Uncertainty quantification of ocean parameterizations: application to the K-Profile-Parameterization for penetrative convection

A. N. Souza[1], G. L. Wagner[1], A. Ramadhan[1], B. Allen[1], V. Churavy[1], J. Schloss[1], J. Campin[1], C. Hill[1], A. Edelman[1], J. Marshall[1], G. Flierl[1], R. Ferrari[1]

[1]Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, United States

Key Points:

- A Bayesian methodology can be applied to turbulence parameterizations to probe parameterizations and better understand their biases and uncertainties.
- We can train parameterizations to match large eddy simulations.
- We can better understand the physics of parameterizations by applying a Bayesian methodology.

Corresponding author: A. N. Souza, `andrenogueirasouza@gmail.com`

Abstract

Parameterizations of unresolved turbulent processes in the ocean compromise the fidelity of large-scale ocean models used in climate change projections. In this work we use a Bayesian approach for evaluating and developing turbulence parameterizations by comparing parameterized models with observations or high-fidelity numerical simulations. The method obtains optimal parameter values, correlations, sensitivities, and more generally, likely distributions of uncertain parameters. We demonstrate the approach by estimating the uncertainty of parameters in the popular 'K-profile parameterization', using an ensemble of large eddy simulations of turbulent penetrative convection in the ocean surface boundary layer. We uncover structural deficiencies and discuss their cause. We conclude by discussing the applicability of the approach to Earth system models.

Plain Language Summary

Climate projections continue to be marred by large uncertainties, which originate in the poor representation of physical processes that occur at scales too small for climate models to properly simulate them, like clouds in the atmosphere and turbulent swirls in the ocean. We propose to develop more accurate representations of small physical ocean processes (parameterizations) trained with high resolution numerical simulations of small ocean patches. A Bayesian methodology is used to calibrate the parameterizations with the high resolution numerical simulations, to assess their fidelity and to identify improvements. Most importantly this approach provides estimates of the uncertainties in the parameterizations which can then be used to quantify uncertainties of climate models. While the approach is illustrated for a parameterization of ocean turbulence, it can be applied to any parameterization in climate models.

1 Introduction

The ocean components of Earth system models are complex systems that couple the resolved ocean circulation with a myriad of unresolved, parameterized and important physical processes. Parameterizations of unresolved physical processes often involve many uncertain parameters which are used to tune the model in an attempt to obtain a desired outcome (Hourdin et al., 2017). Moreover, each component, whether resolved or parameterized, interacts with all the others in nonlinear ways that lead to complex behavior which is sometimes difficult to understand and characterize.

Upper ocean turbulent mixing is a key parameterized process in ocean circulation models. The detailed fluid dynamics of upper ocean turbulent mixing are highly complex, involving surface boundary layer turbulence driven by buoyancy loss or winds, bottom boundary layer turbulence, lateral mixing due to baroclinic effects, and so forth. However, at least in principle, the governing fluid dynamical equations are known. The problem is that the computational resources required to resolve them and, at the same time, the global scale circulation, are not available and will not be for the foreseeable future (Schneider, Teixeira, et al., 2017).

A goal of this paper is to outline and illustrate a Bayesian framework to assess and improve parameterizations. We present a way forward which employs an ensemble of mixing process resolving simulations to train a chosen parameterization. The core idea is that the parameterization must represent the collective effect of sub-grid scale physics faithfully for all relevant relevant external forcings and mean climate states. This contrasts approaches that attempt to diagnose parameters directly from high-resolution simulations or to estimate values that perform well only in a particular experiment. It should be noted, however, that by restricting ourselves to understanding parameterizations in the context of sub-grid scale physics, we may miss out on important interactions with the rest of the climate system, e.g., the interaction of resolved lateral fluxes from the global ocean circulation with parameterized turbulent vertical mixing in the ocean. Nevertheless, studying one subgrid-scale process at a time is not an exercise in futility since it is a necessary first step to optimize a parameterization before considering the interactions with all other components of the full system.

We take a Bayesian perspective in our optimization of parameterizations. There are many ways in which a Bayesian framework can be used. Here we will explore one particular approach: characterizing the parameters of a parameterization via probability distributions. Thus, we will go beyond finding a point estimate for parameters. These probability distributions capture the notion of uncertainty and nonlinear correlations between parameters. Furthermore, they can then be used as prior distributions for parameter sensitivity studies in full climate models. This partially addresses a present deficiency in the current approach used to tune parameters in climate models. "Manual" tuning is done to obtain agreement between models and observations (Hourdin et al., 2017). Since parameters are often correlated, a parameter may be tuned to offset biases introduced by another parameter, resulting in parameterizations that no longer respect the subgrid-scale physics. The Bayesian framework automates parameter search in a way that ensures it respects the underlying physics of a parameter. The calibration of parameterization schemes in climate models has the potential to reduce biases as well as quantify the uncertainty of key climate variables, such as ocean heat content or climate sensitivity; however, innovation is required to make the Bayesian method practical and computationally feasible in the global model. One step towards this is to calculate prior distributions for parameters in a simplified setting, such as the local studies performed here, and then use computationally efficient methods for obtaining posterior distributions in the global climate model such as those proposed in (Schneider, Lan, et al., 2017; Albers et al., 2019; Cleary et al., 2020).

The focus here is to calculate prior distributions for parameters in ocean climate models. We do so by matching parameterizations to large eddy simulations, a philosophy similar in spirit to that which has been done in the atmospheric context for cloud parameterizations (Golaz et al., 2007). To make our discussion concrete we focus on the representation of convectively-driven turbulence in the upper ocean.

Our paper is organized as follows: In section 2 we describe the physical scenario in which we run our Large Eddy Simulations (LES) and parameterization. In section 3 we introduce Bayesian parameter estimation for the parameters in the K-Profile Parameterization (KPP) and perform the parameter estimation in the regime described by section 2. Finally, we end with a discussion in section 4.

2 Large eddy simulations and K-profile parameterization of penetrative convection

During the onset of winter at high latitudes, cooling at the ocean surface generates convective plumes that descend and mix the ocean surface boundary layer, see Marshall and Schott (1999) for a review. Near-surface mixing by convection generates a surface layer of uniform temperature and salinity called the 'mixed layer' which can reach depths of hundreds of meters.

At the base of the mixed layer, convective plumes penetrate further into a strongly-stratified region called the 'entrainment layer', where plume-driven turbulent mixing between the mixed layer and the ocean interior further cools the boundary layer. This process, in which the surface layer is cooled both at the surface and by turbulent mixing in the entrainment layer, is called penetrative convection. Penetrative convection is a crucial oceanic process for storing heat and carbon as well as setting the density structure of the deep ocean. Parameterizations of ocean surface boundary layer mixing must describe penetrative convection accurately. In this paper we evaluate the accuracy of the

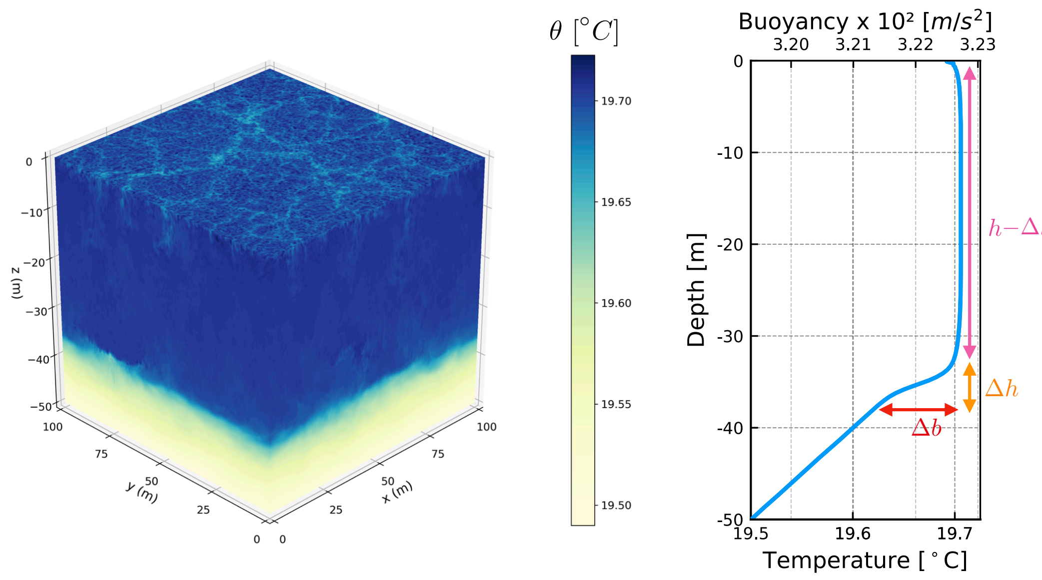

Figure 1. A 3D simulation of the LES model of the Boussinesq equations and its horizontal average at $t = 2$ days.

K-profile parameterization (Large et al., 1994) against large eddy simulations (LES) of idealized penetrative convection scenarios.

2.1 Idealized penetrative convection scenario

Our idealized scenarios impose a constant surface cooling $Q_h > 0$ to a resting, linearly stratified boundary layer with the initial state

$$\boldsymbol{u}|_{t=0} = 0 \text{ and } b|_{t=0} = N^2 z, \tag{1}$$

where $\boldsymbol{u} = (u, v, w)$ is the resolved velocity field simulated by LES, b is buoyancy, and N^2 is the initial vertical buoyancy gradient. The surface buoyancy flux Q_b is related to the imposed surface cooling Q_h, which has units $\mathrm{W\,m^{-2}}$, via

$$Q_b = \frac{\alpha g}{\rho_{\text{ref}} c_p} Q_h, \tag{2}$$

where $\alpha = 2{\times}10^{-4}\,(^\circ\mathrm{C})^{-1}$ is the thermal expansion coefficient, $g = 9.81\,\mathrm{m\,s^{-2}}$ is gravitational acceleration, $\rho_{\text{ref}} = 1035\,\mathrm{kg\,m^{-3}}$ is a reference density, and $c_p = 3993\,\mathrm{J/(kg\,^\circ C)}$ is the specific heat capacity. Our software and formulation of the large eddy simulations is discussed in section Appendix A.

Output of a large eddy simulation of turbulent penetrative convection in a domain $L_x = L_y = L_z = 100$ meters is in Figure 1. The left panel in Figure 1 visualizes the three-dimensional temperature field $\theta = \theta_0 + b/\alpha g$ associated with the buoyancy b, where $\theta_0 = 20^\circ$C is the surface temperature at $z = 0$. The right panel of Figure 1 shows the horizontally averaged buoyancy profile

$$\bar{b}(z,t) \equiv \frac{1}{L_x L_y} \int_0^{L_x} \int_0^{Ly} b(x,y,z,t)\mathrm{d}x\mathrm{d}y. \tag{3}$$

The visualization reveals the two-part boundary layer produced by penetrative convection: close to the surface, cold and dense convective plumes organized by surface cool-

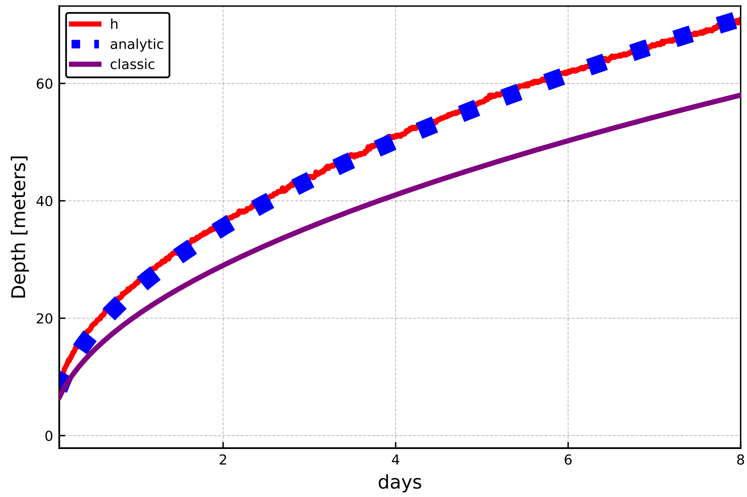

Figure 2. Mixed layer depth and its evolution in time after initial transients. The blue squares are the analytic scaling 4, the red line is an estimate of the boundary layer depth directly from the LES (described in the text), and the purple line is the classic scaling which ignores the entrainment layer 8.

ing sink and mix ambient fluid, producing a well-mixed layer that deepens in time. Below the mixed layer, the momentum carried by sinking convective plumes leads them to overshoot their level of neutral buoyancy (nominally, the depth of the mixed layer), 'penetrating' the stably stratified region below the surface mixed layer and generating the strongly stratified entrainment layer. The total depth of the boundary layer is h and includes the mixed layer and the entrainment layer of thickness Δh. Turbulent fluxes are negligible below $z = -h$ for our purposes.

In figure 2 we show the evolution of $h(t)$ defined as the first depth from the bottom where the stratification is equal to a weighted average of the maximum stratification and the initial stratification[1]. The dotted line confirms that the evolution after an initial transient is best fit by the formula,

$$h \simeq \sqrt{3.0 \frac{Q_b}{N^2} t}, \tag{4}$$

where N^2 is the initial stratification.

This result is easily explained by considering the horizontally averaged buoyancy equation,

$$\partial_t \overline{b} = -\partial_z \left(\overline{wb} + \overline{q^{(z)}} \right), \tag{5}$$

where $\overline{b}$ is the horizontally averaged buoyancy, $\overline{wb}$ is the horizontally averaged vertical advective flux and $\overline{q^{(z)}}$ is the horizontally averaged vertical diffusive flux. Integrating the equation in time between $t' = 0$ and some later time $t' = t$, and in the vertical between the surface, where $q^{(z)} = -Q_b$, and the base of the entrainment layer where all

[1] The weights are 2/3 for the initial stratification N^2 and 1/3 for the maximum stratification N_m^2 so that h satisfies $\partial_z \overline{b}(-h) = 2N_b^2/3 + N_m^2/3$. This guarantees that h is a depth where the local stratification lies between the background stratification and the maximum stratification since it is defined as the *first* depth starting from the bottom that satisfies such a criteria.

turbulent fluxes vanish, one finds,

$$\int_{-h}^{0} \left[\bar{b}(z,t)\ \mathrm{d}z - \bar{b}(z,0)\right]\ \mathrm{d}z = -Q_b t. \tag{6}$$

Substituting $\bar{b}(z,0) = b_0 + N^2(z+h)$ and $\bar{b}(z,t) = b_0 + \Delta b$, which is an appropriate approximation of the profile shown in Fig. 1b except at very early times in the simulation, yields

$$\tfrac{1}{2} N^2 h^2 - h \Delta b = Q_b t. \tag{7}$$

The first term on the left of equation 7 describes boundary layer deepening due to buoyancy extraction at the surface, while the second term corresponds to the further cooling caused by turbulent mixing in the entrainment layer. Ignoring turbulent mixing in the entrainment layer yields the deepening rate

$$h = \sqrt{2.0 \frac{Q_b}{N^2} t}, \tag{8}$$

which differs by roughly 20% from the best fit expression 4 due to the effects of turbulent mixing in the entrainment layer. The scaling in equation 8 is the deepening rate associated with a convective adjustment parameterization. The K-profile parameterization of penetrative convection, on the other hand, introduces a model for entrainment layer mixing in an attempt to describe equation 4.

2.2 The K-profile parameterization of penetrative convection

In penetrative convection in a horizontally-periodic domain, the K-profile parameterization models the evolution of the horizontally averaged temperature profile, $\bar{\theta}(z,t)$, and the boundary layer depth with

$$\partial_t T = -\partial_z F\left(T, h; \boldsymbol{C}\right) \tag{9}$$
$$0 = \mathcal{D}(T, h; \boldsymbol{C}), \tag{10}$$

where T is the temperature profile produced by the K-profile parameterization, h is the boundary layer depth, $\boldsymbol{C} = \{C^S, C^N, C^D, C^H\}$ is a set of free parameters for representing dimensionless proportionality constants following various scaling laws, $F(T, h; \boldsymbol{C})$ is a 'temperature flux function', and $\mathcal{D}(T, h; \boldsymbol{C})$ is a nonlinear-integral constraint to determine the boundary layer depth. We emphasize that the goal of the parameterization is not limited to just getting the mixed layer depth correct or correctly predicting the jump in buoyancy. Rather, the goal is to obtain correct heat exchanges with the atmosphere, entrainment of nutrients for the biology, and flux rates of passive scalars. Thus, it is important to faithfully capture the dynamics of the entire temperature profile.

The K-profile parameterization (KPP) represents F through the sum of a downgradient flux and a non-local flux term (Large et al., 1994),

$$F = -\underbrace{C^D w_\star h \, \tfrac{z}{h} \left(1 + \frac{z}{h}\right)^2}_{\equiv K} \partial_z T + \underbrace{C^N Q^\theta \, \tfrac{z}{h} \left(1 + \tfrac{z}{h}\right)^2}_{\equiv \Phi}, \tag{11}$$

for $-h \le z \le 0$ and 0 otherwise. Here $w_\star = (Q_b h)^{1/3}$ is the convective turbulent velocity scale, h is the boundary layer depth, $\frac{z}{h}\left(1 + \frac{z}{h}\right)^2$ is the 'K-profile' shape function—K is the namesake downgradient diffusivity of the K-profile parameterization—and Φ is a 'non-local' flux term that models convective boundary layer fluxes not described by downgradient diffusion.

In penetrative convection, the KPP model estimates the boundary layer depth h with an implicit nonlinear equation. To motivate the functional form of this criteria, first see figure 1 for reference. The jump in buoyancy, Δb, is the difference between the buoyancy in the mixed layer and the base of the entrainment region. Equivalently we can write $\Delta b = N_e^2 \Delta h$, where N_e the stratification in the entrainment region. From plume theory, see section Appendix B, we obtain $\Delta h \propto w_\star / N_e$ so that

$$C^H = \frac{\Delta b}{w_\star N_e} \tag{12}$$

for some universal proportionality constant C^H, which we call the 'mixing depth' parameter. KPP posits that the mixed layer depth h is the first such depth from the surface at which equation 12 holds. For numerical stability reasons equation 12 is generally formulated as

$$C^H = \frac{h\left[\frac{1}{C^{\mathrm{S}} h}\int_{-C^{\mathrm{S}} h}^{0} B(z)\mathrm{d}z - B(-h)\right]}{(hQ_b)^{1/3}\, h\sqrt{\max\left[0, \partial_z B(-h)\right] + 10^{-11}\mathrm{m}^2\mathrm{s}^{-2}}}, \tag{13}$$

where $B = \alpha g T$. The numerator approximates the jump in buoyancy times the mixed layer depth, $h\Delta b$. The term, $\frac{1}{C^{\mathrm{S}} h}\int_{-C^{\mathrm{S}} h}^{0} B(z)\mathrm{d}z$, serves as an approximation to the buoyancy in the mixed layer. The denominator evaluates the product $h w_\star N_e$ with $w_\star = (hQ_b)^{1/3}$ and $N_e = \sqrt{\max\left[0, \partial_z B(-h)\right]}$ at a given depth, while adding a dimensional term 10^{-11} to prevent division by zero. In section Appendix B we go in further detail about the rationale behind the implicit equation for the boundary layer depth, equation 13, for the case of penetrative convection.

The mixing depth parameter, C^H, is often referred to as the critical bulk Richardson number in the KPP literature (Large et al., 1994), because in mechanically forced turbulence, the denominator is replaced by the mean shear squared times h. In penetrative convection there is no mean shear and C^H is no longer related to a bulk Richardson number.

The K-profile parameterization for penetrative convection has four free parameters: the surface layer fraction C^{S}, the flux scalings C^N and C^D in equation 11, and the mixing depth parameter C^H in equation 13. Their default values, reported in (Large et al., 1994), are

$$(C^S, C^N, C^D, C^H) = (0.1, 6.33, 0.77, 0.95). \tag{14}$$

Our objective is to calibrate the free parameters $\boldsymbol{C} = (C^S, C^N, C^D, C^H)$ by comparing KPP temperature profiles $T(z, t; \boldsymbol{C})$ with the LES output $\bar{\theta}(z, t)$.

3 Model Calibration

We outline a Bayesian method for optimizing and estimating the uncertainty of the four free parameters through a comparison of solutions $T(z, t; \boldsymbol{C})$ to equation 9 to the output $\overline{\theta}(z, t)$ of our large eddy simulations. For this we define a loss function by

$$\mathcal{L}(\boldsymbol{C}) = \max_{t \in [t_1, t_2]} \left\{ \frac{1}{L_z} \int_{-L_z}^{0} \left[T(z, t; \boldsymbol{C}) - \overline{\theta}(z, t)\right]^2 dz \right\}. \tag{15}$$

We choose the square error in space to reduce the sensitivity to vertical fluctuations in the temperature profile. In time we take the maximum value of the squared error to guarantee that the temperature profile never deviates too far from the LES simulation at each instant.

Notably we do not use the boundary layer depth in the definition of the loss function. Firstly, it should be stressed that getting the entire temperature profile correct is a more stringent requirement and would also imply a correct boundary (and mixed) layer depth. We prefer not to use a boundary layer depth directly because it leads to noisy loss functions and depends too much on the precise definition used. In the literature, there are several different kinds of "depth" parameters based on, for example, the KPP definition as per equation 13, the location of the minimum buoyancy flux, the location of the maximum temperature gradient, or the first depth at which the temperature decreases by some ΔT of the surface value (Kara et al., 2000; Van Roekel et al., 2018). It is simpler (albeit more ambitious) to target the entire temperature profile. We prefer not to use the horizontally averaged temperature fluxes or gradients for practical reasons. Fluxes tend to be noisier than the horizontally averaged temperature profile and one would have to apply a smoothing filter. In summary, these other metrics introduce additional sources of systematic bias for little gain in the present circumstance.

A natural way to extend the definition of loss functions in order to take into account parameter sensitivities is to define probability distributions for parameters. Similar to how the functional form of the loss function is critical to the estimation of optimal parameters, the functional form of a probability density is critical for estimating the uncertainties of a parameter. A probability distribution quantifies what we mean by "good" or "bad" parameter choices, (similar to a loss function), but in terms of uncertainties and likelihoods. It is often the case that one has a good feel for how to define meaningful loss functions, but less so for probability distributions. Here we report our choices, but in section Appendix C we provide guidance on criteria to be used when constructing as well as sampling from the probability distribution. It is worth keeping in mind that, just like loss functions, the true test is "after-the-fact"; we inspect results and confirm that they indeed correspond to our intuition. Just like the definition of a loss function implicitly determines a choice of optimal parameters, a choice of probability distribution implicitly determines parameter sensitivities[2]. Both are arbitrary, but that does not mean that loss functions or parameter sensitivities are meaningless.

We adopt the same definition as in (Schneider, Lan, et al., 2017) for the probability distribution:

$$\rho\left(\mathbf{C}\right) \propto \rho^0(\mathbf{C}) \exp\left(-\frac{\mathcal{L}(\mathbf{C})}{\mathcal{L}_0}\right) \tag{16}$$

where ρ^0 is the prior distribution of the parameter values, $\mathcal{L}$ is a loss function, and $\mathcal{L}_0 > 0$ is a *hyperparameter*[3].

The loss function $\mathcal{L}$ has dimensions and the parameter $\mathcal{L}_0$ makesthe quantity in the exponent dimensionless. $\mathcal{L}_0$ could have been absorbed into the loss function, but it has a probabilistic interpretation that is worth emphasizing. We chose the parameter $\mathcal{L}_0$ as the minimum of the loss function $\mathcal{L}(\boldsymbol{C})$–the minimum is found using a modified[4] simulated annealing procedure to compute the minimum of $\mathcal{L}$ (Kirkpatrick et al., 1983). With this choice the likelihood of any other parameter choice, say $\boldsymbol{C}^1$, is determined by the amount by which it increases the minimum of the loss function, i.e.,

$$\rho(\boldsymbol{C}^1)/\rho(\boldsymbol{C}^*) = \exp\left(\frac{\mathcal{L}^0 - \mathcal{L}(\boldsymbol{C}^1)}{\mathcal{L}^0}\right), \tag{17}$$

where $\boldsymbol{C}^*$ denotes the optimal parameter choice with $\mathcal{L}^0 = \mathcal{L}(\boldsymbol{C}^*)$. For example, if the choice $\boldsymbol{C}^1$ increases the minimum of the loss function by a factor of two, i.e. $\mathcal{L}(\boldsymbol{C}^1) = 2\mathcal{L}^0$, then it is $1/e$ less likely.

---

[2] Parameter sensitivities are inversely related to parameter uncertainties. A more sensitive parameter is one that produces larger changes to the loss function. In the context of this paper a more uncertain parameter is one that produces small changes to the loss function. See section **??** for a simple example.

[3] A hyperparameter is a parameter associated with the probability distribution as opposed to a parameter in the parameterization.

[4] The main differences is that we take the minimum "artificial temperature associated with the simulated annealing procedure" to be the best known minimum of the loss function $\mathcal{L}$ rather than 0.

Once $\mathcal{L}_0$ is determined, we use the Random Walk Markov Chain Monte Carlo (RW-MCMC) algorithm (Metropolis et al., 1953), described further in section C2, to sample the probability distribution.

It is worth mentioning that equation 16 is the continuous analogue of Bayes formula

$$\mathbb{P}(\boldsymbol{C}|\text{data}) \propto \mathbb{P}(\boldsymbol{C})\mathbb{P}(\text{data}|\boldsymbol{C}) \tag{18}$$

where $\mathbb{P}$ is a probability distribution. In our context we interpret the formula as follows: We update our prior belief of the distribution of parameters $\boldsymbol{C}$ based on the data (in this case the LES experiment). $\mathbb{P}(\boldsymbol{C})$ is our prior probability for the parameters $\boldsymbol{C}$, while $\mathbb{P}(\text{data}|\boldsymbol{C})$ is the probability that the parameter choices $\boldsymbol{C}$ explain the data. Choosing to model parameters as probability distributions has the consequence that the output of the parameterization is also inherently probabilistic. In particular, the output of KPP will no longer be just a point estimate for temperature at each depth and each moment in time, but rather a probability distribution.

For all the uncertainty quantification that follows, we use resolution and timesteps typical of state of the art ocean models used for climate studies: a resolution of 100 m/16 = 6.25 m and a timestep of ten minutes. The temporal window used to compute the loss function is from $t_1 = 0.25$ days to the final simulation day. We apply the Bayesian parameter estimation procedure to KPP using data from one LES simulation in section 3.1 and from multiple LES simulations using different initial stratifications in section 3.2. We use a uniform prior for the parameters in KPP over the following ranges:

$$0 \leq C^S \leq 1,\ 0 \leq C^N \leq 8,\ 0 \leq C^D \leq 6,\ \text{and } 0 \leq C^H \leq 5. \tag{19}$$

The surface layer fraction C^S, being a fraction, must stay between zero and one. The other parameter limits were chosen to correspond to "reasonable" ranges around the default values, equation 14.

3.1 Calibration of KPP parameters against one LES simulation

In this section we apply the Bayesian calibration method to the LES simulation of penetrative convection described in section 2.1 and quantify uncertainties in parameters of KPP, section 2.2. The horizontal averages from the LES simulations are compared with predictions from solutions of the KPP diffusion scheme. The boundary and initial conditions for KPP are taken to be the same as those for the LES simulation, i.e., 100 W/m^2 cooling at the top, $\partial_z T = 0.01°\text{C m}^{-1}$ at the bottom, and an initial profile $T_p(z,0) = 20°\text{C} + 0.01°\text{C m}^{-1} z$.

We use the RW-MCMC algorithm with 10^6 iterations to sample the probability distributions of the four KPP parameters (C^S, C^N, C^D, C^H). This lead to roughly 10^4 statistically independent samples as estimated using an autocorrelation length, see Sokal (1997). The RW-MCMC algorithm generates the entire four dimensional PDF, equation 16, but visualizing this object is challenging. Instead we look at the marginal distributions, e.g.,

$$\rho_M(C^H) \equiv \iiint \rho(\boldsymbol{C})\ \mathrm{d}C^S \mathrm{d}C^D \mathrm{d}C^N, \tag{20}$$

and similarly for the other parameters. (Constructing the marginal distributions only requires constructing histograms of the trajectories generated by the RW-MCMC algorithm.) Parameter correlations are washed away by focusing on marginal distributions. Nevertheless, marginal distributions give the range of parameter values that yield little change to the loss function and are shown in figure 3. The marginal distribution of the mixing depth parameter C^H is much more compact than that of the other three parameters suggesting that it is the most sensitive parameter. The mixing depth parameter's importance stems from its control over both the buoyancy jump across the entrainment layer and the rate-of-deepening of the boundary layer. (Once again it may be useful to remember that C^H is often referred to the bulk Richardson number in the KPP literature, even though it take a different meaning in convective simulations.) The parameters C^K and C^N set the magnitude of the local and nonlocal fluxes and their specific value is not too important as long as they are large enough to maintain a well-mixed layer. The value of the regularization C^S is quite irrelevant.

The parameter distribution can be used to choose an optimal set of KPP parameters. Of the many choices, we choose the most probably value of the four dimensional probability distribution, the mode, because they minimize the loss function as explained in section Appendix C. (These values do not necessarily correspond to the individual modes of the marginal distributions. For example C^H is set to ≈ 2.0 rather than 1.5.) In figure 4a we show the area averaged temperature profile after 8 days from the LES simulation (continuous line) and the temperature profiles obtained running the KPP parameterization with default and optimal parameters (squares and dtots). The optimized temperature profiles are more similar to the LES simulation than the default value especially in the entrainment region. figure 4b confirms that the square root of the loss function, the error, grows much faster with the default parameters. The oscillations in the error are a consequence of the coarseness of the KPP model: only one grid point is being entrained at any given moment.

The improvement in boundary layer depth through optimization of the parameters is about 10%, or 10 m over 8 days. As we discussed in section 2.1, the rate of deepening can be predicted analytically within 20% by simply integrating over time and depth the buoyancy budget and assuming that the boundary layer is well mixed everywhere, i.e. ignoring the development of enhanced stratification within an entrainment layer at the base of the mixed layer. KPP improves on this prediction by including a parameterization for the entrainment layer. The default KPP parameters contribute a 10% improvement on the no entrainment layer prediction, and the optimized parameters contribute another 10%. While these may seem like modest improvements, they can result into large biases in boundary layer depth when integrated over a few months of cooling in winter rather than just 8 days. We will return to this point in the next section when we discuss structural deficiencies in the KPP formulation.

The probability distributions of the parameters can be used to predict the probability distributions of all variables, for example temperature at each depth and time, predicted by KPP. To do this, we subsample the 10^6 parameter values down to 10^4 and evolve KPP forward in time for each set of parameter choices. We construct histograms for the temperature field at the final time for each location in space individually. We then stack these histograms to create a visual representation of the model uncertainty. This uncertainty quantifies the sensitivity of the parameterization with respect to parameter perturbations as defined by the parameter distributions.

The histogram of temperature profiles at time $t = 8$ days as calculated by both our prior distribution (uniform distribution) and the posterior distribution (as obtained from the RW-MCMC algorithm) is visualized in figure 5. We see that there is a reduction of the uncertainty in the temperature profile upon taking into account information gained from the LES simulation. The salient features of the posterior distribution temperature uncertainty are

1. 0-10 meter depth: There is some uncertainty associated with the vertical profile of temperature close to the surface.
2. 20-60 meter depth: The mean profile of temperature in the mixed layer is very well predicted by KPP.
3. 60-70 meter depth: The entrainment region contains the largest uncertainties.

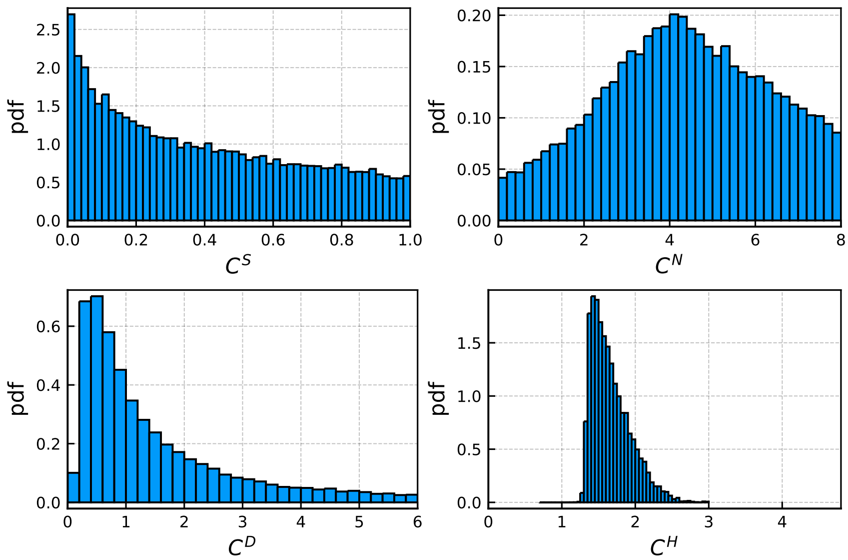

Figure 3. Parameter marginal posterior probability distributions. Marginal probability correspond to parameters parameters: C^S Surface Layer Fraction, C^N nonlocal diffusivity amplitude, C^D diffusivity amplitude, C^H mixing depth parameter. The probability distributions capture the notion of what parameter values are "good" and which ones are "bad". For example, in the pdf for C^H we see that a value of 2.5 is probable but a value of 5 would be not be. This intuitively corresponds to saying that a value of 2.5 would be a "reasonable" choice whereas 5 would be "unreasonable". The width of the C^S and C^N parameters suggest that KPP is quite insensitive to their values. A similar consequence holds for C^D, but there also seems to be a preference for values around one.

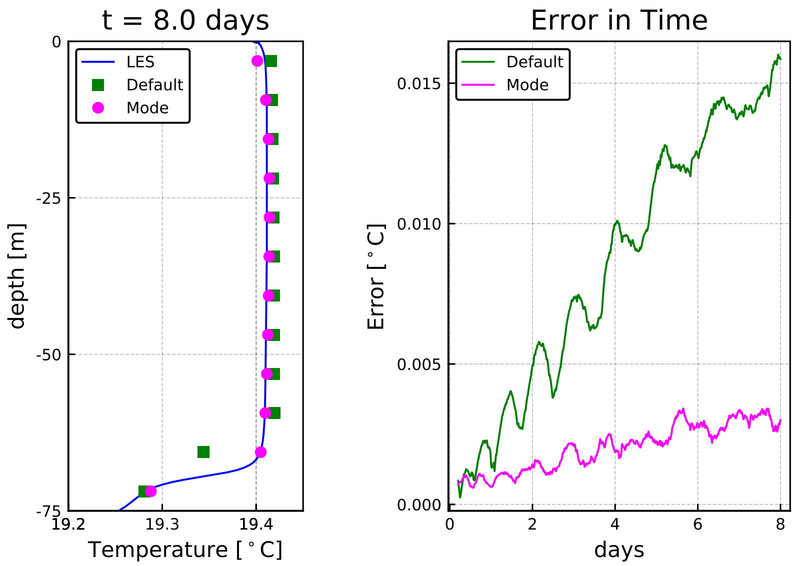

Figure 4. KPP and horizontally averaged LES temperature profiles for different point estimates of parameters at $t = 8$ days as well as the error in time. In the left plot, the squares correspond to default parameter choices, the circles correspond to the optimized parameterization (using the mode of the probability distribution), and the blue line to the horizontally averaged LES solution, all at time $t = 8$ days. On the right plot we show the instantaneous error at each moment in time. We see that the "optimal" parameter does indeed reduce the bias over the time period. The loss function is the largest square of the error over the time interval.

4. 70-100 meter depth: There is virtually no uncertainty. The unstratified region below the boundary layer does not change from its initial value.

Now that we have applied the Bayesian methodology to one LES simulation and explored its implications, we are ready to apply the method to multiple LES simulations covering different regimes in the following section. We focus on the optimization and uncertainty quantification of C^H for the remainder of the paper, since it is the most sensitive parameter. In the background, we are estimating *all* parameters.

3.2 Calibration of KPP parameters from multiple LES simulation

There are many possible directions that one could take at this point. We present an example of how we can use the methodology to explore bias in the KPP model. To this end we investigate what happens when we change the initial stratification in penetrative convection simulations. This is an informed decision motivated by recent work on mixed layer depth biases in the Southern Ocean (DuVivier et al., 2018; Large et al., 2019). In those studies, KPP failed to simulate deep mixed layer in winters when the subsurface summer stratification was strong.

We perform 32 large eddy simulations and calculate parameter distributions for each case. We kept the surface cooling constant at 100 W/m^2 for all regimes, and only varied the initial stratification. The integration time was stopped when the boundary layer depth filled about 70% of the domain in each simulation. We used 128^3 grid points, $\approx$

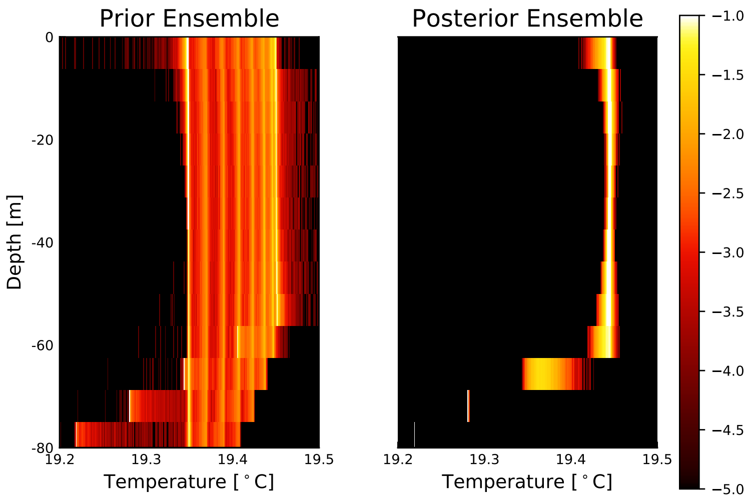

Figure 5. Uncertainty propagation of the temperature profile with respect to the prior and posterior probability distributions. The use of probability distributions for parameters has the consequence that the temperature field is no longer a point estimate, but rather a probability distribution at each moment in space and time. By sampling from the parameter probability distributions and evolving the parameterization forward in time, we obtain a succinct representation of what it means to "fiddle" with parameters. The legend on the right shows what the colors correspond to in terms of the base 10 logarithm of the probability distributions.

0.8 meter resolution in each direction[5]. Each one of the probability distributions used 10^5 iterations of RW-MCMC, leading to effective sample size on the order of 10^3.

The result, which is visualized in figure 6, shows that the parameter C^H depends on the background stratification, N^2. The blue dots are the median values of the probability distributions and the stars are the modes (minimum of the loss function). The error bars correspond to 90% probability intervals, meaning that 90% of parameter values fall between the error bars. The default KPP value is plotted as a dashed line for reference.

The median values and optimal values increase monotonically with the initial stratification value. Given that the parameter is supposed to be dimensionless, this reveals a systematic bias. Furthermore, it exposes *where* the systematic bias somes from: the boundary layer depth criterion in equation 13. No single value of C^H can correctly reproduce the deepening of the boundary layer for all initial stratifications.

The failure of the depth criterion can be understood by going back to the buoyancy budget in equation 7. Using the KPP estimate for the buoyancy jump across the entrainment layer,

$$\Delta b \equiv \frac{1}{C^{\mathrm{S}} h} \int_{-C^{\mathrm{S}} h}^{0} B(z) \mathrm{d}z - B(-h), \tag{21}$$

and introducing $N_h^2 \equiv \partial_z B(-h)$ for the stratification at the base of the entrainment layer to distinguish it from the interior stratification N^2, we find that the boundary layer

[5] Although the parameter estimates will vary upon using less resolution, the qualitative trends are expected to be robust.

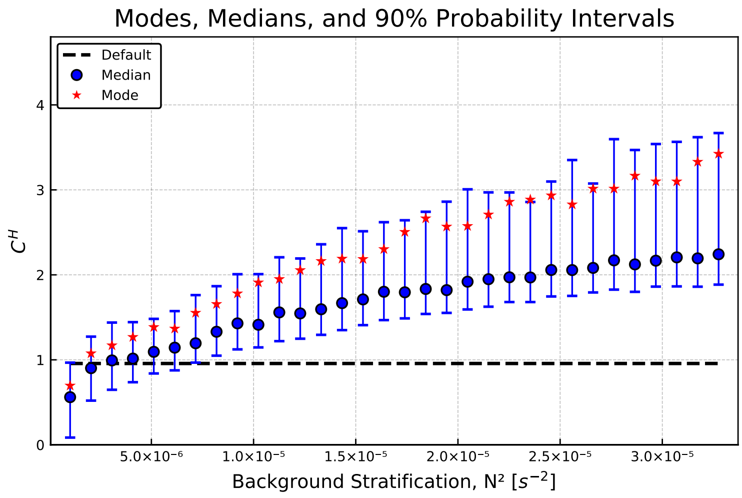

Figure 6. Mixing depth parameter optimized across various background stratification. The dots are the median values, the stars are the mode, and the error bars correspond to 90% probability intervals. The horizontal dashed line is the default value of the mixing depth parameter for reference. Here one can see that there mixing depth parameter when estimated across various regimes produces different results. This is a signature of a systematic bias in the parameterization.

depth criterion, equation 13, implies,

$$h\Delta b \simeq C^H h^{4/3} (Q_b)^{1/3} N_h. \tag{22}$$

Substituting this expression in the buoyancy budget, equation 7, one obtains an implicit equation for the evolution of the boundary layer depth h,

$$\left(\frac{1}{2}N^2 - C^H (Q_b)^{1/3} h^{-2/3} N_h\right) h^2 \simeq Q_b t. \tag{23}$$

The LES simulation described in section 2.1, and many previous studies of penetrative convection, e.g. (Van Roekel et al., 2018; Deardorff et al., 1980), show that the boundary layer depth grows as $\sqrt{t}$. N_h would have to scale as $h^{2/3}$ for KPP to correctly predict that deppening rate of the boundary layer, but this scaling is not observed in the LES simulations nor supported by theory.

3.3 Modification

From the multi-regime study of the previous section we found that there is no optimal KPP mixing depth parameter C^H that works for arbitrary initial stratification. This prompted us to look for an alternative formulation of the depth criterion which satisfies the well known empirical result that the boundary layer depth deepens at a rate,

$$h \simeq \sqrt{c\frac{Q_b}{N^2}t}, \tag{24}$$

where c is a dimensionless constant found to be close to 3.0 with the LES simulation in section 2.1. Furthermore, c was found to be close to 3.0 across all the numerical experiments from section 3.2. Substituting this expression in the buoyancy budget, equation 7, we find that,

$$\frac{\Delta b}{hN^2} \simeq \left(\frac{1}{2} - \frac{1}{c}\right). \tag{25}$$

This expression can then be used as a new boundary layer depth criterion that replaces equation 13,

$$C^\star = \frac{h\left(\frac{1}{C^S h}\int_{-C^S h}^{0} B(z)dz - B(-h)\right)}{N^2h^2 + 10^{-11}\mathrm{m}^2\mathrm{s}^{-2}}, \tag{26}$$

where $C^\star$ replaces C^H as the dimensionless parameter whose value sets the boundary layer depth. Based on equation 25, we expect

$$C^\star \simeq \left(\frac{1}{2} - \frac{1}{c}\right) \simeq \frac{1}{6}, \tag{27}$$

based on the LES result. The relation equation 26 is an implicit equation for h which guarantees that equation 24 holds. Once again, it may be useful to point out that $C^\star$ takes the place of what is generally referred to as the bulk Richardson number in the KPP literature, but that nomenclature is inappropriate for the case of penetrative convection where $C^\star$ parameterizes the effect of convective entrainment rather than shear mixing at the base of the mixed layer.

We now repeat the model calibration in section 3.2 with this new boundary layer depth criterion to test whether there is an optimal value of $C^\star$ that is independent of initial stratification. We estimate all KPP parameters and show the new mixing depth parameter for simulations with different initial stratifications in figure 7. There is no obvious trend in the optimal values of $C^\star$ and the error bars overlap for all cases. This validates the new criterion in that parameters estimated in different regimes are now consistent with one another. The uncertainties in $C^\star$ translate into an uncertainty in boundary layer depth prediction. In particular, values between $0.05 \leq C^\star \leq 0.2$ imply a boundary layer depth growth in the range $\sqrt{2.22tQ_b/N^2} \leq h \leq \sqrt{3.33tQ_b/N^2}$.

Additionally one can check if the constants estimated with the methodology of section 3 are consistent with an *independent* measure directly from the diagnosed LES simulation. In particular the LES simulations suggest that $C^\star \simeq 1/6$ as per equation 27. From figure 7 we see that the optimal $C^\star$ is somewhat smaller than $1/6 = 0.167$ (the dashed black line). A reason for this discrepancy is the neglect of curvature in the buoyancy budget, since we assumed a piece-wise linear buoyancy profile. Another one is the finite resolution in the model. A systematic source of error is how we diagnose the boundary layer depth: a different definition, such as the depth of maximum stratification, would yield a different scaling law (but still proportional to $\sqrt{t}$). At any rate the Bayesian parameter estimation bypasses these ambiguities / inconsistencies by direct comparison with the LES data.

We do not explore other modifications to the boundary layer depth criterion as this would greatly expand the scope of this article. The criterion described in this section assumes a constant initial stratification and a constant surface heat loss, which leads to the $\sqrt{t}$ growth of the boundary layer depth. It would be interesting to extend the criterion to arbitrary initial stratification, variable surface heat fluxes, not to mention the interaction with wind-driven mixing. The goal here was not to derive a new parameterization, but rather to introduce a methodology for obtaining meaningful parameterizations for climate models.

4 Discussion

In this work we have used a Bayesian methodology for estimating parameters in parameterizations of subgrid-scale physics as a first step towards parameter sensitivity studies for Earth Systems Models. We have calculated parameter probability distributions for parameters in the K-profile parametrization (KPP) by comparing with very high resolution simulations of ocean convection.

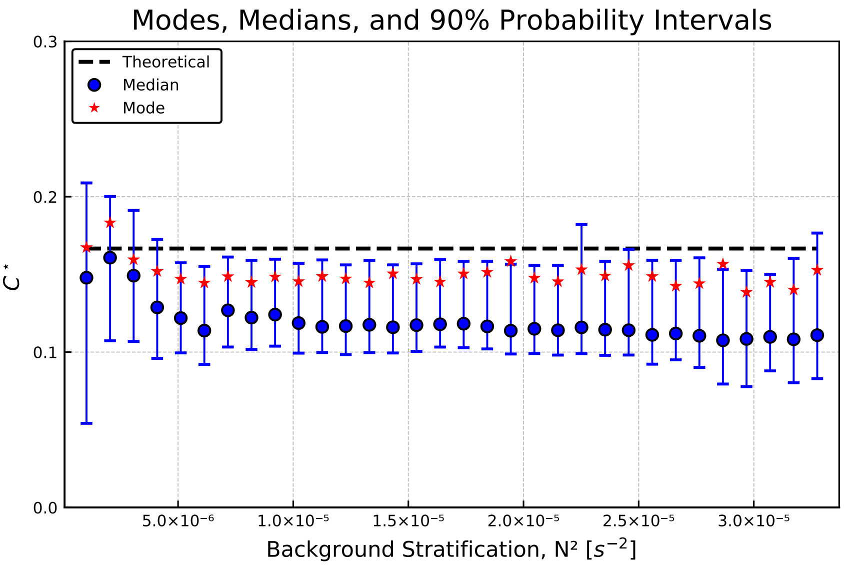

Figure 7. The modified mixing depth parameter optimized across various background stratification. The dots are the median values, the stars are the mode, and the error bars correspond to 90% probability intervals. The dashed line corresponds to 1/6, the theoretical expectation based on equation 27. This is similar to figure 6, but using the modification from section 3.3. Here one can see that there mixing depth parameter when estimated across various regimes produces similar results. This is a desirable feature in a parameterization.

This approach differs from current practice in ocean and climate modelling. Standard practice is to estimate parameters from a high resolution simulation or field campaign, or to tune parameters to reduce biases in global simulations. For example, the nondimensional amplitude of the KPP eddy diffusivity would be estimated as the ratio of the flux to the gradient from a single or a few high resolution simulations, (Van Roekel et al., 2018), or field campaigns, (Large et al., 1994). This assumes at the outset that the parameters calibrated for one test case will apply to all other scenarios, something that should be demonstrated rather than assumed. The other common approach is to tune the parameters in global models to reduce biases in climate relevant variables like ocean heat uptake or sea surface temperature (Menemenlis et al., 2005; Sraj et al., 2016). This can result in parameter choices that are inconsistent with the subgrid-scale physics they are supposed to parameterize. Our approach, instead, relies on a suite of high resolution simulations that span all the scenarios the parameterization is supposed to capture. Applying a Bayesian methodology, we then estimate the probability distributions for parameters which are consistent with the whole suite of high resolution simulations. It is worth pointing out that the methodology is computationally trivial once one has the LES solutions. The intellectual effort goes into identifying appropriate forms for the cost functions and probability distributions to guide the quantification of parameter values and their uncertainty.

We illustrated our approach to estimating KPP parameters for convection in a stratified ocean. We found that no unique set of parameters could capture the deepening of convection for different initial stratifications. We showed that a reformulation of the criterion to estimate the penetration depth of convection allowed us to find parameters that agreed well with the whole set of high resolution simulations. This shows the Bayesian approach is not only useful to estimate the probability distributions of parameter values in a parameterization, but it can also be used to identify and eliminate potential biases in parameterizations.

Ultimately, the hope is that parameter probability distributions estimated in local regimes will be useful for estimating uncertainties in global climate models; however, when coupling different components together, new parameters are introduced and nonlinear interactions between parameters can arise. Thus additional optimization is required for the full system and this requires innovation, because the methodologies described in this paper are not computationally feasibly when applied to larger systems. A promising approach for the global climate system is the Calibrate, Emulate, and Sample (CES) philosophy as outlined in (Cleary et al., 2020). In the CES approach, one uses a prior distribution for parameters (such as the ones calculated here for KPP) in a climate model in order to generate a preliminary ensemble of parameters. One then evolves this ensemble according to a loss function (appropriate to the global model) to generate a set of points that serve as a good “nodes” for interpolation of the loss function (or, alternatively, the “forward map”). A model, also called the emulator/surrogate model, is then chosen as an interpolator: this can be, for example, a Bayesian Neural Network or a Gaussian Process. Then one uses the interpolated function to calculate the probability of the posterior distribution using classic algorithms like the RW-MCMC method. In this way, one avoids rerunning the climate model and instead leverages as much information as possible from limited data. The surrogate model can then be used to update the prior distribution and improve predictions of the global model.

Stated differently we do not allow for arbitrary parameter perturbations when trying to match a climate model to data, parameter perturbations must take into account prior information. We propose obtaining this prior information by using highly resolved local simulations of turbulence. These experiments must be carefully designed and take into account suites of subgrid scale processes that one might expect to encounter in a global ocean model: vertical mixing, baroclinic effects, Langmuir turbulence, surface wave effects, bottom boundary layer turbulence, etc. If the global problem still exhibits significant biases after using all available prior information then this suggests that there is a fundamental deficiency in our understanding of how the different components of the climate system interact with one another. In this way one can start decoupling where biases in climate models come from. When the physics of local processes are well understood, then the additional uncertainties induced by coupling different regimes can be isolated and scrutinized. One no longer allows for biases to compensate for one another.

Appendix A Oceananigans.jl

Oceananigans.jl is open source software for ocean process studies written in the Julia programming language (Bezanson et al., 2017; Ramadhan et al., 2020; Besard et al., 2019). For the large eddy simulations (LESs) reported in this paper, Oceananigans.jl is configured to solve the spatially-filtered, incompressible Boussinesq equations with a temperature tracer equations. Letting $\boldsymbol{u} = (u, v, w)$ be the three-dimensional, spatially-filtered velocity field, θ be the conservative temperature, p be the kinematic pressure, f be the Coriolis parameter, and $\boldsymbol{\tau}$ and $\boldsymbol{q}$ be the stress tensor and temperature flux due to subfilter turbulent diffusion, the equations of motion are A1–A3,

$$\partial_t \boldsymbol{u} + (\boldsymbol{u} \cdot \nabla) \boldsymbol{u} + f \hat{z} \times \boldsymbol{u} + \nabla p = b \hat{z} - \nabla \cdot \boldsymbol{\tau}, \tag{A1}$$

$$\partial_t \theta + \boldsymbol{u} \cdot \nabla \theta = -\nabla \cdot \boldsymbol{q}, \tag{A2}$$

$$\nabla \cdot \boldsymbol{u} = 0. \tag{A3}$$

The buoyancy b appearing in A1 is related to conservative temperature by a linear equation of state,

$$b = \alpha g \left(\theta_0 - \theta \right), \tag{A4}$$

where $\theta_0 = 20°$ C is a reference temperature, $\alpha = 2 \times 10^{-4}\,(°\mathrm{C})^{-1}$ is the thermal expansion coefficient, and $g = 9.81\,\mathrm{m\,s^{-1}}$ is gravitational acceleration at the Earth's surface.

A1 Subfilter stress and temperature flux

The subfilter stress and momentum fluxes are modeled with downgradient closures, such that

$$\tau_{ij} = -2\nu_e \Sigma_{ij} \qquad \text{and} \qquad \boldsymbol{q} = -\kappa_e \nabla\theta, \tag{A5}$$

where $\Sigma_{ij} \equiv \frac{1}{2}\left(\partial_i u_j + \partial_j u_i\right)$ is the strain rate tensor, and ν_e and κ_e are the eddy viscosity and eddy diffusivity of conservative temperature. The eddy viscosity ν_e and eddy diffusivity κ_e in equation A5 are modeled with the anisotropic minimum dissipation (AMD) formalism introduced by (Rozema et al., 2015) and (Abkar et al., 2016), refined by (Verstappen, 2018), and validated and described in detail for ocean-relevant scenarios by (Vreugdenhil & Taylor, 2018). AMD is simple to implement, accurate on anisotropic grids (Vreugdenhil & Taylor, 2018), and relatively insensitive to resolution (Abkar et al., 2016).

A2 Numerical methods

To solve equations A1–A3 with the subfilter model in equation A5 we use the software package '`Oceananigans.jl`' written in the high-level Julia programming language to run on Graphics Processing Units, also called 'GPUs' (Bezanson et al., 2017; Besard et al., 2019; Besard et al., 2019). `Oceananigans.jl` uses a staggered C-grid finite volume spatial discretization (Arakawa & Lamb, 1977) with centered second-order differences to compute the advection and diffusion terms in equation A1 and equation A2, a pressure projection method to ensure the incompressibility of $\boldsymbol{u}$, a fast, Fourier-transform-based eigenfunction expansion of the discrete second-order Poisson operator to solve the discrete pressure Poisson equation on a regular grid (Schumann & Sweet, 1988), and second-order explicit Adams-Bashforth time-stepping. For more information about the staggered C-grid discretization and second-order Adams Bashforth time-stepping, see section 3 in (Marshall et al., 1997) and references therein. The code and documentation are available for perusal at `https://github.com/climate-machine/Oceananigans.jl`.

Appendix B Plume Model Derivation of the Mixing Layer Depth Criteria

We begin by considering the vertical momentum equation for a parcel punching through the transition layer,

$$w' \frac{dw'}{dz} \simeq -(b' - \bar{b}) \tag{B1}$$

where b' is the buoyancy of the parcel, assumed to be equal to the mixed layer value and $\bar{b}$ is the area mean buoyancy profile in the transition layer. This equation holds if the mean buoyancy profile is in hydrostatic balance and the area occupied by sinking plumes is small compared to the total area (Deardorff et al., 1980). Integrating from $z = -h + \Delta h$, where $w' \equiv w_e$, to $z = -h$, where the turbulence and particle descent vanish and hence $w' = 0$, gives,

$$(w_e)^2 \simeq N_e^2 \Delta h^2, \tag{B2}$$

assuming that the background stratification N_e^2 is constant in the entrainment layer. Introducing Δb as the difference between the buoyancy in the mixed layer and that at the base of the transition layer, we have, $\Delta\bar{b} = N_e^2 \Delta h$, and hence,

$$\Delta\bar{b} \propto w^* N_e, \tag{B3}$$

where we assumed that $w_e \propto w^*(-h + \Delta h)$. The criterion for diagnosing the mixing layer depth follows from this relationship; h is defined as the first depth z below the ocean surface where,

$$\frac{\Delta b_p(z)}{w^*(z) N_e(z)} = C^H, \tag{B4}$$

for some universal constant C^H.

Appendix C A Primer on on Probability Distributions

In defining a probability distribution, there are a few desirable features that enable us to make a direct connection to a loss function, $\mathcal{L}$:

1. In the limit of no uncertainty, a probability distribution should collapse to a delta function centered at optimal parameter values of the loss function.
2. The most probable value of the distribution should correspond to the optimal parameter in a loss function.
3. The uncertainty of a parameter value should be determined in terms of its effect on the loss function.

There are many probability distributions that satisfy the above criteria, but we choose

$$\rho(\boldsymbol{C}) \propto \rho^0(\boldsymbol{C}) \exp\left(-\mathcal{L}(\boldsymbol{C})/\mathcal{L}_0\right), \tag{C1}$$

where ρ^0 is a uniform prior distribution, $\mathcal{L}$ is a loss function, and $\mathcal{L}_0$ is a hyperparameter. In the following subsections we hope to elucidate why we made our particular choices. In section C1 we describe why we chose $\mathcal{L}_0$ as we did. In section C2 we go into detail on how to sample the probability distribution via the RW-MCMC algorithm as well as intuition for what it is doing.

C1 Explanation for our choice of $\mathcal{L}_0$

We now describe what the parameter $\mathcal{L}_0$ means in more detail. The limit $\mathcal{L}_0 \to 0$ corresponds to no uncertainty. In this limit, the probability distribution collapses to a delta function centered around the global optimal parameters of the loss function. An easy way to see this is to interpret the definition of the probability distribution as a Boltzmann-Gibbs distribution where the loss function corresponds to the energy of the system, and the constant $\mathcal{L}_0$ is analogous to the temperature of the system, kT where k is the Boltzmann constant, and T is the temperature. In the limit of zero temperature, the system collapses to the lowest energy state, in this case, the minimum of the loss function. The alternative limit $\mathcal{L}_0 \to \infty$ corresponds to an uncertainty that reduces to the prior distribution, ρ^0. In this case, information gained from loss function evaluations are uninformative. In analogy with the Boltzmann-Gibbs distribution this corresponds to infinite temperature and every energy state becomes equally likely (hence uninformative).

The maximum of the probability distribution is the mode. This value will be independent of $\mathcal{L}_0$ if the prior distribution is uniform. Indeed, letting $\mathbf{C}^*$ denote the (global) minimum of the loss function and $\mathbf{C}$ denote any other value, we get

$$\mathcal{L}(\mathbf{C}^*) \leq \mathcal{L}(\mathbf{C}) \;\Rightarrow\; \exp\left(-\mathcal{L}(\mathbf{C})/\mathcal{L}_0\right) \leq \exp\left(-\mathcal{L}(\mathbf{C}^*)/\mathcal{L}_0\right) \Rightarrow\; \rho(\mathbf{C}) \leq \rho(\mathbf{C}^*). \tag{C2}$$

Hence the minimum of the loss function is the most probable value of the probability distribution independent of $\mathcal{L}_0$ for a uniform prior distribution.

As mentioned in section 3, we choose the hyperparameter $\mathcal{L}_0$ to be the minimum of the loss function $\mathcal{L}$. Let us take a step back and explain *why* we used this definition.

Whatever the choice of $\mathcal{L}_0$, we would like the uncertainties of parameters to be independent of the units for which we use to evaluate the loss function. Furthermore, we would like for it to become smaller when there is less model bias and greater when there is more model bias. In other words, there is more uncertainty when the parameterization does a poor job of matching the "truth". This second criteria suggestions that $\mathcal{L}_0$ should be a monotonic function of the global minimum of the loss function (the bias) $\mathcal{L}(\mathbf{C}^*)$ where $\mathbf{C}^*$ denotes the optimal parameter values. The first criteria, coupled with the fact that a perfect model would output a value of 0 for the loss function, yields a number of choices. We use the simple form $\mathcal{L}_0 \propto \mathcal{L}(\mathbf{C}^*)$, as it is consistent with the previous sentence. Stated differently, we take $\mathcal{L}_0$ to be proportional to model bias. It is perhaps more correct to think of $\mathcal{L}_0$ as corresponding to *differences* in the loss function. Here we are using the difference between the "best" parameter choice and a (perhaps) non-existent parameter choice that would correspond to a perfect model whose loss function value is zero.

In the case that we have a *perfect* model and *perfect* data this would correspond to no uncertainty in the parameters and we would go back to having a point estimate for parameter values. Our choice here naturally assumes that both our data and our loss function are "perfect". If we have an idea of how imperfect our data or loss function may be, then this additional uncertainty should be taken into account in the choice of $\mathcal{L}_0$, or more generally, in the functional form of the probability distribution. We do not consider this additional source of uncertainty here.

The choice that we have made for this parameter may be thought of as how far we are willing to deviate from optimal values while still producing small changes in the loss function. In other words it is an e-folding length defined by the minimum of the loss function. In this sense, we do not prescribe the uncertainty of a parameter a priori, but instead, implicitly determine it from a choice of how far from optimal parameters that we are willing to deviate. One could spend a lifetime arguing about the finer details of a choice for this parameter, but ultimately this would detract from the real goal: to gain an understanding of what happens when one perturbs parameters away from optimal values. Any choice of $\mathcal{L}_0 > 0$ would do this. The important part is to make clear how such a choice is made and why.

Admittedly, in practice it is seldom possible to find the true global optimum of $\mathcal{L}$ and the best one could hope for is some approximate value that is the "best known" optimal value $\tilde{\mathbf{C}}$ to get an approximate $\tilde{\mathcal{L}}_0 \equiv \mathcal{L}(\tilde{\mathbf{C}})$. Since $\mathcal{L}_0 = \mathcal{L}(\mathbf{C}^*) \leq \mathcal{L}(\tilde{\mathbf{C}}) = \tilde{\mathcal{L}}_0$, our uncertainty is a conservative estimate (recall that a smaller $\mathcal{L}_0$ corresponds to *less* uncertainty).

C2 Random Walk Markov Chain Monte Carlo

We use standard methods to sample values from the probability distribution. The algorithm that we describe here is the Random Walk Markov Chain Monte Carlo Method (RW-MCMC), first used in (Metropolis et al., 1953). This method is most appropriate for loss functions where gradient information is either unavailable or prohibitively expensive to calculate. If it is possible to differentiate the loss function, then other methods may be more efficient at sampling from the distribution, such as Hamiltonian Monte Carlo. This alternative method is especially relevant when one wants to estimate a large number of parameters; however, here we are estimating four parameters, and there is no need for additional complexity.

The RW-MCMC algorithm, as the name suggests, performs a random walk in parameter space $\mathbf{C}$. It stays in regions of high probability more often, thereby allowing one to take the trajectories of the random walk and construct histograms that are directly related to the empirical distribution of underlying probability distribution function.

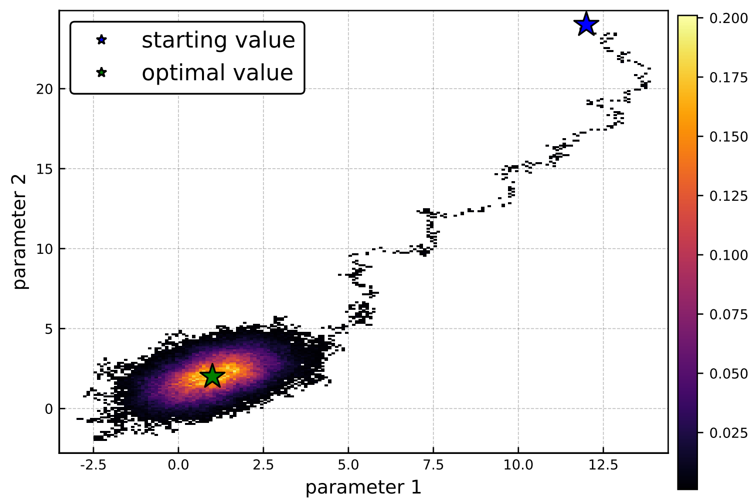

Figure C1. A histogram plot of RW-MCMC output for the target probability distribution equation C3 starting from a suboptimal value $\boldsymbol{C}^0 = [12.0\ 24.0]^T$ and using 10^5 iterations. The white space signifies regions that the algorithm did not explore. The dark regions correspond to places that were rarely visited whereas the red and yellow regions correspond to places that were visited more often. The trail from the starting value to the optimal value illustrates how the random walk is biased towards regions of ever increasing probability.

We further illustrate what the algorithm does by considering the following toy loss function:

$$\mathcal{L}(\boldsymbol{C}) = \frac{1}{2}(\boldsymbol{C} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{C} - \boldsymbol{\mu}) \ , \ \boldsymbol{\mu} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \ , \text{ and } \Sigma = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}, \tag{C3}$$

with the choice $\mathcal{L}_0 = 1$ in the probability distribution. In figure C1 we show a typical output from an RW-MCMC algorithm for this probability distribution. Note that the optimal value of the loss function is $\boldsymbol{\mu} = [1\ 2]^T$. Starting from a poor initial guess, here $\boldsymbol{C}^0 = [12.0\ 24.0]^T$, the algorithm goes towards regions of higher probability (lower cost function) by randomly choosing which direction to go. Once a region of high probability is found, in this case parameter values close to $\boldsymbol{\mu}$, the parameters hover around the minima of the loss function in a way that is consistent with the target probability distribution.

We now go into detail on RW-MCMC algorithm. Ratios of the probability distribution play a prominent role in the RW-MCMC algorithm; however, due to finite arithmetic considerations it is actually better to work with the logarithms of the probability distribution. A convenient way to do this is to use the negative log-likelihood function as $\ell = -\ln \rho$. This function can be thought of as being essentially the same as the loss function 15, but shifted and scaled. Denote elements of a sequence of parameter values by $\mathbf{C}_i$. The RW-MCMC algorithm is:

1. Choose initial parameter values $\boldsymbol{C}_0$. One choice for this parameter is the best known minimizer of the log-likelihood function using standard minimization techniques.
2. Calculate a "proposal parameter" $\tilde{\boldsymbol{C}}_1$. This will be described in more detail later.
3. Calculate $\Delta\ell = \ell(\boldsymbol{C}_0) - \ell(\tilde{\boldsymbol{C}}_1)$. This is a measure of how much more likely $\tilde{\boldsymbol{C}}_1$ is relative to $\boldsymbol{C}_0$.

4. Draw a random uniform random variable from the interval $[0, 1]$, e.g, calculate $u = \mathcal{U}(0, 1)$. This is used to determine whether or not to accept $\tilde{\boldsymbol{C}}_1$ as a new parameter.
5. If $\log(u) < \Delta\ell$ set $\boldsymbol{C}_1 = \tilde{\boldsymbol{C}}_1$. Otherwise set $\boldsymbol{C}_1 = \boldsymbol{C}_0$. This is the "accept / reject" step. Note that if $\Delta\ell > 0$, i.e. the proposed parameter produces a smaller output in the negative log-likelihood function, the proposal is always accepted.
6. Repeat steps 2-5 for $\boldsymbol{C}_i$, replacing $\boldsymbol{C}_0 \to \boldsymbol{C}_i$ and $\boldsymbol{C}_1 \to \boldsymbol{C}_{i+1}$, to generate a sequence (or chain) of parameter values.

Interpreting the negative log-likelihood function as a potential function, the algorithm may be succinctly stated as "always go downhill, sometimes go uphill". The sequence of parameter values generated by this algorithm can then be used to construct any statistics of the probability distribution 16, including empirical distributions, marginal distributions, and joint distributions. In the context of KPP this can also generate the uncertainty of a temperature at a given point in space and time as well as the uncertainty of the mixed layer depth at a given time.

This random walk is different from a random number generator in that successive samples are not independent of one another but are instead correlated. The random walk must be run for enough time to generate a sufficient number of statistically independent samples. The proposal step is crucial to ensure this feature. Thus, we will now describe how to choose a proposal in more detail. If there are no restrictions on the range of parameter values, then one can perturb each parameter by a Gaussian random variable with mean zero and covariance matrix Σ, i.e.

$$\tilde{\boldsymbol{C}}_{i+1} = \boldsymbol{C}_i + \mathcal{N}(0, \Sigma) \tag{C4}$$

Interestingly, the algorithm is guaranteed to work *independent* of the choice of Σ as long as the covariance matrix Σ is nonzero and the same proposal is used throughout the random walk[6]; however, suitable choices of Σ can speed up convergence to the probability distribution. At the end of an RW-MCMC run one can diagnose the "number of independent samples" by using approximations of the correlation length, see Sokal (1997). If Σ is too small then the acceptance rate will be too large since each proposal parameter is barely any different from the original parameter. Too large of a proposal often yields too low acceptance rates since it is typically easier to propose a parameter associated with a region of low probability than high probability (thereby making it likely that one is choosing a point that is "uphill" more often than "downhill"). One option is to take Σ to be a diagonal matrix whose diagonal elements are proportional to the square 10% of the default parameter values, i.e., the standard deviation of the proposal of each was about 10% and that each parameter component proposal was independent. A common option is to choose Σ according to the covariance matrix of the prior distribution. Yet, another option is to perform a preliminary random walk to estimate the covariance of the target distribution and then use this estimated covariance matrix in a new random walk. In general, there is no rule that will always speed up convergence, but we found that the last method to gave the best results.

If we would like to restrict the parameters to be in a finite range it is as simple as making the random walk take place in a periodic domain in parameter space. Another option is to redefine the loss function so that it outputs infinity if one plugs in a value outside a specified range. We opt for the former since it does not "waste" function evaluations.

[6] If one decides to change the proposal one needs to start the random walk over and cannot reuse data generated from another proposal.

Acknowledgments
The authors would like to thank Carl Wunsch, Tapio Schneider, Andrew Stuart, and William Large for numerous illuminating discussions. Our work is supported by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program, and by the National Science Foundation under grant AGS-6939393.

References

Abkar, M., Bae, H. J., & Moin, P. (2016). Minimum-dissipation scalar transport model for large-eddy simulation of turbulent flows. *Physical Review Fluids*, *1*(4), 041701. doi: 10.1103/PhysRevFluids.1.041701

Albers, D. J., Blancquart, P.-A., Levine, M. E., Seylabi, E. E., & Stuart, A. (2019). Ensemble kalman methods with constraints. *Inverse Problems*, *35*(9), 095007. doi: 10.1088/1361-6420/ab1c09

Arakawa, A., & Lamb, V. R. (1977). Computational design of the basic dynamical processes of the ucla general circulation model. In J. Chang (Ed.), *General circulation models of the atmosphere* (Vol. 17, p. 173 - 265). Elsevier. doi: 10.1016/B978-0-12-460817-7.50009-4

Besard, T., Churavy, V., Edelman, A., & Sutter, B. D. (2019). Rapid software prototyping for heterogeneous and distributed platforms. *Advances in Engineering Software*, *132*, 29 - 46. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0965997818310123` doi: https://doi.org/10.1016/j.advengsoft.2019.02.002

Besard, T., Foket, C., & De Sutter, B. (2019, April). Effective extensible programming: Unleashing julia on gpus. *IEEE Transactions on Parallel and Distributed Systems*, *30*(4), 827-841. doi: 10.1109/TPDS.2018.2872064

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, *59*(1), 65–98. doi: 10/f9wkpj

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., & Stuart, A. M. (2020). Calibrate, emulate, sample. *arXiv:2001.03689 [stat.CO]*. Retrieved from `https://arxiv.org/abs/2001.03689`

Deardorff, J. W., Willis, G. E., & Stockton, B. H. (1980). Laboratory studies of the entrainment zone of a convectively mixed layer. *Journal of Fluid Mechanics*, *100*(1), 41–64. doi: 10.1017/S0022112080001000

DuVivier, A. K., Large, W. G., & Small, R. J. (2018). Argo observations of the deep mixing band in the southern ocean: A salinity modeling challenge. *Journal of Geophysical Research: Oceans*, *123*(10), 7599-7617. doi: 10.1029/2018JC014275

Golaz, J.-C., Larson, V. E., Hansen, J. A., Schanen, D. P., & Griffin, B. M. (2007). Elucidating model inadequacies in a cloud parameterization by use of an ensemble-based calibration framework. *Monthly Weather Review*, *135*(12), 4077-4096. doi: 10.1175/2007MWR2008.1

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., ... Williamson, D. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*(3), 589-602. doi: 10.1175/BAMS-D-15-00135.1

Kara, A. B., Rochford, P. A., & Hurlburt, H. E. (2000). An optimal definition for ocean mixed layer depth. *Journal of Geophysical Research: Oceans*, *105*(C7), 16803-16821. doi: 10.1029/2000JC900072

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. doi: 10.1126/science.220.4598.671

Large, W. G., McWilliams, J. C., & Doney, S. C. (1994). Oceanic vertical mixing: A review and a model with a nonlocal boundary layer parameterization. *Reviews of Geophysics*, *32*(4), 363-403. doi: 10.1029/94RG01872

Large, W. G., Patton, E. G., DuVivier, A. K., Sullivan, P. P., & Romero, L. (2019). Similarity theory in the surface layer of large-eddy simulations of the wind-, wave-, and buoyancy-forced southern ocean. *Journal of Physical Oceanography*, *49*(8), 2165-2187. doi: 10.1175/JPO-D-18-0066.1

Marshall, J., Adcroft, A., Hill, C., Perelman, L., & Heisey, C. (1997). A finite-volume, incompressible navier stokes model for studies of the ocean on parallel computers. *Journal of Geophysical Research: Oceans*, *102*(C3), 5753-5766. doi: 10.1029/96JC02775

Marshall, J., & Schott, F. (1999). Open-ocean convection: Observations, theory, and models. *Reviews of Geophysics*, *37*(1), 1-64. doi: 10.1029/98RG02739

Menemenlis, D., Fukumori, I., & Lee, T. (2005, 05). Using green's functions to calibrate an ocean general circulation model. *Monthly Weather Review*, *133*, 1224-1240. doi: 10.1175/MWR2912.1

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087-1092. doi: 10.1063/1.1699114

Ramadhan, A., Wagner, G. L., Hill, C., Campin, J.-M., Churavy, V., Besard, T., … Ferrari, R. (2020). Oceananigans.jl: Fast and friendly geophysical fluid dynamics on GPUs. *The Journal of Open Source Software*, *4*(44), 2018. Retrieved from `https://joss.theoj.org/papers/aeb75d5b99d40ae6e0c8a3a4f09f3285` (Under review.)

Rozema, W., Bae, H. J., Moin, P., & Verstappen, R. (2015). Minimum-dissipation models for large-eddy simulation. *Physics of Fluids*, *27*(8), 085107. doi: 10.1063/1.4928700

Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, *44*(24), 12,396-12,417. doi: 10.1002/2017GL076101

Schneider, T., Teixeira, J., Bretherton, C., Brient, F., Pressel, K., Schär, C., & Siebesma, A. (2017, 01). Climate goals and computing the future of clouds. *Nature Climate Change*, *7*, 3-5. doi: 10.1038/nclimate3190

Schumann, U., & Sweet, R. A. (1988). Fast fourier transforms for direct solution of poisson's equation with staggered boundary conditions. *Journal of Computational Physics*, *75*(1), 123 - 137. doi: https://doi.org/10.1016/0021-9991(88)90102-7

Sokal, A. (1997). Monte carlo methods in statistical mechanics: Foundations and new algorithms. In C. DeWitt-Morette, P. Cartier, & A. Folacci (Eds.), *Functional integration: Basics and applications* (pp. 131–192). Boston, MA: Springer US. doi: 10.1007/978-1-4899-0319-8_6

Sraj, I., Zedler, S. E., Knio, O. M., Jackson, C. S., & Hoteit, I. (2016). Polynomial chaos–based bayesian inference of k-profile parameterization in a general circulation model of the tropical pacific. *Monthly Weather Review*, *144*(12), 4621-4640. Retrieved from `https://doi.org/10.1175/MWR-D-15-0394.1` doi: 10.1175/MWR-D-15-0394.1

Van Roekel, L., Adcroft, A. J., Danabasoglu, G., Griffies, S. M., Kauffman, B., Large, W., … Schmidt, M. (2018). The kpp boundary layer scheme for the ocean: Revisiting its formulation and benchmarking one-dimensional simulations relative to les. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2647-2685. doi: 10.1029/2018MS001336

Verstappen, R. (2018). How much eddy dissipation is needed to counterbalance the nonlinear production of small, unresolved scales in a large-eddy simulation of turbulence? *Computers & Fluids*, *176*, 276–284. doi: 10.1016/j.compfluid.2016.12.016

Vreugdenhil, C. A., & Taylor, J. R. (2018). Large-eddy simulations of stratified plane Couette flow using the anisotropic minimum-dissipation model. *Physics* *of Fluids*, *30*(8), 085104. doi: 10.1063/1.5037039