L. J. (1947) "Planning experiments seeking maxima", in *Techniques of Statistical Analysis*, eds Eisnenhart, C., Hastay, M. W., Wallis, W. A. New York: McGraw-Hill.

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, Zhang, M. (2011), "The Community Climate System Model Version 4", *Journal of Climate*, 24, 4973–4991.

Gladstone, R. M., Lee, V., Rougier, J. C., Payne, A. J., Hellmer, H., Le Brocq, A., Shepherd, A., Edwards, T. L., Gregory, J., Cornford, S. L. (2012). "Calibrated prediction of Pine Island Glacier retreat drink the 21st and 22nd centuries with a coupled flow line model", *Earth and Planetary Science Letters*, 333-334, 191–199.

Goldstein, M and Rougier, J. C. (2009), "Reified Bayesian modelling and inference for physical systems", *J. Stat. Plan. Inference*,139, 1221–1239.

Gordon, C., Cooper, C. Senior, C. A., Banks, H., Gregory, J. M., Johns, T. C., Mitchell, J. F. B., and Wood, R. A. (2000), "The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments," *Clim. Dyn.*, 16, 147–168.

Hargreaves, J. C., Annan, J. D., Edwards, N. R., Marsh, R. (2004), "A efficient climate forecasting method using an intermediate complexity Earch System Model and the ensemble Kalman filter", *Climate Dynamics*, 23, 745–760.

Haylock, R. and O'Hagan, A. (1996), "On inference for outputs of computationally expensive algorithms with uncertainty on the inputs," in *Bayesian Statistics 5*, eds. Bernado, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford University Press, pp. 629–637.

Ingleby, B., and Huddleston, M., 2007: "Quality control of ocean temperature and salinity profiles - historical and real-time data." *Journal of Marine Systems*, 65, 158-175, doi:10.1016/j.jmarsys.2005.11.019.

Johns et al. (2006) "The New Hadley Centre Climate Model (HadGEM1): Evaluation of Coupled Simulations", *Journal of Climate*, 19,1327–1353.

Johns et al. (2008) "Variability of Shallow and Deep Western Boundary Currents off the Bahamas during 2004?05: Results from the 26N RAPID?MOC Array", *Journal of Physical Oceanography*, 38, 605–623.

Joshi, M. M., Webb, M. J., Maycock, A. C., Collins, M. (2010), "Stratospheric water vapour and high climate sensitivity in a version of the HadSM3 climate model", *Atmos. Chem. Phys.*, 10, 7161-7167.

Kalnay et al. (1996), "The NCEP/NCAR 40-year reanalysis project", *Bull. Amer. Meteor. Soc.*, 77, 437–470.

Kennedy, M. C. and O'Hagan, A. (2000), "Predicting the Output from a Complex Computer Code when Fast Approximations are available," *Biometrika*, 87.

Kennedy, M. C. and O'Hagan, A. (2001), "Bayesian Calibration of Computer Models," *J. R. Stat. Soc.. Ser. B*, 63, 425–464.

Kraus, E. B. and Turner, J. (1967), "A one dimensional model of the seasonal thermocline II. The general theory and its consequences," *Tellus*, 19, 98106.

Le Gratiet, L. (2014) "Bayesian analysis of hierarchical multifidelity codes", *SIAM J. Uncertainty Quantification* 1, 244-269.

Lee, L. A., Carslaw, K. S., Pringle, K. J., Mann, G. W., Spracklen, D. V., (2011) "Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters", *Atmospheric Chemistry and Physics*, 11, pp.12253-12273. doi: 10.5194/acp-11-12253-2011.

Loeppky, J. L., Sacks, J., Welch, W. J. (2009), "Choosing the sample size of a computer experiment: a practical guide", *Technometrics*, 51(4), 366–376.

Martin, G. M., Milton, S. F., Senior, C. A., Brooks, M. E., Ineson, S., Reichler, T., Kim, J. (2010), "Analysis and reduction of systematic errors through a seamless approach to modelling weather and climate", *Journal of Climate*, 23, 5933–5957.

Martin, G. M. and Coauthors (2011), "The HadGEM2 family of Met Office Unified Model Climate configurations", *Geosci. Model Dev.*, 4, 723–757.

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., Tomassini, L. (2012), "Tuning the climate of a global model", *Journal of advances in modeling Earth systems*, 4, M00A01, doi:10.1029/2012MS000154.

McNeall, D. J., Challenor, P. G., Gattiker, J. R., Stone, E. J. (2013). "The potential of an observational data set for calibration of a computationally expensive computer model", *Geosci. Model Dev.* 6 1715–1728.

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., Stouffer, R. J., and Taylor, K. E. (2007), "The WCRP CMIP3 multi-model dataset: a new era in climate change research," *Bull. Am. Meteorol. Soc.*, 88, 1383–1394.

Megann, A. P. et al. (2010) "The Sensitivity of a Coupled Climate Model to Its Ocean Component", *Journal of Climate*, 23, 5126–5150, doi: 10.1175/2010JCLI3394.1.

Meijers, A. J. S., Shuckburgh, E., Bruneau, N., Sallee, J. B., Bracegirdle, T. J., Wang, Z. (2012) "Representation of the Atarctic Circumpolar Current in the CMIP5 climate models and future changes under warming scenarios". *Journal of Geophysical Research*, 117, C12008, doi:10.1029/2012JC008412.

Meinen, C. S. and Garzoli, S. L. (2014) "Attribution of Deep Western Boundary Current variability at 26.5 N". *Deep Sea Research I*, 90, 81–90, http://dx.doi.org/10.1016/j.dsr.2014.04.016.

Murphy, J. M., Sexton, D. M. H., Jenkins, G. J., Booth, B. B. B., Brown, C. C., Clark, R. T., Collins, M., Harris, G. R., Kendon, E. J., Betts, R. A., Brown, S. J., Humphrey, K. A., McCarthy, M. P., McDonald, R. E., Stephens, A., Wallace, C., Warren, R., Wilby, R., Wood, R. (2009), "UK Climate Projections Science Report: Climate change projections." *Met Office Hadley Centre*, Exeter, UK. `http://ukclimateprojections.defra.gov.uk/images/stories/projections_pdfs/UKCP09_Projections_V2.pdf`

Pope, V.D. and M.L. Gallani and P.R. Rowntree and R.A. Stratton (2000), "The impact of new physical parameterizations in the Hadley Centre climate model: HadAM3.", *Clim. Dyn.*, 16, 123–146.

Pukelsheim, F. (1994), "The three sigma rule", *Am. Stat.*, 48, 88–91.

Randall, D. A. and Coauthors, (2007), "Climate models and their evaluation". *Climate Change 2007: The Physical Science Basis*, S. Solomon et al. Eds., Cambridge University Press, 589–662.

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

Rougier, J. C. (2008), "Efficient emulators for multivariate deterministic functions," *Journal of Computational and Graphical Statistics*, 17, 827 – 843.

Rougier, J. C., Sexton, D. M. H., Murphy, J. M., and Stainforth, D. (2009), "Emulating the sensitivity of the HadSM3 climate model using ensembles from different but related experiments," *J. Clim.*, 22, 3540–3557.

Rougier, J. C. (2013), "?Intractable and unsolved?: some thoughts on statistical data assimilation with uncertain static parameters" *Phil. Trans. R. Soc. A*, 371, 20120297. (doi:10.1098/rsta.2012.0297) .

Rowlands, D. J., Frame, D. J., Ackerley, D., Aina, T., Booth, B. B. B., Christensen, C., Collins, M., Faull, N., Forest, C. E., Grandey, B. S., Gryspeerdt, E., Highwood, E. J., Ingram, W., J., Knight, S., Lopez, A., Massey, N., McNamara, F., Meinshausen, N., Piani, C., Rosier, S., M., Sanderson, B., J., Smith, L. A., Stone, D. A., Thurston, M., Yamazaki, K., Yamazaki, Y., H., Allen, M. R. (2012), "Broad range of 2050 warming from an observationally constrained large climate model ensemble", *Nat. Geosci.*, published online, doi:10.1038/NGEO1430.

Russell, J. L., Stouffer, R. J., Dixon, K. W. (206) "Intercomparisons of the Southern Ocean circulations in IPCC coupled model control simulations", *Journal of Climate*, 19, 4560–4575.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), "Design and Analysis of Computer Experiments," *Stat. Sci.*, 4, 409–435.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The design and analysis of computer experiments*, Springer-Verlag New York.

Schmittner, A., Urban, N. M., Shakun, J. D., Mahowald, N. M., Clark, P. U., Bartlein, P. J., Mix, A. C., Rosell-Mele, A. (2011). "Climate sensitivity estimated from temperature reconstructions of the last glacial maximum", *Science*, 334, 1385.

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

Sexton, D. M. H., J. M. Murphy, and M. Collins (2011) "Multivariate probabilistic projections using imperfect climate models part 1: outline of methodology", *Clim. Dyn.*, doi:10.1007/s00382-011-1208-9.

Shaffrey, L. et al. (2009) "UK-HiGEM: The New UK High Resolution Global Environment Model. Model description and basic evaluation", *Journal of Climate*, 22 (8), 1861–1896, doi:10.1175/2008JCLI2508.1.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L. (eds.) (2007), *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 2007*, Cambridge University Press.

Uppala et al. (2005), "The ERA-40 re-analysis", *Q. J. R. Meteorol. Soc.*, 131, 2961–3012.

Vernon, I., Goldstein, M., and Bower, R. G. (2010), "Galaxy Formation: a Bayesian Uncertainty Analysis," *Bayesian Anal.* 5(4), 619–846, with Discussion.
Watanabe, M., Suzuki, T., O'Ishi, R., Komuro, Y., Watanabe, S., Emori, S., Takemura, T., Chikira, M., Ogura, T., Sekiguchi, M., Takata, K., Yamazaki, D., Yokohata, T., Nozawa, T., Hasumi, H., Tatebe, H., Kimoto, M. (2010), "Improved climate simulation by MIROC5: Mean states, variability and climate sensitivity", *Journal of Climate*, 23, 6312-6335.
Williamson, D. (2010), "Policy making using computer simulators for complex physical systems; Bayesian decision support for the development of adaptive strategies," Ph.D. thesis, Durham University.
Williamson, D., Goldstein, M. and Blaker, A. (2012), "Fast Linked Analyses for Scenario based Hierarchies," *J. R. Stat. Soc. Ser. C*, 61(5), 665–692.
Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P. Jackson, L., Yamazaki, K., (2012b), "History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble", *Climate Dynamics*, To Appear.
Williamson, D., Blaker, A. T. (2014) "Evolving Bayesian emulators for structurally chaotic time series with application to large climate models", *SIAM/ASA J. Uncertainty Quantification*, 2(1) 1-28.
Williamson, D., Goldstein, M. (2013) "On the use of evolving computer models in Bayesian decision support", *Journal of Statistical Planning and Inference*, In Submission.
Williamson, D., Vernon, I. R. (2013) "Implausibility driven Evolutionary Monte Carlo for efficient generation of uniform and optimal designs for multi-wave computer experiments", *Journal of the American Statistical Association*, In Submission.
Xie, P., and Arkin, P. A. (1997), "Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs.", *Bull. Amer. Meteor. Soc.*, 78, 2539–2558.
Yamazaki, K., Rowlands, D. J., Aina, T., Blaker, A., Bowery, A., Massey, N., Miller, J., Rye, C., Tett, S. F. B., Williamson, D., Yamazaki, Y. H., Allen, M. R. (2012), "Obtaining diverse behaviours in a climate model without the use of flux adjustments", *Journal of Geophysical Research - Atmospheres*, Accepted.

A Building emulators for history matching

What follows is a brief description of the methods we used to construct emulators for the constraints described in this paper. An emulator for element i of $f(x)$ might typically be fitted as

$$f_i(x) = \sum_j \beta_{ij} g_j(x) + \epsilon_i(x) \tag{2}$$

where $g(x)$ is a vector of specified functions of x, β is a matrix of coefficients, and $\epsilon(x)$ is a stochastic process with a specified covariance function. As discussed in section 2 there are many ways to build emulators and the way that is chosen will depend on the size of the PPE available, the type of constraint we wish to emulate and the relationships between the data and the parameters that we find. In this study we had access to large ensembles, and each of our constraints was a univariate quantity and so required less sophisticated modelling than a spatial field or time series might. Hence we fit the emulator mean functions, $\beta g(x)$ in equation (2) using a stepwise regression procedure described below.

The functions we consider adding to $g(x)$ were linear, quadratic and cubic terms in each of the parameters with up to third order interactions between all parameters considered. Switch parameters were treated as factors (variables with a small number of distinct possible "levels") and interactions between factors and all continuous parameters were permitted. For a list of the parameters varied in the ensemble see appendix B.

Our fitting procedure begins with a "forward selection", where we permit each allowed term to be added to $g(x)$ in its lowest available form. For example, if the linear term for x_1 is not yet in $g(x)$, x_1 is available for selection but x_1^2 is not. If x_1 is already in $g(x)$ then all first order interactions with the other linear parameters in $g(x)$ are included and then x_1^2 is available for selection. So, suppose $g(x)$ is $(1, x_2)$, then the selection of x_1 implies that $g(x)$ will become $(1, x_1, x_2, x_1 * x_2)$. If x_1 is selected, at the next iteration we may select any of the other parameters but we may also include quadratic terms x_2^2 andx_1^2. We add the interactions in this way, and do similar for third order interactions when quadratic terms have been included, so that the resulting emulator will be robust to changes of scale (see Draper and Smith, 1998, for discussion). The term that is added to $g(x)$ at each iteration is the term of those available that reduces the residual sum of squares the most after fitting by ordinary least squares.

When it becomes clear that adding more terms is not improving the predictive power of the emulator (a judgement made by the analyst based on looking at the proportion of variability explained by the emulator and at plots of the residuals from the fit) we begin a backwards elimination algorithm. This removes terms from $g(x)$, strictly one at a time, with the least contribution to the sum of squares explained by the fit without compromising the quality of the fit. Lower order terms are not permitted to be removed from $g(x)$ whilst higher order terms remain. We stop when removing the next term chosen by the algorithm leads to a poorer statistical model. For more details on stepwise methods such as these see Draper and Smith (1998).

We allow $\epsilon(x)$ in equation (2) to be mean zero error with variance specified by the residual variability from the fits and no correlation between $\epsilon(x)$ and $\epsilon(x')$ for $x \neq x'$. Though this lack of correlation might not be appropriate if we had smaller ensembles or, perhaps, if we had completed a number of waves of history matching and were focussing on a densely sampled subset of parameter space, it is computationally efficient and a reasonable enough approximation to the data here to be adopted for pragmatism. Including a more complex correlation would reduce our emulator uncertainty and likely lead to more parameter space being ruled out, though at a computational cost. Note that, with zero correlation between any points, the emulator will not interpolate the ensemble members and will have non-zero variance at each of them. Though the model is deterministic (in the sense that running it twice for the same values of the model inputs returns the same answer), it also displays sensitive dependence to initial conditions, hence the fitted variance at the design points represents the model's internal variability. This form of emulator effectively assumes that internal variability is constant throughout parameter space, and hence can be estimated from the ensemble as part of the fitting procedure.

Following the fitting of each emulator we validate its quality using 10% of the ensemble that was chosen randomly and reserved from the training data prior to the fit. This procedure involves checking that the emulator accurately predicts each of the unseen ensemble members to within the accuracy specified by emulator uncertainty. If the emulators pass this diagnostic check, we then use them in our history matching.

Table A.1 A table indicating which terms are in $g(x)$ for our emulator of ACC in equation (2). The column and row labels refer to the parameter names as given by the matching labels in table B.1. The upper triangle labels which interaction pairs are present. The diagonal indicates the order of the highest order term in that variable. The lower triangle indicates which three way interactions are included.

	f	b	d	e	c	a	u	j	p	g	v	w	t	s	l	k
f	2	1	0	1	1	1	1	0	0	0	0	1	0	1	1	1
b		1	1	0	0	1	0	0	1	1	0	0	0	0	0	0
d			1	1	0	0	0	0	1	1	1	0	0	0	0	0
e	**f**			1	1	0	0	0	0	1	0	0	1	0	0	0
c					1	0	1	0	0	0	0	0	0	0	0	0
a						1	0	0	0	0	0	0	0	1	0	0
u	**f**				**u**		2	0	0	1	0	0	1	0	0	0
j								1	1	0	0	1	0	1	0	1
p								**u**	1	1	0	0	0	0	0	0
g										1	0	0	0	0	0	0
v											1	1	0	1	0	0
w	**f**											1	0	1	0	0
t													1	0	0	0
s	**a**													1	0	0
l															1	0
k																1

We give details of our emulator for the ACC strength in HadCM3 to illustrate the complexity of the mean function and the performance of the predictions. The terms selected in $g(x)$ are displayed in table A.1. Each header corresponds to the label given to each of the parameters in

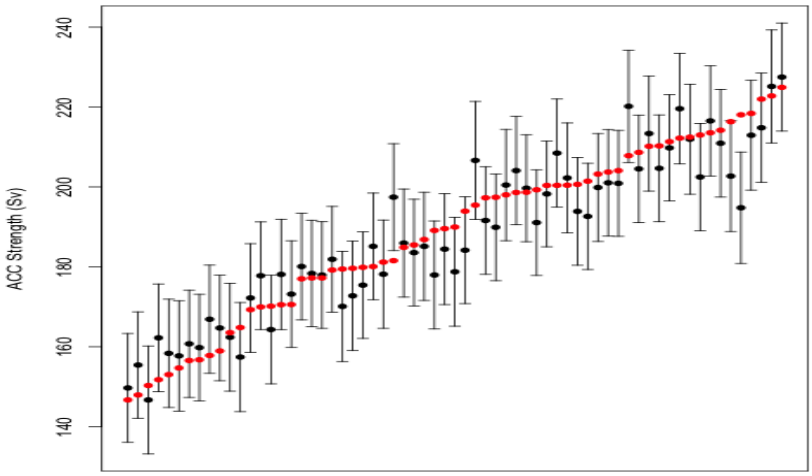

Fig. 13 Predicted ACC strength (black points) with error bars showing approximately 2 standard deviations of the emulator uncertainty for each of the withheld PPE members (red points).

table B.1. Numbers on the diagonal of the table refer to the order of the parameter included in the emulator. For example, the number 2 implies that both quadratic and linear terms in that parameter were included in $g(x)$. Numbers on the upper triangle refer to the inclusion (1) or not (0) of interactions between the two relevant parameters in $g(x)$. So, reading from the first row of the table, the term (entcoef $*$ ct) is included in $g(x)$, but the term (entcoef $*$ rhcrit) is not. Variables in bold on the lower triangle indicate the inclusion of the given third order interaction. For example, the table indicates that the term (AH1_SI^2*SWland) is in $g(x)$. In addition to the terms in the table, the factor r_layers and a linear term in parameter charnock are included, as is 1 so that an intercept is fitted.

Figure 13 shows a validation plot for the ACC emulator. For 65 PPE members, chosen randomly, that were reserved from the emulator at the fitting stage, the data are sorted by ACC strength and plotted in red. We overlay the emulator predictions (black points) and the uncertainty on those predictions (error bars). The uncertainty represents approximately 2 standard deviations for each prediction. We can see that the predictions are generally good with most unseen PPE members laying within the uncertainty on the prediction. In fact, our uncertainty specification may be too conservative, in that we have allowed for more uncertainty in the predictions than is required. If this is the case, that would lead to less space ruled out by history matching, not more, and it is our preference to remain conservative when ruling out regions of parameter space.

B Tables

Tables B.1, B.2 and B.3 give descriptions and ranges for the parameters and the settings of switches used in our ensemble. Some parameters have relationships with other model parameters that were given to us by the Met Office so that a change in one leads to a derivable value for the other. CWland also determines CWsea, the cloud droplet to rain threshold over sea (kg/m^3), MinSIA also determines dtice (the ocean ice diffusion coefficient) and k_gwd also determines kay_lee_gwave (the trapped lee wave constant for surface gravity waves m$^{3/2}$).

Table B.1 Parameter descriptions and model section. CWland determines CWsea, MinSIA determines dtice and k_gwd determines kay_lee_gwave. The label column represents the labels that represent the parameters in table A.1.

Parameter	Description	Section	Label
vf1	Ice fall speed (m/s)	Cloud	a
ct	Cloud droplet to rain conversion rate (/s)	Cloud	b
CWland	Cloud droplet to rain threshold over land (kg/m^3)	Cloud	c
CWsea	Cloud droplet to rain threshold over sea (kg/m^3)	Cloud	
RHCrit	Relative humidity threshold for cloud formation	Cloud	d
eacfbl	Boundary layer cloud fraction at saturation	Cloud	e
entcoef	Convective cloud entrainment rate coefficient	Convection	f
MinSIA	Albedo at ice melting point	Sea Ice	g
dtice	Ocean ice diffusion coefficient	Sea Ice	
Icesize	Ice particle size (μm)	Radiation	h
k_gwd	Surface gravity wavelength (m)	Dynamics	i
lay_lee_gwave	Trapped lee wave constant for surface gravity waves ($m^{3/2}$)	Dynamics	
start_level_gwdrag	First level for gravity wave drag	Dynamics	
dyndiff	Diffusion efolding time (hours)	Dynamics	j
dyndel	Order of diffusion operator	Dynamics	k
asym_lambda	Asymptotic neutral mixing length parameter	Boundary	l
charnock	Charnock constant	Boundary	o
cnv_rl	Free convective roughness length over sea (m)	Boundary	
flux_g0	Boundary layer flux profile parameter	Boundary	p
r_layers	No. of soil levels for evaporation	Land Surface	q
L	SO_2 wet scavenging rate (/s)	Sulphur Cycle	
volsca	Scaling for volcanic SO_2 emissions	Sulphur Cycle	
anthsca	Scaling for anthropogenic SO_2 emissions	Sulphur Cycle	
so2_high_level	Model level for SO_2 emissions	Sulphur Cycle	
vb	Background vertical viscosity (m^2/s)	Ocean	r
kb	Background vertical diffusivity (m^2/s)	Ocean	s
dkb/dz	Background vertical diffusivity gradient (m/s)	Ocean	t
AH1_SI	Isopycnal diffusivity (m^2/s)	Ocean	u
lambda	Wind mixing energy scaling factor	Ocean	v
delta_si	Wind mixing energy decay depth (m)	Ocean	w

C Another NROY ACC model

In the main text we present the behaviour of one of the NROY ACC models, arguing that correcting the ACC strength seems to improve the ocean circulation. Though we do not reproduce all of the figures from the main text, in order to save space, we show the BSF of another of these models in figure 14 to indicate that the chosen model was not a "one-off". This model has a slightly more physical looking sub polar gyre, at the expense of a more diffuse gulf stream. The cold bias in the North Atlantic (not shown) was also greater in this model.

D Anomalies from two additional precipitation climatologies

In the main text we present precipitation anomalies from the ERA40 climatology. However, we caution the reader against interpreting improvements seen in the improved ACC run as robust. We present two alternative precipitation climatologies, the CPC Merged Analysis of Precipitation (CMAP) (Xie and Arkin , 1997) and NCEP/NCAR reanalysis (Kalnay et al. , 1996) in figure 15. These plots indicate that the standard model does have a tendency towards higher than observed precipitation along the ITCZ and over the maritime continent, and the improved ACC run we present tends to exhibit lower than observed precipitation over the western equatorial Pacific. The improved ACC run does perform better than the standard run

Table B.2 Ranges for each of the continuous parameters varied in our ensemble. $*$ indicates that we don't change the standard range in the exploratory sub ensemble. We don't give values for dependent parameters CWsea, dtice and kay_lee_gwave as these are calculated from CWland, MinSIA and k_gwd respectively via a one to one mapping.

Parameter	Section	Standard lower	Standard higher	New lower	New Higher
vf1	Cloud	0.5	2	0.15	2.35
ct	Cloud	5×10^{-05}	4×10^{-04}	$*$	5.625×10^{-04}
CWland	Cloud	1×10^{-04}	2×10^{-03}	$*$	$*$
RHCrit	Cloud	0.6	0.9	$*$	$*$
eacfbl	Cloud	0.5	0.8	$*$	$*$
entcoef	Convection	0.6	9	$*$	$*$
MinSIA	Sea Ice	0.5	0.65	$*$	$*$
Icesize	Radiation	2.5×10^{-05}	4×10^{-05}	2×10^{-05}	8×10^{-05}
k_gwd	Dynamics	$1 \times 10^{+04}$	$2 \times 10^{+04}$	$*$	$*$
dyndiff	Dynamics	6	24	$*$	$*$
asym_lambda	Boundary	0.05	0.5	0.01	0.61
charnock	Boundary	0.012	0.02	0.012	0.024
cnv_rl	Boundary	2×10^{-04}	5×10^{-03}	2×10^{-04}	6.2×10^{-03}
flux_g0	Boundary	5	20	2.5	22.5
L	Sulphur Cycle	0.33	0.33	$*$	$*$
volsca	Sulphur Cycle	1	3	0.5	3.5
anthsca	Sulphur Cycle	0.5	1.5	0.25	1.75
vb	Ocean	5×10^{-06}	8×10^{-05}	1×10^{-06}	1.1×10^{-04}
kb	Ocean	5×10^{-06}	2×10^{-05}	1×10^{-06}	3.1×10^{-05}
AH1_SI	Ocean	200	2000	100	2500
dkb/dz	Ocean	7×10^{-09}	9.8×10^{-08}	$*$	$*$

Table B.3 Switch parameters and their settings in our ensemble. $*$ indicates that there are only 2 settings of a switch.

Parameter	Section	Setting 1	Setting 2	Setting 3
so2_high_level	Sulphur Cycle	3	5	$*$
start_level_gwdrag	Dynamics	3	4	5
r_layers	Land Surface	[2,1]	[3,2]	[4,3]
dyndel	Dynamics	4	6	$*$
lamda/delta_si	Mixed Layer	[0.3,100]	[0.5,50]	[0.7,100]

everywhere outside the tropics, where the standard run has a tendency towards higher than observed precipitation.

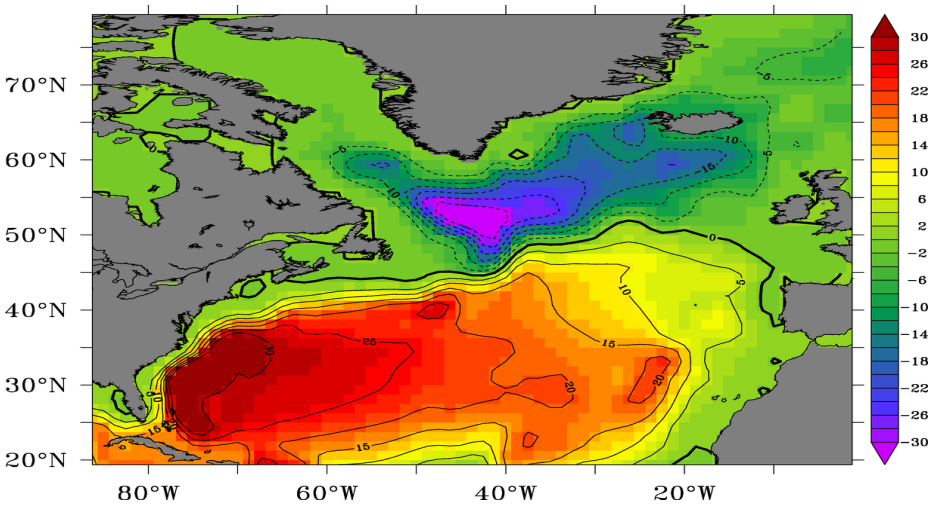

Fig. 14 The barotropic streamfunction (BSF, Sv) for a different ensemble member with realistic ACC strength (another of the blue dots from figure 1).

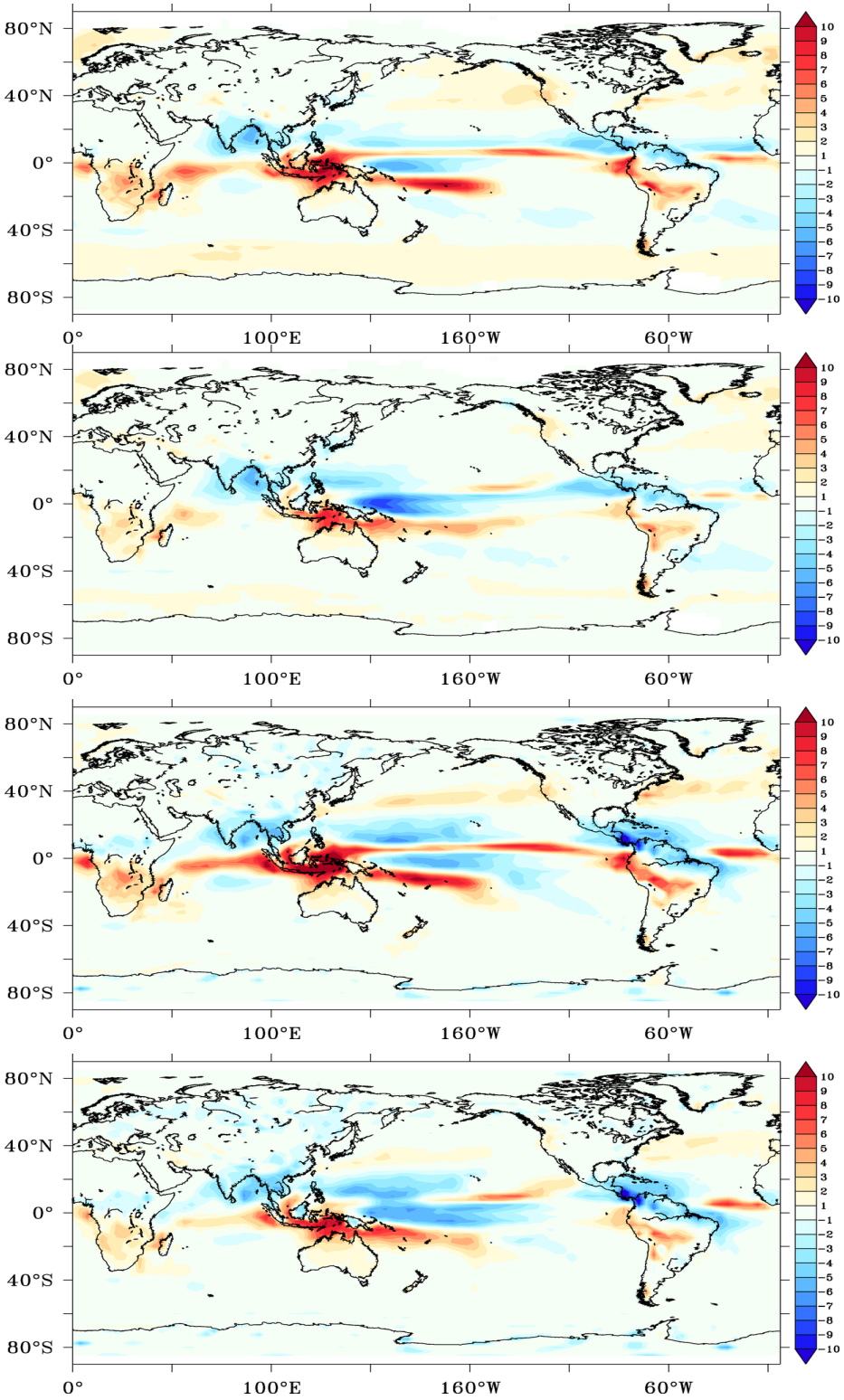

Fig. 15 The top two panels show differences from the CPC Merged Analysis of Precipitation (CMAP) climatology (standard above improved ACC model). The bottom two panels show differences from the NCEP reanalysis precipitation climatology (standard above improved ACC model).