JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2022MS003245

Key Points:

- We calibrate tropical parameters in a gravity wave parameterization to obtain selected properties of the Quasi-Biennial Oscillation
- We use a Gaussian process to emulate an intermediate complexity climate model and then learn a distribution of gravity wave parameters
- We explore the gravity wave parametric uncertainty of the Quasi-Biennial Oscillation period and amplitude in a double CO_2 scenario

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

L. A. Mansfield,
lauraman@stanford.edu

Citation:

Mansfield, L. A., & Sheshadri, A. (2022). Calibration and uncertainty quantification of a gravity wave parameterization: A case study of the Quasi-Biennial Oscillation in an intermediate complexity climate model. *Journal of Advances in Modeling Earth Systems*, *14*, e2022MS003245. https://doi.org/10.1029/2022MS003245

Received 10 JUN 2022
Accepted 21 OCT 2022

Calibration and Uncertainty Quantification of a Gravity Wave Parameterization: A Case Study of the Quasi-Biennial Oscillation in an Intermediate Complexity Climate Model

L. A. Mansfield[1] and A. Sheshadri[1]

[1]Department of Earth System Science, Stanford University, Stanford, CA, USA

Abstract The drag due to breaking atmospheric gravity waves plays a leading order role in driving the middle atmosphere circulation, but as their horizontal wavelength range from tens to thousands of kilometers, part of their spectrum must be parameterized in climate models. Gravity wave parameterizations prescribe a source spectrum of waves in the lower atmosphere and allow these to propagate upwards until they either dissipate or break, where they deposit drag on the large-scale flow. These parameterizations are a source of uncertainty in climate modeling which is generally not quantified. Here, we explore the uncertainty associated with a non-orographic gravity wave parameterization given an assumed parameterization structure within a global climate model of intermediate complexity, using the Calibrate, Emulate and Sample (CES) method. We first calibrate the uncertain parameters that define the gravity wave source spectrum in the tropics, to obtain climate model settings that are consistent with properties of the primary mode of tropical stratospheric variability, the Quasi-Biennial Oscillation (QBO). Then we use a Gaussian process emulator to sample the calibrated distribution of parameters and quantify the uncertainty of these parameter choices. We find that the resulting parametric uncertainties on the QBO period and amplitude are of a similar magnitude to the internal variability under a $2xCO_2$ forcing.

Plain Language Summary Atmospheric gravity waves are excited in the lower atmosphere by disturbances such as mountains, convection and fronts. They travel upwards and break in the upper atmosphere, thus modifying the mean flow. This has large effects on the circulation, including driving a tropical oscillation. Gravity waves have a wide range of spatial scales and a large portion of these are smaller than the grid size of a climate model. This means they cannot be resolved and instead, they are represented through approximations called "parameterizations", which introduce a source of uncertainty in climate model output. In this study, we tune a parameterization so that the model produces an oscillation in the tropical middle atmosphere, with a defined period and amplitude, which is one of the main features of the climate driven primarily by gravity waves. We also explore uncertainties associated with the parameterization.

1. Introduction

1.1. Atmospheric Gravity Waves

Atmospheric gravity waves or buoyancy waves, which owe their existence to the restoring force of gravity in a stratified flow, play a substantial role in the exchange of momentum between the Earth's surface and the free atmosphere. They are forced by a range of processes including flow over orography, convection and frontogenesis in the lower atmosphere. Gravity waves propagate primarily upwards and grow in amplitude until they break and deposit their momentum. This influences the large-scale flow, and affects the circulation, temperature, structure, chemistry and composition of the middle and upper atmosphere (Alexander & Dunkerton, 1999).

The horizontal length scale of gravity waves ranges from tens to thousands of kilometers. While the larger scale gravity waves are resolved explicitly by the numerical scheme in climate models, waves smaller than 2x the horizontal resolution cannot be resolved, leading to an underestimate of gravity wave drag from the dynamical core. At this time, current climate models designed for CMIP6 have resolutions of 1°–2.8°, equivalent to ∼100–300 km spacing at the equator (Priestley et al., 2020; Richter & Tokinaga, 2020). At these resolutions, the majority of gravity wave drag is not resolved and is instead represented through both orographic and non-orographic gravity wave parameterizations (e.g., Alexander & Dunkerton, 1999; Scinocca, 2003; Warner & McIntyre, 1999). These aim to describe the large-scale effect that subgrid-scale gravity waves have on the flow and are often necessary to

obtain realistic circulation patterns, for example, to reduce model biases (e.g., Palmer et al., 1986) and to induce a spontaneous Quasi-Biennial Oscillation (QBO) (Bushell et al., 2020). Parameterized gravity waves are required even at the higher resolution end of the spectrum of models, for instance, HighResMIP, which have resolutions higher than 50 km but typically still include some parameterized subgrid-scale gravity waves (e.g., Kodama et al., 2021). Subgrid-scale parameterizations make several assumptions about the nature of gravity waves which become a source of uncertainty in climate projections. Several recent studies harness machine learning methods to learn data-driven gravity wave parameterizations, which may be faster and/or more accurate (e.g., Chantry et al., 2021; Espinosa et al., 2022; Matsuoka et al., 2020). This study makes use of machine learning methods, but rather than replacing traditional parameterizations, we leverage statistical methods to systematically calibrate an existing gravity wave parameterization and quantify uncertainties associated with it.

1.2. Gravity Wave Parameterizations and Associated Uncertainties

A common type of parameterization is the Lindzen-type parameterization, based on Lindzen (1981), which assumes gravity waves are launched at a fixed source level in the troposphere and propagate in the vertical column until they reach saturation. At this point, it is assumed that breaking occurs, depositing gravity wave drag. These have been further developed into spectral parameterizations, in which a spectrum of waves is launched, leading to a spectrum of breaking levels rather than a single level (Alexander & Dunkerton, 1999). In this type of parameterization, there are several parameter choices to be made, for instance, the phase speeds, amplitudes and location of launched gravity waves. These all influence the magnitude and spatial structure of gravity wave drag deposited by the parameterization.

The parameters should ideally be chosen so that the parameterization output (here the unresolved gravity wave drag) is consistent with observations. However, obtaining observations of gravity wave drag caused by unresolved gravity wave breaking is not trivial (Alexander et al., 2010). Observations from stratospheric superpressure balloon and aircraft flights provide estimates of gravity wave properties, such as their phase speeds and momentum fluxes of individual gravity wave packets (Alexander & Pfister, 1995; Alexander & Rosenlof, 2003; Boccara et al., 2008; Hertzog et al., 2008). Satellite measurements can be used to estimate large scale averages of absolute gravity wave momentum fluxes in the stratosphere (Geller et al., 2013). However, all methods of estimating momentum fluxes are limited by resolution and we cannot easily extract momentum fluxes due to subgrid-scale gravity waves (Alexander et al., 2010), nor can we easily decouple convective gravity waves from orographic gravity waves (Corcos et al., 2021; Grimsdell et al., 2010; Jewtoukoff et al., 2015). Importantly, the main goal of parameterizations is to obtain climate model output consistent with the macrophysical climate state (i.e., large-scale circulation and variability), rather than the microphysical (i.e., gravity wave drag). Therefore, the typical approach is to tune the parameterization to obtain a consistent climate state (e.g., Barton et al., 2019; Couvreux et al., 2021; Donner et al., 2011; Dunbar et al., 2021; Scaife et al., 2002).

Calibration of parameters traditionally involves manual tuning of parameter values until a reasonable output is obtained (e.g., Donner et al., 2011; Kodama et al., 2021), but in recent years has been automated with statistical methods such as Bayesian optimization (Kennedy & O'Hagan, 2001), iterative refocusing/history matching (Williamson et al., 2013) and ensemble Kalman methods (Cleary et al., 2021). These methods typically calibrate the parameters by minimizing a loss function that describes the difference between the climate model output and the observations.

Even after calibration, subgrid-scale parameterizations are a substantial source of uncertainty in climate model output that is generally not considered in model analysis. Uncertainty quantification is a growing field for parameterizations including clouds (Pathak et al., 2021), convection (Dunbar et al., 2021), aerosol microphysics (Lee et al., 2012), and ocean processes (Souza et al., 2020), but has not yet been applied to gravity wave parameterizations. In this paper, we combine calibration and uncertainty quantification methods to explore the importance of parameter choices in a non-orographic gravity wave parameterization within an idealized moist atmospheric model. Specifically, we use the Calibrate-Emulate-Sample framework developed in Cleary et al. (2021) to first estimate the optimal parameters that give model output consistent with observed properties of stratospheric phenomena and to further assess the uncertainty of the output associated with the derived distribution of gravity wave parameters.

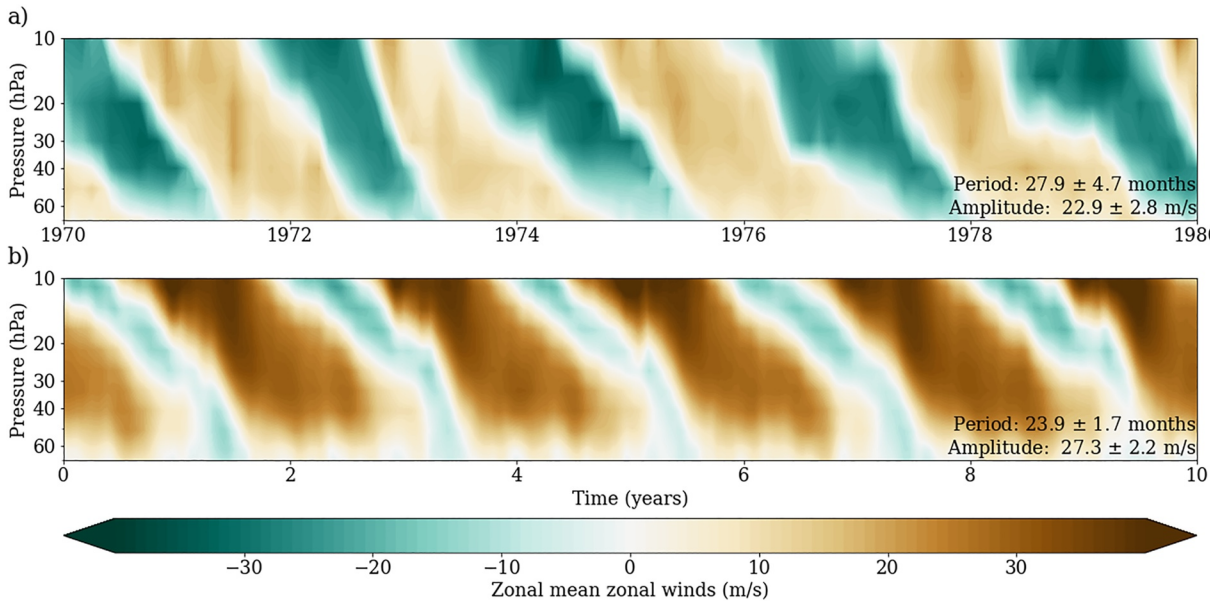

Figure 1. Zonal mean zonal winds at 5°S–5°N over a 10-year segment from (a) global radiosonde observations (Freie Universität Berlin, 2007) and (b) the model used in this study (MiMA2.0 (Garfinkel et al., 2020; Jucker & Gerber, 2017),). In the bottom right corner are the period and amplitude, shown as the mean and 1 standard deviation estimated from (a) the 68 year period of observations and (b) a 50-year control simulation of MiMA.

In the remainder of this section, we describe the QBO, a large-scale oscillation in the tropical stratosphere, realistic simulation of which has depended critically on the choices made in the gravity wave parameterization. Section 2 describes the model and gravity wave parameterization used and Section 3 outlines the CES framework. The results of this are discussed in Section 4, in which we explore CES under the perfect model setting, assuming the "truth" to be a long integration of our model with the parameterization scheme. In Section 4.2, we explore the sensitivity of the QBO to gravity wave parameters and in Section 4.3, we quantify uncertainties of the QBO due to the parameter choices for a control climate and 2xCO_2 scenario. Section 5 contains a summary and discussion of the work.

1.3. Quasi-Biennial Oscillation

The Quasi-Biennial Oscillation (QBO) is the dominant mode of variability in the equatorial stratosphere, occurring in the vertical range of 5–100 hPa (Gray, 2010). The QBO consists of alternating westerly and easterly winds with a period of ∼28 months, descending at ∼1 km/month, as shown in Figure 1a, which shows a cross-section of the zonal mean zonal winds at 5°S–5°N from global radiosonde observations (Freie Universität Berlin, 2007).

The QBO is driven by a broad spectrum of waves, including large-scale Kelvin and Rossby-gravity waves, mesoscale inertia-gravity and high frequency small-scale gravity waves (Baldwin et al., 2001; Lindzen & Holton, 1968). The latter are the gravity waves with zonal wavenumber >40, corresponding to zonal wavelengths between 10 and 1,000 km, that is, mostly subgrid-scale in climate models. Drag due to these contribute significant forcing to the QBO, without which climate models cannot produce a spontaneous QBO. Specifically, only 10 out of 47 CMIP5 models included a non-orographic gravity wave parameterization and of these, only five displayed a QBO-like signal (Schenzinger et al., 2017). Based on more recent models that obtain a spontaneous QBO, at least half of the forcing required is contributed from non-orographic gravity wave parameterizations (Holt et al., 2020). This makes the QBO a useful phenomenon to consider when calibrating the gravity wave parameterization (Anstey et al., 2016; Barton et al., 2019; Scaife et al., 2002).

Simulating a realistic QBO in climate models is important not just for accurately reproducing the tropical stratosphere, but also for tropical convection (Rao et al., 2020), the subtropical jet (Garfinkel & Hartmann, 2011) and the stratospheric polar vortices. The westerly (easterly) QBO phase is associated with a stronger (weaker) polar vortex and fewer (more) sudden stratospheric warmings (the Holton-Tan relationship, Holton & Tan, 1980).

Studies also indicate the QBO influences the transport of aerosols and other atmospheric constituents into and out of the polar vortex (Strahan et al., 2015).

The QBO is defined by a variety of metrics. The first order properties are the period and amplitude of the QBO, which are usually defined in terms of the equatorial zonal mean zonal winds, often at a fixed reference level in the atmosphere. Throughout this paper, we will follow the transition time definition (e.g., Bushell et al., 2020; Richter et al., 2020; Schenzinger et al., 2017) and consider the reference level 10 hPa, where the QBO amplitude is generally a maximum (Bushell et al., 2020). The zonal mean zonal winds between 5°S and 5°N at 10 hPa, $\overline{u}_{eq}$, are first smoothed using a 5-month binomial filter to remove fast fluctuations. Following Schenzinger et al. (2017), a single QBO cycle is determined based on the times at which $\overline{u}_{eq}$ transitions from westward to eastward. The period is defined as the time between subsequent transitions and the amplitude is defined as the maximum amplitude of the zonal mean zonal winds, that is, $\max|\overline{u}_{eq}|$. This gives a period and amplitude for each cycle of the QBO, from which the mean and standard deviation can be estimated.

2. Model Setup

2.1. Model

In this study, we explore the uncertainty of a climate model with respect to the Lindzen-type spectral parameterization introduced in Alexander and Dunkerton (1999), hereafter AD99. We explore uncertainties related to 99 CE parameters that describe the spectrum of gravity waves at the source level. For the climate model, we use the Model of an idealized Moist Atmosphere version 2.0 (MiMA2.0; see Garfinkel et al., 2020; Jucker & Gerber, 2017). This is chosen because it is of intermediate complexity and results in reasonable atmospheric variability, including obtaining a realistic QBO and stratospheric polar vortex but at a lower computational cost than more complex coupled GCMs. We run MiMA at 2.8° resolution (or ∼300 km at equator), which corresponds to T42 spectral resolution, that is, resolving waves only with wavenumber smaller than 42. This leaves the small-scale gravity waves noted as influential for the formation of the QBO (wavenumber >40 (Baldwin et al., 2001)) to be parameterized. These gravity waves are instead captured by the AD99 parameterization, described below.

2.2. Gravity Wave Parameterization

AD99 is a gravity wave parameterization that does not separate the source of gravity waves and treats both orographic and non-orographic gravity waves in the same way. Instead, it launches gravity waves with a fixed phase speed for orographic waves and a spectrum of gravity waves for non-orographic gravity waves. We focus on the non-orographic gravity waves for this study.

2.2.1. Gravity Wave Source

The non-orographic component of AD99 assumes a spectrum of gravity waves with discretized phase speeds centered at $c_0 = 0$ m/s from the source level (315 hPa). The width of this spectrum is defined by the half-width, c_w, which is chosen to be 35 m/s in the default setting, but is not easily constrained by observations. The spectrum of wave momentum flux at phase speed c is given by

$$B_0(c) = \frac{F_{P0}(c)}{\overline{\rho}_0} = \text{sign}\left(c - \overline{u}_0\right) B_m \exp\left[-\left(\frac{c - c_0}{c_w}\right)^2 \ln 2\right] \quad (1)$$

where $F_{P0}(c)$ is the gravity wave stress and $\overline{\rho}_0$ is the mean flow density at the source level. B_m is the momentum flux amplitude of waves with phase speed c_0 and can be constrained by observations of local wave events (e.g., fluctuations in observed wind speed, (u', v', w') and gravity wave phase speeds estimated from superpressure balloon measurements can be used with the polarization relations to derive momentum fluxes ($\rho\overline{u'w'}$, $\rho\overline{v'w'}$) locally (Alexander et al., 2010)). $B_0(c)$ is the momentum flux amplitude in active times and determines when the wave will break, along with the mean flow profile.

The total momentum flux depends not just on $B_0(c)$, but also on the intermittency of the gravity waves. With time, the intermittency reduces the total momentum flux compared to $B_0(c)$ (the momentum flux in active times) and is modeled in AD99 with an intermittency scaling factor,

$$\varepsilon = \frac{\boldsymbol{F_{S0}}\Delta c}{\overline{\rho}_0 \sum_c |B_0(c)|\Delta c} \tag{2}$$

where F_{S0} is the total gravity wave stress at the source level, Δc is the phase speed resolution of the spectrum and $\overline{\rho}_0$ is the mean density at the source level. This equation describes the ratio between the total time-averaged momentum flux to the total momentum flux averaged over all phase speeds of the spectrum.

Although long-term averages of observed $\overline{u'w'}$ and $\overline{v'w'}$, for example, from superpressure balloons can be used to estimate the observed total momentum flux (Geller et al., 2013; Jewtoukoff et al., 2015), climate models typically require the total momentum flux to be smaller than observed values by a factor of 3–5 in order to obtain realistic large-scale flow (Plougonven et al., 2020). Furthermore, gravity wave momentum fluxes are generally estimated in the stratosphere, rather than the source level at 315 hPa (Alexander et al., 2010). This means F_{S0} is not easily constrained by observations and must instead by calibrated to obtain a realistic macrophysical climate state. This gives two uncertain parameters to be calibrated in this study: c_w and F_{S0} (highlighted in bold in Equations 1 and 2 respectively).

2.2.2. Gravity Wave Breaking

Given these properties of gravity waves at the source level, AD99 allows gravity waves to propagate upwards (Alexander & Dunkerton, 1999). At each level, the parameterization checks if the intrinsic frequency magnitude is less than the reflection frequency, and if so, the waves undergo total internal reflection and are eliminated. A stability criterion is also checked at each level, for all phase speeds. The portion of the wave spectrum with phase speeds that do not satisfy the stability criteria undergo breaking and are removed from the spectrum. On breaking, the mean-flow forcing and eddy diffusion coefficients are estimated and fed back into the large-scale flow. For waves that break, indexed by j, between level z_{n-1} and z_n, the forcing on the mean flow is:

$$X\left(z_{n-1/2}\right) = \frac{\epsilon}{\overline{\rho}\left(z_{n-1/2}\right)\Delta z} \sum_j F_{P0}(c_j)$$

and the eddy diffusion coefficient is assumed to be:

$$D\left(z_{n-1/2}\right) = \frac{\epsilon}{\overline{\rho}\left(z_{n-1/2}\right)\Delta z} \frac{1}{N^2\left(z_{n-1/2}\right)} \sum_j \left(c_j - \overline{u}\left(z_{n-\frac{1}{2}}\right)\right) F_{P0}(c_j)$$

where N is the Brunt-Väisälä frequency and $F_{P0}(c_j)$ is the discretized momentum flux carried by waves with phase speed c_j at the source level (Alexander & Dunkerton, 1999). Note this relates to F_{S0}, the total momentum flux at the source level, as $F_{S0} = \sum_{i=1}^{N_c} F_{P0}(c_i)$. The parameters that define the source spectrum affect the forcing and eddy diffusion coefficient through the intermittency scaling factor (Equation 2) and any uncertainty in parameters such as c_w and F_{S0} propagate through to affect the mean flow.

2.2.3. Latitude Dependence of Source Terms

Alexander and Dunkerton (1999) introduced this parameterization for a single vertical column with the intention that it could be applied to global climate models with one-dimensional calculations based on the wind and stability profiles at each geographic point in the model, that is, for each longitude and latitude. Alexander and Rosenlof (2003) find that gravity wave sources in the tropics can differ significantly from those in the extratropics in observations. This can be included in the parameterization by providing latitude-dependent source parameters for c_w and F_{S0}.

The AD99 implementation in MiMA allows c_w to be defined in the tropics (10°S to 10°N) independently of its value outside this region. This means tropical values of c_w can be varied, for example, to explore its effects on the QBO (Garfinkel et al., 2022), while keeping the extratropical value of c_w fixed in order to maintain the stratospheric polar vortices. In this study, we only consider c_w in the tropics, with c_w in the extratropics kept fixed at 35 m/s.

F_{S0} is also latitude dependent. It is typical for GCMs to prescribe a peak in F_{S0} in the tropics due to tropical precipitation (e.g., the Canadian Middle Atmosphere Model (CMAM, Anstey et al. (2016) and MERRA reanalysis/Fortuna version of the Goddard Earth Observing System Mode (GEOS-5) (Molod et al., 2012))) and/or

Table 1
Description of the Two Parameters Calibrated in This Study

Parameter	Description	Control value
c_w	Half-width of phase speed in tropics (10°S to 10°N)	35 m/s
Bt_{eq}	Total gravity wave stress in tropics (10°S to 10°N)	0.0043 Pa

additional stress in extratropical storm track regions, in some cases with a larger value of F_{S0} in the northern hemisphere compared to the southern hemisphere to improve the simulation of the stratospheric polar vortices (e.g., AM3/4, the atmospheric components of the global model from Geophysical Fluid Dynamics Laboratory; see Donner et al., 2011; Zhao et al., 2018). We include the latter, by setting a base of 0.0043 Pa in the extratropics, with an additional 0.0035 Pa in the northern hemisphere that appears to provide roughly the correct number of sudden stratospheric warmings (Equation A3 of Garfinkel et al., 2022). In the tropics (10°S to 10°N), we define $F_{S0} = Bt_{eq}$ as the parameter of interest, responsible for modulating properties of the QBO. Table 1 shows the two parameters calibrated and assessed in this study and their values chosen for the control run setting.

Garfinkel et al. (2022) assessed the sensitivity of the QBO in MiMA to c_w and Bt_{eq}. They found that the QBO amplitude is significantly more sensitive than the period. Increasing Bt_{eq} leads to a faster and stronger QBO. While increasing c_w also leads to a faster and stronger QBO, the period is not affected significantly when c_w is increased beyond 25 m/s.

3. Calibrate, Emulate, and Sample Method

The goal of uncertainty quantification is to obtain a distribution of model outputs, given a distribution of model parameters. To do this, we need samples from the optimal distribution of model parameters that produce model outputs in agreement with an observed dataset. We employ the Calibrate, Emulate and Sample (CES) method (Cleary et al., 2021; Dunbar et al., 2021; Howland et al., 2022). This involves (a) calibration of model parameters so that the model output agrees with the observed dataset, (b) emulation of the expensive model given model parameters to allow for quick evaluations and (c) sampling from the calibrated distribution of model parameters with the emulator.

3.1. Calibration

The first step of CES is the calibration, for which we use Ensemble Kalman Inversion (EKI). Following Cleary et al. (2021), we define the problem as

$$\boldsymbol{y} = \mathcal{G}(\boldsymbol{\theta}) + \eta \tag{3}$$

where $\boldsymbol{\theta}$ are the unknown model parameters (in this case, parameters that define the gravity wave spectrum at the source level, c_w and Bt_{eq}); $\mathcal{G}(\boldsymbol{\theta})$ is the forward model (in this case, MiMA with the AD99 gravity wave parameterization); $\boldsymbol{y}$ is the observable (in this case, long-term averages of stratospheric phenomena); and η is the internal noise in the system. For simplicity, this noise is assumed to be Gaussian with variance Γ, which we write as $\eta \sim N(0, \Gamma)$ (Cleary et al., 2021). Calibration is concerned with solving the inverse problem, to learn optimal model parameter values $\boldsymbol{\theta}$ that produce desired values of $\boldsymbol{y}$.

We take a probabilistic approach to calibration, where the goal is to learn probability distributions of $\boldsymbol{\theta}$ rather than point estimates and we use Bayesian statistics to do this. Here, $p(\cdot)$ indicates probability distributions, for example, $p(\boldsymbol{\theta})$ is the prior probability distribution that represents our prior knowledge of values $\boldsymbol{\theta}$ may take. In calibration, we seek the optimal probability distribution of $\boldsymbol{\theta}$ given the observed data, denoted $p(\boldsymbol{\theta}|\boldsymbol{y})$. This is linked to the likelihood, which describes the probability of the data given a parameter value $\boldsymbol{\theta}$, that is, $p(\boldsymbol{y}|\boldsymbol{\theta})$, and the prior through Bayes' theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{4}$$

From this, we can see that that one can learn the optimal $p(\boldsymbol{\theta}|\boldsymbol{y})$ by optimizing $p(\boldsymbol{y}|\boldsymbol{\theta})$, that is, maximum likelihood. It is standard to choose a Gaussian likelihood (e.g., Cleary et al., 2021; Dunbar et al., 2021; Howland et al., 2022):

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\Gamma}}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\mathcal{G}(\boldsymbol{\theta}))^T\Gamma^{-1}(\boldsymbol{y}-\mathcal{G}(\boldsymbol{\theta}))\right).$$

Where superscript T denotes the transpose. This equation is the probability that the data $\boldsymbol{y}$ originates from $\mathcal{G}(\boldsymbol{\theta})$, allowing for the Gaussian noise with variance Γ as described by Equation 3. Maximizing this is equivalent to minimizing a misfit function which describes a distance between the data, $\boldsymbol{y}$, and the forward model, $\mathcal{G}(\boldsymbol{\theta})$:

$$\Phi(\boldsymbol{\theta},\boldsymbol{y}) = \frac{1}{2}(\boldsymbol{y}-\mathcal{G}(\boldsymbol{\theta}))^T\Gamma^{-1}(\boldsymbol{y}-\mathcal{G}(\boldsymbol{\theta})). \tag{5}$$

This is the Mahalanobis distance. Various optimization methods can be used to minimize $\Phi(\boldsymbol{\theta},\boldsymbol{y})$. Here, we use EKI (Iglesias et al., 2013), a derivative-free optimization method based on Ensemble Kalman filtering which is extensively used in numerical weather prediction to estimate a model state of atmospheric variables given observations. EKI uses the same concepts to solve the inverse problem (Equation 3), but with two fundamental differences to Ensemble Kalman filtering used in data assimilation: (a) rather than finding atmospheric state variables, EKI aims to find the model parameters $\boldsymbol{\theta}$ given observations $\boldsymbol{y}$, removing dependence on the atmospheric state variables by integrating these out with long simulations and (b) the inversion is done offline, without updating the data at each iteration (i.e., no time dependence).

In EKI, we take an ensemble of model parameters, denoted by subscript $m = 1,\cdots,M$, initially drawn from the prior, denoted by $\theta_m^{(0)} \sim p^{(0)}(\cdot)$. At each iteration, denoted by superscript (n), the forward model gives $\mathcal{G}\left(\boldsymbol{\theta}_m^{(n)}\right)$ which is used to update each ensemble member at the next iteration with

$$\boldsymbol{\theta}_m^{(n+1)} = \boldsymbol{\theta}_m^{(n)} + C_{\theta\mathcal{G}}^{(n)}\left(\Gamma + C_{\mathcal{G}\mathcal{G}}^{(n)}\right)^{-1}\left(\boldsymbol{y} - \mathcal{G}\left(\boldsymbol{\theta}_m^{(n)}\right)\right)$$

where $C_{\mathcal{G}\mathcal{G}}^n$ is the covariance matrix of the ensemble output and $C_{\theta\mathcal{G}}$ is the cross-covariance matrix between the ensemble parameters and ensemble outputs. Note that $C_{\theta\mathcal{G}}^{(n)}\left(\Gamma + C_{\mathcal{G}\mathcal{G}}^{(n)}\right)^{-1}$ is the Kalman gain where $\left(\Gamma + C_{\mathcal{G}\mathcal{G}}^{(n)}\right)$ is the innovation covariance, describing the covariance matrix of the differences between $\boldsymbol{y}$ and $\mathcal{G}\left(\boldsymbol{\theta}_m^{(n)}\right)$.

3.1.1. Parameters and Priors

In this study, the model parameters are

$$\boldsymbol{\theta} = (c_w, Bt_{eq})$$

with units [m/s, Pa], described in Table 1, and the model outputs are

$$\boldsymbol{y} = (T_{QBO}, A_{QBO})$$

where T_{QBO} is the QBO period in months at 10 hPa and A_{QBO} is the QBO amplitude in m/s at 10 hPa.

When defining the priors on the model parameters, we first consider physical constraints that total gravity wave stress and the half-width of the phase speeds must be positive everywhere, that is, $Bt_{eq} > 0$ and $c_w > 0$.

We enforce these hard constraints by imposing log-normal priors on all parameters, which equates to transforming the parameters to

$$\hat{\boldsymbol{\theta}} = (\exp(c_w), \exp(Bt_{eq}))$$

and carrying out the calibration on $\hat{\boldsymbol{\theta}}$ with normal priors. We use domain knowledge to inform the choice of the mean and variance of these prior distributions. Observations from stratospheric balloon flights show that gravity waves can have phase speeds generally around 20 m/s, with values up to around 120 m/s (e.g., Boccara et al., 2008; Hertzog et al., 2008) and the half-width of phase speeds in the tropics, c_w in Equation 1, could range from 5 to 80 m/s (Alexander & Rosenlof, 2003). Measurements of gravity wave stress imposed at the source level are not readily available, although models with various non-orographic parameterizations have established

values between 0.001 and 0.01 mPa give realistic gravity wave momentum fluxes in the stratosphere (Geller et al., 2013). Previous studies using 99 CE within MiMA assume values of c_w between 20 and 60 m/s and of Bt_{eq} between 0.003 and 0.005 Pa (Alexander & Dunkerton, 1999; Garfinkel et al., 2022; Jucker & Gerber, 2017). Based on these ranges, we calculate the mean and variance on $\hat{\theta}$ by transforming a normal distribution with means $\mu = (35, 0.0043)$ and variances $\sigma^2 = \left(10^2, 0.001^2\right)$ through the exponential map.

3.2. Emulation

The calibration step allows us to learn the distribution of optimal parameters given the observations. For uncertainty quantification of the model output, we would next sample from this distribution, for example, with a Monte Carlo method such as Markov chain Monte Carlo (MCMC). However, since this requires many expensive model evaluations (e.g., $O\left(10^5\right)$ (Geyer, 2011)), we build an emulator that can be evaluated cheaply. The emulator can be trained with the samples obtained through the EKI calibration step above. These are ideal as the later iterations of EKI sample the posterior distribution, which is ultimately the region of interest for the emulator, and also the early iterations include samples from the prior distribution which helps constrain the emulator at the edges of the posterior distribution.

3.2.1. Gaussian Processes

The emulator we use here is a Gaussian process (GP) emulator, which is a popular Bayesian emulation tool in the calibration and uncertainty quantification community (e.g., Couvreux et al., 2021; Kennedy & O'Hagan, 2001; Williamson et al., 2017). These are used because they model the distribution of functions that satisfies a given dataset, meaning they can produce a mean function and a measure of uncertainty around this (e.g., the standard deviation or confidence intervals). A GP is a type of stochastic process, which is defined as a collection of random variables (i.e., observations) indexed by an index set, x, such as space or time. A "Gaussian" process refers to the case where any finite number of these random variables has a multivariate normal distribution (Rasmussen & Williams, 2006). GPs are specified by a mean function, $m(x)$, and covariance function (also known as a kernel), $C(x, x')$, and are denoted

$$f(x) \sim GP\left(m(x), C\left(x, x'\right)\right).$$

A GP can be viewed as a probability distribution over a function $f(x)$, where the index set represents the x-axis of the function, $m(x)$ are the mean values over the function and $C(x, x')$ describes the correlation between $f(x')$ and $f(x)$, given two values x and x'.

In Section 3.1, we showed how Bayes' theorem relates the posterior distribution to the prior distribution and the likelihood of the data (Equation 4). Bayes' theorem can also be applied to GPs, where the user defines a prior GP which is combined with the data in Bayes' theorem to derive a posterior GP. To define the prior GP, the user specifies $m(x)$ and $C(x, x')$. It is typical to assume $m(x) = 0$ and define the GP's structure through $C(x, x')$ entirely (Rasmussen & Williams, 2006). Domain knowledge can be used to inform $C(x, x')$ (e.g., to include known lengthscales or periodicity) and covariance functions can be combined through linear operations to include multiple features. For more information on covariance function choices, see Chapter 4 of Rasmussen and Williams (2006).

If we denote the prior GP as $f(x)$, the posterior GP can be denoted $f(x)|D_N$ where D_N is the data in the form of N input-output pairs obtained from the expensive model (Gramacy, 2020). Using Bayes theorem, one can derive the posterior GP

$$f(x)|D_N \sim GP\left(m^*(x), C^*\left(x, x'\right)\right)$$

where $m^*(x)$ and $C^*(x, x')$ are the new mean and covariance functions respectively, which can be written in terms of $m(x)$ and $C(x, x')$ entirely. The derivation for this is involved (see e.g., Rasmussen & Williams, 2006) but tractable because the definition of a GP states that any finite number of random variables have a multivariate

normal distribution. $f(x)|D_N$ is also called the GP predictive distribution, since it can be used to predict the probability distribution function at new, unseen values of x (Gramacy, 2020).

3.2.2. Gaussian Process Emulator of QBO Properties Given Gravity Wave Parameters

In this study, the index set is the gravity wave parameters, $\boldsymbol{\theta}$, which is a two-dimensional (2d) vector containing c_w and Bt_{eq}. The Gaussian process $f(\boldsymbol{\theta})$ emulates the climate model output $\mathcal{G}(\boldsymbol{\theta})$, which is also a 2d vector containing the properties of the QBO, (T_{QBO}, A_{QBO}). For each dimension, we define a prior GP as

$$f(\boldsymbol{\theta}) \sim GP\left(m(\boldsymbol{\theta}), C\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right)\right).$$

We choose $m(\boldsymbol{\theta}) = 0$ and for the covariance function, we use a squared exponential kernel (also called a radial basis function). This is a popular choice when little domain knowledge is available, as it provides a smooth covariance that falls exponentially as the distance between points increases (Rasmussen & Williams, 2006):

$$C_{SE}\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) = \sigma^2 \exp\left(\frac{\left(\boldsymbol{\theta} - \boldsymbol{\theta}'\right)^2}{2\boldsymbol{l}}\right).$$

σ and $\boldsymbol{l}$ are both length scale hyperparameters, where σ describes the distance between $f(\boldsymbol{\theta})$ and $f\left(\boldsymbol{\theta}'\right)$ and $\boldsymbol{l}$ describes the distance in $\boldsymbol{\theta}$ needed for the $c\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right)$ to fall by $1/e$. Note that $\boldsymbol{l}$ is also a 2d vector with independent lengthscales for each parameter dimension (this is called automatic relevance determination). This choice of covariance function leads to a smooth $f(\boldsymbol{\theta})$ which will become beneficial in the sample step of CES, since a smooth function has better convergence properties for sampling methods such as MCMC (Cleary et al., 2021).

We also include a white noise covariance function which represents the internal variability in the output,

$$C_{WN}\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) = \sigma_{WN}^2$$

where σ_{WN}^2 is the internal noise assumed to be consistent across all values of $\boldsymbol{\theta}$. This choice is made because we are approximating properties of the system, defined on an infinite time horizon, with finite time averages following Dunbar et al. (2021). We assume that finite time averaged data is a noisy approximation of the infinite time average, where the noise is assumed to be Gaussian given large enough timescales, due to the central limit theorem.

Including both the squared exponential and white noise covariance functions gives us the covariance function

$$C\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) = \sigma^2 \exp\left(\frac{\left(\boldsymbol{\theta} - \boldsymbol{\theta}'\right)^2}{2\boldsymbol{l}}\right) + \sigma_{WN}^2$$

where σ, $\boldsymbol{l}$, and σ_{WN} are hyperparameters that are learned to ensure the posterior GP gives the best fit to the data (Bishop, 2006). Here they are optimized with type II maximum likelihood using Scikit-learn, a collection of machine learning software in Python (Pedregosa et al., 2011).

Since we are emulating a 2d vector, (T_{QBO}, A_{QBO}), we define two prior GPs with two sets of hyperparameters to optimize, independent of each other. However, we do not necessarily expect the white noise variance σ_{WN} for each dimension to be uncorrelated (in fact, here they are positively correlated in observations, i.e., a QBO cycle with a longer period is likely to have also have a larger amplitude (Freie Universität Berlin, 2007)). To avoid this issue, before we build the GP emulator, we transform the outputs to the decorrelated space by performing Singular Value Decomposition (SVD, also called Principal Component Analysis; PCA). Note, this is an optional modeling choice, but is important when the output $\mathcal{G}(\boldsymbol{\theta})$ is likely to contain correlations between output dimensions, which is often the case in climate modeling applications. SVD is also a useful dimension reduction technique for high dimensional $\mathcal{G}(\boldsymbol{\theta})$, as one can emulate only the singular vectors that explain most (e.g., 95%) of the variance in the output (e.g., Howland et al., 2022).

3.3. Sample

Ultimately, we want to know the posterior distribution $p(\boldsymbol{\theta}|y)$, which we can approximate with a large sample of values of $\boldsymbol{\theta}$, where the larger the sample, the better the estimation (Kruschke, 2015). To do this, we use a sampling method built on Bayesian principles, called Markov chain Monte Carlo (MCMC). MCMC is an iterative method that uses a proposal distribution to propose new samples that are accepted with a probability relating to the posterior probability distribution, that is, if they are deemed to be a sample from the posterior distribution. Here, we use a Metropolis random walk MCMC (Metropolis et al., 1953), meaning the proposal distribution is simply a random walk.

We start the MCMC with an initial sample, $\boldsymbol{\theta}^{(0)}$, drawn from the prior distribution, $p(\boldsymbol{\theta})$ defined in Section 3.1. We propose a new sample $\boldsymbol{\theta}^* \sim p\left(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(\mathbf{0})}\right) = N\left(\boldsymbol{\theta}^{(\mathbf{0})}, \Delta^2\right)$ where Δ^2 is a matrix containing the step sizes of the random walk along the diagonal. We must then decide whether to replace the current sample $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^*$, by comparing their relative posterior probabilities. To do this, we evaluate the probability that $\boldsymbol{\theta}^*$ could be a posterior sample, using Bayes theorem (Equation 4), where we assume the likelihood is Gaussian that is,

$$p\left(\boldsymbol{y}|\boldsymbol{\theta}^*\right) = \frac{1}{\sqrt{det(\boldsymbol{\Gamma})}} exp\left(-\frac{1}{2}\left(\left(\boldsymbol{y} - f\left(\boldsymbol{\theta}^*\right)\right)^T \boldsymbol{\Gamma}^{-1}\left(\boldsymbol{y} - f\left(\boldsymbol{\theta}^*\right)\right)\right)\right) \tag{6}$$

where $f\left(\boldsymbol{\theta}^*\right)$ is an evaluation of the GP emulator, to approximate $\mathcal{G}\left(\boldsymbol{\theta}^*\right)$. $\boldsymbol{\theta}^*$ is accepted with probability

$$\frac{p\left(\boldsymbol{\theta}^*|\boldsymbol{y}\right)}{p\left(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(\mathbf{0})}\right)} \frac{p\left(\boldsymbol{\theta}^{(\mathbf{0})}|\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}^{(\mathbf{0})}|\boldsymbol{y}\right)} = \frac{p\left(\boldsymbol{\theta}^*\right) p\left(\boldsymbol{y}|\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(\mathbf{0})}\right)} \frac{p\left(\boldsymbol{\theta}^{(\mathbf{0})}|\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}^{(\mathbf{0})}\right) p\left(\boldsymbol{y}|\boldsymbol{\theta}^{(\mathbf{0})}\right)}$$

known as the acceptance probability. The right-hand side is derived using Bayes' theorem (Equation 4) where the constant of proportionality is the same for $p\left(\boldsymbol{\theta}^{(\mathbf{0})}|\boldsymbol{y}\right)$ and $p\left(\boldsymbol{\theta}^*|\boldsymbol{y}\right)$ and therefore cancels out. Here, $\frac{p\left(\boldsymbol{\theta}^*|\boldsymbol{y}\right)}{p\left(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(\mathbf{0})}\right)}$ is the ratio of the posterior probability to the proposal probability and the acceptance probability compares this ratio to the same ratio for $\boldsymbol{\theta}^{(\mathbf{0})}$, to effectively decide on whether to replace sample $\boldsymbol{\theta}^{(\mathbf{0})}$ with proposed sample $\boldsymbol{\theta}^*$. Details on why this acceptance probability ensures samples are from the posterior probability distribution can be found in, for example, Robert and Casella (2004).

If accepted, we set $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^*$, otherwise, we set $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)}$. We repeat these steps, that is, by proposing a new sample through the random walk $\boldsymbol{\theta}^* \sim N\left(\boldsymbol{\theta}^{(1)}, \Delta^2\right)$, evaluating the acceptance probability and deciding whether or not the new sample is accepted. This is repeated until we have a chain of N samples, $\boldsymbol{\theta}^{(\mathbf{0})}, \boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}$. The first portion of samples are close to the prior distribution, $p(\boldsymbol{\theta})$ and so we discard these as "burn-in". The number of samples to be discarded depends on the specific task and requires some user judgment as to when the samples have converged to a stationary state, but can require $O\left(10^5\right)$ iterations (e.g., Geyer, 2011). This is where we see the benefit of the GP emulator as all instances of $f\left(\boldsymbol{\theta}^*\right)$ in Equation 6 are used to rapidly approximate $\mathcal{G}\left(\boldsymbol{\theta}^*\right)$.

We run the Metropolis random walk MCMC for 10^5 iterations (after 10^4 burn-in iterations) to obtain the posterior distribution. The random walk step sizes (Δ^2) are determined to ensure an acceptance rate close to 25%, deemed to be optimal in Roberts and Rosenthal (2004). Here, one can choose to carry out the MCMC in either the original space or in the decorrelated space, after performing SVD. We run the MCMC in the decorrelated space, to improve efficiency (since we use a random walk with independent step sizes in both directions, i.e., Δ^2 is a diagonal matrix). All results are presented after transforming back into the original parameter space.

4. Results

4.1. Calibrate, Emulate and Sample in the Perfect Model Setting

We explore the results of CES with the "perfect model" setting, as done in Dunbar et al. (2021), where we define the "truth" to be a long 50-year integration of MiMA, with known model parameters, here $c_w = 35$ m/s and $Bt_{eq} = 0.0043$ Pa. Figure 1b shows 10 years of this simulation. Note compared to the observed QBO (Figure 1a), MiMA produces a QBO with westerly phases that are too long and strong relative to the easterly phases. Also, the QBO exists slightly too high up in the stratosphere, as the pattern of alternating winds appear to vanish at around 60 hPa, unlike the observed winds in Figure 1a. For the purpose of demonstrating the CES method, we

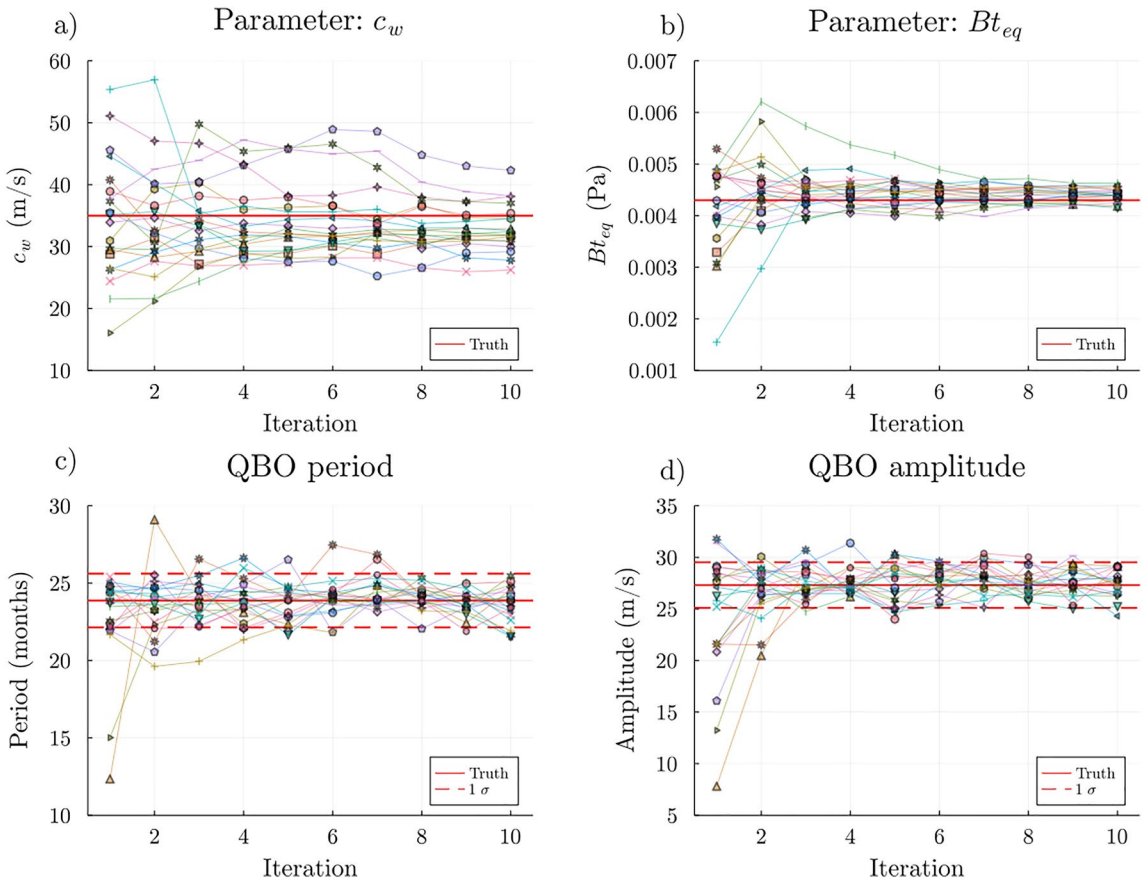

Figure 2. (a–b) Parameter and (c–d) model output values for all iterations of EKI for the perfect model setting, where iteration 1 consists of parameter values drawn from the prior. Each line/marker represents a single ensemble member. The red line denotes in (a–b) the "truth" that is, the known parameter values (Table 1) and in (c–d) the model output obtained in one long MiMA simulation with these parameter values, with the dashed red line showing 1 standard deviation across the simulation.

focus only on the first order properties of the QBO, namely the period and amplitude in the upper stratosphere, at 10 hPa. This means we can only validate the method on these properties. Ultimately, for operational purposes, one may wish to calibrate more properties of the QBO (e.g., period and amplitude in easterly/westerly phase) at all pressure levels, which would provide additional constraints on the gravity wave parameters. For a given GCM with a limited number of tuning parameters, it is possible that there does not exist a solution to $\boldsymbol{\theta}$ in the inverse problem (Equation 3), where $\mathcal{G}(\boldsymbol{\theta}) = \boldsymbol{y}$ is satisfied exactly. However, in this case, EKI remains a suitable tool as it produces the optimal values by minimizing $|\boldsymbol{y} - \mathcal{G}(\boldsymbol{\theta})|$.

At 10 hPa, MiMA produces a QBO period of 23.9 ± 1.7 years and amplitude 27.3 ± 2.2 m/s where the uncertainties here are 1 standard deviation across all QBO cycles in the 50-year integration. The calibration step learns the posterior distribution of parameter values that gives a QBO consistent with this. It allows us to test the method on a simpler problem while developing an understanding of how the model parameters relate to each other.

The first step of CES is to calibrate c_w and Bt_{eq} to the QBO metrics for period and amplitude. EKI is run with an $M = 20$ ensemble. Figure 2 shows the EKI for 10 iterations, where the top two panels show the gravity wave parameters c_w and Bt_{eq} and the bottom two panels show the model output. The parameters move toward convergence after around 6–8 iterations.

Considering each ensemble member at each iteration, EKI gives a total of 200 input-output pairs. These data are used to train the GP emulator in the emulation stage of CES. First, the validity of the emulator is tested by training the GP emulator on 170 input-output pairs, which include all data from the first three iterations and the rest selected at random from the last seven iterations. This leaves aside 30 samples for testing, randomly selected from the last seven iterations. We do this to test the emulator while keeping in mind that the goal is to predict

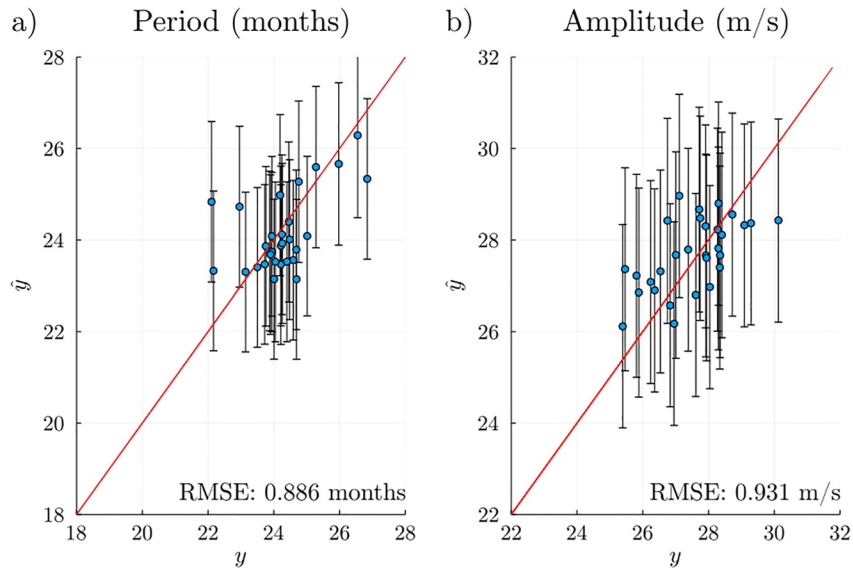

Figure 3. Plots of emulator performance on example test data points, selected at random from the last seven iterations of EKI for (a) period and (b) amplitude of the QBO. The test data values are plotted on the x-axis (y) and the Gaussian process emulator predictions are plotted on the y-axis ($\hat{y}$), where the error bars indicate the Gaussian process 1σ levels. The red line shows where $\hat{y} = y$, indicating a perfect prediction.

regions of the parameter space close to the posterior distribution, avoiding extrapolation to other regions of the parameter space (including the prior). Figure 3 shows these test data, y, against the GP prediction $\hat{y}$, where a perfect prediction would be these points lying on the $\hat{y} = y$ line shown in red. The error bars indicate the 1 σ uncertainty predicted by the GP emulator. The $\hat{y} = y$ line falls within 1 σ of the GP prediction for the majority of test data points, as required for an accurate emulator. We test the emulator repeatedly with different test sets in Figure S1 in Supporting Information S1 and find the emulator performance to be fairly consistent, with RMSEs between 0.7 and 1.0 months for the period and between 0.8 and 1.4 m/s for the amplitude, both of which are within the average 1 σ levels predicted by the emulator. Table S1 in Supporting Information S1 also confirms the emulator outperforms linear regression, albeit only slightly for the QBO period, indicating a linear relationship describes most of the relationship between the period and the gravity wave parameters.

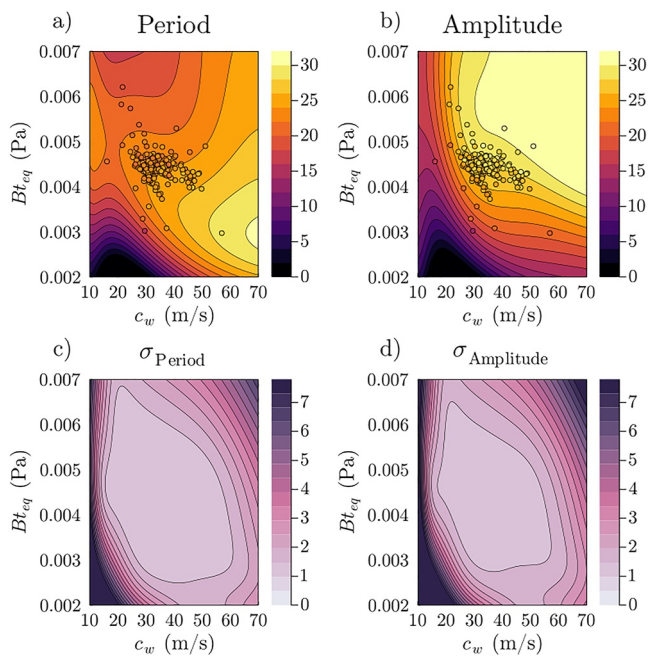

Figure 4. Gaussian process emulator predictions over a sweep across parameter values ($c_w = 10 - 70$ m/s, $Bt_{eq} = 0.002 - 0.007$ Pa) learned from the EKI in the perfect model setting for (a) QBO period and (b) QBO amplitude. The scatter points indicate the training data from MiMA simulations obtained through EKI. The 1 σ uncertainty associated with these predictions is shown in (c) for the period and (d) for the amplitude.

To maximize accuracy, the final emulator used is trained on all 200 samples. A sweep across the parameter space is carried out by varying c_w from 10 to 70 m/s and Bt_{eq} from 0.002 to 0.007 Pa. Figure 4 shows contour plots of a) the QBO period and b) the QBO amplitude for this parameter sweep across c_w and Bt_{eq}. The points indicate the training data values, showing an agreement with the GP emulator. Note that the training points are fairly crowded within the region where the misfit function is minimized ($25 \lesssim c_w \lesssim 40$ m/s and $0.004 \lesssim Bt_{eq} \lesssim 0.005$ Pa). Outside this region, the GP emulator is extrapolating to new regions of the parameter space and therefore is less trustworthy. The 1 σ level predicted by the GP emulator also highlights this in Figures 4c and 4d for the period and amplitude respectively.

The contour plot in Figure 4a estimates a maximum in QBO period for relatively high c_w (50–70 m/s) when Bt_{eq} is chosen to be relatively low (0.002–0.003 Pa). Increasing Bt_{eq} and decreasing c_w leads to a faster QBO. This is expected for Bt_{eq} following the idealized models of Holton and Lindzen (1972) and Plumb (1977), since increased gravity wave stress leads to increased deceleration of winds and therefore more rapidly descending

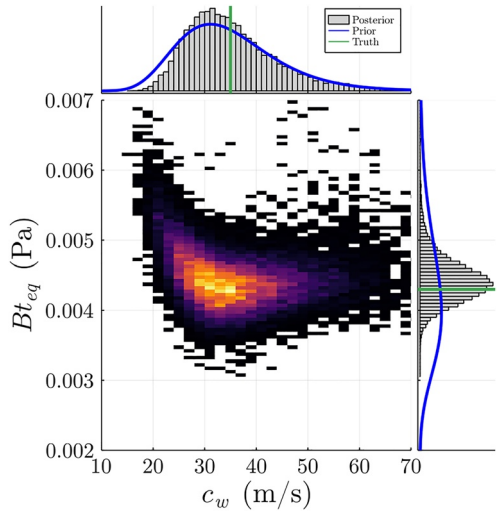

Figure 5. Samples from the posterior distribution of c_w and Bt_{eq} generated by the MCMC in the final stage of CES. The marginal distributions are shown on the corresponding axis, with the prior distributions shown in blue and the known "truth" in green.

westerly/easterly shear zones (Dunkerton, 1997; Schirber et al., 2015). The emulator predicts a faster period with decreasing c_w, consistent with Garfinkel et al. (2022), possibly due to the weaker QBO present under slower phase speeds. However, the period is fairly stable to changes in Bt_{eq} and c_w in the region of the posterior distribution. Note the breakdown in a reasonable QBO at $Bt_{eq} < 0.003$ and $c_w < 30$, with estimated periods of less than 5 months and with Figure 3c showing large uncertainties exceeding this, highlighting where the emulator predictions are not trustworthy.

Figure 4b shows a peak in QBO amplitude when both c_w and Bt_{eq} are relatively high. Increasing c_w increases the QBO amplitude since the higher phase speeds contribute to the faster westerlies and easterlies in the QBO (Holton & Lindzen, 1972; Plumb, 1977; Schirber et al., 2015) but only up until c_w reaches around 30 m/s. Beyond this, increasing c_w has minimal effect, also seen in Garfinkel et al. (2022). This could be because waves with sufficiently large c do not reach breaking levels in the stratosphere and instead continue propagating upwards, without depositing momentum until reaching the sponge layer. For $c_w \gtrsim 30$ m/s, the amplitude is more sensitive to Bt_{eq}, where increasing the gravity wave stress will increase the drag deposited and therefore lead to a stronger QBO.

In the last stage of CES, we sample from the posterior distribution using an MCMC (see Movie S1 in Supporting Information S1). After removing 10,000 iterations for burn-in, 80,000 samples from the posterior distribution are shown in Figure 5, where the 2D histogram is shown in the center with the marginal posterior distributions for c_w and Bt_{eq} shown on the corresponding axis. The prior distribution is also shown in blue, with the known truth in green. The 2D histogram shows a correlation between c_w and Bt_{eq} when $c_w < 35$, indicating that a sample with a larger value of c_w can still produce a QBO with a realistic period and amplitude if Bt_{eq} is decreased appropriately. The narrower posterior distribution for Bt_{eq} indicates this is more crucial for obtaining a correct QBO, while the posterior distribution for c_w more closely follows the prior distribution chosen. Sampling the parameters from this histogram gives a QBO consistent with the "truth" selected here.

The prior distribution should always be chosen to be wider than we expect the posterior distribution to be, since by definition the MCMC cannot sample points outside of the prior. Here, the posterior distribution extends to high phase speeds, following the prior distribution. A wider prior on c_w may produce a posterior distribution with extended tails at high phase speeds. This highlights the importance of choosing a suitable prior that is sufficiently wide, particularly when there is little domain knowledge available to influence the choice. Here, we used domain knowledge to constrain the prior to physical values observed of gravity wave phase speeds (Alexander & Rosenlof, 2003; Boccara et al., 2008; Hertzog et al., 2008), consistent with values used in previous AD99 studies (Alexander & Dunkerton, 1999; Garfinkel et al., 2022; Jucker & Gerber, 2017). This acts as an additional constraint on parameter values when the outputs are less sensitive to the parameter values, for example, Figure 4 shows the QBO is less sensitive to c_w beyond 50 m/s, once Bt_{eq} is constrained to 0.004–0.005 Pa. In general, we expect phase speeds <60 m/s to be of main importance in the stratosphere, while higher phase speed gravity waves continue propagating and break at higher altitudes (Alexander & Rosenlof, 2003). This example shows how the choice of prior matters and how we can leverage both physical understanding and statistical relationships to constrain parameter values.

4.2. Global Sensitivity Analysis

We carry out Global Sensitivity Analysis (GSA) to measure the sensitivity of the climate model output to the gravity wave parameters through variance-based sensitivity indices that describe how much of the variance in the output can be attributed to the variance in each input parameter for a given input parameter distribution (Saltelli et al., 2007). This method averages over all possible values for all other parameters ("global" sensitivity analysis) rather than keeping them fixed at the default values ("local" sensitivity analysis). This requires a large number of samples of the model, so the availability of the emulator to obtain inexpensive samples is crucial for this analysis.

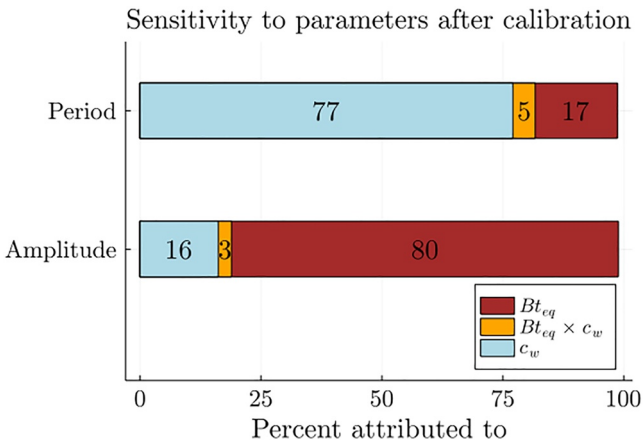

Figure 6. Sensitivity indices as a percentage, describing the proportion of variance in the QBO period and amplitude attributed to the variance in the parameters, c_w and Bt_{eq}.

The first order sensitivity index describes the variance in an output variable, Y, due to a single parameter, θ_i, and is given by

$$SI_i = \frac{Var\,(\theta_i)\left(E_{\theta_{\sim i}}(Y|\theta_i)\right)}{Var(Y)}$$

where $Y|\theta_i$ denotes the estimated output due to parameter θ_i and $E_{\theta_{\sim i}}(\cdot)$ indicates the average over all other parameters except for θ_i. The Sobol' method (Sobol', 2001) approximates this by estimating $Var\,(\theta_i)$ (see Saltelli et al., 2010). Higher order sensitivity indices can be estimated to attribute the interaction between multiple parameter values.

We estimate first order sensitivity indices in the decorrelated space (applying SVD to remove correlations between c_w and Bt_{eq}). After transforming these back into the real space, the sensitivity indices in percentages of the QBO period and amplitude are shown in.

The QBO period is most sensitive to c_w, while the QBO amplitude is most sensitive to Bt_{eq}. This is in agreement with the contour plots in Figure 4 in the region of the calibration. We expect that, before calibration, the QBO period is primarily controlled by Bt_{eq} and therefore after calibration, the remaining uncertainties are due to uncertainties in c_w. Similarly, before calibration, the QBO amplitude is mostly governed by c_w, which pushes QBO wind speeds toward the phase speeds. During the calibration stage, c_w is constrained so that remaining uncertainties in the QBO amplitude are caused mostly by Bt_{eq}. Note that the interaction terms are small, since the analysis is carried out in the decorrelated space (Figure 6).

4.3. Uncertainty Quantification in New Scenario

Understanding the uncertainty in climate model output due to the gravity wave parameterization is one of the main motivations for this analysis. In this section, we explore the parametric uncertainty in a climate change projection, meaning the uncertainty in model output that is due to the possible values that c_w and Bt_{eq} could take. This can be assessed through a perturbed parameter ensemble, where an ensemble of simulations is run with parameter values sampled from their distribution in Figure 5 (Murphy et al., 2014). Here we run a perturbed parameter ensemble for a 2xCO_2 integration. We use this ensemble of simulations to quantify parametric uncertainty for both scenarios.

We run a perturbed parameter ensemble of 50 simulations for 10 years each, initialized with a spun-up climate (Wan et al., 2014), obtained through a 200 years 2xCO_2 integration with fixed model parameters. Each 10-year simulation provides around 4–5 QBO cycles per ensemble member, after allowing 1 year for spin-up (a total of 140 QBO cycles). The QBO periods and amplitudes are plotted in red in Figure 7 and compared against a single long simulation in blue, which was run for 300 years to give roughly the same number of QBO cycles (142 cycles). Note that several QBO disruptions occurred in both the long simulation (3 disruptions) and the ensembles (4 disruptions). These were removed manually from the dataset before the analysis, as they resulted in QBO periods that were either unusually short (<10 months) or unusually long (>38 months). All QBO cycles for both the long simulation and the ensemble members are shown in Figures S2-3 in Supporting Information S1.

The larger variance in the ensembles (red) in Figure 7 compared to the long simulation (blue) is due to the uncertainty in parameter values. The internal variability can be estimated as the standard deviation across the 300-year simulation, denoted σ_{int} in Figure 7. The difference between the standard deviation in the ensemble, σ_{ens}, and the internal variability can be used to estimate the parametric uncertainty, σ_θ, by assuming a Gaussian distribution of QBO periods and amplitudes across all cycles so that $\sigma^2_{ens} = \sigma^2_{int} + \sigma^2_\theta$.

This gives parametric uncertainty estimates in the period of 1.53 months and in the amplitude of 2.14 m/s under 2xCO_2 forcing, when the parameter values are sampled from the distribution in Figure 5. Here we have tuned the parameters to a long integration of a present-day climate, but the natural extension would be to calibrate parameters to observations, which would introduce further uncertainties. Therefore we may expect the

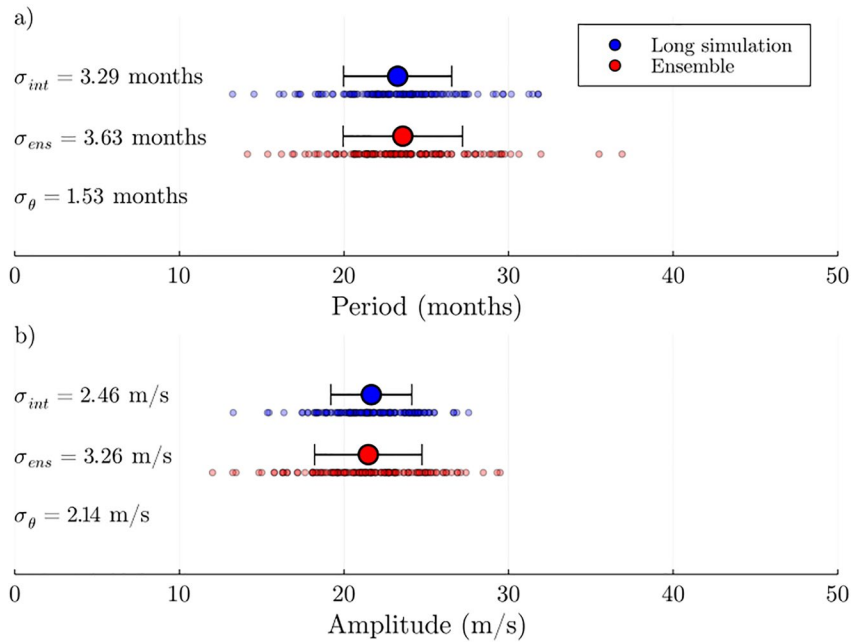

Figure 7. Range of values of QBO (a) period and (b) amplitude for a 2xCO$_2$ scenario for a long simulation of 300 years in blue, where parameter values are fixed at $c_w = 35$ m/s, $Bt_{eq} = 0.0043$ Pa, compared against an ensemble in red (50 simulations, each of 10 years) where parameter values are drawn from the distribution in Figure 5. The large markers show the mean across the long simulation/ensemble and the error bars show 1 standard deviation. The smaller markers show the period and amplitude for all QBO cycles. Note that QBO disruptions are removed before analysis. The internal variability estimated from the long simulation is shown as σ_{int}, the ensemble variability is σ_{ens}, and the parametric uncertainty is σ_{θ}.

parametric uncertainties presented here to be a lower bound on uncertainties associated with the gravity wave parameterization.

5. Discussion

This study demonstrates how the Calibrate, Emulate and Sample (CES) method can be applied to tune parameters and quantify uncertainties associated with a gravity wave parameterization within an intermediate complexity climate model. We have explored the application of CES under the perfect model setting, where we prescribe the "truth" as a long model simulation with known parameter values. However, in future studies this will be extended to a more realistic setting, using observational data from global radiosonde measurements as the "truth" (Freie Universität Berlin, 2007).

The CES method allows us to learn the optimal distribution of parameter values for the half-width of the phase speeds, c_w, and the total gravity wave stress, Bt_{eq}, both of which define the gravity wave spectrum at the source level. We find that these parameters have an anti-correlated distribution, that is, a higher value of Bt_{eq} can be compensated with a lower value of c_w to achieve the same QBO period and amplitude.

A global sensitivity analysis highlighted that after calibration the QBO period is most sensitive to c_w, since it has been constrained mainly by Bt_{eq}, which directly influences the deceleration of easterly/westerly winds. Similarly, the QBO amplitude is more sensitive to Bt_{eq}, as wind speeds are constrained predominantly by gravity wave phase speeds c_w (Dunkerton, 1997; Lindzen & Holton, 1968).

We have quantified parametric uncertainties in MiMA associated with the gravity wave parameterization under a 2xCO$_2$ forcing as 1.53 months for the QBO period and 2.14 m/s for the amplitude. We expect these to be a lower bound on the parametric uncertainty, since we calibrated the parameters to a long model integration, in the absence of realistic QBO variability and measurement error. These are of a similar order of magnitude to the internal variability, highlighting their relevance to climate change projections. Note that parametric uncertainty does not account for uncertainty in the structure of the parameterization itself, rather the uncertainty in the

parameter values of c_w and Bt_{eq} alone. Here, the parameter values are tuned based on the QBO in the present day climate, isolating the effect of the gravity wave parameters from any changes in the source, such as convection, which is likely to change under a warming climate.

The results presented here rely on the Gaussian process emulator. We find the emulator to provide an accurate approximation to the period and amplitude of the QBO within MiMA, as shown in Figure 3 and Figure S1 in Supporting Information S1. However, the limitation here is the presence of internal variability, which manifests as noise in the relationship between gravity wave parameters and the simulated QBO. This noise would be present even if full MiMA runs were used in place of the emulator and has long been known to present difficulties in climate model analyses (e.g., Deser et al., 2012; Tebaldi & Knutti, 2018), including in emulation studies (Castruccio et al., 2019; Watson-Parris, 2021; Williamson et al., 2017). This should be considered when designing emulators for climate model output, for instance, by including a white noise kernel to model the internal noise, as done here (Dunbar et al., 2021; Williamson & Blaker, 2014). Note that this example is fairly simple with only two input parameters and with outputs that can be modeled reasonably with a linear regression (Table S1 in Supporting Information S1). More complicated problems may find that more training simulations are required to emulate the outputs to the desired accuracy, so as not to impact the reliability of the conclusions.

In this study, we calibrated to the QBO period and amplitude at 10 hPa, since these are the first order properties of the QBO. Further extensions of this would be to explore other properties of the QBO such as the period and amplitude at different (or all) levels of the stratosphere or the westerly and easterly amplitudes (e.g., to reduce the westerly bias in MiMA in Figure 1b). This may be more complicated as Giorgetta et al. (2006) find that both the QBO in the lower stratosphere and the westerly phase of the QBO are controlled more by resolved waves, rather than subgrid-scale parameterizations.

Calibrating the gravity wave parameterization in the tropics aims to produce a realistic QBO, but does not directly address model errors at higher latitudes (Anstey et al., 2016; Garcia & Richter, 2019). It is known that non-orographic gravity waves contribute to the breakdown of the polar vortices, influencing the frequency and properties of sudden stratospheric warmings (Siskind et al., 2007, 2010; Wright et al., 2010) and the timing of the stratospheric final warming (Gupta et al., 2021). The effect of varying extratropical gravity wave parameters has not yet been explored in MiMA. Calibrating extratropical gravity wave parameters to properties of the stratospheric polar vortex in both hemispheres is a topic of future research.

Ultimately, one may wish to carry out CES on more than two parameters for gravity wave parameterizations and/or other subgrid-scale processes. Scaling this up introduces challenges for all three steps of CES. For the calibration stage, optimizing the posterior distribution in a higher dimensional setting increases the chance that the parameters cannot be constrained to produce model output consistent with the observations. This is by definition if the number of parameters exceeds the number of observed outputs. Regularization methods can be used to remedy this (e.g., Iglesias, 2016; Iglesias et al., 2013). Aside from this, ensemble Kalman methods scale fairly well with dimension when $O(100)$ ensemble members are used (Dunbar et al., 2021; Kalnay, 2002; Ott et al., 2004). The GP emulator, however, does not scale well with an increasing number of input parameters. The number of simulations required for training is generally assumed to be around 10 times the number of input parameters (Loeppky et al., 2009) which further leads to poor scaling. GP emulators are generally suitable for $O(10)$ parameters, but Dunbar et al. (2021) suggest that alternative emulators that do scale well, such as neural networks, could be used in place of the GP emulator for higher dimensional problems. In the sampling stage of CES, MCMC scales reasonably well with high dimensional problems. Although increasing dimensions can increase the chance of the chain becoming "stuck" in local minima, we can run multiple MCMC simulations in parallel, initialized independently to mitigate this (Brooks et al., 2011). Overall, we can expect the version of CES described here to deal well with calibration and uncertainty quantification of $O(10)$ parameters, and higher dimensional problems can be approached with variations on the emulator.

Overall, the introduction of automated methods such as Ensemble Kalman Inversion allows us to calibrate subgrid-scale parameterizations in GCMs, as far fewer climate model integrations are required ($O(100)$ compared to $O(10^5)$. However, for high complexity GCMs, even running 100 model integrations is highly costly, which is why these are typically tuned crudely (e.g., Kodama et al., 2021). Applying EKI to intermediate complexity climate models, such as MiMA, provides useful insights into how EKI can be best leveraged for higher

complexity climate models, for example, by building a more informed prior probability distribution, in order to reduce the total number of expensive EKI iterations.

Conflict of Interest

The authors have no conflicts of interest.

Data Availability Statement

The code used in this analysis, including scripts to run MiMA and reproduce all results presented here can be found at https://doi.org/10.5281/zenodo.6629730. The codebase for Calibrate, Emulate, Sample and Ensemble Kalman Inversion are both maintained by the Climate Modeling Alliance (Clima) group and can be found at https://github.com/CliMA/CalibrateEmulateSample.jl and https://github.com/CliMA/EnsembleKalmanProcesses.jl. The Model of an idealized Moist Atmosphere (MiMA) (Garfinkel et al., 2020; Jucker & Gerber, 2017) is available at https://github.com/mjucker/MiMA.

Acknowledgments
This research was made possible by Schmidt Futures, a philanthropic initiative founded by Eric and Wendy Schmidt, as part of the Virtual Earth System Research Institute (VESRI). AS acknowledges support from the National Science Foundation through grant OAC-2004492. We thank Oliver Dunbar and Tapio Schneider for useful discussions. We are grateful to the two anonymous reviewers whose insightful comments have improved the manuscript.

References

Alexander, M. J., & Dunkerton, T. J. (1999). A spectral parameterization of mean-flow forcing due to breaking gravity waves. *Journal of the Atmospheric Sciences*, *56*(24), 4167–4182. https://doi.org/10.1175/1520-0469(1999)056<4167:aspomf>2.0.co;2

Alexander, M. J., Geller, M., McLandress, C., Polavarapu, S., Preusse, P., Sassi, F., et al. (2010). Recent developments in gravity-wave effects in climate models and the global distribution of gravity-wave momentum flux from observations and models. *Quarterly Journal of the Royal Meteorological Society*, *136*(650), 1103–1124. https://doi.org/10.1002/qj.637

Alexander, M. J., & Pfister, L. (1995). Gravity wave momentum flux in the lower stratosphere over convection. *Geophysical Research Letters*, *22*(15), 2029–2032. https://doi.org/10.1029/95GL01984

Alexander, M. J., & Rosenlof, K. H. (2003). Gravity-wave forcing in the stratosphere: Observational constraints from the upper atmosphere research satellite and implications for parameterization in global models. *Journal of Geophysical Research (Atmospheres)*, *108*, 4597. https://doi.org/10.1029/2003JD003373

Anstey, J. A., Scinocca, J. F., & Keller, M. (2016). Simulating the QBO in an atmospheric general circulation model: Sensitivity to resolved and parameterized forcing. *Journal of the Atmospheric Sciences*, *73*(4), 1649–1665. https://doi.org/10.1175/JAS-D-15-0099.1

Baldwin, M. P., Gray, L. J., Dunkerton, T. J., Hamilton, K., Haynes, P. H., Randel, W. J., et al. (2001). The quasi-biennial oscillation. *Reviews of Geophysics*, *39*(2), 179–229. https://doi.org/10.1029/1999RG000073

Barton, C. A., McCormack, J. P., Eckermann, S. D., & Hoppel, K. W. (2019). Optimization of gravity wave source parameters for improved seasonal prediction of the quasi-biennial oscillation. *Journal of the Atmospheric Sciences*, *76*(9), 2941–2962. https://doi.org/10.1175/JAS-D-19-0077.1

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Boccara, G., Hertzog, A., Vincent, R. A., & Vial, F. (2008). Estimation of gravity wave momentum flux and phase speeds from quasi-Lagrangian stratospheric balloon flights. Part I: Theory and simulations. *Journal of the Atmospheric Sciences*, *65*(10), 3042–3055. https://doi.org/10.1175/2008JAS2709.1

Brooks, S., Gelman, A., Jones, G. L., & Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo* (3rd ed.). Chapman & Hall/CRC. Retrieved from https://www.mcmchandbook.net/

Bushell, A. C., Anstey, J. A., Butchart, N., Kawatani, Y., Osprey, S. M., Richter, J. H., et al. (2020). Evaluation of the Quasi-Biennial oscillation in global climate models for the SPARC QBO-initiative. *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.3765

Castruccio, S., Hu, Z., Sanderson, B., Karspeck, A., & Hammerling, D. (2019). Reproducing internal variability with few ensemble runs. *Journal of Climate*, *32*(24), 8511–8522. https://doi.org/10.1175/JCLI-D-19-0280.1

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I., & Palmer, T. (2021). Machine learning emulation of gravity wave drag in numerical weather forecasting. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002477. https://doi.org/10.1029/2021MS002477

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., & Stuart, A. M. (2021). Calibrate, emulate, sample. *Journal of Computational Physics*, *424*, 109716. https://doi.org/10.1016/j.jcp.2020.109716

Corcos, M., Hertzog, A., Plougonven, R., & Podglajen, A. (2021). Observation of gravity waves at the tropical tropopause using superpressure balloons. *Journal of Geophysical Research: Atmospheres*, *126*(15), e2021JD035165. https://doi.org/10.1029/2021JD035165

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., et al. (2021). Process-based climate model development harnessing machine learning: I. A calibration tool for parameterization improvement. *Journal of Advances in Modeling Earth Systems*, *13*(3), e2020MS002217. https://doi.org/10.1029/2020MS002217

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: The role of internal variability. *Climate Dynamics*, *38*(34), 527–546. https://doi.org/10.1007/s00382-010-0977-x

Donner, L. J., Wyman, B. L., Hemler, R. S., Horowitz, L. W., Ming, Y., Zhao, M., et al. (2011). The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component AM3 of the GFDL global coupled model CM3. *Journal of Climate*, *24*, 3484–3519. https://doi.org/10.1175/2011JCLI3955.1

Dunbar, O. R. A., Garbuno-Inigo, A., Schneider, T., & Stuart, A. M. (2021). Calibration and uncertainty quantification of convective parameters in an idealized GCM. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2020MS002454. https://doi.org/10.1029/2020MS002454

Dunkerton, T. J. (1997). The role of gravity waves in the quasi-biennial oscillation. *Journal of Geophysical Research*, *102*(D22), 26053–26076. https://doi.org/10.1029/96JD02999

Espinosa, Z. I., Sheshadri, A., Cain, G. R., Gerber, E. P., & DallaSanta, K. J. (2022). Machine learning gravity wave parameterization generalizes to capture the QBO and response to increased CO_2. *Geophysical Research Letters*, *49*(8), e2022GL098174. https://doi.org/10.1029/2022GL098174

Freie Universität Berlin. (2007). The quasi-biennial-oscillation (QBO) data serie. *The Quasi-Biennial-Oscillation (QBO) Data Serie*. Retrieved from https://www.geo.fu-berlin.de/en/met/ag/strat/produkte/qbo/index.html

Garcia, R. R., & Richter, J. H. (2019). On the momentum budget of the Quasi-biennial oscillation in the whole atmosphere community climate model. *Journal of the Atmospheric Sciences*, *76*(1), 69–87. https://doi.org/10.1175/JAS-D-18-0088.1

Garfinkel, C. I., Gerber, E. P., Shamir, O., Rao, J., Jucker, M., White, I., & Paldor, N. (2022). A QBO cookbook: Sensitivity of the quasi-biennial oscillation to resolution, resolved waves, and parameterized gravity waves. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002568. https://doi.org/10.1029/2021MS002568

Garfinkel, C. I., & Hartmann, D. L. (2011). The influence of the quasi-biennial oscillation on the troposphere in winter in a hierarchy of models. Part I: Simplified dry GCMs. *Journal of the Atmospheric Sciences*, *68*(6), 1273–1289. https://doi.org/10.1175/2011JAS3665.1

Garfinkel, C. I., White, I., Gerber, E. P., Jucker, M., & Erez, M. (2020). The building Blocks of northern hemisphere wintertime stationary waves. *Journal of Climate*, *33*(13), 5611–5633. https://doi.org/10.1175/JCLI-D-19-0181.1

Geller, M. A., Alexander, M. J., Love, P. T., Bacmeister, J., Ern, M., Hertzog, A., et al. (2013). A comparison between gravity wave momentum fluxes in observations and climate models. *Journal of Climate*, *26*(17), 6383–6405. https://doi.org/10.1175/JCLI-D-12-00545.1

Geyer, C. (2011). Introduction to Markov chain Monte Carlo. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (Vol. 20116022). Chapman and Hall/CRC. https://doi.org/10.1201/b10905

Giorgetta, M. A., Manzini, E., Roeckner, E., Esch, M., & Bengtsson, L. (2006). Climatology and forcing of the quasi-biennial oscillation in the MAECHAM5 model. *Journal of Climate*, *19*(16), 3882–3901. https://doi.org/10.1175/JCLI3830.1

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design and optimization for the applied sciences*. Chapman Hall/CRC.

Gray, L. J. (2010). Stratospheric equatorial dynamics. In *The stratosphere: Dynamics, transport, and chemistry* (pp. 93–107). American Geophysical Union (AGU). https://doi.org/10.1029/2009gm000868

Grimsdell, A. W., Alexander, M. J., May, P. T., & Hoffmann, L. (2010). Model study of waves generated by convection with direct validation via satellite. *Journal of the Atmospheric Sciences*, *67*(5), 1617–1631. https://doi.org/10.1175/2009JAS3197.1

Gupta, A., Birner, T., Dörnbrack, A., & Polichtchouk, I. (2021). Importance of gravity wave forcing for springtime southern polar vortex breakdown as revealed by ERA5. *Geophysical Research Letters*, *48*(10), e2021GL092762. https://doi.org/10.1029/2021GL092762

Hertzog, A., Boccara, G., Vincent, R. A., Vial, F., & Cocquerez, P. (2008). Estimation of gravity wave momentum flux and phase speeds from quasi-Lagrangian stratospheric balloon flights. Part II: Results from the vorcore campaign in Antarctica. *Journal of the Atmospheric Sciences*, *65*(10), 3056–3070. https://doi.org/10.1175/2008JAS2710.1

Holt, L. A., Lott, F., Garcia, R. R., Kiladis, G. N., Cheng, Y.-M., Anstey, J. A., et al. (2020). An evaluation of tropical waves and wave forcing of the QBO in the QBOi models. *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.3827

Holton, J. R., & Lindzen, R. S. (1972). An updated theory for the quasi-biennial cycle of the tropical stratosphere. *Journal of the Atmospheric Sciences*, *29*(6), 1076–1080. https://doi.org/10.1175/1520-0469(1972)029<1076:autftq>2.0.co;2

Holton, J. R., & Tan, H.-C. (1980). The influence of the equatorial Quasi-Biennial Oscillation on the global circulation at 50 mb. *Journal of the Atmospheric Sciences*, *37*(10), 2200–2208. https://doi.org/10.1175/1520-0469(1980)037<2200:tioteq>2.0.co;2

Howland, M. F., Dunbar, O. R. A., & Schneider, T. (2022). Parameter uncertainty quantification in an idealized GCM with a seasonal cycle. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002735. https://doi.org/10.1029/2021MS002735

Iglesias, M. A. (2016). A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems. *Inverse Problems*, *32*(2), 025002. https://doi.org/10.1088/0266-5611/32/2/025002

Iglesias, M. A., Law, K. J. H., & Stuart, A. M. (2013). Ensemble Kalman methods for inverse problems. *Inverse Problems*, *29*(4), 045001. https://doi.org/10.1088/0266-5611/29/4/045001

Jewtoukoff, V., Hertzog, A., Plougonven, R., De la Cámara, A., & Lott, F. (2015). Comparison of gravity waves in the southern hemisphere derived from balloon observations and the ECMWF analyses. *Journal of the Atmospheric Sciences*, *72*(9), 3449–3468. https://doi.org/10.1175/JAS-D-14-0324.1

Jucker, M., & Gerber, E. P. (2017). Untangling the annual cycle of the tropical tropopause layer with an idealized moist model. *Journal of Climate*, *30*(18), 7339–7358. https://doi.org/10.1175/JCLI-D-17-0127.1

Kalnay, E. (2002). *Atmospheric modeling, data assimilation and predictability*. Higher Education from Cambridge University Press; Cambridge University Press. https://doi.org/10.1017/CBO9780511802270

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, *63*(3), 425–464. https://doi.org/10.1111/1467-9868.00294

Kodama, C., Ohno, T., Seiki, T., Yashiro, H., Noda, A. T., Nakano, M., et al. (2021). The Nonhydrostatic ICosahedral Atmospheric Model for CMIP6 HighResMIP simulations (NICAM16-S): Experimental design, model description, and impacts of model updates. *Geoscientific Model Development*, *14*(2), 795–820. https://doi.org/10.5194/gmd-14-795-2021

Kruschke, J. K. (2015). Chapter 7—Markov chain Monte Carlo. In J. K. Kruschke (Ed.), *Doing Bayesian data analysis* (2nd ed., pp. 143–191). Academic Press. https://doi.org/10.1016/B978-0-12-405888-0.00007-6

Lee, L. A., Carslaw, K. S., Pringle, K. J., & Mann, G. W. (2012). Mapping the uncertainty in global CCN using emulation. *Atmospheric Chemistry and Physics*, *12*(20), 9739–9751. https://doi.org/10.5194/acp-12-9739-2012

Lindzen, R. S. (1981). Turbulence and stress owing to gravity wave and tidal breakdown. *Journal of Geophysical Research*, *86*(C10), 9707. https://doi.org/10.1029/JC086iC10p09707

Lindzen, R. S., & Holton, J. R. (1968). A theory of the quasi-biennial oscillation. *Journal of the Atmospheric Sciences*, *25*(6), 1095–1107. https://doi.org/10.1175/1520-0469(1968)025<1095:atotqb>2.0.co;2

Loeppky, J. L., Sacks, J., & Welch, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, *51*(4), 366–376. https://doi.org/10.1198/TECH.2009.08040

Matsuoka, D., Watanabe, S., Sato, K., Kawazoe, S., Yu, W., & Easterbrook, S. (2020). Application of deep learning to estimate atmospheric gravity wave parameters in reanalysis data sets. *Geophysical Research Letters*, *47*(19), e2020GL089436. https://doi.org/10.1029/2020GL089436

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. https://doi.org/10.1063/1.1699114

Molod, A., Takacs, L., Suarez, M., Bacmeister, J., Song, I.-S., & Eichmann, A. (2012). *The GEOS-5 atmospheric general circulation model: Mean climate and development from MERRA to fortuna*. (GSFC.TM.01153.2012). Retrieved from https://ntrs.nasa.gov/citations/20120011790

Murphy, J. M., Booth, B. B. B., Boulton, C. A., Clark, R. T., Harris, G. R., Lowe, J. A., & Sexton, D. M. H. (2014). Transient climate changes in a perturbed parameter ensemble of emissions-driven Earth system model simulations. *Climate Dynamics*, *43*(9), 2855–2885. https://doi.org/10.1007/s00382-014-2097-5

Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E. J., Corazza, M., et al. (2004). A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A: Dynamic Meteorology and Oceanography*, *56*, 415–428. https://doi.org/10.3402/tellusa.v56i5.14462

Palmer, T. N., Shutts, G. J., & Swinbank, R. (1986). Alleviation of a systematic westerly bias in general circulation and numerical weather prediction models through an orographic gravity wave drag parametrization. *Quarterly Journal of the Royal Meteorological Society*, *112*, 1001–1039. https://doi.org/10.1002/qj.49711247406

Pathak, R., Dasari, H. P., El-Mohtar, S., Subramanian, A. C., Sahany, S., Mishra, S. K., et al. (2021). Uncertainty quantification and Bayesian inference of cloud parameterization in the NCAR single column community atmosphere model (SCAM6). *Frontiers in Climate*, *3*. https://doi.org/10.3389/fclim.2021.670740. Retrieved from https://www.frontiersin.org/article/10.3389/fclim.2021.670740

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Plougonven, R., De la Cámara, A., Hertzog, A., & Lott, F. (2020). How does knowledge of atmospheric gravity waves guide their parameterizations? *Quarterly Journal of the Royal Meteorological Society*, *146*(728), 1529–1543. https://doi.org/10.1002/qj.3732

Plumb, R. A. (1977). The interaction of two internal waves with the mean flow: Implications for the theory of the Quasi-Biennial Oscillation. *Journal of the Atmospheric Sciences*, *34*(12), 1847–1858. https://doi.org/10.1175/1520-0469(1977)034<1847:tiotiw>2.0.co;2

Priestley, M. D. K., Ackerley, D., Catto, J. L., Hodges, K. I., McDonald, R. E., & Lee, R. W. (2020). An overview of the extratropical storm tracks in CMIP6 historical simulations. *Journal of Climate*, *33*(15), 6315–6343. https://doi.org/10.1175/JCLI-D-19-0928.1

Rao, J., Garfinkel, C. I., & White, I. P. (2020). How does the Quasi-Biennial Oscillation affect the boreal winter tropospheric circulation in CMIP5/6 Models? *Journal of Climate*, *33*(20), 8975–8996. https://doi.org/10.1175/JCLI-D-20-0024.1

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Richter, I., & Tokinaga, H. (2020). An overview of the performance of CMIP6 models in the tropical Atlantic: Mean state, variability, and remote impacts. *Climate Dynamics*, *55*(9), 2579–2601. https://doi.org/10.1007/s00382-020-05409-w

Richter, J. H., Anstey, J. A., Butchart, N., Kawatani, Y., Meehl, G. A., Osprey, S., & Simpson, I. R. (2020). Progress in simulating the quasi-biennial oscillation in CMIP models. *Journal of Geophysical Research: Atmospheres*, *125*(8), e2019JD032362. https://doi.org/10.1029/2019JD032362

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods*. Springer. https://doi.org/10.1007/978-1-4757-4145-2

Roberts, G. O., & Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, *1*, 20–71. https://doi.org/10.1214/154957804100000024

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, *181*(2), 259–270. https://doi.org/10.1016/j.cpc.2009.09.018

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., et al. (2007). *Global sensitivity analysis. The primer*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470725184

Scaife, A. A., Butchart, N., Warner, C. D., & Swinbank, R. (2002). Impact of a spectral gravity wave parameterization on the stratosphere in the Met office unified model. *Journal of the Atmospheric Sciences*, *59*(9), 1473–1489. https://doi.org/10.1175/1520-0469(2002)059<1473:ioasgw>2.0.co;2

Schenzinger, V., Osprey, S., Gray, L., & Butchart, N. (2017). Defining metrics of the Quasi-Biennial Oscillation in global climate models. *Geoscientific Model Development*, *10*(6), 2157–2168. https://doi.org/10.5194/gmd-10-2157-2017

Schirber, S., Manzini, E., Krismer, T., & Giorgetta, M. (2015). The quasi-biennial oscillation in a warmer climate: Sensitivity to different gravity wave parameterizations. *Climate Dynamics*, *45*(3), 825–836. https://doi.org/10.1007/s00382-014-2314-2

Scinocca, J. F. (2003). An accurate spectral nonorographic gravity wave drag parameterization for general circulation models. *Journal of the Atmospheric Sciences*, *60*(4), 667–682. https://doi.org/10.1175/1520-0469(2003)060<0667:aasngw>2.0.co;2

Siskind, D., Eckermann, S., McCormack, J., Coy, L., Hoppel, K., & Baker, N. (2010). Case studies of the mesospheric response to recent minor, major, and extended stratospheric warmings. *Journal of Geophysical Research*, *115*, 0–3. https://doi.org/10.1029/2010JD014114

Siskind, D., Eckermann, S. D., Coy, L., McCormack, J. P., & Randall, C. E. (2007). On recent interannual variability of the Arctic winter mesosphere: Implications for tracer descent: Mesospheric interannual variability. *Geophysical Research Letters*, *34*(9). https://doi.org/10.1029/2007GL029293

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, *55*(1), 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6

Souza, A. N., Wagner, G. L., Ramadhan, A., Allen, B., Churavy, V., Schloss, J., et al. (2020). Uncertainty quantification of ocean parameterizations: Application to the K-profile-parameterization for penetrative convection. *Journal of Advances in Modeling Earth Systems*, *12*, e2020MS002108. https://doi.org/10.1029/2020MS002108

Strahan, S. E., Oman, L. D., Douglass, A. R., & Coy, L. (2015). Modulation of Antarctic vortex composition by the quasi-biennial oscillation. *Geophysical Research Letters*, *42*(10), 4216–4223. https://doi.org/10.1002/2015GL063759

Tebaldi, C., & Knutti, R. (2018). Evaluating the accuracy of climate change pattern emulation for low warming targets. *Environmental Research Letters*, *13*(5), 055006. https://doi.org/10.1088/1748-9326/aabef2

Wan, H., Rasch, P. J., Zhang, K., Qian, Y., Yan, H., & Zhao, C. (2014). Short ensembles: An efficient method for discerning climate-relevant sensitivities in atmospheric general circulation models. *Geoscientific Model Development*, *7*(5), 1961–1977. https://doi.org/10.5194/gmd-7-1961-2014

Warner, C. D., & McIntyre, M. E. (1999). Toward an ultra-simple spectral gravity wave parameterization for general circulation models. *Earth Planets and Space*, *51*(7), 475–484. https://doi.org/10.1186/BF03353209

Watson-Parris, D. (2021). Machine learning for weather and climate are worlds apart. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *379*(2194), 20200098. https://doi.org/10.1098/rsta.2020.0098

Williamson, D., & Blaker, A. T. (2014). Evolving Bayesian emulators for structured chaotic time series, with application to large climate models. *SIAM/ASA Journal on Uncertainty Quantification*, *2*(1), 1–28. https://doi.org/10.1137/120900915

Williamson, D., Blaker, A. T., & Sinha, B. (2017). Tuning without over-tuning: Parametric uncertainty quantification for the NEMO ocean model. *Geoscientific Model Development*, *10*, 1789–1816. https://doi.org/10.5194/gmd-10-1789-2017

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*(7), 1703–1729. https://doi.org/10.1007/s00382-013-1896-4

Wright, C. J., Osprey, S. M., Barnett, J. J., Gray, L. J., & Gille, J. C. (2010). High resolution dynamics Limb sounder measurements of gravity wave activity in the 2006 Arctic stratosphere. *Journal of Geophysical Research*, *115*. https://doi.org/10.1029/2009JD011858

Zhao, M., Golaz, J.-C., Held, I. M., Guo, H., Balaji, V., Benson, R., et al. (2018). The GFDL global atmosphere and land model AM4.0/LM4.0: 2. Model description, sensitivity studies, and tuning strategies. *Journal of Advances in Modeling Earth Systems*, *10*(3), 735–769. https://doi.org/10.1002/2017MS001209