

Finding Ocean States That Are Consistent with Observations from a Perturbed Physics Parameter Ensemble

SARAH SPARROW,^a RICHARD J. MILLAR,^b KUNIKO YAMAZAKI,^c NEIL MASSEY,^d ADAM C. POVEY,^e
ANDY BOWERY,^a ROY G. GRAINGER,^e DAVID WALLOM,^a AND MYLES ALLEN^{b,f}

^a *Oxford e-Research Centre, University of Oxford, Oxford, United Kingdom*

^b *Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, United Kingdom*

^c *Met Office Hadley Centre, Exeter, United Kingdom*

^d *National Centre for Atmospheric Science, Centre for Environmental Data Analysis, Science and Technology Facilities Council Rutherford Appleton Laboratory, Harwell Oxford, Didcot, United Kingdom*

^e *National Centre for Earth Observation, Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, United Kingdom*

^f *Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, United Kingdom*

(Manuscript received 31 July 2017, in final form 2 March 2018)

ABSTRACT

A very large ensemble is used to identify subgrid-scale parameter settings for the HadCM3 model that are capable of best simulating the ocean state over the recent past (1980–2010). A simple particle filtering technique based upon the agreement of basin mean sea surface temperature (SST) and upper 700-m ocean heat content with EN3 observations is applied to an existing perturbed physics ensemble with initial conditions perturbations. A single set of subgrid-scale parameter values was identified from the wide range of initial parameter sets that gave the best agreement with ocean observations for the period studied. The parameter set, different from the standard model parameters, has a transient climate response of 1.68 K. The selected parameter set shows an improved agreement with EN3 decadal-mean SST patterns and the Atlantic meridional overturning circulation (AMOC) at 26°N as measured by the Rapid Climate Change (RAPID) array. Particle filtering techniques as demonstrated here could have a useful role in improving the starting point for traditional model-tuning exercises in coupled climate models.

1. Introduction

Simulations forced by historical estimates of external climate drivers are common tools used to evaluate coupled general circulation models (GCMs) of the atmosphere and ocean (Flato et al. 2013; Jones et al. 2013). Despite a continual improvement of the fidelity of coupled atmosphere–ocean GCMs over the past decades, significant biases in the representation of the present-day observed ocean properties remain. Individual GCMs can have biases in their simulations of sea surface temperatures of over 2°C or more in higher latitudes (Flato et al. 2013). The future evolution of important large-scale oceanic circulations, such as the Atlantic meridional overturning circulation (AMOC), is uncertain, at least in part because of the substantial

spread in mean-state simulations apparent in GCMs (Gregory et al. 2005). Because of the ocean's important role in determining the time scales of the climate response to radiative forcing, as well as being an important sink of anthropogenic CO₂ emissions, improving the representation of ocean mean states could have substantial value in helping to constrain estimates of future climate changes.

Representations of the climate mean state and/or transient climate changes over the historical period can be improved either by alterations to the parameterization schemes that represent unresolved climate processes in terms of resolved variables on the finite grid of the GCM, or by changes to the numerical value of specific coefficients within these schemes. The approach of sampling large numbers of alternative realizations of subgrid-scale parameters to assess their impact on the climate representation is commonly known as a perturbed physics [or perturbed parameter ensemble

Corresponding author: Sarah Sparrow, sarah.sparrow@oerc.ox.ac.uk

(PPE)] approach. Many previous studies (Williamson et al. 2013; Yamazaki et al. 2013; Shiogama et al. 2012; Sanderson 2011; Collins et al. 2007; Stainforth et al. 2005; Murphy et al. 2004) have used PPEs to explore the effect of changes in atmospheric and ocean parameters on the quality of the model's simulated climate, largely within the Met Office family of models. PPEs can be used in a variety of ways, ranging from establishing the sensitivity of the model response to changes in different parameters to identifying parts of parameter space that are incapable of producing a realistic present-day simulated climate. Williamson et al. (2013) investigate the latter of these two applications and use a large PPE in which multiple parameters were varied simultaneously in order to span a multidimensional parameter space. They applied a history matching technique, using statistical emulators with the simulation of a number of atmosphere climate variables as constraints, to rule out areas of parameter space that could not match observational metrics within their uncertainty ranges. They were able to rule out around half of parameter space by using this technique and suggest this technique can be used to remove complex and irrelevant behavior in unrealistic parts of parameter space.

In this paper, parameter sets are assessed using a PPE (Murphy et al. 2004; Stainforth et al. 2005) of the Hadley Centre Coupled Model, version 3 (HadCM3), GCM (Gordon et al. 2000), run without flux adjustments (Yamazaki et al. 2013), for their ability to accurately simulate properties of the oceanic climate mean state over recent decades. The PPE that was used to ensure reasonable TOA balance in Yamazaki et al. (2013) forms the starting point for this study, and is conceptually similar to that of Shiogama et al. (2012) and uses the spinup control experiment of Johns et al. (2011). Thirty-three model parameters are perturbed simultaneously such that each parameter set is well separated in parameter space and initial results have been filtered to ensure each parameter set produces a reasonable top-of-atmosphere radiative balance. Very large ensembles are used to sample both initial condition uncertainty and uncertainty in the many different HadCM3 subgrid-scale parameters (in both the atmosphere and ocean). For each ensemble member, we assess the compatibility of the simulated oceanic heat content and sea surface temperature (SST) with observational datasets on a bi-annual basis, with the aim of identifying new parameter combinations that are capable of best simulating the entirety of the recent observational record.

Ocean parameter perturbation studies have been performed previously with versions of the Hadley Centre GCM (Forest et al. 2006; Collins et al. 2007), and in general have been found to have limited impact on

global quantities such as the transient climate response (TCR) and equilibrium climate sensitivity (ECS). However, it remains unclear how important oceanic/atmospheric parameter uncertainty is to the ability of a GCM to achieve consistency with observed changes in the oceanic state on nonglobal scales. Understanding these relationships will be key to understanding the implications (if any) of the recent warming hiatus to climate sensitivity, as recent work has indicated that the warming hiatus may be linked with changes in the pattern of oceanic heat uptake relative to earlier periods (Nieves et al. 2015; Liu et al. 2016). This study looks to find potential ocean parameter sets that yield a better agreement of the absolute values of sea surface temperature and upper 700-m ocean heat content in the recent past by principally identifying parameter adjustments that lead to an improvement of the mean state of the model.

This paper is structured as follows. Section 2 describes the model, and section 3 describes the sulfate emissions used as forcing over the historical integration. Section 4 details the methodology used to refine parameter sets based upon their fit to observed ocean metrics and details the resulting TCR values of the selected parameter sets. Section 5 compares the selected parameter sets to the standard HadCM3 configuration and is followed by a general discussion of results.

2. The model

Because of the large number of possible perturbations to coupled GCM parameters and the potentially complex interactions between them, very large ensembles are required to identify suitable parameter sets that best represent ocean mean states. This is achieved via the use of the distributed computing project climateprediction.net (CPDN), where large numbers of public volunteers donate resources by agreeing to run simulations during idle time on their home computers. The model used in this study, the HadCM3 (Gordon et al. 2000), has been used previously by CPDN in studies such as Yamazaki et al. (2013), from which the starting conditions for this experiment are taken. The parameter perturbations used here are equivalent to those in Yamazaki et al. (2013), and parameter definitions are included in Tables 1–3 for easy reference. Within this ensemble, atmospheric and oceanic parameters are varied simultaneously, with the initial ranges chosen by use of a Latin hypercube so as to fill the available parameter space. Hence, the climate representation of each of the initial parameter sets may be substantially different from each other because of the nature of the design of the perturbed parameter sampling.

TABLE 1. Atmospheric physics and dynamics parameter definitions.

Parameter name	Description
VF1	Ice particle fall speed
CT	Cloud liquid water–precipitation conversion rate
CW_LAND	Threshold cloud liquid water content over land
CW_SEA	Threshold cloud liquid water content over sea
RHCRT	Critical relative humidity for cloud formation
EACF	Empirically adjusted cloud fraction
ENTCOEF	Atmospheric entrainment rate coefficient
ALPHAM	Albedo at sea ice melting point
DTICE	Temperature range over which ice albedo varies
ICE_SIZE	Ice particle size
START_LEVEL_ GWDRAG	Lowest model level for gravity wave drag
KAY_GWAVE	Surface gravity wave drag wavelength
KAY_LEE_GWAVE	Surface gravity wave trapped lee-wave constant
ASYM_LAMBDA	Vertical distance before air parcels mix with surround
CHARNOCK	Constant in Charnock formula (momentum transport over sea)
ZOFSEA	Sea surface roughness (heat, moisture transport)
G0	Used in calculation of boundary layer stability function
R_LAYERS	No. of soil levels from which water can be extracted
DIFF_COEF	Horizontal diffusion coefficient
DIFF_COEF_Q	Horizontal diffusion coefficient for water vapor
DIFF_EXP	Order of horizontal diffusion
DIFF_EXP_Q	Order of horizontal diffusion for water vapor

The atmospheric component of the model is run at $3.75^\circ \times 2.5^\circ$ longitude/latitude resolution with 19 vertical levels (N48L19). The vertical levels use hybrid vertical coordinates. This version of the model includes an interactive sulfur cycle as described in Jones et al. (2001). Sulfate aerosol concentrations are simulated from prescribed emission fields of SO₂, natural dimethyl sulfide (DMS), and tropospheric sulfur arising from quasi-regular volcanic eruptions. In this experiment, the natural fields are left unaltered and historical anthropogenic emissions are extended with RCP4.5. SO₂ emissions data are used with a 2% yr⁻¹ increase applied to 1995 values in East Asia (see section 3a for details). A further modification has been made to the model so that volcanic forcings [taken from Sato (2013)] are applied in 73 (as opposed to 4) latitude bands. All other forcing fields are consistent with the CMIP5 RCP4.5 scenarios.

TABLE 2. Atmospheric chemistry parameter definitions.

Parameter name	Description
L0	Sulfate mass scavenging parameter L0
L1	Sulfate mass scavenging parameter L1
SO2_HIGH_LEVEL	Model level for high-level sulfur dioxide emissions
VOLSCA	Scaling factor for volcanic emissions
ANTHSCA	Scaling factor for anthropogenic sulfate aerosols

The ocean model component [as in Cox (1984)] includes a thermodynamic sea ice model and has a horizontal resolution of 1.25° in both directions and 20 vertical levels, spaced nonuniformly to give greater resolution near the surface.

3. Methodology

a. Sulfate emissions

As one of the aims of this study is to try to find new parameter combinations that are capable of simulating the recent ocean state in a model that is not flux adjusted, it is important that the external climate forcings used to drive the model are represented as accurately as possible.

A large source of uncertainty in net climate forcing over the recent past relates to uncertainties in both sulfate aerosol emissions in East Asia (Hong et al. 2017) and their net radiative effect on atmospheric energy flows (Myhre et al. 2013). As emissions of sulfate aerosols could induce a potentially very large negative (cooling) radiative forcing, which would offset a substantial fraction of the positive (warming) radiative forcing of greenhouse gases, underestimates or overestimates of East Asian sulfate emissions provided as climate model drivers could have a substantial effect on the simulation of regional climate mean states over the past several decades.

In this study we use the Advanced Along Track Scanning Radiometer (AATSR) and the second Along Track Scanning Radiometer (ATSR2) satellite retrievals of aerosol optical thickness (AOT) using the Optimal Retrieval of Aerosol and Cloud (ORAC; Thomas et al. 2009) to estimate the observed trend in sulfate aerosol as a percentage trend per year over the western North Pacific Ocean (WNPO). That range of trends (in percent per year) is then applied to East Asian emissions that are used to drive HadCM3 and the impact on global surface temperature assessed. AATSR and ATSR2 retrievals of AOT over the western North Pacific Ocean (defined as 20°–60°N, 100°–180°E) are sectioned into 10° × 10° bins. Trends from 1995 to present

TABLE 3. Ocean dynamics parameter definitions.

Parameter name	Description
ISOPYC	Surface tracer isopycnal diffusion
VERTVISC	Background momentum vertical diffusion
VDIFFSURF	Background vertical diffusion of tracer at surface
VDIFFDEPTH	Increase of tracer background diffusion with depth
MLLAM	Wind-mixing energy scaling factor
MLDEL	Wind-mixing energy rate of decay with depth

are calculated and result in the distribution shown in Fig. 1. Fitting a Gaussian (Fig. 1, gray line) to this distribution and computing the standard deviation (Fig. 1, range denoted by dark blue shading) gives an estimate of the plausible range of values of the observed trend. Rounding these values up to the nearest integer value (for convenience) gives a range of a $\pm 4\%$ yr^{-1} trend in sulfate aerosol to be considered in this study (denoted by light blue shaded range in Fig. 1).

The SO_2 CMIP5 RCP4.5 emissions fields supplied to the model are then adjusted with each integer value between a $\pm 4\%$ yr^{-1} trend over the East Asian region (EAS). In this case, the EAS boundaries are taken to be the Giorgi and Francisco (2000) EAS region (20° – 50°N , 100° – 145°E). There is a slight discrepancy between the region used to measure the sulfate aerosol trends and the Giorgi and Francisco (2000) region where SO_2 emissions are altered. Sulfate emissions are mainly from land sources, whereas the sulfate aerosol trends are measured by satellite retrievals over the ocean. Satellite retrievals are only considered over the ocean and become noisier nearer the land, and so the WNPO region is larger than EAS in an attempt to minimize the noise. Most of the emissions occur over the land area, and this represents the area in which the emissions are most increased. Figure 2 shows the implications of mean SO_2 emissions in the EAS region (right panel) and globally (left panel). The solid and dotted yellow lines represent the unaltered CMIP5 RCP4.5 and RCP8.5 emissions scenarios, respectively. Each of the perturbed EAS emissions used in this experiment is denoted by the other colored lines. This highlights that the range of plausible emissions studied here (gray shading) is much larger than the difference between CMIP5 RCP scenarios. For each of these sulfate emission scenarios, an initial conditions ensemble with 256 members is run for HadCM3 with standard physics parameters.

For each of the nine different sulfate emissions perturbations and the unaltered CMIP5 RCP4.5 sulfate emissions, 256 model simulations are performed. The EAS and global-mean surface temperatures for each of

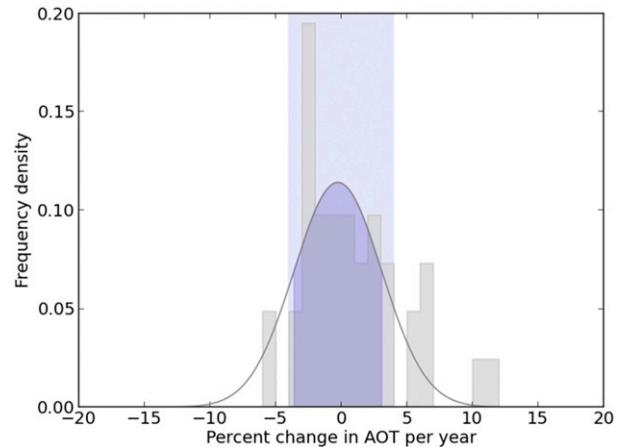


FIG. 1. Distribution of the percentage change in AOT from 1995 to present from AATSR and ATSR2 measurements over the North Pacific Ocean (gray shaded histogram). The gray curve shows a Gaussian fit with the dark blue shaded region corresponding to ± 1 std dev. The light blue shaded region represents the $\pm 4\%$ yr^{-1} variations used in this study.

the EAS sulfate adjustments and the unaltered CMIP5 RCP4.5 emissions scenario are shown in Fig. 3. Colored lines represent the ensemble mean for each of the nine different EAS sulfate perturbations along with the CMIP5 RCP4.5 sulfate emissions simulations. The shaded regions denote the standard deviation of ensemble members within each emissions set. It is clear that EAS sulfate emissions have little impact on either the global or regional surface temperature and are indistinguishable from both each other and the standard RCP emissions scenario. Johns et al. (2003) suggest that it is possible that the apparent insensitivity to SO_2 emissions may be linked to too-low aerosol loading arising from too-fast wet and dry deposition rates in this version of HadCM3. This lack of sensitivity to sulfate emissions within HadCM3 suggests that the precise choice of sulfate emissions will not substantially affect the model simulations, and hence the selection process to find new plausible parameter sets that are capable of reproducing recent oceanic observations.

b. Filtering process

Starting conditions for 1980, 1990, and 2000 taken from existing RAPID–ChAAOS (Change in the Atlantic Atmosphere–Ocean System) simulations, detailed in Yamazaki et al. (2013), are used to initialize the experiment for each decade of simulation. In Yamazaki et al. (2013) a larger ensemble was filtered to ensure that ensemble members give good top-of-atmosphere radiation balance in their simulated control climates. Forcings have since been updated over the recent period to include small volcanic eruptions, and a 2% yr^{-1} East Asian

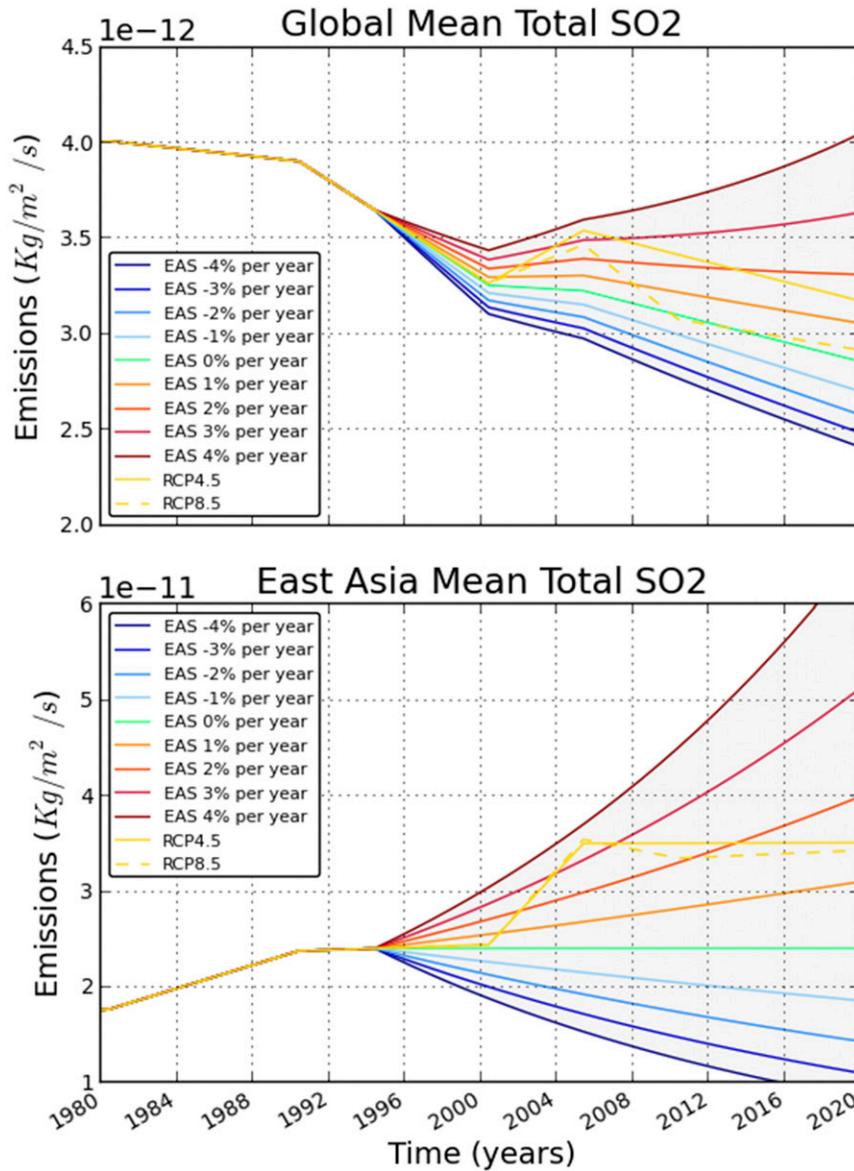


FIG. 2. (top) The global-mean and (bottom) EAS-mean total SO_2 emissions. The solid and dotted yellow lines show the CMIP5 RCP4.5 and RCP8.5 SO_2 forcing fields, respectively. The gray region denotes a $\pm 4\% \text{ yr}^{-1}$ trend starting in 1995 applied to the EAS region with the line denoting the trend to emissions applied in the simulations.

emission growth was chosen with RCP4.5 data used elsewhere. This is consistent with observational data trends published in Yoon et al. (2016; section 6). In this study, results are split into eight ocean basins, namely, the North Atlantic, South Atlantic, North Pacific, South Pacific, north Indian, south Indian, Southern Ocean, and Arctic Ocean. During the initial setup, ocean restart files with identical ocean parameter settings are swapped with alternative atmospheric restart files (in practice a very small number, since the parameter sets were initially designed to fill parameter space). The standard physics parameter sets

are included in the initial ensemble. After each 2-yr simulation period, for each ocean basin (except the Arctic) the Euclidean distance between the modeled SST and upper 700-m ocean heat content (OHC) and the EN3 v2a gridded objective analysis observations (Ingleby and Huddleston 2007) of these quantities are calculated for each ensemble member and ranked. This dataset includes bias correction for expendable bathythermograph (XBT) fall rates as detailed in Ingleby and Huddleston (2007). The basin mean SST and OHC were chosen to constrain the ensemble, as they give a broad

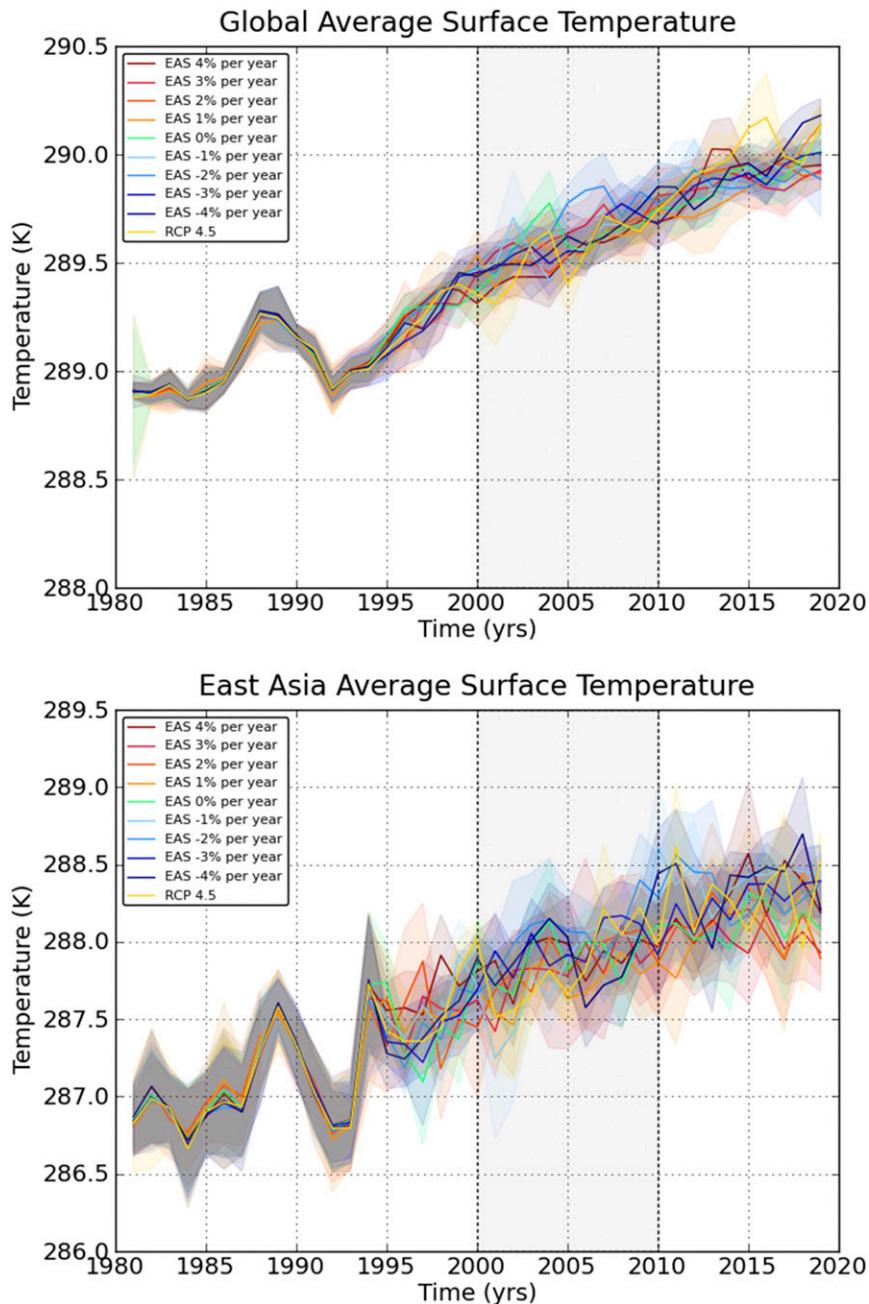


FIG. 3. (top) The global-mean and (bottom) and EAS-mean 1.5-m temp for each of the EAS SO_2 emissions adjustments starting in 1995. For each scenario the ± 1 std dev values for the 256 members is indicated by the light shaded areas. The unaltered RCP4.5 scenario is also included for comparison.

indication of the surface and upper-ocean states without matching finer features. They are also quantities where ocean datasets are more reliable for comparison, since deep ocean measurements are sparse. The basin ranks of SST and OHC are summed and then reranked to form a combined ranking over both constraining quantities for each model in each ocean basin. These ocean basin

ranks are then summed across all ocean basins and re-ranked again to find an ordering of models globally. The top 15% of these models are then selected for each year and only those that are in the top 15% across all three filtering periods (1980s, 1990s, and 2000s) are retained at each stage. Initial condition perturbations are then applied to these model states to maintain the size of the

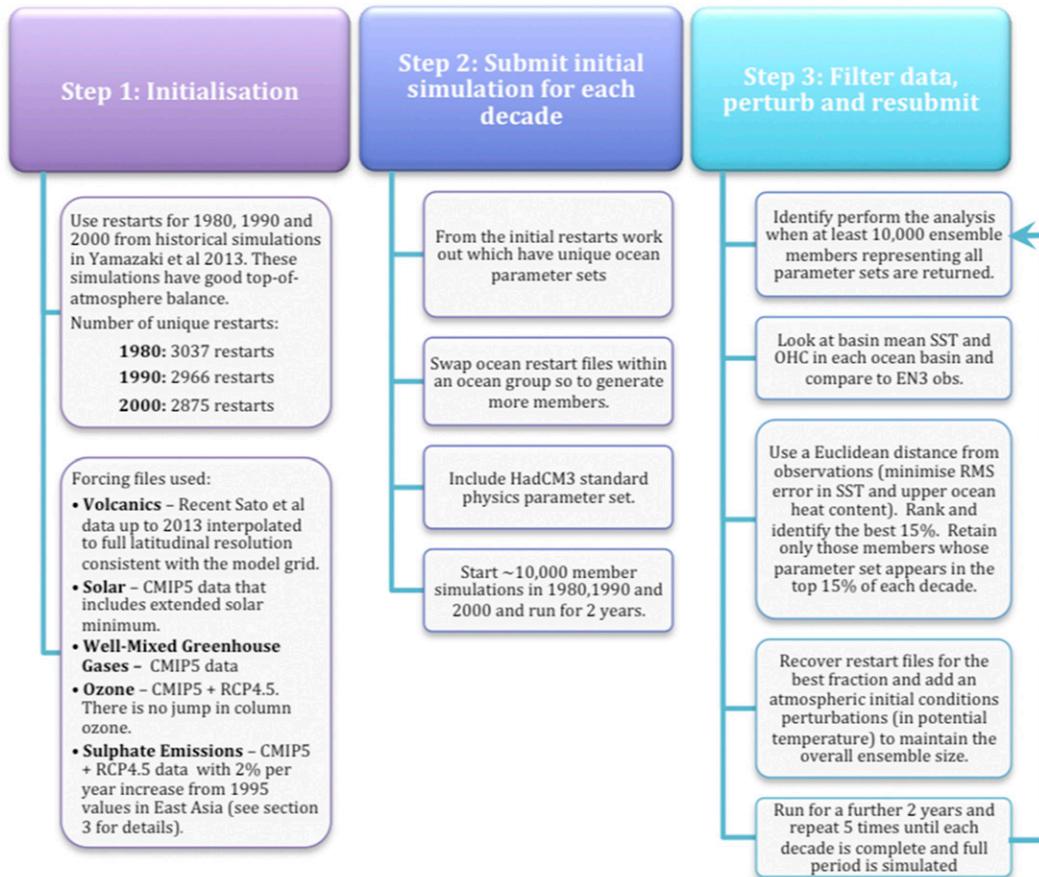


FIG. 4. Summary of the experiment configuration, design, and ensemble selection process.

ensemble, and they are run forward for a further two years. By using this method (summarized in Fig. 4), if initial conditions were dominating the variability in a given ocean basin, models from a large number of the available parameter sets will be retained. However, if variability in an ocean basin is primarily governed by parameter perturbations, then it will act as a strong filter on the model variants selected to continue. At no point in this procedure are any flux adjustments applied to ensure good agreement with observations. It is worth noting that since more initial conditions perturbations are added for each surviving parameter set at every 2-yr filtering stage, by the end of the 10-yr selection period the ensemble will be sampling the variability that the model is capable of producing relatively comprehensively. This in itself will not improve the model's variability, but will ensure that by the end of the selection process the variability is well sampled. Typically the difference between parameter sets is larger than the difference between initial conditions perturbations, so the selection will principally screen for these different parameter sets

based on their associated simulation of the mean state.

4. Results

a. Validation of methodology

Figures 5 and 6 show basin mean time series of SST and OHC for the ensemble mean (gray) and EN3 observations (yellow) for all ocean basins (except the Arctic). This shows how the ensemble develops through the filtering process and, by positive selection, becomes closer to observed values. There are basins that show a persistent bias compared to the observed values, such as North Pacific, north Indian Ocean, and Southern Ocean SST values, and North Atlantic and North Pacific OHC. As some basins display a temporal trend in bias, caution should be taken in using these results for understanding climate change on a regional scale. The mean of ensemble members selected to be in the top 15% across all decades and ocean basins (hereafter “the selected ensemble”) are shown in blue (annual means) to illustrate that there are no major discontinuities between one

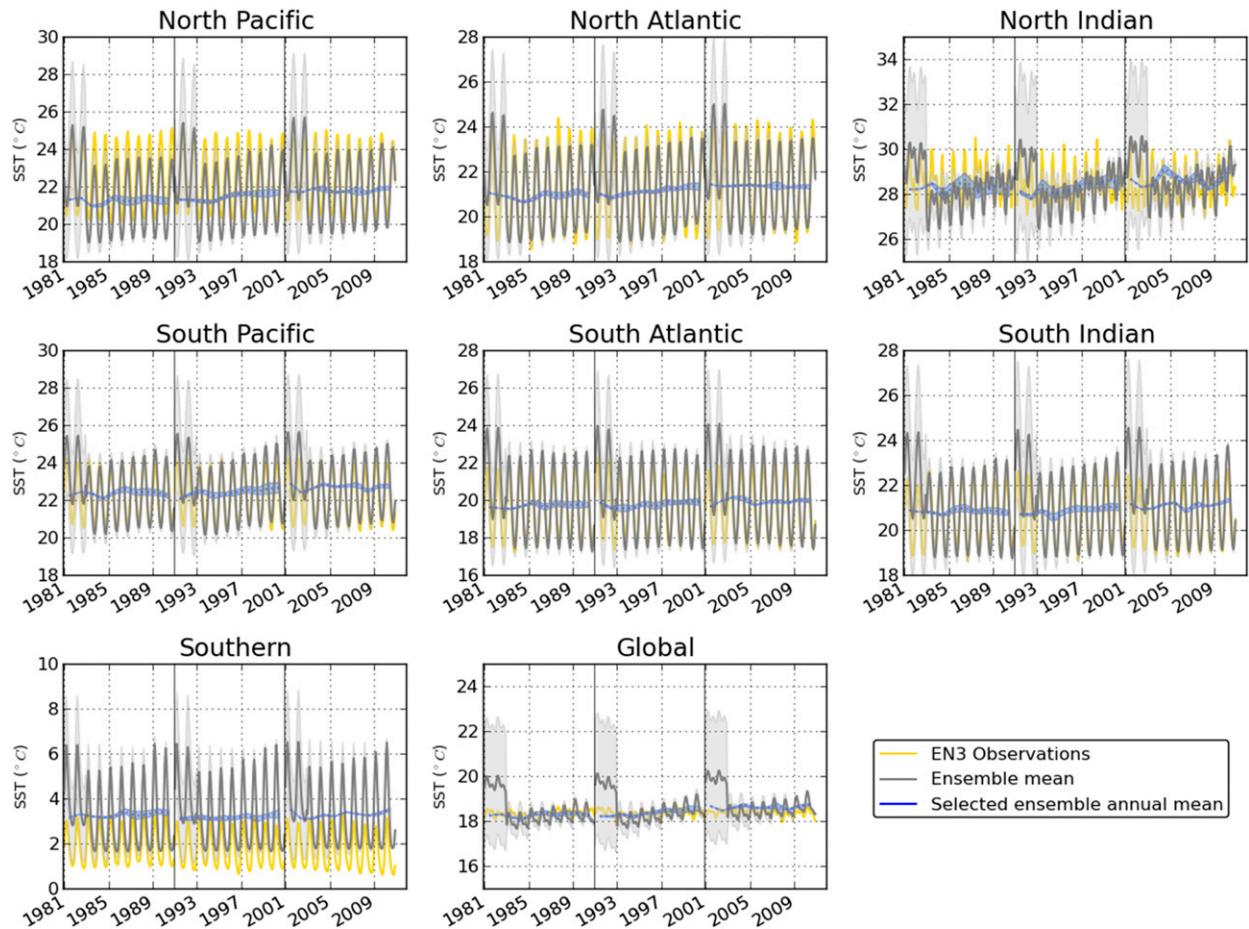


FIG. 5. SST time series for each ocean basin. EN3 observations are shown in yellow and the simulated ensemble mean is shown in gray. Shading denotes the 1 std dev spread in the ensemble. Annual-mean values for the selected ensemble are shown in blue.

decade and the next. The basin mean differences from the EN3 observations are shown in Figs. 7 and 8. The selected ensemble mean is shown in blue with darker colors denoting that more ensemble members contribute to the mean value (because of the reinflation of the ensemble after each filtering step). This shows that ensemble members close to the EN3 observations are being selected correctly and these ensemble members are close to observations for the whole period. Individual ocean basin biases exist for either the SST or OHC but, with the exception of the North Pacific, not usually in both. In the case of the North Atlantic OHC, there is a consistent 1×10^{23} J positive bias compared to EN3 observations. This bias, although marked, is a large improvement on the standard physics configuration (Figs. 7 and 8, purple), and demonstrates the benefits of using such an ensemble filtering technique. The SSTs in the Southern Ocean are biased warm and show a large difference when compared to observations in the annual cycle, which is

likely to be related to differences in modeled sea ice. However, some of the biases reduce (e.g., North Pacific SST) or increase over the simulation period (e.g., North Pacific OHC), which indicates some long-term drift in this ocean basin when using the selected model parameter settings. The biases in SST and OHC over the other ocean basins are smaller, although still show annual cycle differences.

The TCRs for the parameter sets, computed through $1\% \text{ yr}^{-1}$ CO_2 concentration increase experiments, are shown in Fig. 9 for each selection year. The TCR of the standard physics parameter set lies on the median of the parameter set distribution (denoted by the middle red line and curves for each year in Fig. 9). Although parameter sets with similar TCR values were retained until the third filter, the standard physics simulation was eliminated in the initial selection. The TCR of the selected parameter set is calculated to be 1.68 K (which is within the IPCC range) and is toward the lower end of the distribution of TCRs in the initial PPE, although it is

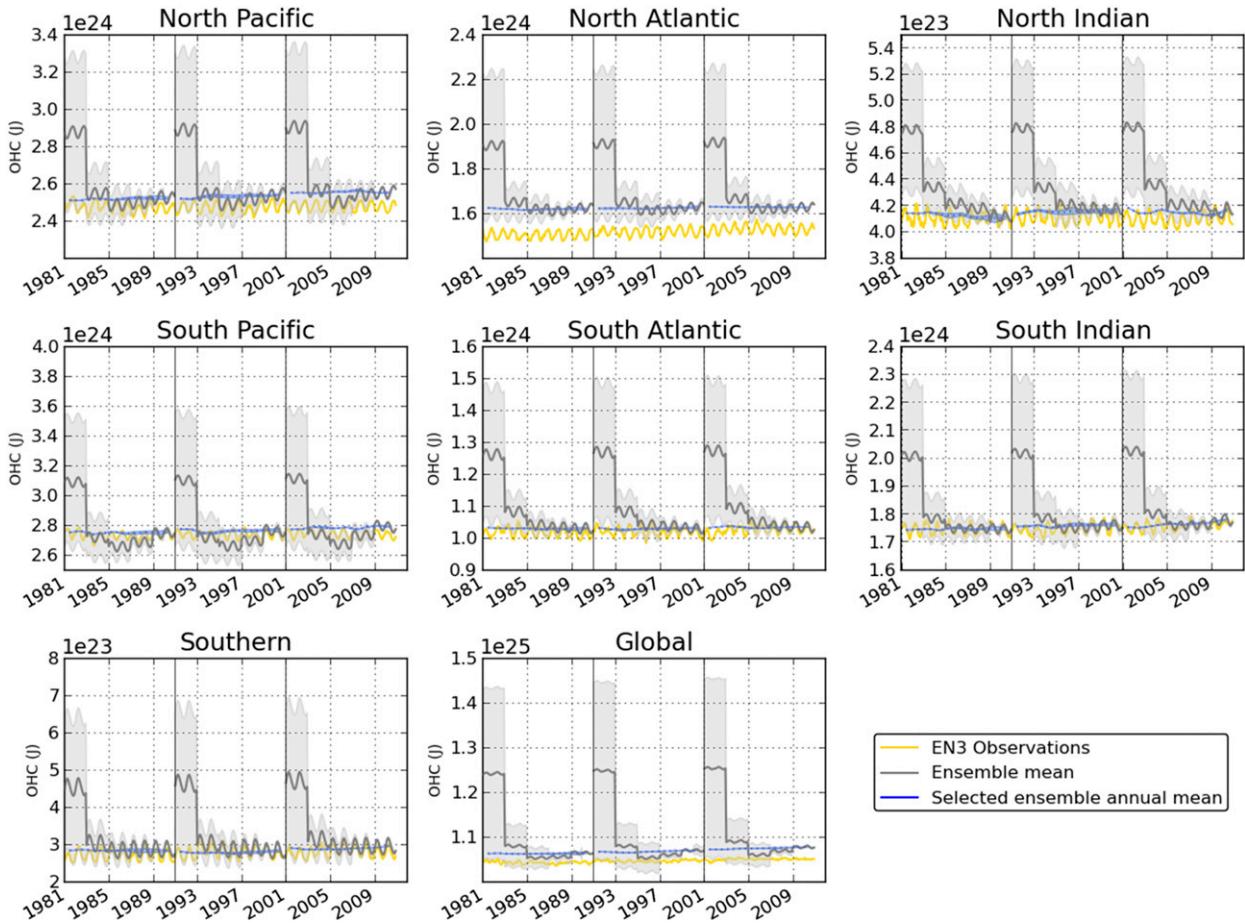


FIG. 6. As in Fig. 5, but for OHC.

by no means the lowest included in this study. The parameter set distributions (violin curves) give an indication of the likelihood that a parameter set of a given TCR value is selected at random. The selection of a TCR for the selected ensemble toward the lower end of initial TCR distribution indicates that the use of such filtering techniques to reduce mean-state biases can potentially have important consequences for the distribution of future transient climate projections, which are strongly mediated by ocean responses, when the ensemble is constrained in this way.

Figure 10 shows the normalized change in parameter values compared to the standard physics settings for the selected parameter set (blue), and six parameter sets that survive until the fourth filtering stage (hereafter “next-best sets”; yellow). The range of initial parameter values are denoted by the gray box. It is worth noting that some parameters, such as “START_LEVEL_GWDRAG” and “R_LAYERS” are only sampled discretely within the range shown. The six next-best parameter sets are included for comparison, as there may be

parameter perturbations in the selected parameter set that are not important in improving the simulation’s agreement with observations. It is clear from Fig. 10 that a diverse combination of parameter sets can yield reasonable climate simulations. This is due to the highly nonlinear and complex way in which the effects of parameters can combine. Common changes in both the best and next-best parameter sets may be more indicative in defining the relevant changes that may be made to standard model values to achieve better agreement. Focusing principally on the common changes in the ocean parameters suggests that increasing the “VERTVISC” parameter, which relates to the background vertical diffusion of momentum (viscosity) of the ocean, “VDIFFDEPTH,” which represents tracer background diffusion with depth, and “ISOPYC,” which is the surface tracer isopycnal diffusion, may be important. Also, looking at nonocean parameters that may be most relevant, an increased “CHARNOCK” parameter (which relates to momentum transport over the sea) may contribute to an

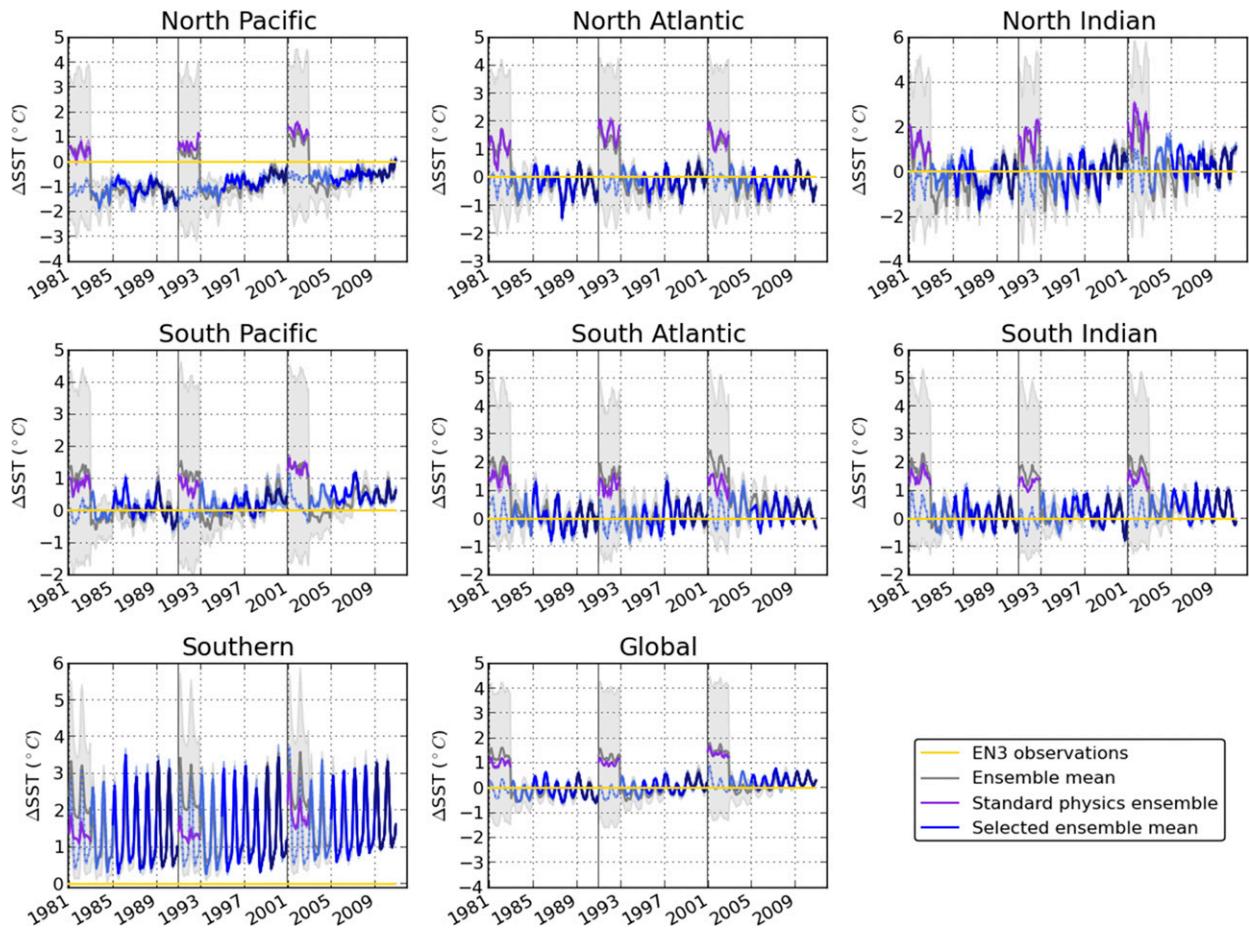


FIG. 7. SST time series differences for each ocean basin from EN3 observations (yellow). The selected ensemble mean is shown in blue (darker blue indicates more ensemble members); the ensemble mean is shown in gray; and the standard physics ensemble mean is shown in purple. Shading shows 1 std dev spread of the ensemble.

improved agreement with observations. It is evident from Fig. 10 that the precise magnitude of the change in these parameters is uncertain, likely because of complex nonlinear interactions of parameters that mean that a wide range of combinations can result in a similarly realistic climate. This high degree of uncertainty is reflected in the history-matching process of Williamson et al. (2013), where regions of parameter space are “ruled out” from yielding realistic climates, leaving the rest of parameter space as “not ruled out.” It is worth noting that there are other parameter values that suggest a common change may yield better results for this model, such as a decrease in “KAY_GWAVE” and “KAY_LEE_GWAVE,” or an increase in “CW_LAND,” “CW_SEA,” “ALPHAM,” and “VOLSCA.”

A recent study by Mulholland et al. (2017) looks at how perturbed parameter ensembles could be used to inform parameter perturbations that may need to be made to

reduce short-term forecast drift. They suggest that regionally varying parameter perturbations may be useful in improving model forecasts and that such an implementation may be useful to improve climate simulations. The presence of different biases in SST and OHC for different ocean basins shown in Figs. 7 and 8 may lend support to this claim, and it is possible that if regionally varying parameters were used in models that better overall agreement with observations could be attained. Parameter changes suggested by Mulholland et al. (2017) (their Fig. 5) show a number of common features with Fig. 10, such as an increase in “ISOPYC” and “CW_LAND”; however, there are also a number inconsistencies. These inconsistencies may be reduced if regional parameter adjustments were possible within the setup used here and may partially account for the ambiguity in the value of the highest ranked parameter sets. In short, a single globally homogeneous parameter value may not be ideal for matching observations in all ocean basins.

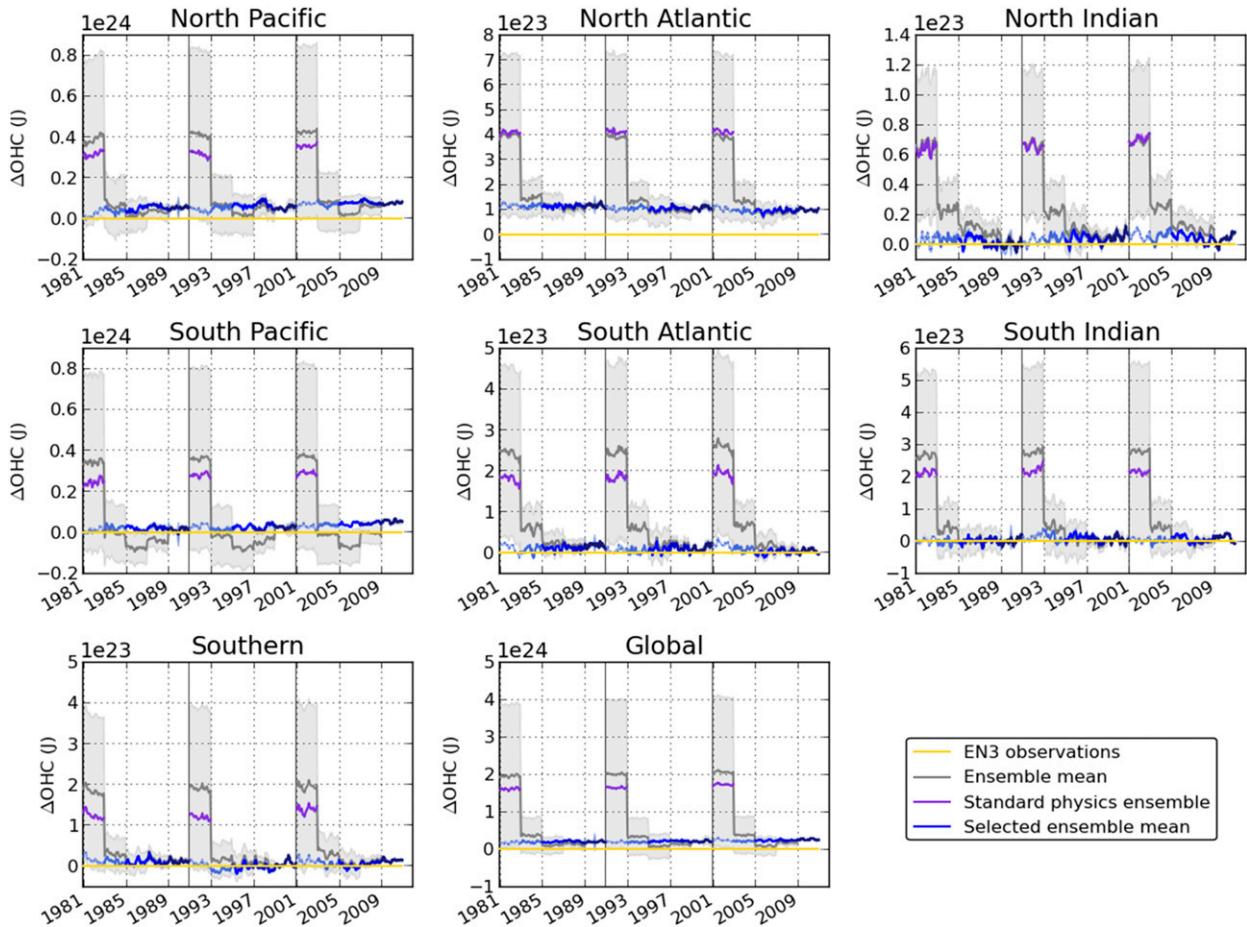


FIG. 8. As in Fig. 7, but for OHC.

b. Comparison to standard physics

A comparison can be made between the decadal-mean SST patterns for the EN3 observations, selected ensemble, and a 230-member standard physics ensemble (Fig. 11). For the selected ensemble (Fig. 11, middle panel), a yearly mean of all members is calculated for each year to give a yearly time series for the selected ensemble. This is then used to calculate a decadal mean. By constructing the decadal mean in this way, the different number of ensemble members between years in a given decade caused by reinflating the ensemble after every selection year is accounted for. The global-mean absolute bias is smaller in the selected ensemble than in the standard physics over all decades. The remaining bias between the selected ensemble and observations displays the same structural pattern over all three decades. Together this highlights that the principal effect of the filtering process is to reduce mean-state bias rather than improve model variability, which is an order of magnitude smaller than the biases shown in this

ensemble. It is clear from Fig. 11 that the major improvement of the selected ensemble over the standard physics ensemble lies in the more realistic representation of the tropical SSTs, where the warm bias seen in the standard physics ensemble is reduced. Figure 11 also shows the north-south SST gradient in Northern Hemisphere Pacific midlatitudes is not as well represented in the selected ensemble as in the standard physics ensemble, with the gradient being steeper in the selected ensemble. In the North Atlantic, this results in a cool bias of the SSTs aligning spatially with the Gulf Stream of around -3° to -6°C . The most recent version of EN4 data (Good et al. 2013) provides uncertainty estimates, which are high in regions of high variability such as the Gulf Stream. Similar uncertainty patterns are likely to have been present in the EN3 v2a observations used in this analysis, such that the discrepancy of the selected ensemble to the EN3 observations in this region may be less significant.

Figure 12 shows the AMOC at 26°N (left) and 50°N (right) from the final two years simulated, traced back in

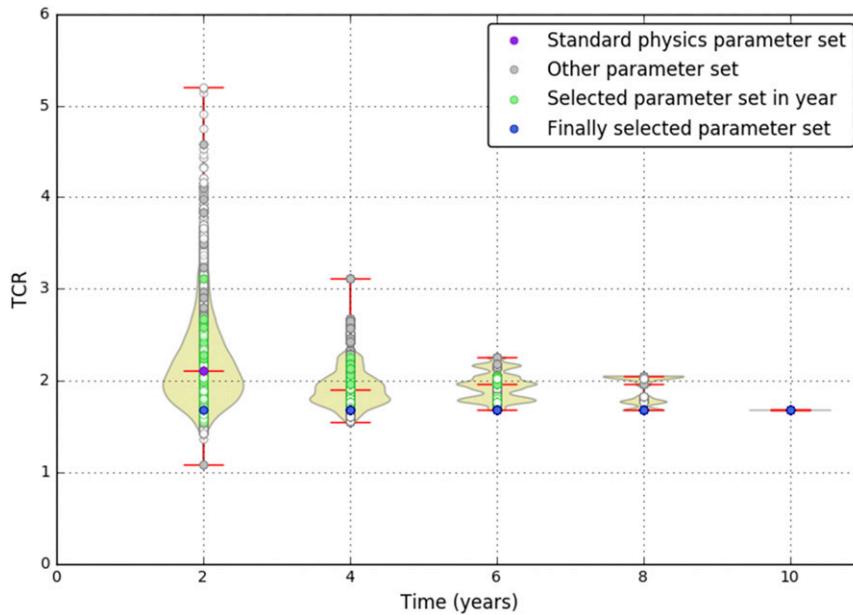


FIG. 9. TCR values for parameter sets retained during each selection year. The selected ensemble is shown in blue; ensemble members that are within the top 15% for a given year are shown in green; ensemble members outside the top 15% are shown in gray; and standard physics ensemble members are shown in purple. Open circles denote TCR values derived from an emulation process. The number of points for a given TCR is denoted by the curved distributions, with red lines denoting median and extreme values.

time through the filtering process to the first years simulated. The selected ensemble mean is shown in blue (darker blue indicates more ensemble members make up the mean value), with shading denoting a 1 standard deviation spread of the ensemble. The ensemble-mean AMOC index calculated from the 230-member initial condition ensemble of the standard physics HadCM3 parameters set is shown in purple in Fig. 12. For all simulations the AMOC index is calculated from annual-mean values. At both 26°N and 50°N the AMOC is stronger in the selected parameter simulations than for standard physics HadCM3. The 26°N AMOC index values for the selected parameter set better match the observations from the RAPID array at 26°N (Fig. 12, orange). This is further evidence that, although there is a systematic bias in the North Atlantic OHC compared to EN3 observations, the ocean transport of the selected ensemble members remains realistic. Similarly, biases in meridional heat transport at 26°N are also reduced in line with the reduction in AMOC mean-state bias (not shown).

To establish how the selected parameter set performs outside of the selection period, a 42-member ensemble was run for the 2000–15 period. Figure 13 shows monthly mean AMOC at 26°N compared to the RAPID array observations. While the model variability is too small, the overall magnitude of the AMOC continues to show a

good agreement in terms of the mean bias, with the observations into the future period. The bottom panel of Fig. 13 shows the power spectra for the RAPID observations and selected parameter set. A Bartlett window is applied with a width corresponding to the length of the observational data. Although the power at lower periods is similar, and there is evidence that the peaks around 1 and 3 years exist within our data, the magnitude of the power is much weaker than in observations. This highlights differences in the variability that remain between the model simulation and observations and indicates the periods that are underrepresented in the model.

5. Conclusions

The purpose of this work was to identify new parameter sets for HadCM3 that are capable of capturing the observed ocean state over the recent past (1980–2010) and reducing the biases present in the standard version of HadCM3's ocean mean state. A particle filtering technique is applied to a very large perturbed physics ensemble to select simulations based on their agreement with basin mean SST and OHC observations. Starting conditions are taken from existing HadCM3 historical simulations [detailed in Yamazaki et al. (2013)], which have been preselected to have a good top-of-atmosphere balance and include both physics and initial condition

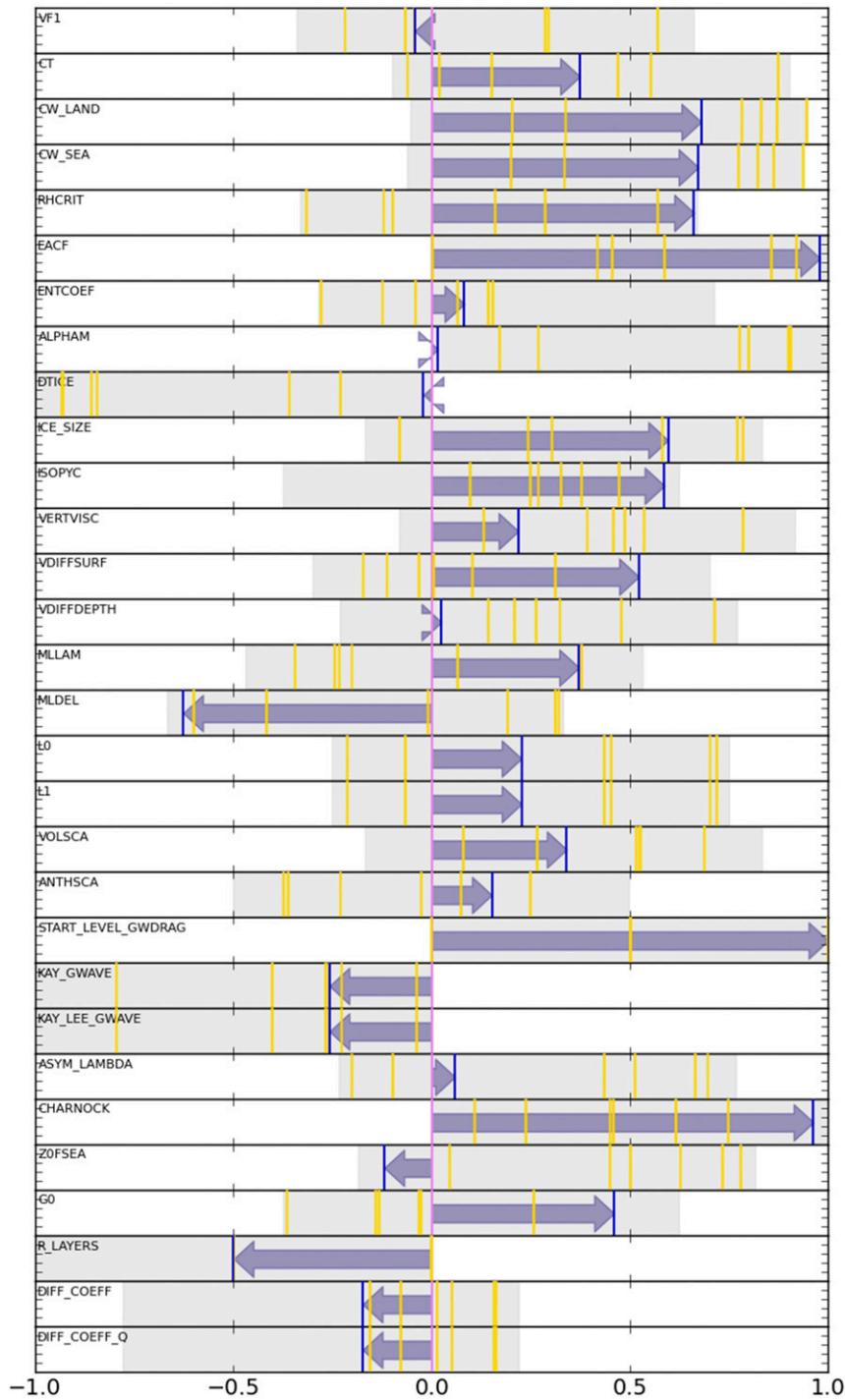


FIG. 10. Comparison of the top selected parameter set (blue line) with the standard physics parameter set (purple line). The gray range shows the range of values for the parameter at the start of the filtering process. Parameter values are normalized between 0 and 1 and shifted so that the standard physics ensemble is centered on zero. The arrow indicates the change in parameter value from standard physics to selected parameter. The six parameter sets that make it as far as the second-to-last filtering step are shown in yellow.

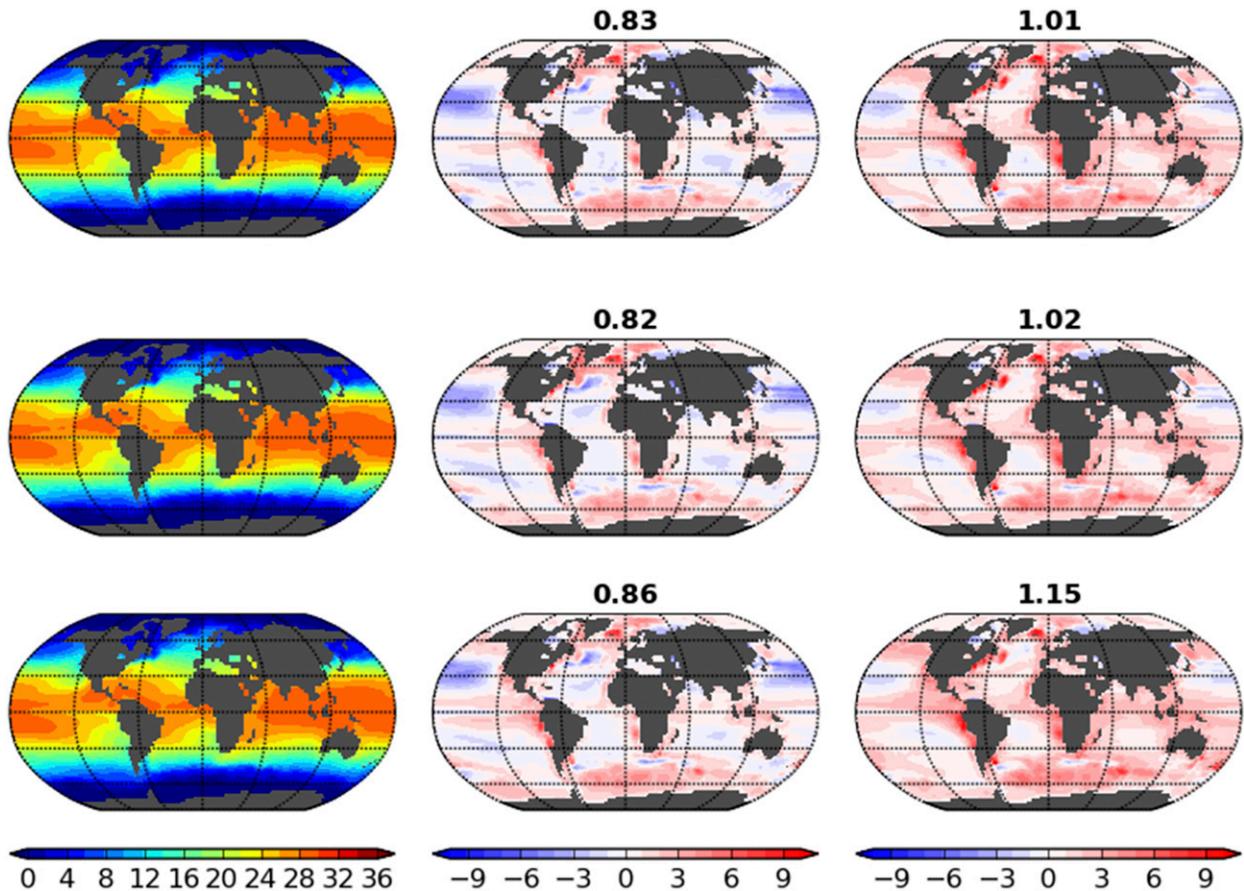


FIG. 11. (left) Decadal-mean SST patterns from EN3 observations; (middle) the mean of the selected ensemble members minus EN3 observations; (right) the mean of the standard physics ensemble minus EN3 observations for each of the three decades [(top) 1980s, (middle) 1990s, and (bottom) 2000s]. Bold numbers denote the global-mean absolute bias.

perturbations. As mentioned in the introduction, the role of internal modes of variability presents a challenge for comparing free-running models over the historical period to observations. In the filtering technique used here, the selection process will principally select to reduce biases in the mean state of a coupled atmosphere–ocean GCM, without flux adjustments. The addition of further initial conditions as the process evolves means that during the final two years simulated, the variability that the model is capable of producing with the surviving parameter set is well sampled. This does not improve the variability that the model is capable of producing for those parameter settings in itself. The simple filtering method used in this paper on such a large ensemble has the advantage of not distinguishing between ensemble members generated by physics perturbations or initial condition uncertainty and only requiring good agreement with observations for an ensemble member to be carried through to the next simulation phase, whether this is the result of internal variability or not.

As part of this study, the impact of East Asian aerosol emissions on global and East Asian surface temperature is considered while trying to ensure that the applied forcings are representative of the recent observed past. Satellite retrievals of AOT over the western North Pacific Ocean were used to determine a plausible range of trends (in percent per year) of sulfate aerosol from 1995 to present. These percent-per-year changes were applied to the 1995 CMIP5 RCP4.5 sulfate emissions fields in East Asia as inputs to the sulfur cycle in the model. A large ensemble of 2560 HadCM3 simulations show that there is no impact on either global or East Asian surface temperature by altering values within this range. Indeed, they are also indistinguishable from the original RCP4.5 scenario. The insensitivity of the model response to the different applied sulfate emissions in East Asia within HadCM3 may be truly representative of the real world or may simply indicate that a more sophisticated aerosol scheme is required [e.g., improving wet and dry deposition

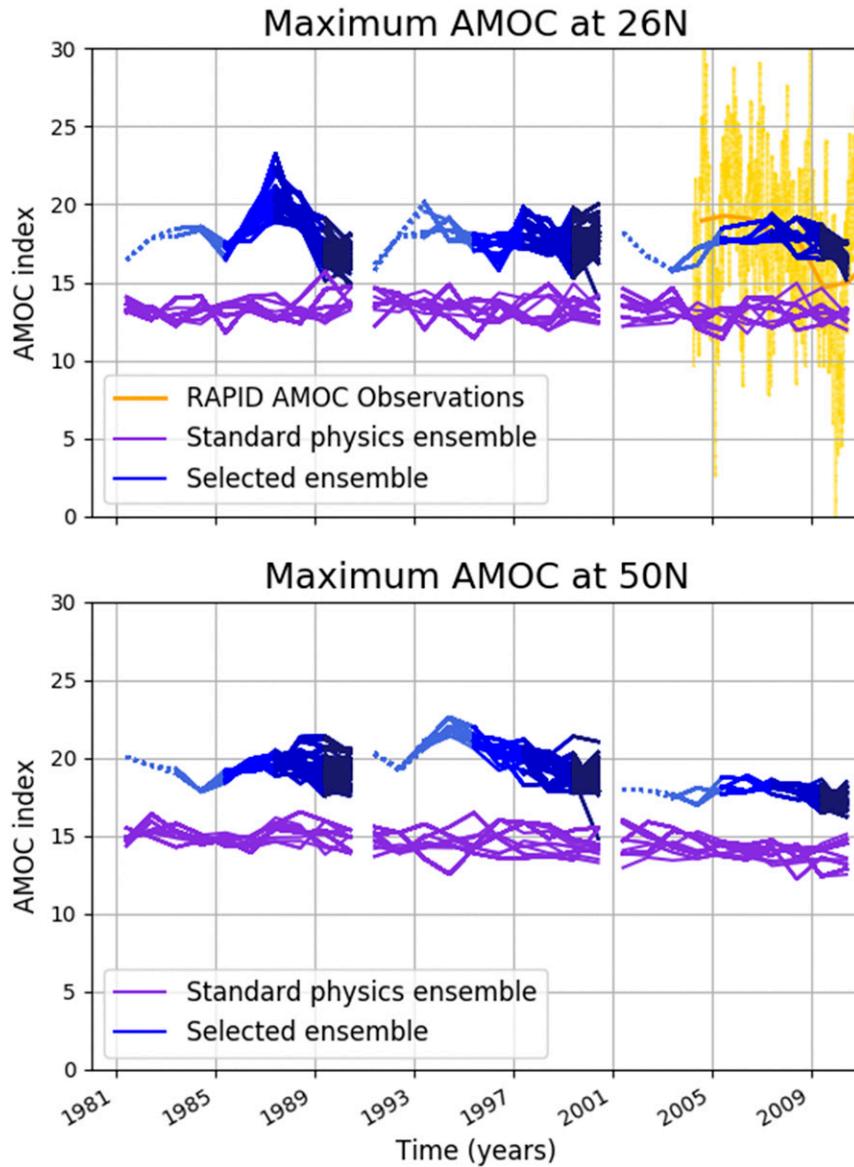


FIG. 12. AMOC (in Sverdrups; $1 \text{ Sv} = 10^6 \text{ m}^3 \text{ s}^{-1}$) at (top) 26°N and (bottom) 50°N from the final two years simulated, traced back in time to the first years simulated. The selected ensemble is shown in blue (darker blue indicates more ensemble members) simulated during that period. The standard physics ensemble, taken from a 230-member initial condition ensemble, is shown in purple. The observations from the RAPID array at 26°N are shown in orange (yearly mean) and yellow (12-hourly data).

rates and hence low-level aerosol loading as in Johns et al. (2003)] to simulate any such effect.

Our results demonstrate that the simulation of ocean properties depends in a complex and nonlinear fashion on perturbations to subgrid-scale parameters. As such, optimizing model performance in these metrics is likely to require large ensembles to enable these interactions to be fully explored. Therefore, where possible, the deployment of a similar technique could add value to the

model development process but comes with computational challenges for state-of-the-art high-resolution ocean models. A possible implementation may involve use of a lower-resolution model for initial screening before a traditional tuning exercise is undertaken with the full-resolution model.

A single-parameter configuration was identified that gave the best agreement with basin mean SST and OHC observations over the whole period. The TCR value

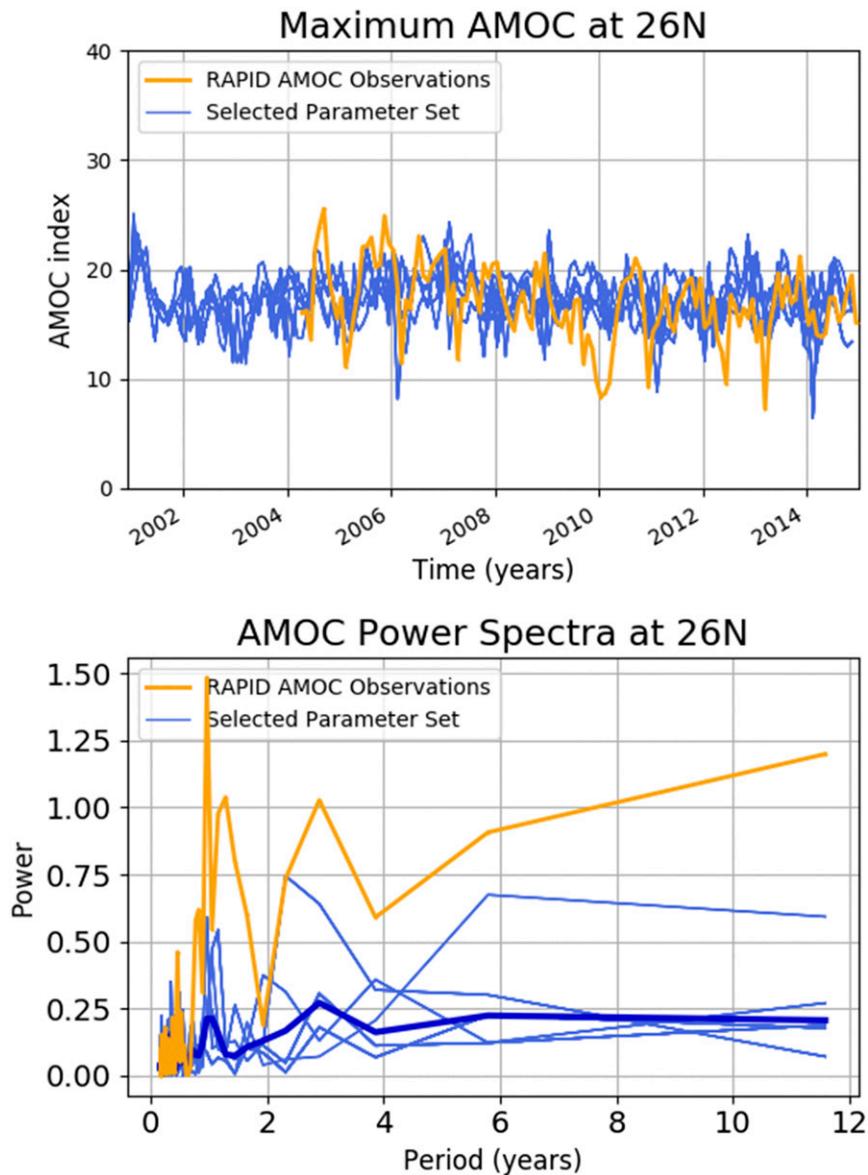


FIG. 13. (top) AMOC (in Sv) at 26°N between 2000 and 2015 for a 42-member ensemble of the selected parameter set (shown in blue and as ensemble mean and std dev). Observations from the RAPID array at 26°N are shown in orange (monthly mean) and yellow (12-hourly data). (bottom) The power spectra for the AMOC time series shown in the top panel with the dark blue line indicating the ensemble-mean power spectrum. A Bartlett window for the full width of the observational time series is applied.

associated with this parameter set is 1.68 K, close to peak of a lognormal distribution fit to the IPCC likely range. Within HadCM3, in order to better match the recent observed changes, a model setup is needed that has a lower TCR than the standard physics configuration. Reducing absolute mean-state bias may therefore have important implications for the range of uncertainty in future climate projections under an observationally constrained perturbed parameter setup. Although the

selection criteria constrain only basin mean SST and OHC compared to observations, this resulted in an improvement of transport diagnostics. For instance, the AMOC circulation at 26°N shows a better agreement with observations from the RAPID array than those captured by the standard physics configuration of HadCM3. Johns et al. (2011) show within a PPE that there is positive relationship between climate sensitivity and a decrease of AMOC strength. The results presented

here seem to indicate a relationship between simulated AMOC and TCR in perturbed physics ensembles. In the selected ensemble, the TCR is decreased relative to the standard physics parameter set and this results in an increased AMOC.

In this study, equal weighting was applied to each of the selection diagnostics and to the results from each ocean basin considered, as a proof of concept. This technique could be further refined by altering these weights to target particular phenomena of interest. It may be possible to identify more parameter sets that can capture the ocean state of the recent past in two ways: 1) by retaining more than 15% of members at each filtering point, or 2) by adding in further parameter sets in the identified region of the parameter-phase space. One advantage of using a method such as this is that a realistic ocean state can be achieved without the use of either nudging or flux adjustments. Such a technique (or a suitable variant) could be useful for model tuning to achieve good mean-state agreement with observations during the development process. The complexity of the dependence of ocean mean state on parameter settings highlights the importance of using large perturbed parameter ensembles to tune ocean mean states. The technique outlined here is more computationally efficient as a means of narrowing the parameter selection compared to a full PPE. This could be used as a precursor to traditional model-tuning techniques to produce a prefiltered set of parameters to tune from.

Acknowledgments. Data are available upon request from the corresponding author. This research has been funded as part of the Natural Environment Research Council (NERC) RAPID-WATCH Programme looking at the ChAAOS (Grant NE/G007799/1). R. G. and A. C. P. were funded as part of NERC's support of the National Centre for Earth Observation (NE/nceo020006). We thank the Met Office Hadley Centre for their technical and scientific support for the development and application of CPDN. We thank David Anderson for his advice as well as Chris Brierley and two anonymous reviewers whose comments have led to improvements in the manuscript. Finally, we thank all of the volunteers who have donated their computing time to CPDN.

REFERENCES

- Collins, M., C. M. Brierley, M. MacVean, B. B. Booth, and G. R. Harris, 2007: The sensitivity of the rate of transient climate change to ocean physics perturbations. *J. Climate*, **20**, 2315–2320, <https://doi.org/10.1175/JCLI4116.1>.
- Cox, M. D., 1984: A primitive equation, three-dimensional model of the ocean. GFDL Ocean Group Tech. Rep. 1, 143 pp.
- Flato, G., and Coauthors, 2013: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 741–866, <https://doi.org/10.1017/CBO9781107415324.020>.
- Forest, C. E., P. H. Stone, and A. P. Sokolov, 2006: Estimated PDFs of climate system properties including natural and anthropogenic forcings. *Geophys. Res. Lett.*, **33**, L01705, <https://doi.org/10.1029/2005GL023977>.
- Giorgi, F., and R. Francisco, 2000: Evaluating uncertainties in the prediction of regional climate change. *Geophys. Res. Lett.*, **27**, 1295–1298, <https://doi.org/10.1029/1999GL011016>.
- Good, S. A., M. J. Martin, and N. A. Rayner, 2013: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *J. Geophys. Res. Oceans*, **118**, 6704–6716, <https://doi.org/10.1002/2013JC009067>.
- Gordon, C., C. Cooper, C. A. Senior, H. Banks, J. M. Gregory, T. C. Johns, J. F. B. Mitchell, and R. A. Wood, 2000: The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Climate Dyn.*, **16**, 147–168, <https://doi.org/10.1007/s003820050010>.
- Gregory, J. M., and Coauthors, 2005: A model intercomparison of changes in the Atlantic thermohaline circulation in response to increasing atmospheric CO₂ concentration. *Geophys. Res. Lett.*, **32**, L12703, <https://doi.org/10.1029/2005GL023209>.
- Hong, C., Q. Zhang, K. He, D. Guan, M. Li, F. Liu, and B. Zheng, 2017: Variations of China's emission estimates: Response to uncertainties in energy statistics. *Atmos. Chem. Phys.*, **17**, 1227–1239, <https://doi.org/10.5194/acp-17-1227-2017>.
- Ingleby, B., and M. Huddleston, 2007: Quality control of ocean temperature and salinity profiles—Historical and real-time data. *J. Mar. Syst.*, **65**, 158–175, <https://doi.org/10.1016/j.jmarsys.2005.11.019>.
- Johns, T. C., and Coauthors, 2003: Anthropogenic climate change for 1860 to 2100 simulated with the HadCM3 model under updated emissions scenarios. *Climate Dyn.*, **20**, 583–612, <https://doi.org/10.1007/s00382-002-0296-y>.
- Johns, W. E., and Coauthors, 2011: Continuous, array-based estimates of Atlantic Ocean heat transport at 26.5°N. *J. Climate*, **24**, 2429–2449, <https://doi.org/10.1175/2010JCLI3997.1>.
- Jones, A., D. L. Roberts, M. J. Woodage, and C. E. Johnson, 2001: Indirect sulphate aerosol forcing in a climate model with an interactive sulphur cycle. *J. Geophys. Res.*, **106**, 20 293–20 310, <https://doi.org/10.1029/2000JD000089>.
- Jones, G. S., P. A. Stott, and N. Christidis, 2013: Attribution of observed historical near-surface temperature variations to anthropogenic and natural causes using CMIP5 simulations. *J. Geophys. Res. Atmos.*, **118**, 4001–4024, <https://doi.org/10.1002/jgrd.50239>.
- Liu, W., S.-P. Xie, and J. Lu, 2016: Tracking ocean heat uptake during the surface warming hiatus. *Nat. Commun.*, **7**, 10926, <https://doi.org/10.1038/ncomms10926>.
- Mulholland, D. P., K. Haines, S. N. Sparrow, and D. Wallom, 2017: Climate model forecast biases assessed with a perturbed physics ensemble. *Climate Dyn.*, **49**, 1729–1746, <https://doi.org/10.1007/s00382-016-3407-x>.
- Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772, <https://doi.org/10.1038/nature02771>.
- Myhre, G., and Coauthors, 2013: Anthropogenic and natural radiative forcing. *Climate Change 2013: The Physical Science*

- Basis, T. F. Stocker et al., Eds., Cambridge University Press, 659–740, <https://doi.org/10.1017/CBO9781107415324.018>.
- Nieves, V., J. K. Willis, and W. C. Patzert, 2015: Recent hiatus caused by decadal shift in Indo-Pacific heating. *Science*, **349**, 532–535, <https://doi.org/10.1126/science.aaa4521>.
- Sanderson, B. M., 2011: A multimodel study of parametric uncertainty in predictions of climate response to rising greenhouse gas concentrations. *J. Climate*, **24**, 1362–1377, <https://doi.org/10.1175/2010JCLI3498.1>.
- Sato, M., 2013: National Aeronautics and Space Administration (NASA) Goddard Institute for Space Studies (GISS) stratospheric aerosol optical thickness. NASA GISS, accessed November 2014, <http://data.giss.nasa.gov/modelforce/strataer/>.
- Shiogama, H., and Coauthors, 2012: Perturbed physics ensemble using the MIROC5 coupled atmosphere–ocean GCM without flux corrections: Experimental design and results. *Climate Dyn.*, **39**, 3041–3056, <https://doi.org/10.1007/s00382-012-1441-x>.
- Stainforth, D. A., and Coauthors, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433**, 403–406, <https://doi.org/10.1038/nature03301>.
- Thomas, G. E., and Coauthors, 2009: The GRAPE aerosol retrieval algorithm. *Atmos. Meas. Tech.*, **2**, 679–701, <https://doi.org/10.5194/amt-2-679-2009>.
- Williamson, D., M. Goldstein, L. Allison, A. Blaker, P. Challenor, L. Jackson, and K. Yamazaki, 2013: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dyn.*, **41**, 1703–1729, <https://doi.org/10.1007/s00382-013-1896-4>.
- Yamazaki, K., and Coauthors, 2013: Obtaining diverse behaviors in a climate model without the use of flux adjustments. *J. Geophys. Res. Atmos.*, **118**, 2781–2793, <https://doi.org/10.1002/jgrd.50304>.
- Yoon, J., and Coauthors, 2016: Trend estimates of AERONET-observed and model-simulated AOTs between 1993 and 2013. *Atmos. Environ.*, **125**, 33–47, <https://doi.org/10.1016/j.atmosenv.2015.10.058>.