# Tuning without over-tuning: parametric uncertainty quantification for the NEMO ocean model

Daniel Williamson[1], Adam T. Blaker[2], and Bablu Sinha[2]

[1]College of Engineering, Mathematics and Physical Sciences , University of Exeter, Exeter , d.williamson@exeter.ac.uk
[2]National Oceanography Centre , Southampton, UK, SO14 3ZH

*Correspondence to:* Daniel Williamson (d.williamson@exeter.ac.uk)

**Abstract.** In this paper we discuss climate model tuning and present an iterative automatic tuning method from the statistical science literature. The method, which we refer to here as *iterative refocussing* (though also known as history matching), avoids many of the common pitfalls of automatic tuning procedures that are based on optimisation of a cost function; principally the over-tuning of a climate model due to using only partial observations. This avoidance comes by seeking to rule out parameter

5    choices that we are confident could not reproduce the observations, rather than seeking the model that is closest to them (a procedure that risks over-tuning). We comment on the state of climate model tuning and illustrate our approach through 3 waves of iterative refocussing of the NEMO ORCA2 global ocean model run at $2^o$ resolution. We show how at certain depths the anomalies of global mean temperature and salinity in a standard configuration of the model exceeds 10 standard deviations away from observations and show the extent to which this can be alleviated by iterative refocussing without compromising model

10    performance spatially. We show how model improvements can be achieved by simultaneously perturbing multiple parameters, and illustrate the potential of using low resolution ensembles to tune NEMO ORCA configurations at higher resolutions.

## 1 Introduction

The development of ocean, atmosphere and coupled climate models represents a huge scientific undertaking that is happening simultaneously and relatively separately throughout the world's modelling centres and within the many universities that collab-

15    orate with them. Recently, with increasing importance placed on the comparison of models through Model Inter-comparison Projects (MIPs) such as the Coupled Model Inter-comparison Projects (CMIP3 and CMIP5 (Meehl et al. (2007); Taylor et al. (2012))), development of these models has entered cycles in which a few years of development of the models lead up to the submission of a set of globally organised experiments wherein the performance of each model can be compared and projections of the models under different future forcing scenarios and in different modes (e.g. atmosphere only) can be studied in order

20    to explore, in some sense, uncertainty in future climate change due to differences in the models (we do not comment on the validity of this practice here). The next such experiment is CMIP6 and will be completed in 2017.

Initially, the model development cycle may involve an increase in resolution and the replacement, improvement or inclusion of new features or sub-gridscale parameterisation schemes that characterise the physical behaviour of the code when it is run. When improvements are made to the model that are based on an improved physical understanding of the world and/or improved

computer power that enables a similar model to be run at a higher resolution, the performance of the model can be assessed by comparing the output to observations or reanalyses. Invariably, the model will not initially perform as well as its predecessor, by which we mean "be as close to" many of the observations that the modellers care about. This is only to be expected as many of the new parameterisations will have "free parameters", numbers that may or may not have a physical interpretation, but are

5    needed in order to run the model and whose values are unknown. Additionally, carefully "optimised" values of free parameters in schemes that have survived the improvements in resolution and process representation are unlikely to still be optimal and may even force the model climate into very unphysical regimes. The next phase of the model development is now increasingly known as a "tuning" phase (Mauritsen et al. (2012); Gent et al. (2011); Hourdin et al. (2016)), whereby changes are made either to the parameters or to schemes in order to bring the model "closer" to observations.

10   Currently, tuning is a highly subjective process and the experiments done in order to tune the model will vary greatly between the different centres. Mauritsen et al. (2012) argued for tuning procedures to be documented and published at the end of a development cycle. What they described as tuning involved many phases, some revisiting and potentially changing parameterisations. As such, any tuning method characterised in this way is and must be subjective, requiring a great deal of physical insight into the processes being parameterised and the limitations of the physical description given. However, there is

15   a part of the procedure that can and should be more automatic. Once suitable physical descriptions have been fixed in the form of the parameterisations, we must fix the free parameters of the model so that, to the extent that fundamental limitations of the parametric description and resolution of the model allow, the model adequately represents the physical processes we know and the observations of them that we have. Stated in this way, tuning of the free parameters is an optimisation problem and, as such, there is no reason that this should be done by hand.

20   It is easy to cast tuning (of the free parameters at least) as an optimisation problem, but in fact viewing it as such has inherent limitations and ignores much of what we know about the capability of the models and the nature of the observations we are using to benchmark them. There are a number of problems with optimisation in ocean, atmosphere and coupled climate model tuning and we will devote more space to the discussion of this in section 3.4. Concisely, the main issues are that firstly, any observational metrics that we would use have uncertainty associated with them so that, for any given metric, we would expect

25   a region of parameter space, rather than a single value of the parameters, will be consistent with the data (and choosing only one representative value risks over-tuning); and secondly, that optimising the model towards one set of observational metrics is likely (perhaps almost certain) to lead to models that have unacceptable values for metrics that have not been used in automatic tuning. This last limitation has meant that modelling centres have been cautious in using optimisation procedures suggested by other academic communities, such as statistics, preferring instead to change a small number (e.g. 1 or 2) of parameters at a

30   time by hand and investigating a large number of metrics "by eye" to ensure no major new biases are introduced (see e.g. Johns et al. (2006); Megann et al. (2014)). Throughout this paper we will argue for automatic tuning methods, but against tuning as an optimisation problem.

Instead, we will argue that in the presence of uncertainty in either observation or process-based metrics and in the presence of inherent limitations (structural errors) in the representation of the physics, tuning should be an exercise in locating regions

35   of parameter space wherein the model is predicted to be consistent with the observations and the relevant uncertainties. The

behaviour of the model throughout this region represents a source of uncertainty in our inference about the real world termed *parametric uncertainty*. This parametric uncertainty can be particularly pertinent for studies of complex problems such as the stability of the Atlantic meridional overturning circulation (MOC) (Williamson et al., 2013) and, as such, it should be quantified and at least representative models reported as the final step in a tuning exercise.

5  Even tuning methods such as Bayesian calibration (Kennedy and O'Hagan, 2001; Rougier, 2007; Sexton et al., 2011), which explicitly quantify parametric uncertainty in the form of a probability distribution for model parameters, can also be described as forms of optimisation and hence suffer from some of the drawbacks stated above. The method we describe in this paper is different in that we do not push the parameter settings towards those that are close to the metrics we are tuning with or weight them with respect to how close they are given the uncertainty in those metrics. We avoid this for the reasons given above.

10 Instead we use the tuning metrics to rule out regions of the parameter space that are too far from these, refocussing our search in the remaining parameter space with new insight into the model behaviour in the key regions and with new metrics.

The method we advocate is known more widely as history matching Craig et al. (1996); Vernon et al. (2010); Williamson et al. (2013), however, here we will refer to it as *iterative refocussing* as this name lends itself more naturally to climate model tuning. We believe iterative refocussing is a very natural and automatic mimic of the way that models are currently tuned

15 by hand. We make the most of physical insight and leave the decision about the final model (if there is only to be one) or representative set of models (more ideally) to the modellers when they have been given a set that have passed the tests they have been submitted to.

Iterative refocussing has other benefits too. Allowing us to formally define and locate structural errors as well as offering the modellers insight into the way the model responds to the perturbation of multiple parameters at the same time. Such insight can

20 lead to focus on improving particular parameterisations or work on particular parts of the model, thus our method represents a tool that can be used within a model development program.

In this paper we present the first application of iterative refocussing (or multi-wave history matching) to a GCM tuning problem and discuss the unique aspects of applying this methodology to such problems within modelling centres. In section 2 we describe the numerical model and experimental protocol we use for this study. Section 3 describes the method, and section

25 4 the application of the method to our chosen numerical model. We then present a comparison of a model representative of the "tuned" parameter space with both observations and the numerical model's standard configuration (section 5), and conclude with a comment on and example of the application of iterative refocussing to high resolution models (section 6) and a discussion (section 7).

## 2 NEMO-ORCA2

30 ### 2.1 Model description

We use NEMO (Nucleus for European Modelling of the Ocean) ORCA2 (Madec, 2008) v3.5 in the global ORCA2 ($2^o$) configuration. The model grid is tripolar isotropic mercator, with enhanced meridional resolution in the tropics and 31 $z$-coordinate vertical levels increasing in thickness from 10 m at the surface to 500 m in the abyssal ocean. It is forced with

the CORE-2 normal year forcing (Large and Yeager, 2004, 2008). Ice is represented by the Louvain-la-Neuve Ice Model version 2 (LIM2) sea-ice model (Timmermann et al, 2005). Climatological initial conditions for temperature and salinity were taken in January from PHC2.1 (Steele et al., 2001) at high latitudes, MEDATLAS (Jourdan et al., 1998) in the Mediterranean, and Levitus et al. (1998) elsewhere. Configurations of ORCA2 have been widely used for scientific studies (e.g. Friocourt

5    et al (2005); Timmermann et al (2005)), and have also participated in coordinated ocean-ice reference experiments (Griffies et al. (2009)). NEMO is the ocean component of a large number of the world's climate models (Hewitt et al., 2011; Dufresne et al., 2013; Fogli et al., 2009; Voldoire et al., 2013; Hazeleger et al., 2012). We obtained the source code from http://www. nemo-ocean.eu, along with an ORCA2 'reference' configuration and namelist containing a default set of values for each parameter and switch.

## 2.2   Parameter space elicitation

Following discussions with Gurvan Madec (pers. comm.) we chose to vary parameters and switches for the numerical ocean model which were of most interest to the community, deliberately avoiding schemes which were at the time under development, known to be a poor choice or to have stability issues, or were soon to be depracated in a future release. We then elicited plausible ranges for each of the parameters of interest (Table 1). A new parameter controlling the shape of the turbulent kinetic energy

15   penetration below the mixed layer, rn_htau, was added to the code to examine the sensitivity of the model to the fixed parameter choice of 30 m.

For the purpose of this study we do not consider uncertainties in the model domain (bathymetry) or initial conditions, the surface forcing data or the bulk formulae through which the surface fluxes are derived, or the parameter choices made for the reference configuration of the LIM2 sea-ice model. As part of a comprehensive tuning of NEMO within a modelling centre,

20   each of these additional uncertainties might be taken into consideration, though this is not current practice. To ensure that it would be possible to complete the study with the available computational resources we chose to consider only the parameter space of the numerical ocean model component.

## 2.3   Ensemble Design and Experimental Protocol

We use a method involving Latin Hypercubes detailed in section 3.3 to construct an initial ensemble of 400 integrations

25   of ORCA2. Each ensemble member was integrated on ARCHER, the UK National High Performance Computing Service. Output was processed and transferred to disk storage at the National Oceanography Centre.

We chose to integrate each ensemble member for 150 years starting from rest. In choosing this length of integration we considered several factors, including the computational cost of the simulations, the desire to achieve a steady state (or at least a state where the effects of spin up, or model drift, are small), and our desire to be able to realistically achieve similar lengths

30   of integration at higher resolutions in future. A 150 year integration is sufficiently long for the upper ocean to reach a quasi-equilibrated state, although the deep ocean will continue to drift for several thousand years, with consequent effects on the upper ocean.

**Table 1.** Parameters varied during the study, the process they are attributed to, a brief description of the physical process they control. Column 5 shows the values assigned to each parameter in the 'standard' namelist. Columns 4 and 6 show the lower and upper bounds which were elicited.

| Parameter | Process | Description | Low | Standard | High |
|---|---|---|---|---|---|
| rn_si1 | Penetr. solar rad. | shortest depth of extinction | 10 | 23 | 30 |
| rn_deds | Surface Bdy Cond. | salinity damping | | -166.67 | |
| rn_shlat | Lat. momentum | boundary slip condition | 0 | 2 | 2 |
| rn_bfri1 | Bottom friction | bottom drag coefficient | 1e-5 | 4e-4 | 1e-3 |
| rn_ahtbbl | Bottom bdy layer | lateral mixing coef | 1000 | 1000 | 10000 |
| rn_gambbl | Bottom bdy layer | advective coef | 0 | 10 | 100 |
| ln_traldf_grif | Lat. diffusion (tr.) | use griffies triads | | TRUE | |
| rn_aeiv_0 | Lat. diffusion (tr.) | 1 in ORCA2/1 (switch) | | 1 | |
| rn_aht_0 | Lat. diffusion (tr.) | Horizontal eddy diffusivity | 400 | 2000 | 4000 |
| rn_ahm_0_lap | Lat. diff. (mom.) | Horizontal eddy diffusivity | 10000 | 40000 | 100000 |
| rn_avm0 | Vert. physics | eddy viscosity ($m^2$/s) | 1e-5 | 1.2e-4 | 1.5e-4 |
| rn_avt0 | Vert. physics | eddy diffusivity ($m^2$/s) | 1e-6 | 1.2e-5 | 1.5e-5 |
| rn_ediff | TKE vert. diff. | eddy coef | 0.05 | 0.1 | 0.5 |
| rn_ediss | TKE vert. diff. | Kolmogorov dissipation coef | 0.1 | 0.7 | 0.7 |
| rn_ebb | TKE vert. diff. | surface input coef | 4.75 | 67.83 | 100 |
| rn_emin | TKE vert. diff. | minimum value | 1e-7 | 1e-6 | 1e-6 |
| nn_mxl | TKE vert. diff. | mixing length scale switch | 2 | switch(2) | 3 |
| rn_mxl0 | TKE vert. diff. | surf. buoy. length scale min value | 0.01 | 0.04 | 0.5 |
| rn_lc | TKE vert. diff. | Langmuir cell coef | 0.05 | 0.15 | 0.5 |
| rn_efr | TKE vert. diff. | fraction of surf. TKE which penetrates below ML | 0 | 0.05 | 0.1 |
| rn_htau | TKE vert. diff. | exponential decrease of TKE below ML | 0.5 | 30 | 50 |
| rn_htmx | Tidal mixing | turbulence decay scale | 100 | 500 | 1000 |
| rn_tfe | Tidal mixing | dissipation efficiency | 0.1 | 0.333 | 0.9 |
| rn_me | Tidal mixing | mixing efficiency | 0.1 | 0.2 | 0.4 |

As part of a tuning procedure based on optimisation, the length of integration is a crucial decision, particularly for ocean only or coupled GCMs. This is because the ocean cannot reach equilibrium in a time-frame compatible with tuning. Hence any optimisation procedure potentially fixes the parameters of a drifting model so that at the exact time we halt the integration (in our application, after 150 years and in high-resolution examples, 30 years (Megann et al., 2014)), the drift has met the

observations. We also note that the real ocean has never been in equilibrium and hence a tuning procedure that works by comparison to observations may not require an equilibrated ocean.
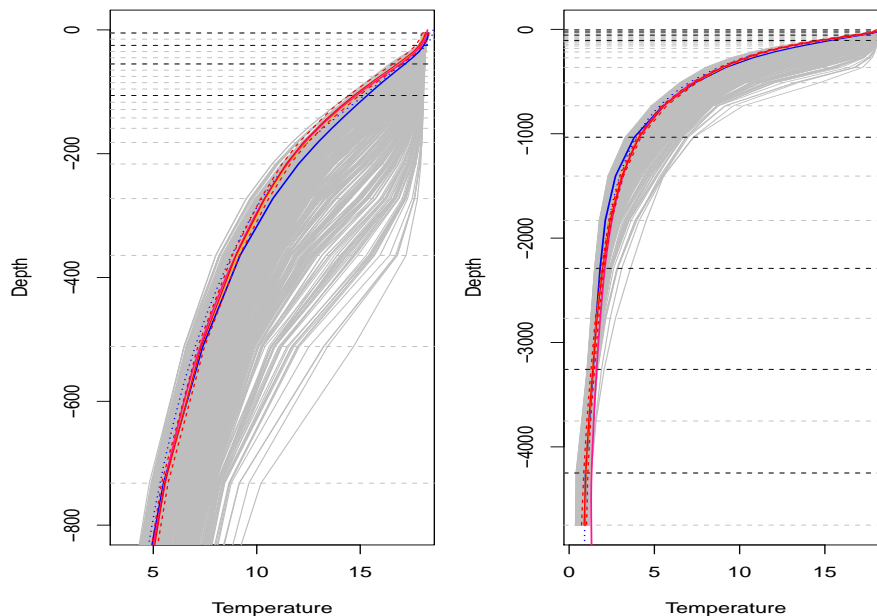
**Figure 1.** Global mean potential temperature as a function of depth from: EN3 (red, with error bounds indicated by red dashed lines); all wave 1 ensemble members (grey); standard ORCA2 (dark blue); World Ocean Atlas (pink, Locarnini et al. (2013)); the initial state (gold); and GO5 averaged over years 1996-2005 (Megann et al. (2014)) (blue dotted). The left panel shows a vertical zoom of the top 800 m, whilst the right panel shows the full depth.

## 3 Tuning with iterative refocussing

The procedure we describe here will be referred to as iterative refocussing. It is most commonly referred to as history matching
5 (Craig et al., 1996; Vernon et al., 2010; Williamson et al., 2013) and has also been called "history matching and iterative refocussing" (Craig et al., 1997) and "precalibration" (Edwards et al., 2011). We prefer to focus on the "iterative refocussing" term rather than history matching when applying these methods to numerical model tuning in this paper, as we want to highlight the importance of the iterative nature of the procedure and how it compliments model tuning. The idea is based on running ensembles in a pre-defined parameter space, using them to train statistical emulators that predict the key metrics from the
10 model output (reporting with it the uncertainty on the prediction), and then using the emulator to rule out regions of parameter space that are "too far" from observations. We formalise the procedure below.

Though history matching has been applied to GCM class models before by Williamson et al. (2013) and Williamson et al. (2015), they only performed this analysis for 1 "wave" due to their ensemble being one of opportunity. The method is most
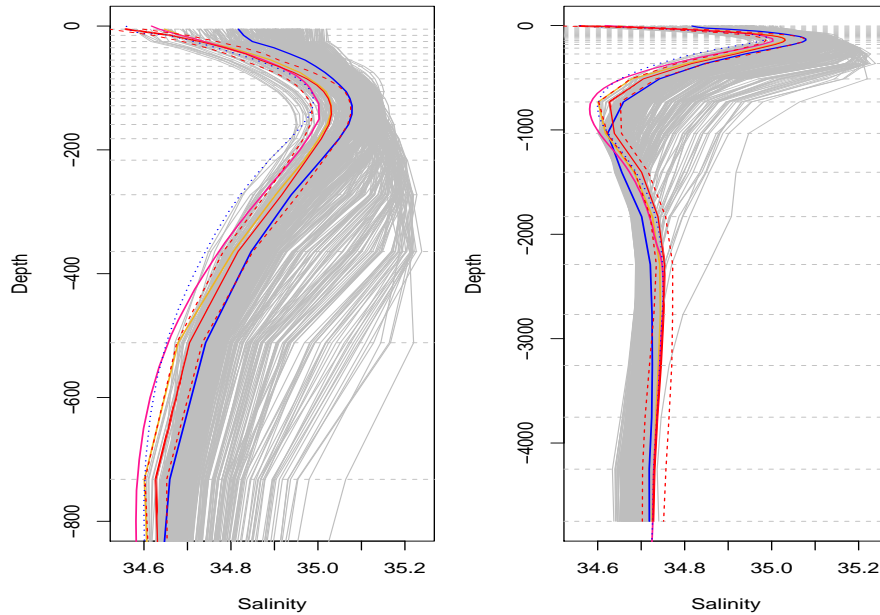
**Figure 2.** Global mean salinity as a function of depth from: EN3 (red, with error bounds indicated by red dashed lines); all wave 1 ensemble members (grey); standard ORCA2 (dark blue); World Ocean Atlas (pink, Zweng et al. (2013)); the initial state (gold); and GO5 averaged over years 1996-2005 (Megann et al. (2014)) (blue dotted). The left panel shows a vertical zoom of the top 800 m, whilst the right panel shows the full depth.

powerful when refocussing steps are taken. Having cut the parameter space down, a new perturbed physics ensemble is run within the remaining parameter space, and the procedure is repeated. This is aptly termed refocussing because with each new ensemble in a reduced space, we increase the density of our ensemble, thus improving the performance of our statistical emulators and refining the search for potentially good models.

## 3.1 Selection of metrics

Typical tuning procedures are examples of optimisation (Yang et al., 2012; Zou et al., 2014; Zhang et al., 2015), where the goal is to find the setting of the model parameters that represents the model that is somehow "closest" to a set of pre-chosen observations or metrics. Tuning by iterative refocussing represents a completely different philosophy and approach to the problem. Instead of looking for the best model, we look to rule out entire regions of parameter space as inconsistent with reproducing the metrics of interest to within an acceptable tolerance to error. This tolerance to error is certainly subjective, in one sense, as tolerance to a model's ability to reproduce certain metrics will depend on the requirements of the modelling centre. For example, a centre concerned with forecasting or climate projections for Europe will be far more intolerant to error in European temperatures and in processes around the North Atlantic than will an Asian modelling centre concerned with

projections of the monsoon. However, the tolerance to error must be bounded below by the uncertainty in the observations, which must be quantified and included in order to avoid over-tuning, and so there is an objective minimum tolerance to error.

When tuning a model or a model sub-component, a suite of diagnostics will be observed by the modellers. The choice of diagnostics is often based on readily available observational datasets, so typically surface quantities, although it can also
5 include more qualitative diagnostics based on expert judgement. Typical types of quantities used will include zonal, area, or volume integrated means, key volume, heat and salt transport metrics and 2D spatial anomaly fields. The most fundamental question when comparing a model to observations in this way is "what does close mean?" For example, how can we judge whether an area integrated mean depth profile is "close" to a similar curve derived from a data product?

Uncertainty in the observations is a minimum starting point. To illustrate this further we take as an example the Atlantic
10 MOC. If we observe the mean MOC to be around 17.5 Sv with a standard error of 1.5 (c.f. Cunningham et al. (2007)), we might use that to construct a 95% confidence interval, say $[14.5, 20.5]$ (using 2 standard errors around the mean as a guide). If we had a perfect model, it would then be reasonable to observe a model result of 22 Sv and to think that this was too far from the observations, hence leading to a requirement for tuning. But a model result of 15 Sv would be perfectly consistent with the data and there could be no justification for tuning to get it closer to 17.5 Sv, because "close" is defined by our uncertainty and
15 we are already close enough. Indeed, the uncertainty stated in this way makes it clear that the observations themselves could be 15 Sv. If they could not, our uncertainty is misspecified and is too large.

Whilst having a quantification of observation uncertainty is a crucial minimum starting point when tuning, we must also have some idea of the structural error present in the model. Structural error, also called model discrepancy, represents the inherent limitations of the model description of the relevant metrics due to resolution, unresolved sub-gridscale processes, misspecified
20 parameterisations, lack of complete physical knowledge and error in numerical solvers. For example, we could not expect to get the position of the gulf stream right in a $2^o$ ocean model for anything other than the wrong physical reasons. Since "structural error" is "real", for any given metric, we might think of this as a random quantity that could be estimated using a combination of expert modeller judgement and information from dynamic observations and process-based high resolution simulations. Despite much reference to structural error in the literature (Kennedy and O'Hagan (2001); Sexton et al. (2011);
25 Bryjarsdottir and O'Hagan (2014)), quantification of the uncertainty for the random quantity within climate science remains as far away as ever, particularly in the modelling centres and at the tuning phase.

In fact, part of the point of the tuning phase is to learn about structural error and to find out whether or not limitations of the current version of the model are due to the parameterisations or to the choice of free parameters. If errors can be "tuned out" with better choices of the free parameters, they are not structural at all, they are parametric and the goal of tuning is to find
30 and remove errors due to poor choices of the free parameters. For this reason, Williamson et al. (2013) suggested that instead of thinking of the underlying structural error, we consider our tolerance to model error. We can then think about the minimum distance a model metric would have to be from the observations before we would be prepared to use it in future projections or as part of a coupled simulation. This is a more natural description of the way models are tuned anyway, with focus given to those metrics or processes that the modellers feel they need to get right (and how near they need to be in order to have
35 confidence in the model) during the tuning.

When selecting metrics for tuning, the following ingredients are crucial:

1. It is judged physically reasonable/desirable and important to use the proposed metric to constrain the model by the developers.

2. We have a quantification of the uncertainty in the metrics. Without this, we don't know how close we are nor when we have succeeded.

3. The metric actually provides sufficient constraint on the parameter space: Certain metrics may be physically important, but do not vary sufficiently as the model parameters are varied to make them useful in tuning (McNeall et al., 2013).

The order in which metrics are applied when refocussing is also important, and we discuss this later.

### 3.2 Emulators

When we have selected the metrics we'd like to use to constrain the model, the principle of our method is to cut any region of parameter space where the model metric is not close enough to the observations. If the model were inexpensive (of the order 1 second to evaluate), we could do this using it directly (Gladstone et al., 2012). However, this would require hundreds of thousands, or even millions of model evaluations, which is not feasible. The solution to this is to run a carefully designed smaller ensemble of the model, changing all parameters simultaneously, and using that ensemble to train an "emulator" that can take the place of the full model when exploring and cutting parameter space.

An emulator is a statistical model that can predict some of the output of our climate model (the metrics we have chosen) as a function of the model parameters, quantifying our uncertainty in the prediction. As such, we can use it as a tool to assist in tuning, using the emulator to point to regions of parameter space that are of more interest so that we can then further interrogate the climate model there.

Let the free parameters or inputs of the climate model be denoted by the vector $\boldsymbol{x}$, where each $\boldsymbol{x}$ is a point in $d$-dimensional parameter space $\mathcal{X}$. Let the climate model itself be $\boldsymbol{f}(\boldsymbol{x})$, a vector-valued function of those inputs. We acknowledge here that numerical models also have forcing inputs, for example, the NEMO ocean GCM we use in this study receives surface fluxes of momentum and buoyancy through a set of bulk formulae, which interpret a dataset of observed quantities such as air temperatures and wind speeds at a distance of a few metres above the ocean. We could include both the bulk formulae and the observational dataset in $\boldsymbol{x}$, or we could consider them to be part of the functional form $\boldsymbol{f}(\cdot)$. In our application, they are considered part of $\boldsymbol{f}(\cdot)$, but we discuss forcing in section 7.

An emulator for $\boldsymbol{f}(\boldsymbol{x})$ can then usually be written as

$$f_i(\boldsymbol{x}) = \sum_j \beta_{ij} g_j(\boldsymbol{x}) + \epsilon_i(\boldsymbol{x}), \qquad \epsilon_i(\boldsymbol{x}) \sim \mathrm{GP}\left(0, C_i(\cdot, \cdot; \boldsymbol{\phi_i})\right) \tag{1}$$

where the vector $\boldsymbol{g}(\boldsymbol{x})$ contains specified basis functions in $\boldsymbol{x}$, the matrix $\beta$ is a set of coefficients to be fitted, GP stands for 'Gaussian Process', the infinite-dimensional extension of the normal distribution, the $C_i$s are pre-specified covariance functions and the $\boldsymbol{\phi_i}$s are their parameters. The basis functions can be anything from simple monomials to complex non-linear

expressions in $\boldsymbol{x}$ and allow us to add physical insight into the emulator where we have it. We can think of the left-hand term in equation (1) as a mean function, capturing 'global' or large-scale relationships (those that occur across the whole parameter space).

The Gaussian process can be thought of as a residual term, capturing 'local' variability around our global mean function. The covariance function and it's parameters specify how much variability there is and how smooth the residual process is as we move through parameter space by quantifying the correlation between the residual from our mean function at any two points in parameter space. A common choice, for example, is the separable exponential power covariance function

$$C(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\phi}) = \sigma^2 \left( \nu \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x'}\} + (1-\nu) \prod_{k=1}^{d} \exp\left\{-\theta_k |x_k - x'_k|^{\kappa_k}\right\} \right), \qquad \boldsymbol{\phi} = \{\sigma, \nu, \boldsymbol{\theta}, \boldsymbol{\kappa}\}. \tag{2}$$

Note that in this formulation, setting the correlation parameters $\boldsymbol{\theta}$ and the nugget term $\nu$ (the proportion of residual variability due to internal variability) to $0$ leads to the more familiar regression equation with uncorrelated independent errors. This is often used as a fast and approximate emulator for climate models and has been seen as "good enough" in a number of studies for calibration and history matching (Rougier et al., 2009; Sexton et al., 2011; Williamson et al., 2013). However, Salter and Williamson (2016) show that retaining correlation between input parameter choices in the Gaussian process is important in iterative refocussing as the amount of space reduction at each wave can be significantly affected, as can the final inference.

An emulator is a Bayesian model. It is completed by specifying a prior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\phi})$ and then updated using an ensemble of runs of the climate model. Let our ensemble be run at $n$ points in $\mathcal{X}$, $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ (collected into matrix $\boldsymbol{X}$) and denote $(\boldsymbol{f}(\boldsymbol{X_1}), \ldots, \boldsymbol{f}(\boldsymbol{X_n}))$ by $\boldsymbol{F}$. We discuss details of emulator parameter estimation for NEMO in section 4, but, we show the update for $\boldsymbol{f}(\boldsymbol{x})$ by $\boldsymbol{F}$ given $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ here to illustrate how model simulations affect our uncertainty. The posterior distribution $\boldsymbol{f}(\boldsymbol{x}) | \boldsymbol{F}, \{\boldsymbol{\beta}, \boldsymbol{\phi}\}$ is

$$f_i(\boldsymbol{x}) | \boldsymbol{F}, \{\boldsymbol{\beta}, \boldsymbol{\phi_i}\} \sim \mathrm{GP}(\boldsymbol{m^*}(\boldsymbol{x}), C^*(\cdot, \cdot; \boldsymbol{\phi_i}))$$

with

$$\boldsymbol{m^*}(\boldsymbol{x}) = \sum_j \beta_{ij} g_j(\boldsymbol{x}) + \boldsymbol{K}(\boldsymbol{x}) V^{-1} \left( \boldsymbol{F} - \sum_j \beta_{ij} g_j(\boldsymbol{X}) \right)$$

$$C^*(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\phi}) = C(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\phi}) - \boldsymbol{K}(\boldsymbol{x}) V^{-1} \boldsymbol{K}(\boldsymbol{x'})^T$$

where $V$ is the $n \times n$ matrix with $ij$th element $C(\boldsymbol{X_i}, \boldsymbol{X_j}; \boldsymbol{\phi})$, and $\boldsymbol{K}(\boldsymbol{x})$ is the vector with $j$th element $C(\boldsymbol{x}, \boldsymbol{X_j}; \boldsymbol{\phi})$. The same equations are used in Kalman filtering and in optimal interpolation for producing data-based reanalyses (Ingleby and Huddlestone, 1997) (because each can be seen as an update of a Gaussian process, but that is beyond the scope of this paper).

There exists a great deal of free software (in R and other platforms) for fitting Gaussian processes (building emulators), and estimating the parameters. However, we include the updating equations for the process to highlight two important features that are relevant for tuning. The first is that the mean function $\boldsymbol{m^*}(\boldsymbol{x})$ will interpolate the ensemble if $\nu = 0$, and will be pulled towards the ensemble members (to within the amount of internal variability specified or estimated) if not.

The second is that the uncertainty as characterised by the posterior variance $C^*(\boldsymbol{x}, \boldsymbol{x}; \boldsymbol{\phi})$ shrinks to the internal variability, $\nu\sigma^2$, at the ensemble design points in $\boldsymbol{X}$. The larger the ensemble the lower our uncertainty in the prediction of the climate model and the closer our prediction will be to the true values of the model output. Though this can be interpreted as an argument for very large ensembles, it is actually an argument for a high density of ensemble members in those regions of parameter space

5  where the model performs most like reality, and this reason forms a principal motivation for our approach to cutting parameter space in waves. We comment further on ensemble design in the next sub-section.

### 3.3 Ensemble design

The design of ensembles for iterative refocussing (or multi-wave history matching) is a relatively new area of research. The general principles are similar to those of one-off design of computer experiments. Namely, attempt to "fill" parameter space as

10  uniformly as possible, and, if possible, aim for minimal correlation between the parameters in the design. The first goal leads to designs that are classed as "space filling" and the second to designs that are "orthogonal". There is a large literature on space filling and orthogonal designs for computer experiments, largely based on the Latin Hypercube (Morris and Mitchell, 1995). A Latin Hypercube for an $N$-member ensemble divides the margins of each model parameter into $N$ intervals and ensures that there is exactly one representative of each interval in the ensemble. Computer experiment design usually then comes down to

15  a question of which "flavour" of Latin Hypercube to use and how large $N$ should be.

The principles of space filling and orthogonal designs are important as they aim to allow us to build emulators that are as accurate as possible throughout parameter space, and thus, in our context, to rule out as much space in one wave as possible. Similarly, guidelines on how large $N$ needs to be are mainly heuristic and aimed at making sure the estimates of the parameters in the emulator (particularly the correlation parameters) are accurate. The principle guideline used in the literature is $N = 10p$

20  where $p$ is the number of model parameters (Loeppky et al., 2009). However, the size of the ensemble can be significantly reduced without impacting emulator accuracy if data from lower resolution models is available (Kennedy and O'Hagan, 2000; Williamson et al., 2012; Le Gratiet, 2014, also see section 6).

Our wave 1 ensemble was designed using an orthogonal maximin 24-extended Latin Hypercube of size 400. This is a Latin Hypercube of size 16 that is extended 24 times, each time adding a Latin Hypercube of size 16 so that the result is also a

25  Latin Hypercube and in a way that maximises coverage and minimises correlation of parameters in the ensemble design. By designing a Latin Hypercube that has a number of component Latin Hypercubes, we have a number of sub-designs that fill space optimally in a way that allows us to validate our emulators and insures against model crashes when the experiments are running. This design element also enables additional smaller future experiments which are consistent with the original ensemble to be conducted, such as initial condition ensembles, or ensembles with different forcing datasets. Note that our

30  ensemble size is larger than the principle guidelines suggest is necessary, but this was chosen deliberately to allow scope for future studies on ensemble design size, for example when linking to ensembles at higher spatial resolution. The design method is developed in Williamson (2015) and specific merits of designs of this type for this ensemble are discussed therein. Code to generate these designs is publicly available to download.

### 3.4 Implausibility

Having constructed an emulator for a computationally expensive climate model, we can now use it to search for values of the model parameters that lead to "close enough" models (as defined by our uncertainties). Arguably, the most obvious approach to this would be to formally define a distance between the model and the observations and to find the choice of parameters that minimised this distance using the emulator. Suppose our vector of metrics is denoted $z$, then this would mean that tuning represented the optimisation problem: find $x^*$ with

$$x^* = \arg\min_{x} ||z - f(x)||_f,$$

where the norm $||\cdot||_f$ represents an appropriate measure that accounts for the sources of uncertainty discussed above. For example, using a malhalanobis distance-type function, a natural choice would be

$$||z - f(x)||_f = (z - f(x))^T \operatorname{Var}[z - f(x)]^{-1} (z - f(x)).$$

However, there are 2 problems with this. Firstly, we cannot observe every single element of the climate model state vector as part of $z$ and, even for those observations that we do have, we do not tune to them all. Hence, by minimising this distance and fixing our climate model at that setting of the model parameters, we may be artificially close to our tuning metrics in a way that unacceptably biases other metrics that we have not included in our set (perhaps because we don't have observations for them, for example). This is an example of what statisticians term over-fitting and what we refer to throughout as "over-tuning". Secondly, we cannot optimise $f(x)$ directly, so would need to use the emulator expectation $\operatorname{E}[f(x)] = m^*(x)$.

Though the distance we must consider will contain the emulator in practice, it becomes clear why this is an issue for optimisation-based approaches to tuning when we look at the distance function. The term $\operatorname{Var}[z - f(x)]^{-1}$, when $f(x)$ is also unknown and hence contributes to our uncertainty, will increase, and thus decrease $||z - f(x)||_f$, when we are very uncertain regarding the model. This happens when the contribution from $C(x, x; \phi)$ is large, which occurs when $x$ is far from any of the design points. This says that perhaps the optimum value of $||z - f(x)||_f$ occurs at a choice of $x$ where we are least sure what the model is doing.

To formalise this we would require a *statistical model* that provides a relationship between the observations $z$ and the climate model $f(x)$. Such a model will enable us to derive $\operatorname{Var}[z - f(x)]$ and thus the scaled distance between the observations and the model output. To illustrate our point we choose a model that leads to a simple, interpretable expression for this variance, though Goldstein and Rougier (2009) and Williamson et al. (2013) present alternative and more complex forms. Our model assumes that the truth, $y$, is observed with independent measurement error $e$ with variance matrix $V_e$ so that $z = y + e$. We then assume that the model, at the "optimally tuned" parameter values, $x^*$ is sufficient for the climate information available from the given parametric description so that $y = f(x^*) + \eta$, with mean-zero structural error $\eta$ independent of the climate model and with variance matrix $V_\eta$.

As we must use the emulator, our distance is $||\boldsymbol{z} - \mathrm{E}\left[\boldsymbol{f}(\boldsymbol{x})\right]||_f$ and, supposing the above form for our distance norm at $\boldsymbol{x}^*$, we would have

$$
\begin{aligned}
||\boldsymbol{z} - \mathrm{E}\left[\boldsymbol{f}(\boldsymbol{x}^*)\right]||_f &= (\boldsymbol{z} - \boldsymbol{m}^*(\boldsymbol{x}^*))^T \, \mathrm{Var}\left[\boldsymbol{z} - \mathrm{E}\left[\boldsymbol{f}(\boldsymbol{x}^*)\right]\right]^{-1} (\boldsymbol{z} - \boldsymbol{m}^*(\boldsymbol{x}^*)). \\
&= (\boldsymbol{z} - \boldsymbol{m}^*(\boldsymbol{x}^*))^T \, \mathrm{Var}\left[(\boldsymbol{z} - \boldsymbol{y}) + (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}^*)) + (\boldsymbol{f}(\boldsymbol{x}^*) - \mathrm{E}\left[\boldsymbol{f}(\boldsymbol{x}^*)\right])\right]^{-1} (\boldsymbol{z} - \boldsymbol{m}^*(\boldsymbol{x}^*)) \\
&= (\boldsymbol{z} - \boldsymbol{m}^*(\boldsymbol{x}^*))^T \, (V_e + V_\eta + C^*(\boldsymbol{x}^*, \boldsymbol{x}^*; \boldsymbol{\phi}))^{-1} (\boldsymbol{z} - \boldsymbol{m}^*(\boldsymbol{x}^*)).
\end{aligned}
$$

If this distance is large for some $\boldsymbol{x}^*$, we are confident that the model output is too far from the observations, even given all of the uncertainties. Put another way, that value of $\boldsymbol{x}^*$ would be inconsistent with our uncertainty specification and our statistical model so that we would find it *implausible* that $\boldsymbol{x}^*$ were the optimal setting of the parameters. However, small values of $||\boldsymbol{z} - \mathrm{E}\left[\boldsymbol{f}(\boldsymbol{x}^*)\right]||_f$ can either occur when $\boldsymbol{z} - \boldsymbol{m}^*(\boldsymbol{x}^*)$ is small, or when $C^*(\boldsymbol{x}^*, \boldsymbol{x}^*; \boldsymbol{\phi})$ is large. In other words, small distances do not necessarily imply good models. Any optimiser of $||\boldsymbol{z} - \mathrm{E}\left[\boldsymbol{f}(\boldsymbol{x})\right]||$ then, might find models that are extremely close (perhaps too close) to our observations if they exist, but will also favour models where our emulator is extremely uncertain, with no guarantee that the model is close to the observations there.

It follows then that whilst the search for models with small distance from the observations does not necessarily correspond to the search for good models and that minimising this distance does not necessarily find a model acceptably close to the observations at all, the search for models with large distance does correspond to the search for 'bad' models. The approach we advocate here is to locate and *rule out* all of the bad models by ruling out regions of the parameter space $\mathcal{X}$ because they correspond to large values of $||\boldsymbol{z} - \mathrm{E}\left[\boldsymbol{f}(\boldsymbol{x})\right]||_f$. To do this we must decide what "too large" means.

### 3.5 'Not ruled out yet' space

Let the implausibility function be $\mathcal{I}(\boldsymbol{x}) = ||\boldsymbol{z} - \mathrm{E}\left[\boldsymbol{f}(\boldsymbol{x})\right]||_f$. We define a threshold $a$ for which a parameter choice $\boldsymbol{x}$ is ruled out of parameter space if $\mathcal{I}(\boldsymbol{x}) > a$ for some value of $a$. If $\boldsymbol{z}$ is a scalar, $\sqrt{\mathcal{I}(\boldsymbol{x})}$ behaves like a standardised distance so that we can use the 3 sigma rule (Pukelsheim (1994)), which says that, for any unimodal distribution, $95\%$ of the probability density is within 3 standard deviations of the mean, to set $a = 9$. If $\boldsymbol{z}$ represents multiple metrics, we may emulate the joint distribution of those metrics from the model and specify covariance matrices for the error on the observations and the structural error. In this case we can compare $\mathcal{I}(\boldsymbol{x})$ to a Chi-squared random variable and set $a$ so that, say, $95\%$ or $99\%$ of the probability density would be less than $a$. The value of $a$ in this case would then depend on the number of metrics. This is particularly appropriate if $\boldsymbol{z}$ is a spatial field, where correlations in the observation error are important. However, if we are unwilling or unable to specify these correlations in the observation error, we can evaluate separate implausibilities for each metric and then rule out any parameter settings that fail to meet either all of these targets or most. This is the most often taken approach and we present an example of it in our refocussing of the ORCA2 parameter space.

Suppose we have set a threshold $a$ and defined our implausibility function $\mathcal{I}(\boldsymbol{x})$. Having chosen the initial ensemble design $X_{[1]} \in \mathcal{X}$ and built an emulator that depends on the data from this design, $F_{[1]}$, we define the subset of $\mathcal{X}$ that is Not Ruled Out Yet (NROY) to be the subset for which $\mathcal{I}(\boldsymbol{x}; F_{[1]}) \le a$. Mathematically, NROY space is

$$
\mathcal{X}^1 = \{\boldsymbol{x} \in \mathcal{X} : \mathcal{I}(\boldsymbol{x}; F_{[1]}) \le a\}.
$$

Finding NROY space $\mathcal{X}^1$ by designing an ensemble, emulating $\boldsymbol{f}(\boldsymbol{x})$ and forming $\mathcal{I}(\boldsymbol{x}; F_{[1]})$ is called *wave 1*. Refocussing is the process of repeating this multiple times, each time, in wave $k$, beginning with the parameter space $\mathcal{X}^{k-1}$. Specifically

$$\mathcal{X}^k = \{\boldsymbol{x} \in \mathcal{X}^{k-1} : \mathcal{I}(\boldsymbol{x}; F_{[k]}) \leq a\},$$

where $\mathcal{I}(\boldsymbol{x}; F_{[k]})$ can be evaluated by running an ensemble $X_{[k]} \in \mathcal{X}^{k-1}$ and using the output $F_{[k]}$ to build an emulator for
5  $\boldsymbol{f}(\boldsymbol{x})$ inside $\mathcal{X}^{k-1}$.

The process of refocussing: running ensembles, building emulators and using implausibility to further cut down parameter space by improving emulators, provides a lot of flexibility of approach. For example, we may not choose to use the same $\mathcal{I}(\boldsymbol{x})$ at every wave. Usually, and in our application to NEMO, this is because we might find that having reduced parameter space with one set of metrics, additional sets become natural secondary metrics to include. We might also feel that the inclusion of
10  complex metrics, such as spatial fields or time series, might wait until some of the very non-physical regions of parameter space have been removed in earlier waves.

The question of how many waves to run when refocussing, or rather, when to stop, is often a pragmatic one. If the entire parameter space is ruled out using a certain metric, a structural error has been located. Williamson et al. (2015) discuss this in more detail. Otherwise, if all metrics of interest have been used and the emulator uncertainty has been reduced to the level of
15  the internal variability in the model, then there is little point continuing to refocus. However, both of these cases are extreme. In fact, either time or computational budget are the limiting factors for the number of waves of refocussing in this way when tuning climate models. It may also be the case that all models are sufficiently similar (say in their transient response to CO2 forcing), that there is little point trying to further reduce the parameter space and refine the parametric uncertainty.

Having completed the exercise, the final NROY space contains all models that could not be ruled out by comparison to
20  observations or to process based knowledge. By rights then, any model within this space is worthy of study and a representative set should be submitted to any MIP type exercise if this is possible, or at least the results of having done that summarised for the benefit of the wider field.

## 3.6   Multi-Wave Ensemble design

The difficulties with multi-wave design when refocussing by history matching arise because, at least after the first wave, it is no
25  longer possible to use Latin Hypercubes as NROY space is not conformable with the unit $p$-dimensional hypercube. Put simply, NROY space has a typically non linear shape and may not even be simply connected. It may also be tiny so that even finding a point within NROY space is very difficult. Williamson and Vernon (2014) developed a way of generating candidate points for multi-wave designs for tiny NROY spaces of order $10^{-6}$ the volume of the original parameter space. However, how one selects which design points to use (and how many) is not yet well studied. In section 4 we present a new method of choosing
30  the location of points within NROY space for each wave.

The number of points one should use and how one divides a budget of runs between multiple waves of analysis are both interesting questions for further research. Beck and Guillas (2015) show how sequential designs (where the optimal next point in parameter space is selected by the algorithm and then that ensemble (size 1) is run and folded into the emulators

before choosing the next point) can improve calibration. When designing ensembles on OGCMs which must be run using supercomputers, this is not practical as it would require too much scientist time and does not take full advantage of parallel computing. Automating the design, run, update iterative process would enable the computational cost of tuning to be minimised but is beyond the scope of this study. For convenience we choose the ensemble size to be the same as wave 1, namely 400.

## 4   Iterative refocussing of NEMO - ORCA2

The metrics we will use will be derived from 1960-1990 climatological mean depth profiles of global mean temperature and salinity computed from the EN3 climatology (Ingleby and Huddlestone (1997)). Global means are computed for the EN3 grid and then interpolated onto the 30 depth levels in ORCA2. Unfortunately, uncertainty on the climatological means is not available as part of the dataset. This is a very common issue in the reporting of uncertainty, meaning that tuning to even very well observed global metrics is challenging. Even in the case of EN4 (Good et al. (2013)), the latest dataset published which includes point standard errors, the correlations in the observation errors that would allow us to accurately construct the required distance function for tuning are not reported even though they are likely derived as part of the derivation of the data.

To obtain observation error variances for each depth level for the climatology, we use the data for the observation standard deviation given as $\bar{\sigma}_o$ in Table 3 of Ingleby and Huddlestone (1997). This is an average observation error for use in data assimilation at each depth level of the EN3 grid and hence is an observation error that would apply to individual observations and not the whole climatology we are deriving. These average estimates in the table are based on a large number of observations from ocean stations, CTDs and Argo floats, where the number of observations varies with depth. We interpolate the number of observations at each depth onto the ORCA grid to give $N_1, \ldots, N_{30}$.

To scale each $\bar{\sigma}_{o,1}, \ldots, \bar{\sigma}_{o,30}$ to be consistent with a climatological estimate, we use Rayner et al. (2003) to estimate the given uncertainty in global mean SST at $0.1^o$. Converting this to a standard deviation $s$, we can derive a scaling factor $\lambda_1$ to apply to $\bar{\sigma}_{o,1}$ so that our estimated standard deviation of the observation error variance at the surface is $\sigma_1 = \bar{\sigma}_{o,1}/\lambda_1$ and is equivalent to $s$. We do not apply the same scaling factor to each depth as we want to account for the fact that the surface is better observed than other depth levels. We use a standard Monte Carlo argument that a standard error estimate is scaled by $1/\sqrt{N}$ where $N$ is the number of observations, to adjust the surface scaling to different depths. We therefore set $\sigma_i = \bar{\sigma}_{o,i}/\lambda_i$ with

$$\lambda_i = \frac{\bar{\sigma}_{o,1}\sqrt{N_i}}{\sqrt{N_1}s}$$

The derived observations and uncertainties (as standard deviations) that we use for tuning are given in table 2. We use the same scaling process for salinity observation errors. Though we acknowledge that our point wise observation error estimates are very unlikely to be accurate, we are insured against over-tuning by two factors. The first is that we add a tolerance to error (which we might interpret as an upper bound on the structural error) and the second is that we will have a separate implausibility for each level and force the model to be "too far" from the observations at at least 3 levels before it is ruled out (see below). This being said, the paucity of reported uncertainty for observations that are used as metrics for comparing models is a major issue for climate science. As argued in section 3, the similarity for a spatial field such as SST as output from a model and from

**Table 2.** Observations or temperature $z_T$, and salinity $z_S$, with observation standard deviations $\sigma_T$ and $\sigma_S$ used for our tuning of NEMO ORCA2.

| Depth (m) | $z_T$ (°C) | $\sigma_T$ | $z_S$ (PSU) | $\sigma_S$ |
|---|---|---|---|---|
| 5 | 18.15 | 0.051 | 34.56 | 0.0131 |
| 15 | 18.04 | 0.057 | 34.66 | 0.0133 |
| 25 | 17.86 | 0.060 | 34.73 | 0.0131 |
| 35 | 17.58 | 0.070 | 34.79 | 0.0143 |
| 45 | 17.20 | 0.072 | 34.83 | 0.0143 |
| 55 | 16.79 | 0.075 | 34.87 | 0.0143 |
| 65 | 16.37 | 0.078 | 34.91 | 0.0145 |
| 75 | 15.95 | 0.081 | 34.94 | 0.0148 |
| 85 | 15.26 | 0.081 | 34.97 | 0.0152 |
| 95 | 15.11 | 0.081 | 35.00 | 0.0158 |
| 106 | 14.70 | 0.082 | 35.01 | 0.0163 |
| 117 | 14.30 | 0.091 | 35.02 | 0.0157 |
| 129 | 13.88 | 0.095 | 35.03 | 0.0152 |
| 142 | 13.43 | 0.082 | 35.03 | 0.0148 |
| 159 | 12.92 | 0.075 | 35.03 | 0.0143 |
| 182 | 12.24 | 0.070 | 35.01 | 0.0136 |
| 217 | 11.42 | 0.067 | 34.97 | 0.0131 |
| 272 | 10.38 | 0.064 | 34.92 | 0.0125 |
| 364 | 8.99 | 0.058 | 34.82 | 0.0113 |
| 512 | 7.31 | 0.054 | 34.70 | 0.0098 |
| 732 | 5.56 | 0.052 | 34.63 | 0.0083 |
| 1033 | 4.09 | 0.038 | 34.64 | 0.0055 |
| 1405 | 3.09 | 0.031 | 34.70 | 0.0051 |
| 1830 | 2.40 | 0.024 | 34.74 | 0.0059 |
| 2290 | 1.97 | 0.024 | 34.75 | 0.0063 |
| 2768 | 1.64 | 0.031 | 34.75 | 0.0072 |
| 3257 | 1.39 | 0.034 | 34.74 | 0.0085 |
| 3752 | 1.15 | 0.036 | 34.74 | 0.0092 |
| 4250 | 0.98 | 0.040 | 34.73 | 0.0090 |
| 4750 | 0.88 | 0.049 | 34.73 | 0.0082 |

observations can only be judged with reference to this uncertainty (by, for example scaling the difference by the observation error variance matrix).

As we do not have correlations between errors on our observations at different depths for temperature and salinity, we define an implausibility function that gives a separate scaled distance for each depth. In wave 1 we will only use temperature to rule out regions of parameter space and observe what happens to model realisations of salinity. Using the notation of section 3, let $\mathcal{I}_i(\boldsymbol{x})$ be the scaled distance at depth level $i$ with $\mathrm{Var}\left[z_i - \mathrm{E}\left[f_i(\boldsymbol{x})\right]\right] = \sigma_{T,i}^2 + V_i + C_i^*(\boldsymbol{x}, \boldsymbol{x})$ with the emulator variance for each depth constructed as described below. We define $\mathcal{I}(\boldsymbol{x})$ for a whole temperature profile to be the third largest $\mathcal{I}_i(\boldsymbol{x})$ for $i = 1, \ldots, 30$. Written mathematically,

$$\mathcal{I}(\boldsymbol{x}) = \max\left\{\max\left\{\mathcal{I}_i(\boldsymbol{x}) \setminus \max\left\{\mathcal{I}_i(\boldsymbol{x})\right\}\right\} \setminus \max\left\{\max\left\{\mathcal{I}_i(\boldsymbol{x}) \setminus \max\left\{\mathcal{I}_i(\boldsymbol{x})\right\}\right\}\right\}\right\}.$$

We then set wave 1 NROY space, to be

$$\mathcal{X}^1 = \{\boldsymbol{x} \in \mathcal{X} : \sqrt{\mathcal{I}(\boldsymbol{x})} \leq 3\},$$

so that if 3 or more metrics are more than 3 standard deviations away from the observations for some parameter choice $\boldsymbol{x}$, that choice is ruled out. Our standard deviation here contains a component from the observations in table 2, a component from emulators of each chosen depth (explained before) and a tolerance to error in the form of $V_i$, indicating the amount of structural error we are prepared to tolerate at depth $i$. We allow the model to be out by as much as the observation error at each depth in the absence of any particular judgment as to structural errors that would lead to a global mean depth level bias. This sets $V_i = \sigma_{T,i}^2$. For wave 1 we will consider only 8 representative depth levels in order to reduce the burden in statistical modelling with emulators. We discuss this further below.

## 4.1 Refocussing NEMO

**Wave 1**

We emulate global mean temperature only at 8 depth levels corresponding to 5 m, 25 m, 55 m, 106 m, 1033 m, 2290 m, 3257 m, and 4250 m. These 8 temperatures will represent our wave 1 metrics. We use the technology described in section 3.2 to build each emulator using the following method. First we select simple functions to regress the model output on by entering them into the vector $\boldsymbol{g}(\boldsymbol{x})$ in equation (1). We use a forwards and backwards stepwise selection routine for this that first searches for the most important model parameters to enter into $\boldsymbol{g}(\boldsymbol{x})$ individually, entering all interactions between newly entering parameters and parameters already in $\boldsymbol{g}(\boldsymbol{x})$ at each step. Reduction in the residual sum of squares (RSS) from a standard least squares fit is used to guide selection. Higher order polynomials in the parameters are also available for selection, with the appropriate interactions included as per standard model selection rules in regression (see, for example Draper and Smith, 1998). Once significant reduction in RSS is no longer available or half of the degrees of freedom have been spent (the number of terms in $\boldsymbol{g}(\boldsymbol{x})$ is greater than 200), backwards elimination, whereby one at a time, the single term in $\boldsymbol{g}(\boldsymbol{x})$ that contributes the least to the RSS is removed, is used to reduce the number of terms in $\boldsymbol{g}(\boldsymbol{x})$ to at most 10% of the number of degrees of freedom. The procedure for this selection is given in further detail in Williamson et al. (2013).

For the Gaussian process covariance function, we only allow parameters that were selected into $\boldsymbol{g}(\boldsymbol{x})$ prior to the backwards elimination step be correlated in the residual, hence setting the $\theta_k$ in (2) for any unselected parameters to be zero. We follow

Bayarri et al. (2007) in setting each $\kappa_k$ to 1.9 as opposed to the more typical 2 that leads to infinite differentiability of the emulator (as this is typically too smooth for climate models). With these choices in place, the only other things required to build the emulator are prior distributions for $\boldsymbol{\beta}, \sigma, \nu$ and $\theta_j$ (for $j$ such that $x_j \in \boldsymbol{g}(\boldsymbol{x})$). We choose the reference prior given by Haylock and O'Hagan (1996),

$$\pi(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}, \nu) \propto \frac{1}{\sigma^2} \pi(\boldsymbol{\theta}) \pi(\nu)$$

and use "half-length correlations" to elicit an informative prior on the correlation lengths $\pi(\boldsymbol{\theta})$. By considering half-length correlations to elicit a distribution for $\theta_k$, we form the prior elicitation question as one of considering judgements for the correlation between $\epsilon(x_1)$ and $\epsilon(x_2)$ when all elements of $x_1$ and $x_2$ are equal aside from that element corresponding to the parameter in question, and where the distance between $x_1$ and $x_2$ is equal to half of the range of that parameter. By considering a prior for the half-length correlation for each parameter, we can derive a prior for $\boldsymbol{\theta}$ (see Williamson and Blaker, 2014, for further details).

We used a Beta prior for the half length correlations of each parameter and used the MATCH elicitation tool, an online tool that effectively allows the user to draw the shape of distribution they require and to obtain the relevant parameters of simple probability distributions (Morris et al., 2014), to set the prior for each half-length correlation to be $\mathrm{Beta}(2.9, 5)$. Our prior for the nugget $\nu$ was obtained using the MATCH tool and was set to be $\mathrm{Beta}(3.8, 1.7)$.

The posterior distribution defined by our prior modelling and by the form of the emulator must be sampled using Markov Chain Monte Carlo, and hence, whenever we evaluate the emulator to explore parameter space using history matching, we would have to use a potentially expensive sampling algorithm. To avoid this, we use an initial sample to fix the correlation parameters $\boldsymbol{\theta}$ and the nugget parameter $\nu$ at their maximum *a posteriori* (MAP) estimates. These parameters are often fixed in computationally expensive applications involving emulators (as suggested by Kennedy and O'Hagan, 2001, for example). We obtain MAP estimates using simulated annealing.

Having fit the emulator, we run diagnostic checks to assess its performance before using it to rule out parameter space. Figure 3 presents such a diagnostic plot for the emulator at 5 m, obtained by leaving each of the sub Latin Hypercubes (LHC) in the design out of the ensemble, refitting the emulator using the preselected $\boldsymbol{g}(\boldsymbol{x}), \nu$ and $\boldsymbol{\theta}$, then predicting the temperature for each sub design. Each panel represents one left out LHC predicted using the emulator. Black points and error bars ($\pm 2$ standard deviation prediction intervals) are from the emulator mean and variance, whilst green/red dots are the true model output coloured and sized by whether or not the truth lies within the error bar. Each panel represents SST plotted against the parameter `rn_ediff`, our most active parameter during wave 1. As 2 standard deviations represents an approximate $95\%$ confidence interval, we would still expect around $5\%$ of our dots to be red if our emulator were good, and we see this here.

The third panel of the 2nd row highlights a potential issue, having too many red dots for one sub-LHC. However, as these are clustered around high values of the eddy diffusivity, we might suspect that these points are also extreme in one or more other crucial parameters and are hence important for tying down the emulator behaviour in that corner of parameter space in the final model. This was the case here as 3 also had the smallest values for the langmuir cells parameter (our second most important parameter) and the other had the most extreme value of the coefficient for langmuir cells. This suggests that the
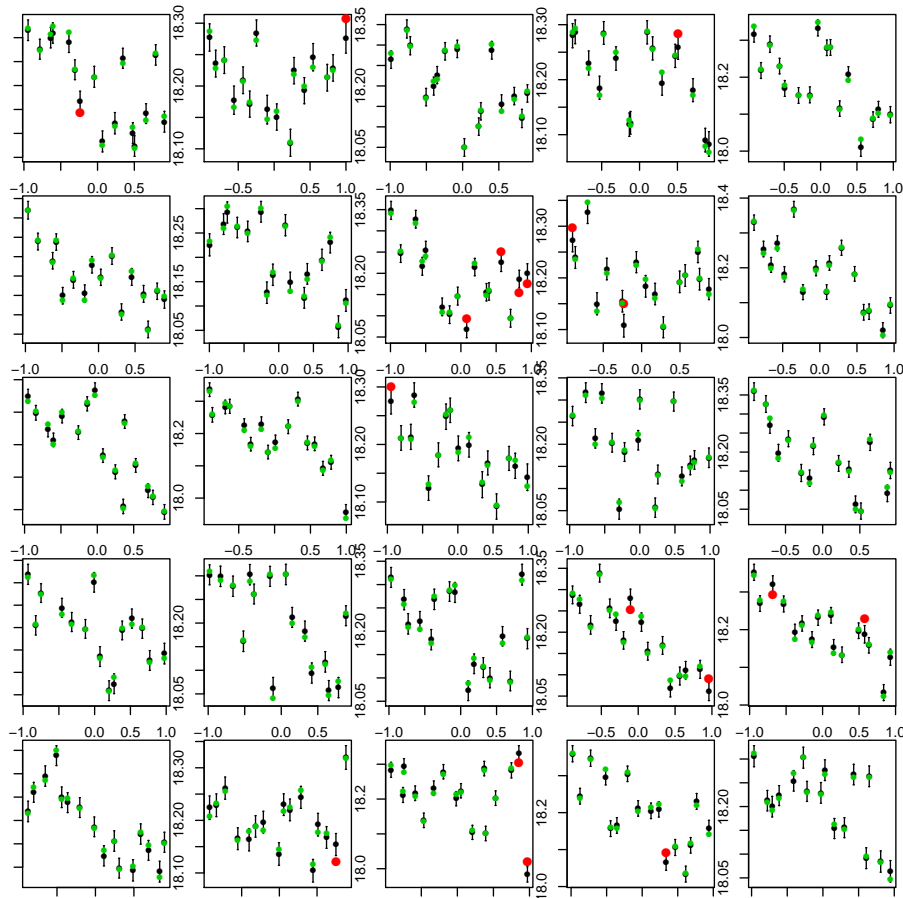
**Figure 3.** LOLHO (Leave One Latin Hypercube Out) diagnostic plots for each of the 25 16-member sub-LHC's that made up our ocean model design. Each panel is constructed by removing a sub LHC from the design, refitting our emulator using the same basis functions and correlation parameters, and predicting the model output for the 16 simulations in the sub-LHC that has not been seen by the emulator. The predictions and 2 standard deviation prediction intervals are in black. The true values are in either green (smaller dots), if they are within 2 standard deviations of the prediction, or red (larger dots) otherwise. The x-axis in each plot is the parameter `rn_ediff`, an eddy mixing parameter in the TKE mixed layer scheme. The y-axis is potential temperature at 5 m.

inclusion of this sub-LHC is crucial to the fit of the emulator, giving us some confidence that the emulator based on the full design represents the model behaviour in the full space.

Applying the implausibility metric described above, we rule out 77.5% of the elicited NEMO ORCA2 model parameter space. We can view the effect of the history match on our ensemble for global mean T and S in figures 4 and 5. Each figure shows the T or S depth profiles for each ensemble member, coloured by whether that ensemble member was ruled out (grey) or is NROY (cyan). We note that though we have constrained the model using temperature only, the global mean salinity profiles
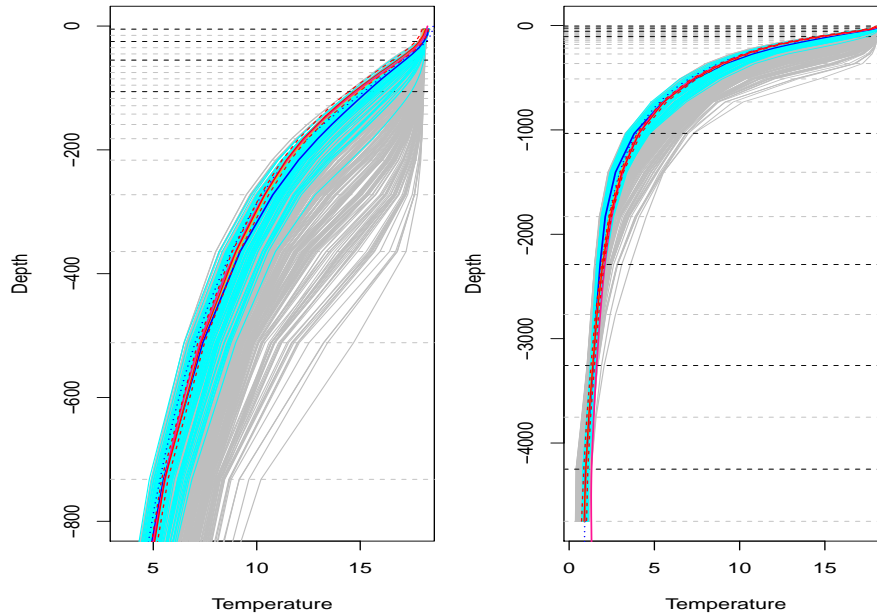
**Figure 4.** Global mean potential temperature as a function of depth from: EN3 (red, with error bounds indicated by red dashed lines); wave 1 NROY ensemble members (cyan) and RO ensemble members (grey); standard ORCA2 (dark blue); WOA (pink); the initial state (gold); and GO5 averaged over years 1996-2005 (Megann et al. (2014)) (blue dotted). The left panel shows a vertical zoom of the top 800 m, whilst the right panel shows the full depth.

are far more consistent with the data in NROY space, suggesting that much of the space ruled out was subject to either density compensated errors or excessive/insufficient mixing. We also note that standard ORCA2 is NROY at this point.

**Wave 2 and wave 3**

We design a further ensemble of 400 runs in NROY space. We can obtain a uniform sample from NROY space by randomly generating points in the original parameter space and using rejection sampling (rejecting those not in NROY) to leave a uniform sample. Whilst this might be considered a reasonable approach, there are two issues with it. Firstly, there is no guarantee that the design will "fill" NROY space, which is the reason for using Latin Hypercubes as opposed to uniform sampling for the first wave. The second is that this procedure will generally return 400 parameter choices with implausibilities over $2 - 2.5$ and near 3, as this space is vast compared to any regions of space with very low implausibility. Though, philosophically, we don't believe at this point that regions of parameter space with low implausibility are necessarily good, because the emulator uncertainty may be large there and hence driving the implausibility low there, we would like to evaluate the model in these regions as this will reduce emulator uncertainty there, hence establishing whether or not the distance between the model and the observations really is low.
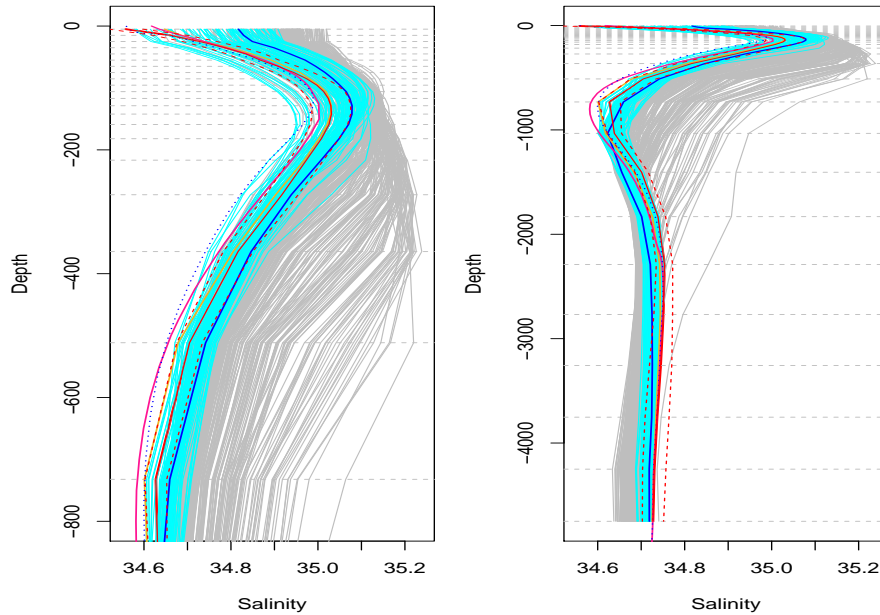
**Figure 5.** Global mean salinity as a function of depth from: EN3 (red, with error bounds indicated by red dashed lines); wave 1 NROY ensemble members (cyan) and RO ensemble members (grey); standard ORCA2 (dark blue); WOA (pink); the initial state (gold); and GO5 averaged over years 1996-2005 (Megann et al. (2014)) (blue dotted). The left panel shows a vertical zoom of the top 800 m, whilst the right panel shows the full depth.

We stratify our sampling of NROY space so that roughly $5\%$ of the design has implausibility less than 1 and that regions with implausibility less than $1.5$, 2 and 3 are sampled according to their relative volumes. These volumes are assessed and large sets of uniformly distributed candidate points from each subspace of NROY space are generated using an algorithm for uniformly sampling of tiny NROY spaces described in Williamson and Vernon (2014). The volume of the subspace with implausibility
5   less than 1 is $0.4\%$ of the size of the original space. Having decided how many points of each subspace are required and having generated a large number of uniformly designed candidates in each space, we use simulated annealing to generate an optimally space filling design from the available candidates by maximising the same coverage criterion maximised during the generation of the first wave design (described in detail in Williamson (2015)).

We add additional metrics to our implausibility criterion for wave 2, including 2 extra temperature depths at 216 m and 1406
10  m, and 5 salinity depths at 106 m, 512 m, 1033 m, 1406 m, and 2290 m. Re-emulating the model in NROY space for each of the temperature depths and using the same implausibility criterion as in wave 1 (where 3 metrics must fail in order for a point to be removed from NROY), we rule out $96\%$ of the original space, including $87\%$ of our wave 2 ensemble members and the standard settings of ORCA2.

We repeat the process for a 3rd wave of history matching, adding no further metrics but designing a new 400 member
15  ensemble in the new NROY space using the same method as for wave 2 and this time, due to the improved performance of our

emulators in this space, allowing models that fail 2 our more of our constraints to be ruled out. This final wave ruled our 75% of our wave 3 ensemble, leaving our final NROY space at 1.5% of the original parameter space.
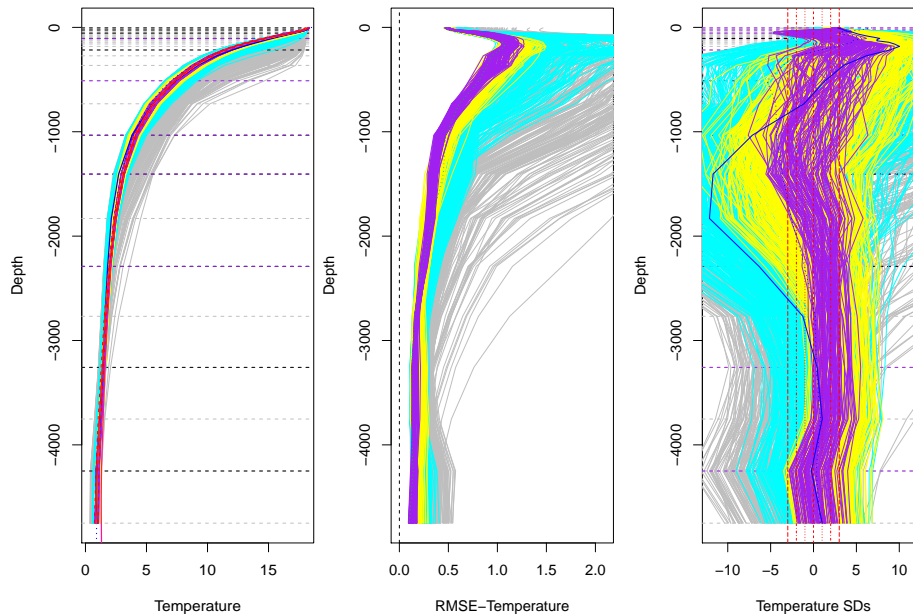
**Figure 6.** Global mean depth profiles of temperature (left), temperature RMSE (centre), and temperature depth profiles cast as departure from the EN3 global mean profile in units of $\sigma$. Colours represent W3 NROY (purple), W2 NROY (yellow), W1 NROY (cyan) and W1 RO (grey). The standard configuration is shown in solid blue, GO5 (ORCA025) as dotted blue, the initial conditions (magenta), WOA (gold).

We plot the depth profiles for all 3 waves as the left-most panel in Figures 6 and 7, with runs that were ruled out in wave 2 coloured in cyan along with the wave 1 NROY ensemble members, wave 2 NROY ensemble members and wave 3 ruled out members in yellow and wave 3 NROY in purple. We describe our final NROY space in some detail in the next section. The centre panels in Figures 6 and 7 show the root mean square error (RMSE) for temperature and salinity respectively, whilst the right-most panels of each plot show the global mean temperature/salinity depth profiles standardised by the observation and structural uncertainties (as given in table 2, so that the observations would lie on the 0 line). The RMSE figures show that improvements to global mean T and S through each refocussing step do not generally come at the price of large compensating spatial biases (as these would increase RMSE). The standardised plots show that by wave 3 most ensemble members perform reasonably well at most depths, though certain biases near the mixed layer remain difficult to remove.

## 5 ORCA 2 NROY space

Whilst calibration at each wave was performed against global mean profiles of T and S, global mean root mean square error (RMSE) provides a sanity check to ensure that plausible global mean values of T and S are not being achieved by averaging
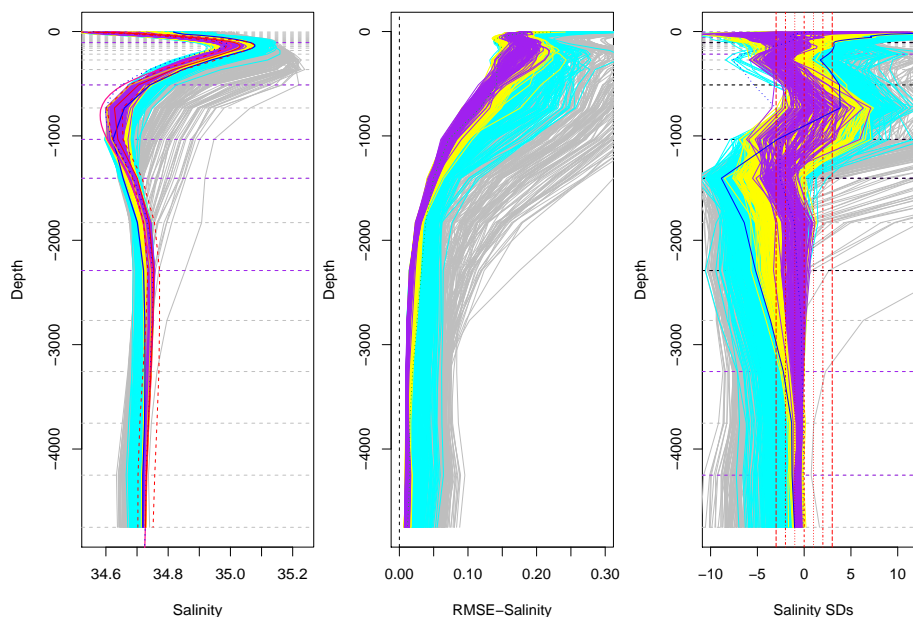
**Figure 7.** Global mean depth profiles of salinity (left), salinity RMSE (centre), and salinity depth profiles cast as departure from the EN3 global mean profile in units of $\sigma$. Colours represent W3 NROY (purple), W2 NROY (yellow), W1 NROY (cyan) and W1 RO (grey). The standard configuration is shown in solid blue, GO5 (ORCA025) as dotted blue, the initial conditions (magenta), WOA (gold).

large biases of opposite sign (e.g. strong positive biases in the tropics balanced by strong negative biases at high latitudes). We stress here that the goal is not to achieve zero RMSE. Uncertainty in the observations arising from measurement error and representativeness error mean that we should accept/expect a certain level of RMSE.

The global mean profiles of temperature and salinity already reveal several interesting features about ORCA2. Starting with
5  temperature (figure 6), we notice immediately that even within the vast parameter space we are searching it is difficult to find models which exhibit a cold bias in the mixed layer (0-300 m depth range). Almost the entire parameter space is biased warm, and the same bias is visible in the ORCA025 GO5 configuration (Megann et al. (2014)). This warm bias is indicative of excessive deepening of the mixed layer, with the standard configuration exceeding 8 $\sigma$ warmer than the EN3 climatological profile at 300 m. The most active parameters for the T emulators in the upper 300 m are `rn_ediff`, `rn_lc`, `rn_ediss`
10  and `rn_ebb`, all of which are part of the TKE mixed layer scheme. This may indicate a structural bias in the model which could be addressed with improvements in the representation of the mixed layer. Figure 8 provides insight into the structure of NROY space, and may indicate which elements of the TKE mixed layer scheme could be targeted for improvement. Choosing values of `rn_ediff`, `rn_lc` and `rn_ebb` towards the lower end of their elicited parameter ranges is more likely to result in acceptable model solutions. In contrast higher values of `rn_ediss` are more likely to yield acceptable solutions.
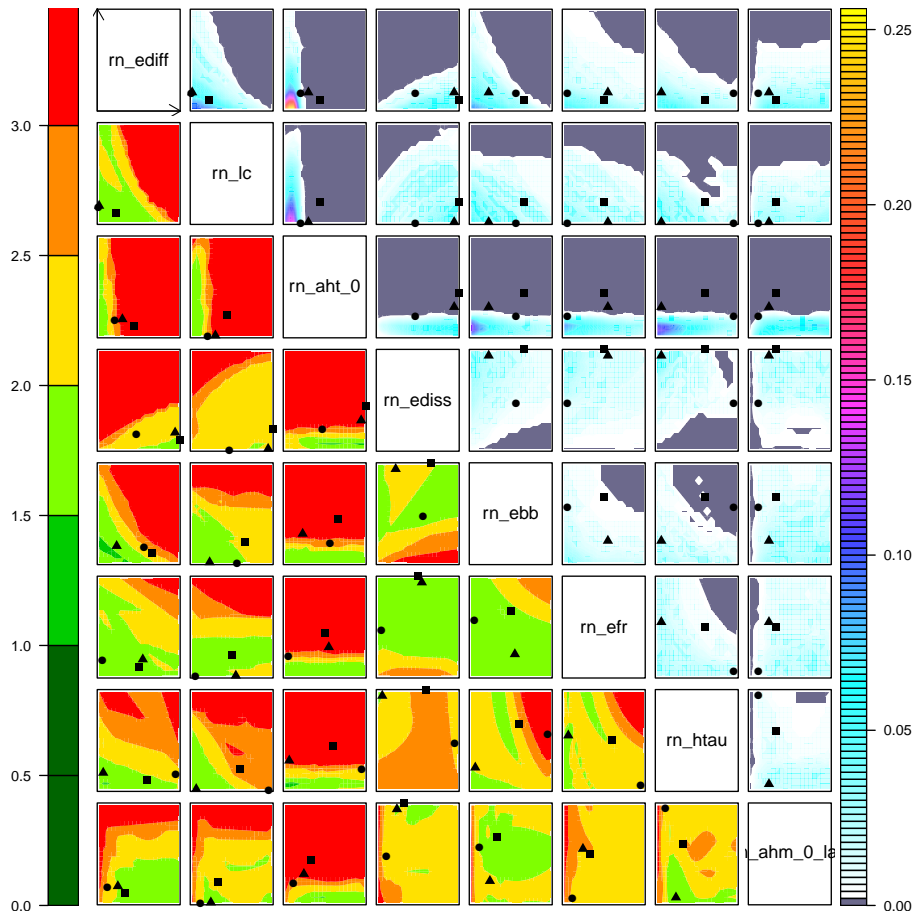
**Figure 8.** NROY density plots (upper triangle) and minimum implausibility plots (lower triangle) for 2D projections of NROY space. Each panel plots either NROY density or minimum implausibility for a pair of NEMO parameters. NROY densities, for each pixel on any panel in the upper triangle, represent the proportion of points behind that pixel in the remaining 19 dimensions of NEMO'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 19 dimensions of the NEMO parameter space. These plots are orientated the same way as those on the upper triangle, for ease of visual comparison. Standard ORCA2 is depicted on each panel as the square point. Two of the ensemble members discussed in the text are depicted with a circle (3i6) and a triangle (3jl).

The model tends to exhibit a cold bias in the 800-2000 m depth range, with the cold bias in the standard configuration peaking at 12 $\sigma$ in the 1500-1800 m depth range. Below 1000 m we are generally able to constrain the temperature bias to within 5 $\sigma$ inside W3 NROY.

Global mean profiles of salinity reveal further interesting characteristics. At the surface we identify a very large range of

5   values. The standard configuration has a salty bias in excess of 12 $\sigma$, but W3 NROY is readily able to identify configurations

with SSS much closer to observed values. Almost all of the model configurations follow a pattern of preferring a neutral to positive salinity bias around 500-1000 m and then a negative to neutral bias below 1300 m (the standard configuration reaches a fresh bias of 7 $\sigma$ at this depth). Below 2000 m we find almost no regions of parameter space able to produce a salty bias. This could, potentially, indicate a structural error in the model since it is unable to achieve solutions in which the ocean could

5   quite plausibly be located. A tendency to freshen the deep ocean will weaken the density structure and reduce vertical density gradients (and therefore transports) within the model.

The dominant parameter for the emulators in the deep ocean is the horizontal eddy tracer diffusivity, `rn_aht_0`. Figure 8 shows that acceptable models are only found for lower values of this parameter. Simultaneously low values of either `rn_ebb` or `rn_htau` are more likely to lead to NROY models.

10   Whilst it is possible to improve significantly on global mean T and S errors this does not equate to improvements everywhere. It is also of interest to examine similarities and differences between the spatial distribution of biases in the ensemble. We present spatial plots of the T and S anomalies for a selection of depth layers for the standard simulation and simulation 3jl, which is representative of the NROY space identified in the Wave 3 ensemble (Figures 9-14).

At the surface the both the standard configuration and 3jl shows cold anomalies of up to 2$^o$C over the North Atlantic,

15   Labrador and GIN seas and to the south of New Zealand. The Southern Ocean contains the strongest warm anomalies, reaching 1.5-2$^o$. There is a fairly uniform weak warm bias in the tropics and weak cold bias in the extratropical and subpolar regions (figure 9). It is worth noting that the geographical distribution and sign of the surface biases are broadly consistent throughout our ensembles, indicating that they are not determined by parameter choice but that they arise either from structural deficiencies in NEMO or from external factors which we have not tested, such as the bulk formulae, surface forcing, and the ice model.

20   The relatively small scale (order 10$^o$) surface temperature biases along the northern flank of the Antarctic Circumpolar Current arise because the models do not represent these details, which are present in the EN3 surface temperature field, adequately. We are able to achieve modest improvements in surface salinity over much of the global domain. Surface salinity on average remains too high, with the strongest biases in the Arctic. The salty surface bias in the Arctic quickly becomes a fresh bias subsurface, indicating that this may be a problem with representation of the near surface mixing in this region.

25   Dynamic features dominate the anomalies in T and S at 216 m (figure 10). Biases at this depth, where the vertical gradients in T and S are large, are particularly sensitive to modest vertical displacements of the water column. Many of the biases present are density compensating, with anomalies appearing as cold and fresh or warm and salty. Again, the general pattern of these anomalies remains fairly consistent throughout our ensemble. All simulations within our wave 3 NROY perform better than the standard configuration for global average T, and for global average S our wave 3 NROY the global mean error is of similar

30   magnitude but fresh instead of salty.

Descending to 732 m (figure 11), where the global mean T in the standard configuration is within 1 standard deviation of EN3, but global mean S is biased salty by around 4 standard deviations, we again see broad agreement in the geographical distribution of errors across our ensemble. The salinity bias in the standard configuration is largest in the northern tropical and eastern north Atlantic. Simulation 3jl shows substantial improvement in both the S and T biases in this region, although the

35   biases in the southern hemisphere T and S worsen slightly.
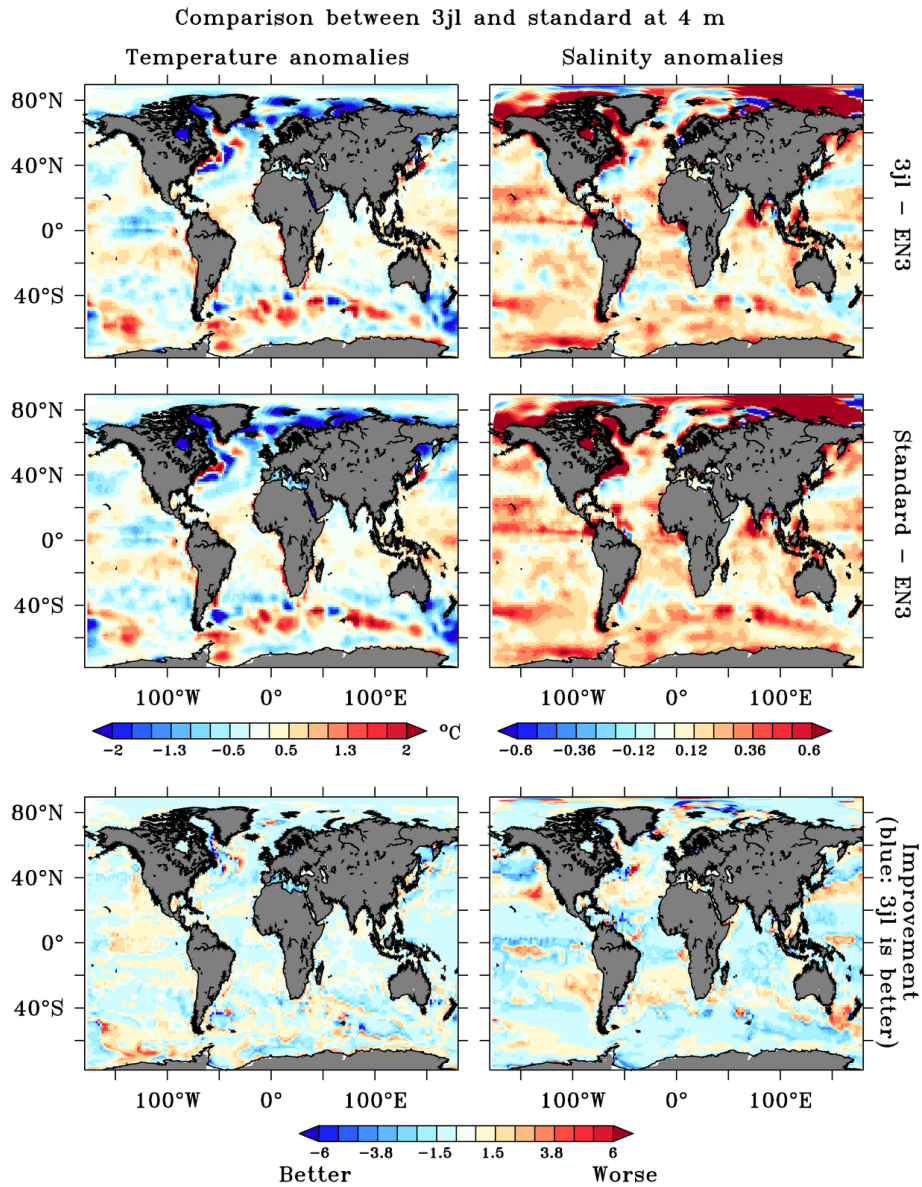
**Figure 9.** T (left) and S (right) anomalies at 4 m depth from EN3 1960-1990 climatology for simulation 3jl (top) and the standard ORCA2 configuration (middle). A metric of improvement is computed for grid cells where both are more than 3 $\sigma$ away from EN3 climatology (bottom). Negative values (blue) indicate the simulation with alternative parameter choices is performing better, whilst positive values (red) indicate the standard parameter choices are performing better. A value of -2 indicates the bias has halved compared with the standard simulation, whilst a value of +2 indicates that the bias has doubled.

We next look at 1405 m depth (figure 13) where in terms of global mean T and S biases the standard configuration is performing particularly poorly. The S bias in the eastern north Atlantic exceeds -0.3 PSU in the standard configuration, with
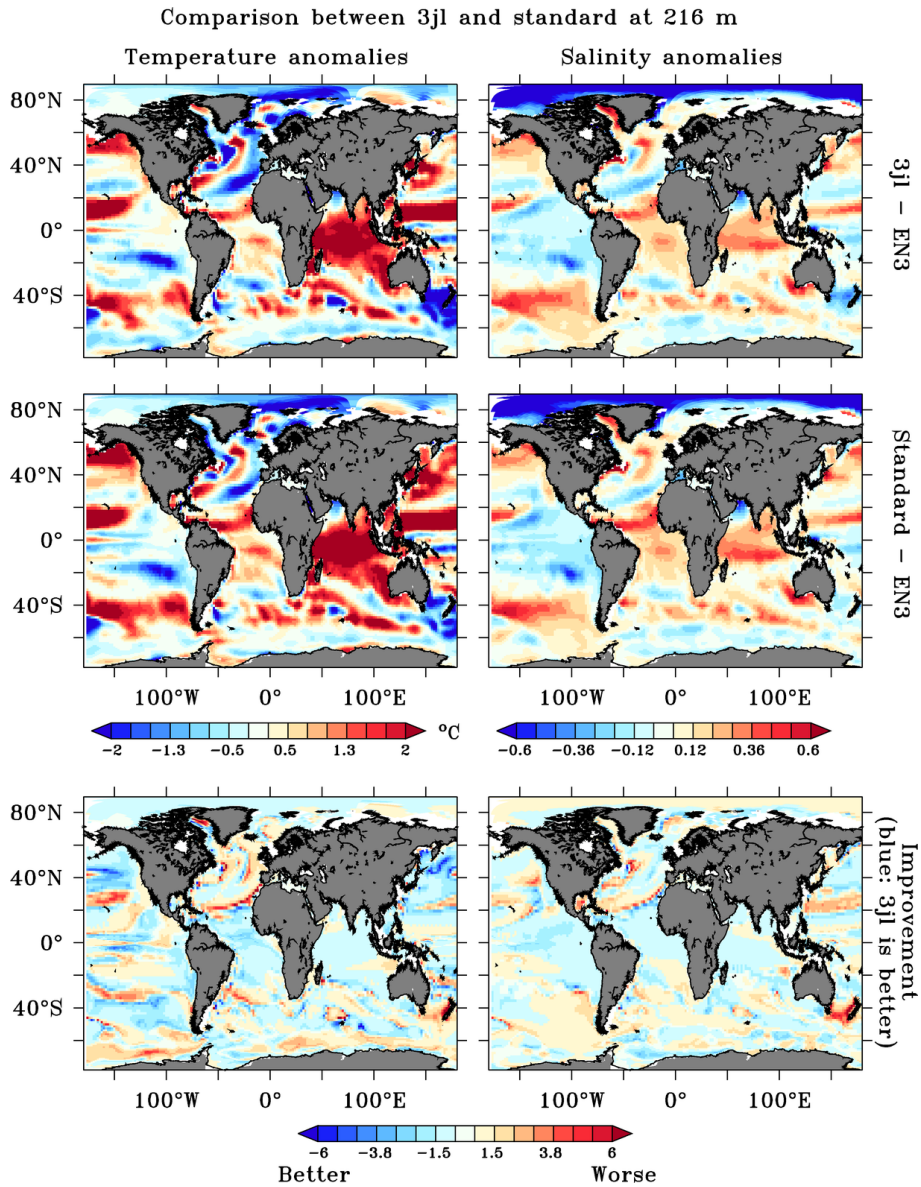
**Figure 10.** As figure 9 but showing T (left) and S (right) anomalies at 216 m depth from EN3 1960-1990 climatology.

a corresponding cold bias exceeding -0.6$^o$C. Both biases are substantially improved in simulation 3jl. Biases in the southern hemisphere are reduced but remain slightly worse than the standard configuration, one exception being an improvement in the warm bias in the Weddell Gyre which extends along the Antarctic coast to 100$^o$E. Anomalies at 1830 m (not shown) are similar to those at 1405 m.

5    At 3 km depth the T and S biases reflect a drift in the water masses indicative of a bias in the circulation. The Atlantic and Southern Ocean show warm and fresh biases, whilst biases in the Pacific are very small and of the opposite sign. Improvements
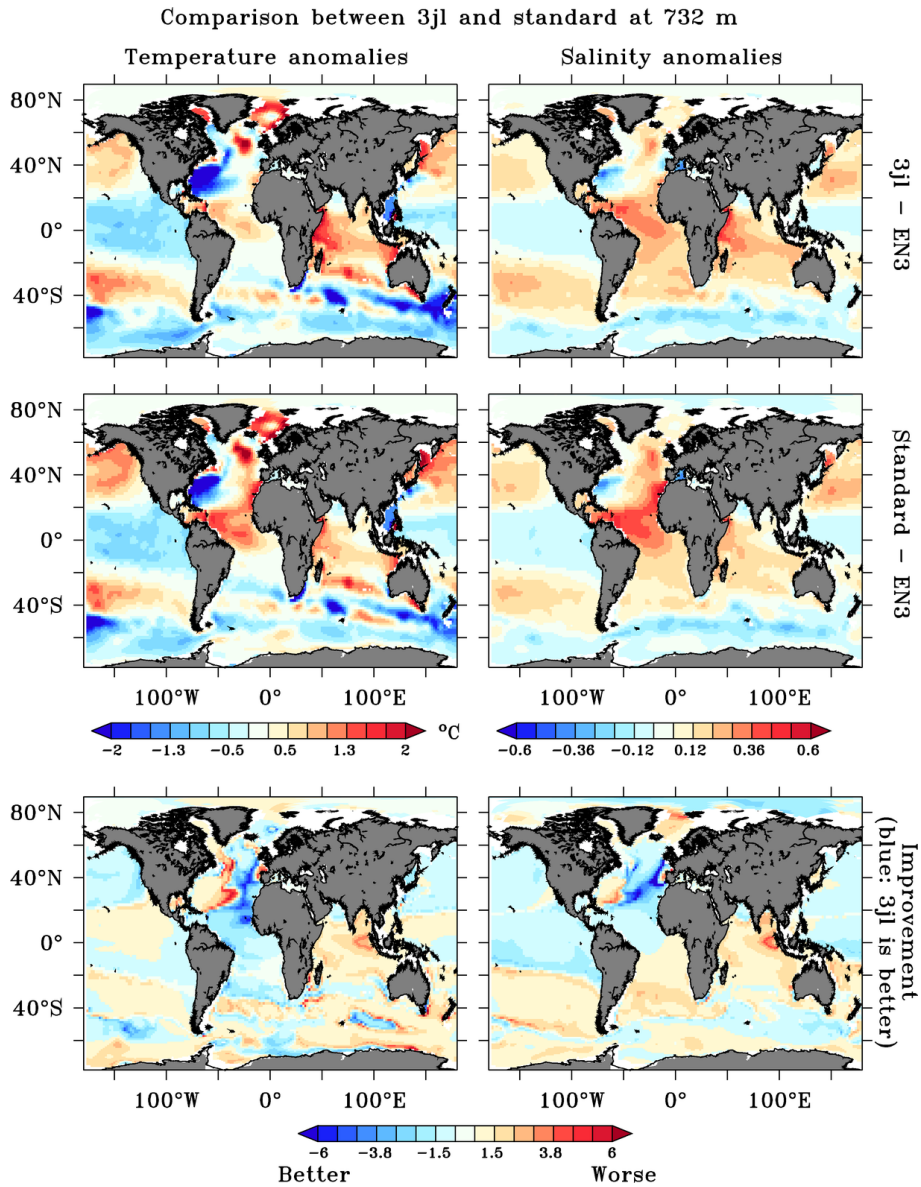
**Figure 11.** As figure 9 but showing T (left) and S (right) anomalies at 732 m depth from EN3 1960-1990 climatology.

made in the biases at 3 km depth (figure 14) also appear to be density compensating, with T(S) biases in simulation 3jl improved (worsened) in the Atlantic and Indian sectors of the Southern Ocean as well as the North Atlantic and parts of the Pacific.

In another NROY wave 3 ensemble member (3i6) we find characteristics very similar to the standard configuration over most of the global ocean. However, this ensemble member shows substantial improvement in T (and to a lesser extent S) 5 throughout the Indian Ocean at 1830 m depth (figure 15), albeit at the cost of an increase in the warm bias around the Atlantic and Indian sectors of the Southern Ocean. It is not known whether it would be possible to obtain the improvements in the
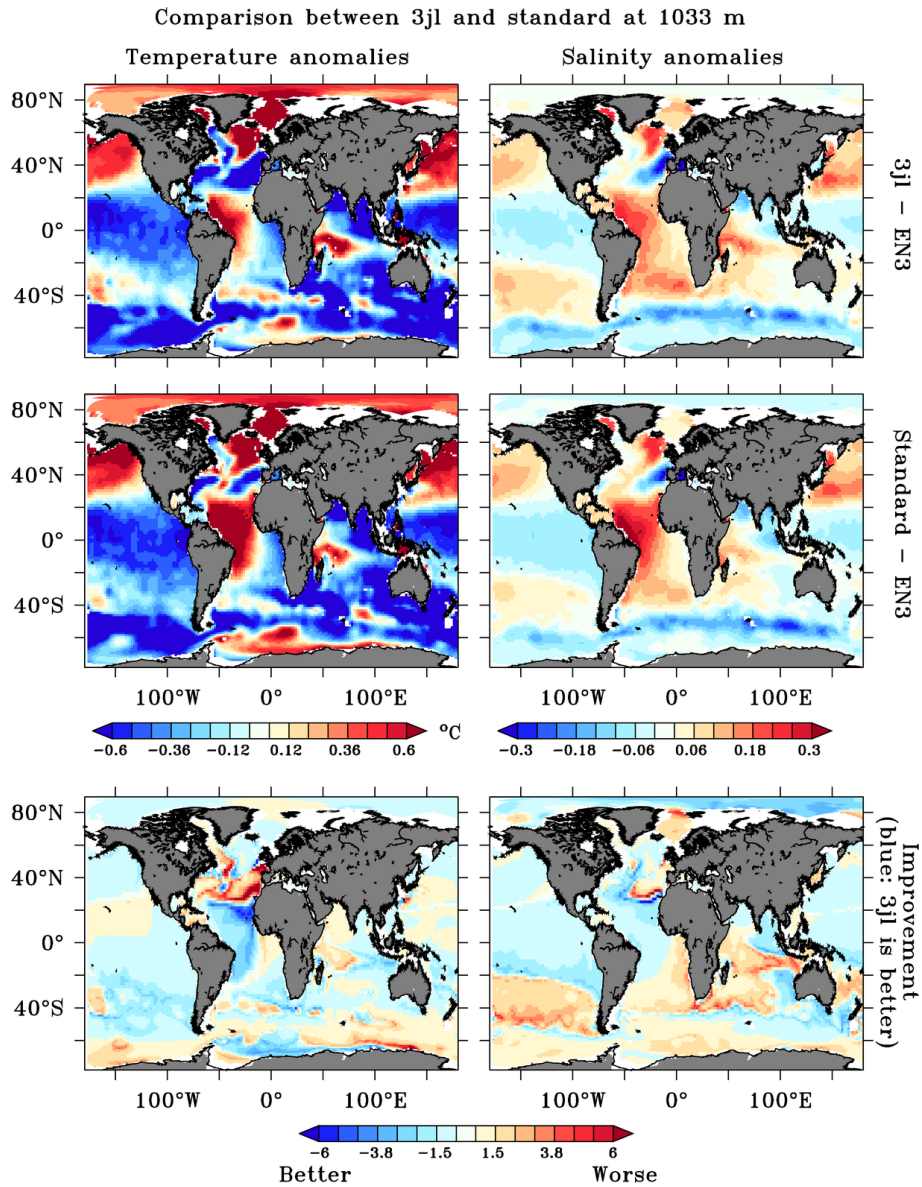
**Figure 12.** As figure 9 but showing T (left) and S (right) anomalies at 1033 m depth from EN3 1960-1990 climatology.

Indian Ocean without incurring the increase in the Southern Ocean warm bias. The most efficient way to investigate this would be to introduce regional average, or two dimensional, metrics, but that is beyond the scope of the current work.

One of the metrics of particular interest in ocean and climate models is that of the Atlantic meridional overturning circulation (AMOC). It is frequently reported at $26^oN$ to align with the RAPID/MOCHA observational array (McCarthy et al., 2015). The

5    maximum value at this latitude is typically close to 1000 m depth. We stress that comparisons of the transport should be made with caution for a number of reasons. One of the main reasons for this is that the RAPID time series is computed as the sum
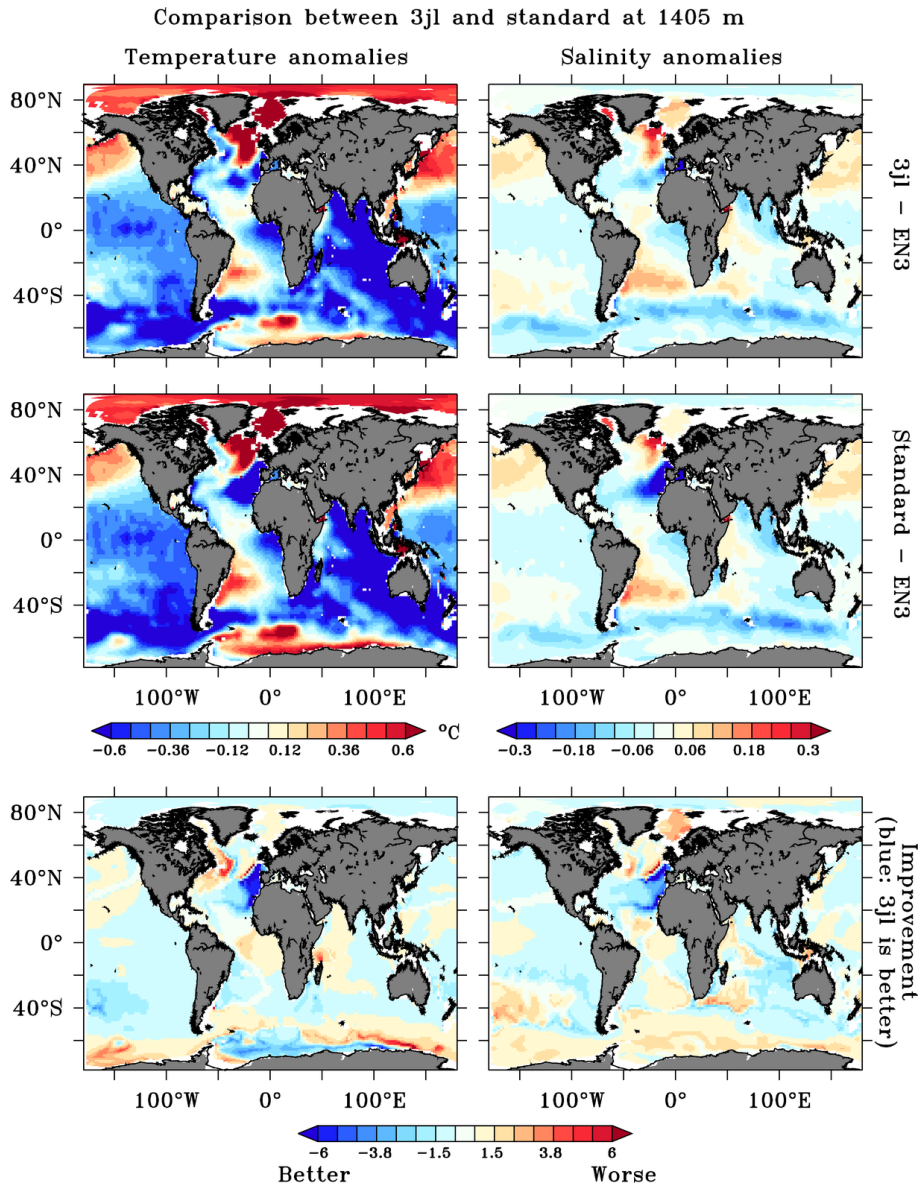
**Figure 13.** As figure 9 but showing T (left) and S (right) anomalies at 1405 m depth from EN3 1960-1990 climatology.

of three transports derived from different observations of wind stress, basin wide density gradients, and cable measurements of the Florida Straits transport. Model AMOC calculations are most commonly reported as the zonal and depth integrated meridional velocity. Sampling a numerical model in a manner consistent with the observational method can be problematic, particularly where the ocean model grid is coarse. Nevertheless we present the AMOC at 26°N and 1000 m from each of the

5   ORCA2 simulations (figure 16).
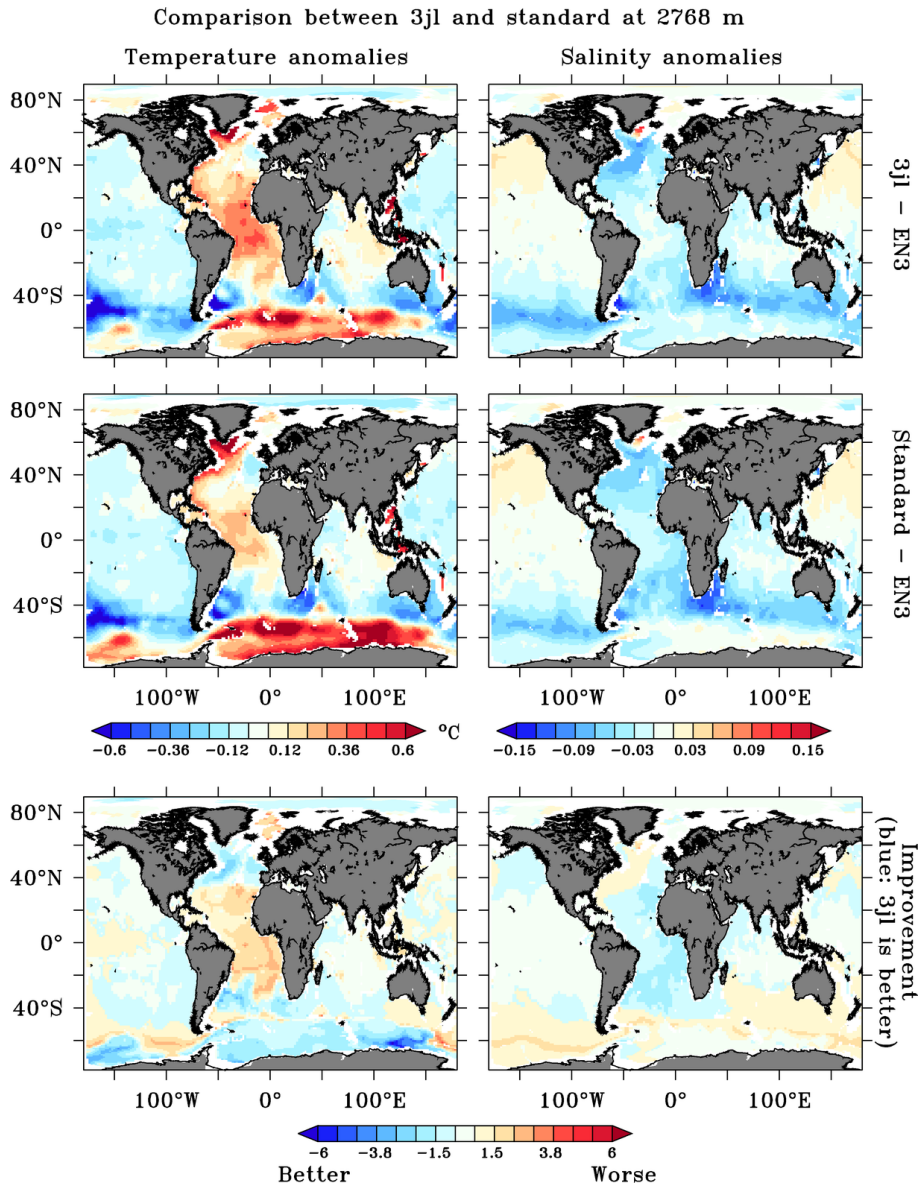
**Figure 14.** As figure 9 but showing T (left) and S (right) anomalies at 2768 m depth from EN3 1960-1990 climatology.

## 6 Higher resolution models: ORCA 1

We present an important and available tool for the tuning of complex ocean and climate models that we believe should be used by model developers developing a wide range of ESM components and by modelling centres developing models for CMIP6. However, a key distinction between models developed for CMIP6 and the $2^o$ version of NEMO we have explored here is 5 the available ensemble size. It might be argued, for example, that we were only able to focus our search for good models on
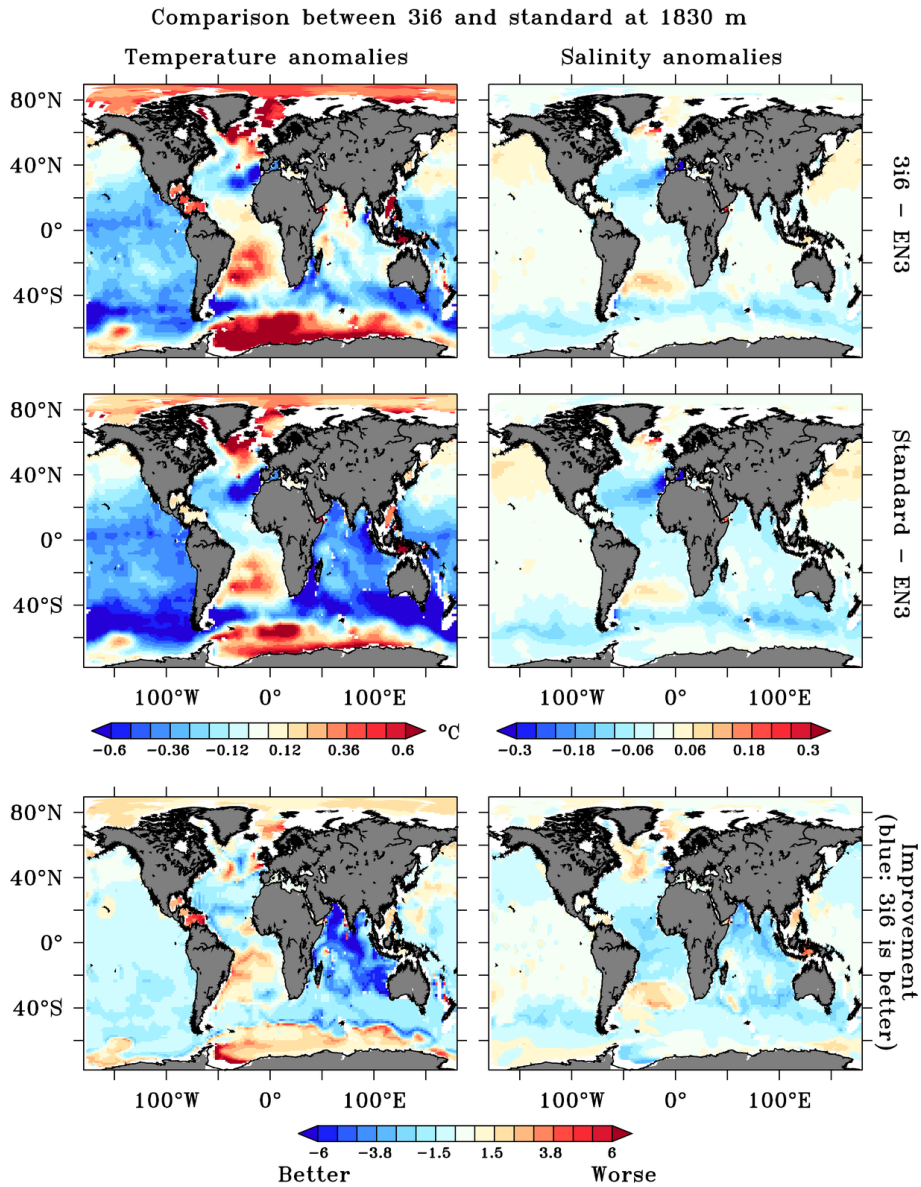
**Figure 15.** As figure 9 but for ensemble member 3i6 and showing T (left) and S (right) anomalies at 1830 m depth from EN3 1960-1990 climatology.

1.5% of the original parameter space because we were able to use 400 member ensembles and 150 year integrations. Such an argument may lead to the dismissal of the method for high resolution models where very few integrations can be done to assist in tuning.

5    However, the method has two key features that make it powerful and applicable at any resolution. The first is that it takes whatever information we do have and uses it to say which parts of parameter space can be ruled out, given all of our uncer-
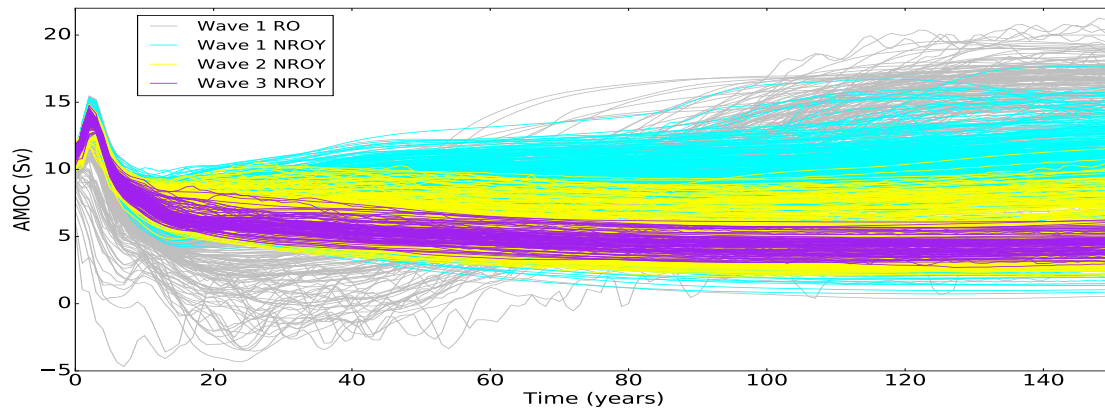
**Figure 16.** Atlantic meridional overturning circulation at $26^o$N and 1000 m depth for waves 1, 2 and 3.

tainties. This contrasts it starkly with optimisation methods that require us to find a good region or, typically, the single best parameter choice. If small ensembles mean that we rule out less parameter space, that does not preclude us from using the method to cut out what parameter space we can and our results will still be valid. The second feature is that the only thing we require is *an emulator* of the model in order to begin cutting out space and large ensembles do not provide the only means of

5  building emulators.

    Two examples of the flexibility are pertinent here. Firstly, dynamic emulators of time series (Conti et al., 2009; Liu and West, 2008; Williamson and Blaker, 2014) allow us to construct emulators for the way a model is evolving in time. This is one method of using ensembles of short integrations to build emulators for long integrations that could be used to refocus parameter space. Secondly, for most models, hierarchies of complexity are available that allow lower resolution versions or versions with

10  simpler physics to be used to run large ensembles that can help develop informative priors for many of the emulator parameters discussed in section 3.2 (Kennedy and O'Hagan, 2000; Cumming and Goldstein, 2009; Williamson et al., 2012; Le Gratiet, 2014). Highly informative priors developed using lower resolutions can dramatically reduce the size of ensemble required to build useful emulators for refocussing.

    We illustrate this in Figure 17 using the $1^o$ version of our model, ORCA1. The left panel shows ORCA1 temperature at

15  216 m depth (green dots) for a 32 member ensemble. The black dots and error bars are the diagnostic plots of a standard emulator for ORCA1 built using the methods we used to build ORCA2 emulators but using only the 32 runs to select the model parameters. The predicted run has been left out of the emulator fit for each error bar in the diagnostic plot (so this is the one point at a time version of Figure 3). The red solid horizontal line and the dashed lines either side represent the global mean temperature in EN3 at 216 m and the relevant uncertainty. What we see from this picture is that we are able to predict ORCA1

20  well with the emulator using an ensemble of only 32 members, in so far as each model run lies within our emulator uncertainty. However, our uncertainty is such that our emulator hardly helps us to rule out parameter space at all. We are so uncertain that
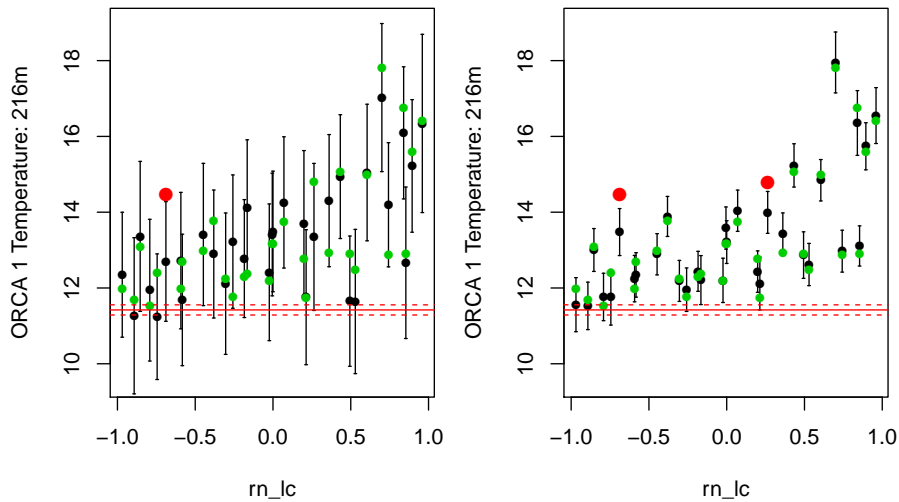
**Figure 17.** Leave one out plots for emulators of ORCA 1 trained using only a 32 member ensemble of ORCA1 (left panel) and using both the wave 1 ORCA2 ensemble and the same 32 member ORCA1 ensemble to build the emulator (right panel).

almost the whole parameter space is NROY. Alone, a 32 member ensemble is insufficient to rule out a considerable amount of the 21 dimensional parameter space.

The panel on the right uses the emulator for ORCA2 as a starting point for an ORCA1 emulator and uses the 32 member ensemble to effectively model differences between the models at the two resolutions. The result is a substantial reduction in our
5   emulator uncertainty for ORCA1. We can see that relative to the size of the errors on the observational constraint, our emulator is sufficiently accurate to enable us to rule out a considerable amount of parameter space using this model.

We build emulators for ORCA1 temperature at all depths used for refocussing in this paper using just the ORCA1 ensemble and using the ORCA2 informed emulators for ORCA1 and compare the reduction in parameter space if we were to treat this analysis as wave 1 of a history match. The history match using only the 32 member ensemble of ORCA1 lead to removing 37%
10   of the ORCA1 parameter space explored, whereas the history match using the ORCA2-informed emulators removed double that at 74% of space cut out. In this example then, we can halve the volume of the parameter space we are searching with a small ensemble if we also use information from a low resolution model to help build emulators, and consequently there is potential to achieve satisfactory levels of calibration with far less computational resource and hence to apply this methodology to higher (i.e. more costly to run) resolutions. Note also that our wave 1 ORCA1 space reduction is similar to our wave 1 space
15   reduction for ORCA2 (77.5%).

We have not presented a full example of tuning of ORCA1 due to resource limitations. We used this test ensemble to illustrate that the tuning method we advocate here is not only applicable if we have access to large ensembles. We do not provide details of tuning using model hierarchies, as we believe this subject is worthy of another paper.

## 7   Discussion

We have described and illustrated iterative refocussing for the ocean model NEMO run at $2^o$ resolution and argued for the method to be used for tuning complex numerical models of the ocean, atmosphere and climate. Iterative refocussing (also referred to as history matching in the statistics literature), is a method of automatic tuning that allows each of the different
5   sources of uncertainty present when comparing climate models to data to define a region of model parameter space that is consistent with the data we are using and allows us to focus the search for good models only in those subspaces. Though it could be used to find just one setting of the model parameters to represent the model in a MIP type experiment (e.g. CMIP6), we argue that it should be used to return a representative set of models that cannot be ruled out by comparison to observations, thus quantifying an important source of uncertainty in climate model inter-comparison.

10   In our application with NEMO we have shown that iterative refocussing is effective as a means to reduce biases in metrics of interest, such as global mean T and S properties. However, is it also apparent that regional biases can be large and persistent across large regions of parameter space, indicating that either there are structural deficiencies in the model, or that they arise as a result of external forcings or untested elements of the model. Some regional biases, for example those arising from the Mediterranean outflow into the northeast Atlantic presented earlier, may be sensitive to parameter choices, and it may fall to
15   the scientist to choose 'preferred' solutions from those within NROY space.

To obtain improved representations of the global ocean additional metrics targeted towards reducing biases in critical regions can be applied. Ideally observations with suitable measures of uncertainty should exist for each metric. Metrics and associated tolerances for error based on expert judgement can be defined but should be used with caution. Without observations the risk of overfitting a model (or worse, fitting it to an unrealistic value) is significant. Whilst we have demonstrated that it is possible
20   to find parameter choices for ORCA2 that substantially improve the representation of global mean T and S, we caution than many features of the ocean properties, dynamics and variability in simulations inside wave 3 NROY have not been examined here. Prior to using models within wave 3 NROY the characteristics of features important to the study should be assessed and where necessary metrics targeted towards reducing biases in key regions of interest should be introduced, but this is beyond the scope of this study. In comparison the reference configuration available through the NEMO website has been extensively
25   tested and studied. We have also demonstrated in section 6 that small ensembles can be used, in tandem with large ensembles of lower resolution models to tune high resolution models.

Though we are advocating for an automatic method of model tuning, we are not arguing that the method is a panacea for the expert judgement of model developers. Not only is the choice of metrics and parameters to vary crucial, but also the tolerance to model error present for each chosen metric and the range of each chosen parameter to search for good models.
30   Each wave of our procedure allows the modellers to assess the impact of the applied constraints and to use emulators to suggest alternative metrics to use or regions of parameter space that might be of special interest for the next wave of the experiment. The emulators themselves may also suggest unforeseen interaction of different parameterisations and potential re-parameterisations that may lead to significant improvement of the model. Tuning currently is a very manual, labour-intensive process and iterative refocussing does not remove that altogether, but does bring important and powerful tools to bear on the problem.

Throughout the paper we have been careful not to say that the NROY parameter space contains 'plausible' climates or even good models of any kind. We are so careful to do this so that we avoid over-tuning (what statisticians might call over-fitting). Tuning to partial observations always risks pushing the model too close to the observations you have at the expense of processes (elements of the climate model state vector) to which you have not tuned and may only partially understand. However, if the

5  model is "too far" from one set of observations, we can say that that model is unacceptable and not risk falling into the trap of over tuning. This is our approach here.

Also crucial to the avoidance of over-tuning is the knowledge of how much tuning is required. "How close is close enough"? The uncertainty in the observations defines a lower bound on this distance, however, for many metrics for which we might have observations that we want to use, these uncertainties are not reported. Routine reporting of these uncertainties across the

10  field would make the task of tuning models more transparent and robust. The quantification of structural uncertainty, even in the form of tolerance to error as we have described it in the paper, remains a challenge for the climate modelling community and for statisticians working with it.

We note that our approach in this paper has been to use only area integrated quantities averaged over time. More discretised forms of the data, for example, constraining parameter space using 2D fields as metrics, would provide larger constraint on

15  parameter space. Certainly emulators for spatial fields are available (Higdon et al., 2008; Sexton et al., 2011), however, the specification of observation and structural uncertainty becomes even further complicated. Tuning to 2D and 3D fields will be one of the next steps we take with the NEMO model.

## 8   Code and Data Availability

The NEMO source code can be obtained from http://www.nemo-ocean.eu. The output required for emulation throughout the

20  paper is provided in R format along with all emulators fitted and the R file "FindNROYandPlots.R" that demonstrates use of the emulators for history matching in R. Code is provided to alter emulator parameters and refit the model and the first 30 lines of the R source file explains what the tuneable statistical parameters are and provides a commented out worked example for interested readers. The R file reproduces Figures 6 and 7 and similar plots for the ensemble by recomputing NROY space using the emulators. 2D and 3D fields for reproducing other figures are available on request from the authors. Figure 8 was constructed using 1.6M emulator evaluations per panel and requiring the submission to the condor cluster at Exeter of a large number of calls to the provided R function ManyImplausibilities() in the same way as demonstrated in the provided code. We don't provide code to run the emulators on a cluster, as each cluster's architecture is different.

There are a number of formal R packages available to download for building and using emulators, for example, the R

5  library DiceKriging. The customised code provided here should not be seen as an exemplar for the optimal fitting of Gaussian Processes (both from an efficiency or a model selection point of view), nor is it in submission for review as such. The fitting of statistical models requires judgement, and the provided emulators represent the judgements/uncertainties of the lead author at the time of analysis. We provide code so that interested readers may explore the methodology and for illustrative purposes.

Readers wishing to fit emulators to their own models are invited to explore the code to see how the authors have implemented
10  the technology, but are advised to use a more robust public package such as DiceKriging in the first instance.

We strongly advise that these emulators are not re-used to tune a custom version of NEMO. The model response is likely
to be sensitive to a great many things in addition to the parameters we have varied here, and a different set up or forcing is
highly likely to lead to model output that is sufficiently different from that which trained the emulators. The authors welcome
enquiries on any aspect of the methods and the analysis in this paper.

15  *Acknowledgements.* Daniel Williamson was funded by EPSRC Fellowship No. EP/K019112/1. Adam Blaker and Bablu Sinha were sup-
ported by NERC's National Capability Programme. This work used the ARCHER UK National Supercomputing Service (http://www.archer.
ac.uk).

# References

Beck, J., Guillas, S. 2015. Sequential design with Mutual Information for Computer Experiments (MICE): Emulation of a Tsunami model. *arXiv*.

Brynjarsdottir, J., O'Hagan, A. 2014. Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, **30**, 114007 (24pp).

Bayarri, M. J., Berger, J. O., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R. J., Paulo, R., Sacks, J., and Walsh, D. 2007. Computer model validation with functional output. *The Annals of Statistics*, **35**, 1874–1906.

Conti, S., Gosling, J. P., Oakley, J. E. and O'Hagan, A. 2009. Gaussian process emulation of dynamic computer codes, *Biometrika*, **96**, 663-676.

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., Seheult, A. H., and Smith, J. A. 1997. Pressure matching for hydrocarbon reservoirs: A case study in the use of Bayes linear strategies for large computer experiments, in *Case studies in Bayesian statistics*, eds. Gatsonis, C., Hodges, J. S., Kass, R. E., McCulloch, R., Rossi, P., and Singpurwalla, N. D., New York: Springer-Verlag, vol. III, pp. 36–93.

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., Kanzow, T., Rayner, D., Barringer, M. O., Johns, W. E., Marotzke, J., Longworth, H. R., Grant, E. M., Hirschi, J. J-M., Beal, L. M., Meinen, C. S., Bryden, H. L., 2007. Temporal Variability of the Atlantic Meridional Overturning Circulation at 26.5N. *Science* **317**, 935, doi:10.1126/science.1141304

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

Dufresne, J. et al. 2013. Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. *Climate Dynamics*, **40**, 2123.

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

Yann Friocourt, Sybren Drijfhout, Bruno Blanke, and Sabrina Speich, 2005: Water Mass Export from Drake Passage to the Atlantic, Indian, and Pacific Oceans: A Lagrangian Model Analysis. J. Phys. Oceanogr., 35, 1206?1222. doi: http://dx.doi.org/10.1175/JPO2748.1

Fogli, P. G., Manzini, E., Vichi, M., Alessandri, A., Patara, L., Gualdi S., Scoccimarro, Masina, S., Navarra, A. 2009. INGV-CMCC Carbon (ICC): A Carbon cycle Earth system model. CMCC Research Papers RP0061. 31.

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.

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, doi:10.1002/2013JC009067.

Griffies, S. M., A. Biastoch, C. Boning, F. Bryan, G. Danabasoglu, E. P. Chassignet, M. H. England, R. Gerdes, H. Haak, R. W. Hallberg, W. Hazeleger, J. Jungclaus, W. G. Large, G. Madec, A. Pirani, B. L. Samuels, M. Scheinert, A. Sen Gupta, Camiel A. Severijns, H. L. Simmons, A.-M. Treguier, M. Winton, S. Yeager and J. Yin, 2009. Coordinated Ocean-ice Reference Experiments (COREs), Ocean Modelling, 26, 1-2, 1-46, doi:10.1016/j.ocemod.2008.08.007

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.

Hazeleger, W. et al. 2012. EC-Earth V2.2: description and validation of a new seamless earth system prediction model. *Climate Dynamics* **39** 2611. doi:10.1007/s00382-011-1228-5.

Hewitt, H. T. Copsey, D., Culverwell, I. D., Harris, C. M., Hill, R. S. R., Keen, A. B., McLaaren, A. J., Hunke, E. C. 2011. Design and implementation of the infrastructure of HadGEM3: the next-generation Met Office climate modelling system. *Geosci. Model Dev.* **4**, 223-253.

Higdon, H., Gattiker, J., Williams, B., and Rightley, M. 2008. Computer model calibration using high-dimensional output, *Journal of the American Statistical Association*, **103**, 570–583.

Hourdin, F., Mauritsen, T., Gettleman, A., Golaz, J.C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., Rauser, F., Rio, C., Tomassini, L., Watanabe, M., Williamson, D. (2015) The art and science of climate model tuning, *BAMS*, Revised Twice.

Ingleby, B., and M. Huddleston, 2007: Quality control of ocean temperature and salinity profiles - historical and real-time data. Journal of Marine Systems, 65, 158-175 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.

Jourdan D, Balopoulos E, Garcia-Fernandez M, Maillard C (1998) Objective analysis of temperature and salinity historical data set over the mediterranean basin. Technical report IEEE.

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.

Large WG, Yeager SG (2004) Diurnal to decadal global forcing for ocean and sea-ice models: The data sets and flux climatologies. Technical Report TN-460+STR(NCAR):105pp

Large WG, Yeager SG (2008) The Global Climatology of an Interannually Varying Air-Sea Flux Data Set. Climate Dynamics Doi:10.1007/s00382-008-0441-3

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

Levitus S, Conkright M, Boyer TP, O'Brian T, Antonov J, Stephens C, Johnson LSD, Gelfeld R (1998) World Ocean Database 1998. Technical report NESDIS 18, NOAA Atlas:346pp.

Liu, F. and West, M. 2008. A dynamic modelling strategy for Bayesian computer model emulation. *Bayesian Analysis*, **4(2)**, 393–412.

Locarnini, R. A., A. V. Mishonov, J. I. Antonov, T. P. Boyer, H. E. Garcia, O. K. Baranova, M. M. Zweng, C. R. Paver, J. R. Reagan, D. R. Johnson, M. Hamilton, and D. Seidov, 2013. World Ocean Atlas 2013, Volume 1: Temperature. S. Levitus, Ed., A. Mishonov Technical Ed.; *NOAA Atlas* **NESDIS 73**, 40 pp.

Loeppky, J. L., Moore, L. M., Williams, B. J. 2009. Batch sequential designs for computer experiments. *Journal of Statistical Planning and Inference* **140(6)**: 1452-1464.

Madec G (2008) NEMO ocean engine. Note du Pole de modélisation, Institut Pierre-Simon Laplace (IPSL), France 27:1288–1619.

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.

McCarthy, G. D., Smeed, D. A., Johns, W. E., Frajka-Williams, E., Moat, B. I., Rayner, D., Baringer, M. O., Meinen, C. S., Collins, J., Bryden, H. L. (2015) Measuring the Atlantic Meridional Overturning Circulation at 26N, *Progress in Oceanography*, **130**, 91–111, doi:10.1016/j.pocean.2014.10.006.

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., Storkey, D., Aksenov, Y., Alderson, S., Calvert, D., Graham, T., Hyder, P., Siddorn, J., Sinha, B. 2014. GO5.0: the joint NERC?Met Office NEMO global ocean model for use in coupled and forced applications. *Geosci. Model Dev.*, **7**, 1?24, doi:10.5194/gmd-7-1-2014

Morris, M. D., Mitchell, T. J. 1995. Exploratory designs for computational experiments. *J. Stat. Plan. Inference* **43**: 381–402.

Morris, D. E., Oakley, J. E., Crowe, J. A. 2014. A web-based tool for eliciting probability distributions from experts. *Environmental Modelling and Software* **52**: 1–4, ISSN 1364-8152, http://dx.doi.org/10.1016/j.envsoft.2013.10.010.

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

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

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E., C., Kaplan, A. 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.* **108**, 4407.

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

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, *Journal of Climate*, **22**, 3540–3557.

Salter, J. M., Williamson, D. 2016. A comparison of statistical emulation methodologies for multi-wave calibration, *Environmetrics*, Accepted.

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.

Smeed, D., McCarthy, G., Rayner, D., Moat, B. I., Johns, W. E., Baringer, M. O., Meinen, C. S. 2016. Atlantic-meridional overturning circulation observed by the RAPID-MOCHA-WBTS (RAPID-Meridional Overturning Circulation and Heatflux Array- Western Boundary Time Series) array at 26N from 2004 to 2015. British Oceanographic Data Centre - Natural Environment Research Council, UK.

Steele M, Morley R, Ermold W (2001) PHC: A global ocean hydrography with a high quality Arctic Ocean. Journal of Climate 14:2079–2087

Taylor, K.E., Stouffer, R.J., Meehl, G.A. (2012) An Overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485-498, doi:10.1175/BAMS-D-11-00094.1.

Timmermann A, Goosse H, Madec G, Fichefet T, Ethe C, Dulire V (2005) On the representation of high latitude processes in the ORCA-LIM global coupled sea-ice ocean model. Ocean Modelling 8:175–201

Vernon, I., Goldstein, M., and Bower, R. G. 2010. Galaxy formation: a Bayesian uncertainty analysis. *Bayesian Analysis* **5(4)**: 619–846, with Discussion.

20  Voldoire, A. et al. 2013. The CNRM-CM5.1 global climate model: description and basic evaluation. *Climate Dynamics* **40**, 2091.

Williamson, D. Goldstein, M. and Blaker, A. T. 2012. Fast Linked Analyses, for Scenario-based Hierarchies. *J. Roy. Stat. Soc. Ser. C* **61(5)**: 665–691.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P. Jackson, L., Yamazaki, K. 2013. History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dynamics* **41**: 1703-1729. doi:10.1007/s00382-013-1896-4.

Williamson, D., Blaker, A. T. 2014. Evolving Bayesian emulators for structurally chaotic time series with application to large climate models.

850  *SIAM/ASA J. Uncertainty Quantification*, **2(1)** 1-28.

Williamson, D., Vernon, I.R., 2014. Efficient uniform designs for multi-wave computer experiments, under review, arXiv:1309.3520.

Williamson, D., Blaker, A. T., Hampton, C., Salter, J. 2015. Identifying and removing structural biases in climate models with history matching. *Climate Dynamics*, **45**, 1299-1324.

Williamson, D. 2015. Exploratory ensemble designs for environmental models using k-extended Latin Hypercubes. *Environmetrics*, **26(4)**.

855  Yang, B., Qian, Y., Lin, G., Leung, L.R., Rasch, P.J., Zhang, G.J., McFarlane, S.A., Zhao, C., Zhang, Y., Wang, H., Wang, M., Liu, X. 2012. Uncertainty Quantification and Parameter Tuning in the CAM5 Zhang-McFarlane Convection Scheme and Impact of Improved Convection on the Global Circulation and Climate, *J. Geophys. Res.*, **118**, 395-415.

Zhang, T., Li, L., Lin, Y., Xue, W., Xie, D., Xu, H., Huang, X. 2015. An automatic and effective parameter optimization method for model tuning. *Geosci. Model Dev.*, **8**, 3579-3591.

860  Zuo, L. W., Qian, Y., Zhou, T. J., Yang, B. 2014. Parameter tuning and calibration of RegCM3 with MIT-Emanuel cumulus parameterization scheme of CORDEX East Asia domain. *J. Climate*, **27**, 7687-7701.

Zweng, M.M, J.R. Reagan, J.I. Antonov, R.A. Locarnini, A.V. Mishonov, T.P. Boyer, H.E. Garcia, O.K. Baranova, D.R. Johnson, D.Seidov, M.M. Biddle, 2013. World Ocean Atlas 2013, Volume 2: Salinity. S. Levitus, Ed., A. Mishonov Technical Ed.; *NOAA Atlas* **NESDIS 74**, 39 pp.