Bayesian History Matching applied to the calibration of a gravity wave parameterization

Robert C King[1], Laura A Mansfield[1], and Aditi Sheshadri[1]

[1]Stanford University

December 27, 2023

Abstract

Breaking atmospheric gravity waves in the tropical stratosphere are essential in driving the roughly two year oscillation of zonal winds in this region known as the Quasi-Biennial Oscillation (QBO). As Global Climate Models (GCM)s are not typically able to directly resolve the spectrum of waves required to drive the QBO, parameterizations are necessary. Such parameterizations often require knowledge of poorly constrained physical parameters. In the case of the spectral gravity parameterization used in this work, these parameters are the total equatorial gravity wave stress and the half width of phase speed distribution. Radiosonde observations are used to obtain the period and amplitude of the QBO, which are compared against values obtained from a GCM. We utilize two established calibration techniques to obtain estimates of the range of plausible parameter values: History Matching & Ensemble Kalman Inversion (EKI). History Matching is found to reduce the size of the initial range of plausible parameters by a factor of 98%, requiring only 60 model integrations. EKI cannot natively provide any uncertainty quantification but is able to produce a single best estimate of the calibrated values in 25 integrations. When directly comparing the approaches using the Calibrate, Emulate, Sample method to produce a posterior estimate from EKI, History Matching produces more compact posteriors with fewer model integrations at lower ensemble sizes compared to EKI; however, these differences become less apparent at higher ensemble sizes.

Bayesian History Matching applied to the calibration of a gravity wave parameterization

Robert C. King[1], **Laura A. Mansfield**[1], **Aditi Sheshadri**[1]

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

Key Points:

- History Matching and Ensemble Kalman Inversion were used to calibrate parameters in a gravity wave parameterization using QBO observations.
- History Matching was found to rapidly and compactly produce an estimate of the plausible space of parameters when compared to EKI.
- EKI was found to be strong at single best estimates of the calibrated parameters at low ensemble sizes requiring few iterations.

Corresponding author: Robert C. King, `robcking@stanford.edu`

Abstract

Breaking atmospheric gravity waves in the tropical stratosphere are essential in driving the roughly two year oscillation of zonal winds in this region known as the Quasi-Biennial Oscillation (QBO). As Global Climate Models (GCM)s are not typically able to directly resolve the spectrum of waves required to drive the QBO, parameterizations are necessary. Such parameterizations often require knowledge of poorly constrained physical parameters. In the case of the spectral gravity parameterization used in this work, these parameters are the total equatorial gravity wave stress and the half width of phase speed distribution. Radiosonde observations are used to obtain the period and amplitude of the QBO, which are compared against values obtained from a GCM. We utilize two established calibration techniques to obtain estimates of the range of plausible parameter values: History Matching & Ensemble Kalman Inversion (EKI). History Matching is found to reduce the size of the initial range of plausible parameters by a factor of 98%, requiring only 60 model integrations. EKI cannot natively provide any uncertainty quantification but is able to produce a single best estimate of the calibrated values in 25 integrations. When directly comparing the approaches using the Calibrate, Emulate, Sample method to produce a posterior estimate from EKI, History Matching produces more compact posteriors with fewer model integrations at lower ensemble sizes compared to EKI; however, these differences become less apparent at higher ensemble sizes.

Plain Language Summary

Atmospheric gravity waves (GWs) are buoyancy driven oscillations which propagate through the atmosphere and deposit momentum where they break. This momentum exchange plays a significant role in setting various large-scale atmospheric phenomena, of which a prominent example is the Quasi Biennial Oscillation, a roughly two year oscillation of winds in the tropical stratosphere. Many of the waves responsible for creating these large scale patterns are too small to be simulated by climate models. Thus, we use parameterizations to estimate their impact on the large scale. These parameterizations have settings that require tuning, to enable the model to produce variability that matches the observed climate. In this work, we utilize and compare two techniques: History Matching and Ensemble Kalman Inversion. These methods are combined with observations of the Quasi Biennial Oscillation to tune the settings for the gravity wave parameterization.

1 Introduction

Global Climate Models (GCM)s are powerful tools for understanding and predicting the evolution of the Earth's climate. For reasons of computational cost, the current generation of climate models has a horizontal resolution of $\mathcal{O}$(100km) resolution in the horizontal. Motions on scales smaller than this model resolution and which vary on time scales smaller than a model time step are not explicitly resolved, but can significantly impact the resolved scales of motion.

One such subgrid-scale process is atmospheric gravity waves (GW)s, which are generated in the atmosphere by a wide range of sources including mountains, deep convective storms and fronts (Fritts & Alexander, 2003). The horizontal scale of these GWs can range from tens to thousands of kilometers (Alexander et al., 2010). GWs are responsible for substantial momentum transport from their source region to higher levels in the atmosphere, where they break and deposit the momentum into the mean flow (Fritts & Alexander, 2003). This breaking of gravity waves in the stratosphere plays a substantial role in driving large scale atmospheric patterns, including the Quasi-Biennial Oscillation (QBO).

The QBO is the dominant mode of variability in the tropical stratosphere and consists of alternating descending westerly and easterly zonal winds with a period of around 28 months. The QBO is forced by a mixture of various tropical waves (Holton & Lindzen, 1972). However, to simulate a spontaneous QBO in models, the impact of small scale GWs (approximated by GW parameterizations), appears crucial (Dunkerton, 1997; Lindzen & Holton, 1968).

In practice, GW parameterizations can be divided into two classes, orographic parameterizations (Lott & Miller, 1997), useful for studying the impact of stationary mountain waves, and non-orographic parameterizations which typically utilize a spectrum of gravity wave phase speeds. We concern ourselves with the latter, specifically the commonly used parameterization developed by Alexander and Dunkerton, henceforth referred to as AD99 (Alexander & Dunkerton, 1999). In practice, default parameter settings are chosen manually based on whether a given parameter produces realistic behavior of large scale, observable patterns known to be driven by GWs.

Whilst these default choices are often sufficient to test the implementation of a parameterization, the choice of parameters is rarely optimal. The task of obtaining an optimal set of parameters based on observations of a related phenomenon is known as *calibration*. Calibration can be formulated as an inverse problem in which a complex model, which is a function of parameterization settings, outputs some estimate of a real world observable. In this work an intermediate complexity GCM implementing AD99 was used to output predictions of the QBO period and amplitude. The root mean squared error (RMS) between the predictions and the observations weighted by the uncertainties was used as the loss function for the calibration. Due to the computational cost of running such a GCM, this loss function cannot practically be optimized by conventional gradient descent methods.

Various classes of methods exist to solve inversion problems. In this work, we will utilize an approach known as Bayesian History Matching. This approach was initially developed to calibrate models for oil exploration (Craig et al., 1997) and has found wide utility in various disciplines. This includes in calibrating models of galactic formation (Williamson et al., 2013), HIV disease transmission (Andrianakis et al., 2015) and recently in calibrating multi timescale dynamical systems (Lguensat et al., 2023).

During each iteration of history matching, the current "plausible" parameter space is sampled and forward model integrations at the sampled points are used to obtain estimates of the observables. An emulator, trained on the results of the model integrations is then used to predict the observables across the space. By comparing these predictions to the true observables we calculate an implausibility statistic which is minimized in regions of space where the predictions agree with the observations or those with high uncertainties. By determining the regions where this implausibility is below a certain threshold we obtain the "Not Yet Ruled Out" (NROY) space (Lguensat et al., 2023), a uniform space of parameters that, relative to the uncertainties, simulate a QBO consistent with observations.

An alternate calibration method known as Ensemble Kalman Inversion (EKI) was investigated on AD99 in a previous study (Mansfield & Sheshadri, 2022); and has also been utilized in the calibration of other parameterization schemes, e.g. (Dunbar et al., 2021). EKI is a gradient free optimization method, which converges upon a singular point that minimizes a loss function (Kovachki & Stuart, 2019; Iglesias et al., 2013). Whilst an emulator is not required for the update step of the calibration in EKI, it is required in order to reconstruct the complete posterior distribution to obtain a structure that is analogous to the NROY space in history matching (Cleary et al., 2021), a process known as Calibrate, Emulate, Sample (CES).

In this paper, we present the results of applying an implementation of History Matching for calibrating the AD99 parameterization and a comparison to the EKI calibration method (Mansfield & Sheshadri, 2022). The method and theory of this technique in addition to the emulator development are described in section 2. The results of the history matching algorithm are then presented in section 3, with a comparison to EKI made in section 3.1. A discussion of the relative ability of EKI and History Matching to calibrate the AD99 parameterization is presented in section 4.

2 Method

2.1 Computational Configuration

In this investigation, we utilize the Model of an Idealized Moist Atmosphere (MIMA) (Jucker & Gerber, 2017; Garfinkel et al., 2020), an intermediate complexity GCM that contains an implementation of the AD99 parameterization. The model is run at a T42 spectral resolution using 40 vertical levels on a 128x64 longitude-latitude grid. This corresponds to a resolution of around 310km at the equator, far too coarse to directly resolve much of the spectrum of gravity waves (Baldwin et al., 2001). In order to capture a sufficient number of complete QBO cycles to characterize the distribution, 20 years of forward integration are performed. A mixture of cold start and hot start integrations are utilized in this investigation, with cold starts initialized with a uniform temperature of 260K and with a spin up period of 20 years. Hot start integrations utilized already initialized MiMA integration states containing a QBO, which only required a 2 year spinup period and are used once the cold start runs were completed.

As we calibrated based on observations of the QBO in this investigation, we focused on the tropical parameters of AD99 following the approach of Mansfield & Sheshadri (Mansfield & Sheshadri, 2022). These are the $c_w^{tropics}$ & B_t^{eq} parameters, henceforth referred to as c_w and B_t. The former of these parameters sets the half width of the half maximum Gaussian spectrum of GW phase speeds that will be utilized by AD99 within the tropics. The B_t factor corresponds to the equatorial gravity wave total momentum stress and is used within AD99 to set the GW intermittency factor via a re-scaling of the GW spectrum. Neither of these parameters are well constrained by observations and as such form the target parameters for our calibration.

2.2 Observations of the QBO

Radiosonde observations, primarily over Singapore, which were collated by the Freie Universität Berlin (Kunze, 2007) are used as reference data for the QBO. Specifically, monthly averaged zonal wind speeds at the 10hPa level are used. A 5 month rolling mean is used to remove noise and high frequency components of the signal that are not due to the QBO. The QBO period is calculated using the Transition Time (TT) method commonly employed by other studies (Bushell et al., 2022; Schenzinger et al., 2017; Richter et al., 2020). In this method, the signal is divided into individual periods based on the transition from the westerly to the easterly phase, which then allows the period to be calculated directly as the time difference between each transition. This yields a sample of the QBO periods from which an estimate of the population mean with an associated error is calculated via the Central Limit Theorem (CLT). The QBO amplitude is calculated from the same smoothed signal of the zonal wind, u, by calculating: $(u_{max} - u_{min})/2$ for each individual QBO cycle calculated via the TT method above. As with the period, we use the CLT to determine an estimate for the QBO mean amplitude, with the associated error calculated as $\sigma/\sqrt{N}$.

Using this method, the mean period of the QBO is calculated to be $\overline{T_{QBO}} = 27.92 \pm 0.86$ months and the mean amplitude is determined to be: $\overline{A_{QBO}} = 22.90 \pm 0.52\, m/s$. When applied to the model output from MiMA, the zonal wind component at the level closest to 10hPa zonally averaged from 5°S to 5°N is used. The same TT method is employed to extract a distribution of the periods and amplitudes of the smoothed signal which are then averaged.

2.3 History Matching

The objective of history matching is to iteratively reduce the size of the not yet ruled out (NROY) space of parameters $\boldsymbol{\theta}$ that go into a model $\boldsymbol{f}(\boldsymbol{\theta})$ that produces as output an estimate of some physical observable $\boldsymbol{y}$:

$$\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{y} + \boldsymbol{\epsilon_f} \tag{1}$$

$$\boldsymbol{\epsilon_f} \sim N(0, \Sigma_f) \tag{2}$$

Where ϵ_f is the model uncertainty in predicting $\boldsymbol{y}$. This error is assumed to be drawn from a zero mean Gaussian distribution with covariance Σ_f. We further assume that errors in the prediction of each component of y are independent and thus Σ_f is a diagonal matrix. History Matching measurements $\boldsymbol{z}$ of observable $\boldsymbol{y}$ in order to determine which inputs $\boldsymbol{\theta}$ give plausible values of the observable. Such measurements $\boldsymbol{z}$ will also contain an error term:

$$\boldsymbol{z} = \boldsymbol{y} + \boldsymbol{\epsilon_z} \tag{3}$$

$$\boldsymbol{\epsilon_z} \sim N(0, \Sigma_z) \tag{4}$$

Where again ϵ_z represents the error in the observations of the physical process $\boldsymbol{y}$ which is also assumed to follow a zero mean Gaussian with each observation being independent of each other. In this investigation $\boldsymbol{f}$ represents forward integrations of MIMA and thus our chosen parameters, θ, correspond to the aforementioned settings of the AD99 parameterization:

$$\boldsymbol{\theta} = (c_w, B_t) \tag{5}$$

Meanwhile the outputs of this model are the mean period and amplitude of the QBO in the zonal wind component at the 10hPa level calculated using the TT method as described above for the radiosonde data:

$$\boldsymbol{f}(\boldsymbol{\theta}) = (\overline{T_{QBO}}, \overline{A_{QBO}}) \tag{6}$$

The History Matching procedure requires the specification of some initial NROY space, typically the largest possible range of plausible values of $\boldsymbol{\theta}$. Based on domain knowledge we determined that the initial plausible range of phase speed half-widths ranged from 5m/s to 80m/s, whilst the plausible maximum equatorial momentum fluxes where chosen to range from 1mPa to 7mPa.

As forward integrations of a GCM are expensive, we wish to minimize the number of required integrations. In the conventional history matching approach this is achieved by developing an *emulator* that is trained on a small number of true integrations and predicts our target vector $\boldsymbol{z}$ across the current NROY space. The points for these integrations are randomly sampled with a space filling objective. To that end, we utilize Minmax Latin Hypercube sampling (LHS) (McKay et al., 1979), which is a computationally efficient method for sampling a uniform unit hypercube. To draw N samples from a k dimensional hypercube space, this method works by subdividing the space into a grid where each axis contains n smaller hypercubes of size $(1/n)^k$. We then pick n of these smaller cubes as our sample points at random, subject to the criterion that along each axis of the grid, each $(1/n)$ subsection contains one and only one smaller hypercube. In the 2-dimensional case this is analogous to the problem of trying to position chess rooks such that no rook directly attacks another (Golomb & Posner, 1964). The additional min-max constraint enforces the constraint that out of all possible valid configurations, we pick configurations such that the minimum distance between any two sub hypercubes is maximized. This eliminates trivial configurations such as selecting samples along any diagonal of the space.

In this work, we investigate the impact in sampling with different number of points at each iteration to determine the optimal tradeoff between computation time and emulator accuracy. Specifically, we determined the impact of sampling 5, 10, 20 & 50 points from the NROY space at each iteration, with more points allowing for a more accurate emulator at the expense of greater computational cost. The number of points sampled at each iteration is denoted as N. After completing the corresponding integrations we obtained a training set $\{\boldsymbol{\theta_i}\}$ of points with associated estimates for the QBO observables $\{\boldsymbol{z_i}\}$ which have an estimated error $\{\boldsymbol{\sigma_i}\}$.

For each iteration, once the parameters were sampled and the forward integrations completed, we follow the approach of Andrianakis (Andrianakis et al., 2015) and develop a Gaussian Process (GP) based emulator to estimate $\boldsymbol{z}$ across the entirety of the current NROY space. Gaussian Processes are useful for building emulators as they are capable of taking a distribution over an infinite range of basis functions that are conditioned on the training dataset. Specifically, we trained one Gaussian Process Regression (GPR) emulator, implemented via the scikit-learn library, per each dimension of the output vector of the model. Thus with a 2 dimensional output vector, 2 independent GPs are trained on the 2D input parameter space. The input parameters are scaled each wave to have zero mean and unit variance along each feature axis. Additionally, each output training label in the GPR is normalized to have a zero mean and unit variance which typically gives the best training performance for the default case where a zero mean, unit variance prior is used in the GPR.

One important pathological case that needed to be considered for the input data to the GPR training was the case where no QBO was present in the output signal, defined as no transition in the zonal wind direction across the entire 20 year window. There are a variety of approaches to deal with these cases, however in this work we decided to exclude data points with no QBO present. This was done because the QBO breakdown results in a non smooth critical transition in the QBO period and amplitude. Thus if these points were included in the emulator training it would likely be captured poorly by the emulator and greatly bias the mean value of the emulator predictions. For history matching, the best way to deal with such anomalous points is to manually exclude regions of the space that are clearly implausible such as those with B_t or c_w values below those needed to drive a QBO.

The choice of kernel in a GPR is also critical for setting the smoothness of the emulated functions as well as for setting the scale of the emulator variance at each point. For this work a Radial Basis Function (RBF) kernel as defined below is used:

$$K(\theta, \theta') = C \exp\left(\frac{-|\theta - \theta'|^2}{2l}\right) \tag{7}$$

where l and C were kernel hyper parameters representing the characteristic length scale and scale factor. The standard scaling of our input and output parameters gives a convenient choice for our length scale and scale factor of $C = l = 1$, as the standard deviation of the input points will by construction be 1.

In addition to the hyperparameter tuning, a "nugget" term is provided in the form of an array of the point wise training data noise estimated during the TT method described above. The nugget is added to the diagonal terms of the above kernel matrix during training and has been shown to be a useful term for emulating climate model output (Williamson et al., 2015).

The final stage of an iteration of history matching is to calculate the *implausibility*, a measure of how likely it is that a given point of the current NROY space is consistent with observations subject to some user defined cutoff. Thus a small implausibility implies that either a parameter configuration is predicted to produce an output very close to the observations or that there is sufficiently large uncertainty in the predicted value at that point that the point must remain in consideration for future iterations. In the standard univariate History Matching case, the implausibility takes the form of the Mahalanobis distance:

$$I = \frac{|\hat{f}(\theta) - z|}{\sqrt{\sigma_z^2 + \sigma_{\hat{f}}^2}} \tag{8}$$

Where σ_z is the observational uncertainty and $\sigma_{\hat{f}}$ is the uncertainty in the emulator prediction. Typically for the univariate case (Andrianakis et al., 2015; Lguensat et al., 2023; Couvreux et al., 2021; Williamson et al., 2013), Pulksheim's rule is invoked which states that for a continuous unimodal distribution 95% of the probability mass lies within 3 standard deviations of the mean value (Pukelsheim, 1994). For the multivariate case, as in this work with a 2-dimensional $\boldsymbol{z}$ there are various definitions of the implausibility. One approach is to calculate for the jth component of $\boldsymbol{z}$, the corresponding univariate implausibility I_j and then define the total implausibility as:

$$I = \max_j I_j \tag{9}$$

A more robust method is to follow the approach of Vernon (Vernon et al., 2010) and calculate a full multivariate implausibility of the form:

$$I^2 = \left(\hat{\boldsymbol{f}} - \boldsymbol{z}\right)^T \left(\Sigma_z + \Sigma_{\hat{f}}\right)^{-1} \left(\hat{\boldsymbol{f}} - \boldsymbol{z}\right). \tag{10}$$

Here Σ_z is the covariance matrix of the observational uncertainty defined previously and $\Sigma_{\hat{f}}$ is the covariance matrix of the emulator at a given point $\boldsymbol{x}$ in the parameter space. As the emulator for each component of $\boldsymbol{f}$ is trained independently of the others, $\Sigma_{\hat{f}}$ will also be a diagonal matrix, however one could consider a more advanced multivariate emulator that outputs a non diagonal matrix for $\Sigma_{\hat{f}}$. Equation 10 demonstrates that the implausibility corresponds to a sum of squared random variables, which will follow a χ^2 distribution with an order equal to the number of dimensions in the output space. Thus we can use a standard χ^2 to reject (Vernon et al., 2010). In the specific case of this investigation a significance level of 1% corresponds to a cut off implausibility squared of $I^2_{max} = 9.21$ a threshold which is very similar to the $I_{max} = 3$ used to invoke Pulkshiem's rule in the univariate case.

Once this cut-off is applied, the next wave NROY space is obtained, from which additional samples can be drawn. As this space is unlikely to be a rectangular space, the LHS sampling approach cannot be utilized. For simplicity, conventional exclusion based sampling was performed in these cases. By running additional MiMA integrations at these points, a new emulator can be trained utilizing the new points alongside the existing ones which allows more of the NROY space to be ruled implausible with each iteration resulting in a chain of GPR emulators being developed (Salter & Williamson, 2016). These iterations can be performed continually until the NROY space is sufficiently converged. In this work, this is defined as when the fractional change in the area of the NROY space between consecutive iterations is below 5%.

2.4 Ensemble Kalman Inversion

In this work we also compare the results of the calibration obtained via History Matching with that obtained by Ensemble Kalman Inversion. The EKI algorithm can be considered as an inverse formulation of the ensemble Kalman filter (EnKF), beginning with some prior set of parameters $\{\boldsymbol{\theta}^{(n)}\}$ which are progressively updated by comparison between the estimates of observables at these parameters with the true observables. This is achieved by performing a global minimization (Kovachki & Stuart, 2019) of the Mahlobonis distance, defined in equation 8. For a prediction of some state $f(\theta)$ as defined in equation 1, the nth ensemble member is updated via the following update equation (Iglesias et al., 2013):

$$\boldsymbol{\theta}_{t+1}^{(n)} = \boldsymbol{\theta}_t^{(n)} + C_{f\theta}(\Gamma + C_{ff})^{-1}(\boldsymbol{y} \ - \boldsymbol{f}(\boldsymbol{\theta})) \tag{11}$$

Where C_{ff} is the empirical covariance between the forward integrations for all ensemble members and Γ is the error covariance matrix representing the uncertainty in observations and predictions of the observations. Meanwhile $C_{f\theta}$, is a cross covariance matrix defined as:

$$(C_{f\theta})_{ij} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{f}(\boldsymbol{\theta}^{(n)})_i - \overline{\boldsymbol{f}_i})(\boldsymbol{\theta}_j^{(n)} - \overline{\boldsymbol{\theta}_j}) \tag{12}$$

At each iteration t, the current best estimate of the calibrated parameter values is taken as the ensemble mean. Iterations are run continually until the ensemble mean converges to a fixed value. Unlike History Matching, EKI is a optimization algorithm which alone does not provide any estimates of the distribution of the plausible parameter and thus it cannot be used for uncertainty quantification. To address this, the Calibrate, Emulate, Sample (CES) approach developed by Cleary et al. (Cleary et al., 2021) can be used to draw samples from the calibrated posterior distribution of parameters. Under CES, a Gaussian Process Emulator, as described in the previous section, is trained on the entire ensemble of all timesteps, in order to predict the observables over the entire parameter space. In contrast to History Matching, where we train a chain of emulators at each iteration to further refine the current NROY space, under CES we need a single emulator that is able to perform well across the entire parameter space.

Several adjustments are therefore made to the GP architecture described in section 2.3, including using a “MinMax” scaler to transform the input parameters, as opposed to a zero mean unit variance standard scaler. This is needed under EKI as in the infinite time limit, ensemble members converge to a single point which under a standard scaler pushes the early timestep points farther away from the origin, resulting in degraded emulator performance away from the converged point. Under “MinMax” scaling each input parameter is re-scaled such that the total range spans in the interval $[0, 1]$. This choice removes the basis for choosing a fixed length scale of 1 in the RBF kernel in equation 7, therefore hyper-parameter tuning is required during the emulator training. In addition, to improve the performance of the emulator over the wider parameter space a white noise kernel is added to account for unresolved noise as defined below:

$$\kappa(\theta, \theta') = \begin{cases} \sigma^2 & \text{if } \theta = \theta' \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

In order to tune the above hyper-parameters, the log marginal likelihood $p(y|\theta; l, C, \sigma)$ is optimized in accordance with the method described by Rasmussen (Rasmussen & Williams, 2005). This optimization approach naturally tends to favour hyper-parameter choices that give models of intermediate complexity, balancing model complexity with quality of model fit.

Once this emulator is obtained we assume a Gaussian function at each point, yielding a likelihood function of the form:

$$p(y|\theta) = \frac{1}{\sqrt{\det \Gamma}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{f}})^T \Gamma(\theta)^{-1} (\boldsymbol{y} - \hat{\boldsymbol{f}})\right) \tag{14}$$

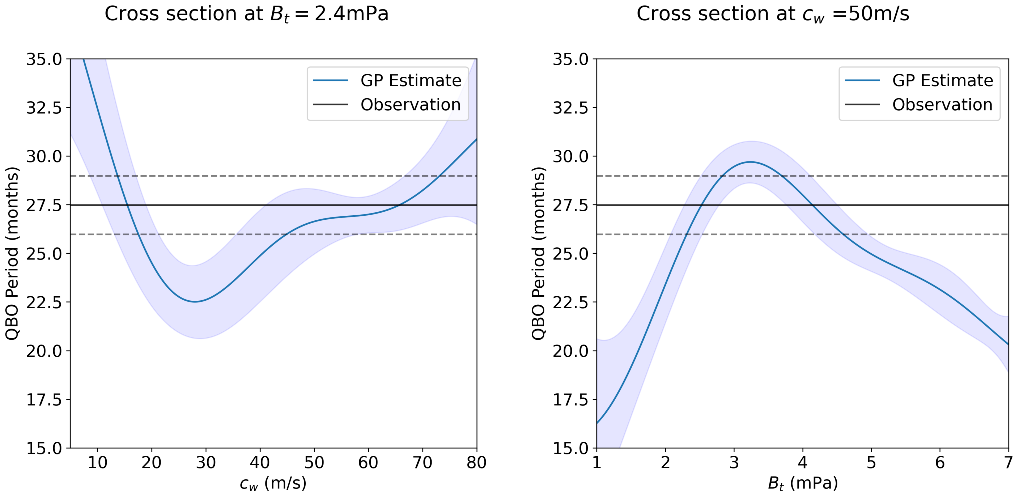
Where in the above, $\hat{\boldsymbol{f}}$ represents the mean value predicted by the GP emulator at point θ. By use of Bayes' law combined with the uniform prior distribution specified previously we may calculate the posterior distribution $p(\theta|y)$. The Metropolis-Hastings algorithm, a Markov Chain Monte Carlo method, is then used to sample this posterior distribution (Metropolis et al., 1953).

For this work, EKI runs are launched with the same ensemble sizes, N, as those used in history matching as the number of sample points per iteration (5, 10, 20 & 50 points). In addition, the LHS samples drawn in the first iteration of history matching were also used as the initial ensemble members for EKI. This setup allowed for a comparison between the convergence characteristics of EKI and History Matching to be conducted, noting that for both approaches the time taken to perform the forward integration, $\boldsymbol{f}(\boldsymbol{\theta})$ on each sample far exceeds the time taken to perform the calibration step.

3 Results

The first test case for History Matching that is investigated utilized $N = 50$ sample points. MiMA integrations are performed to train the GP emulator, and its performance for the QBO period is demonstrated in Figure 1. Two cross sections are indicated where B_t and c_w are kept at constant values in θ space. Indicated in black is the observed value for the QBO period as taken from the radiosonde data along with the associated 95% confidence interval of the observational uncertainty, indicated using dashed black lines. Indicated in solid blue in figure 1 is the mean GP prediction with the shaded blue indicating the 95% confidence interval for the prediction. The viability of the emulator is validated by withholding a single point out of the training set and using it as a validation point. For the case for the single point withheld in the 50 point case, the emulator is capable of producing a prediction compatible with the withheld point. We use a two sided t-test to determine whether the emulated mean value for the QBO period and amplitude is consistent with that of the withheld point. The test statistic value for the QBO period is 0.73 and for the QBO amplitude the test statistic is -1.23. The p-value for both of these statistics lie within a standard significance level of 5%, indicating consistency of the emulator predictions with the MiMA GCM.

The emulator predictions in Figure 2 suggest that the emulator has learnt non trivial relationships between the input parameters and the QBO statistics. For example, for a c_w greater than approximately 20m/s, the QBO amplitude is primarily set by B_t. In the emulator for the QBO period, it is observed that there is a horseshoe shaped region in which the QBO period is predicted to be consistent with observations. Both these features are useful to explain the structure of the implausibility in figure 3, where we see that the gradients in the implausibility space are substantially greater along the B_t axis than the c_w axis, forming a "banana" shaped region. The form of this space resembles that obtained by Mansfield and Sheshadri (Mansfield & Sheshadri, 2022) when an uncertainty quantification analysis was performed on AD99 using EKI.



(a) (b)

Figure 1: Demonstration of the Gaussian Process Emulator trained on 50 samples taken at two distinct cross-sections. Indicated is the mean GP estimate in solid blue with the 95% confidence interval shaded. The solid black line indicates the observed QBO period with the dashed black lines indicating the 95% confidence interval in the observational value.

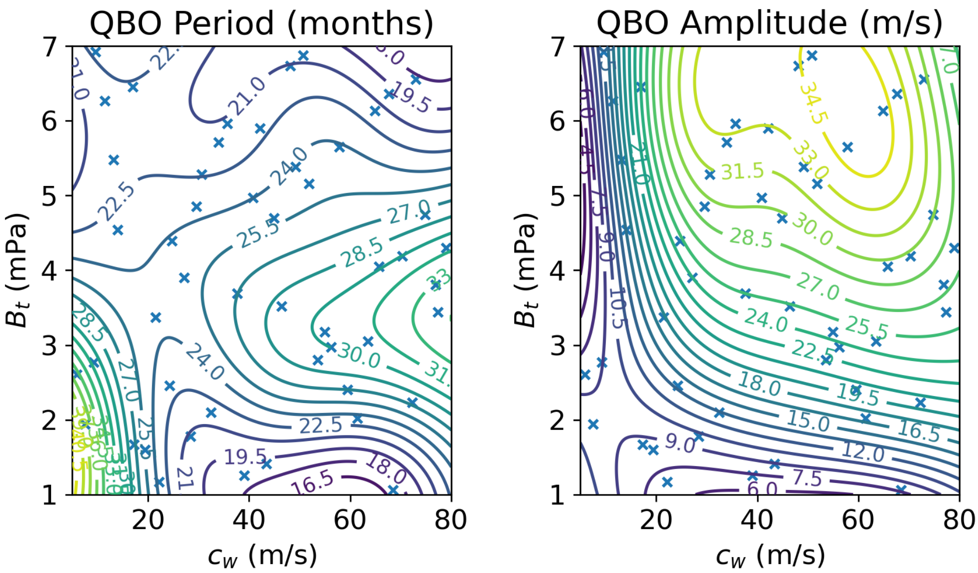


Figure 2: Emulator predictions across the initial NROY space for both the QBO period and the QBO amplitude. Training points for the emulator are indicated with blue crosses.

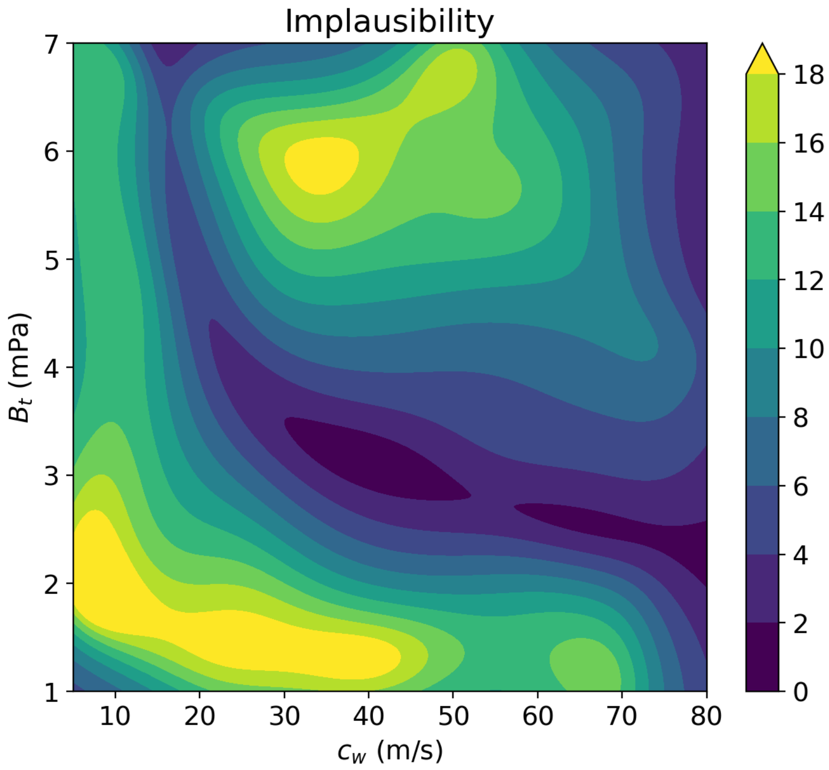


Figure 3: Calculated implausibility between the emulator in figure 2 and the QBO observations. Lower values imply regions of space that either agree more with observations or have higher uncertainties.

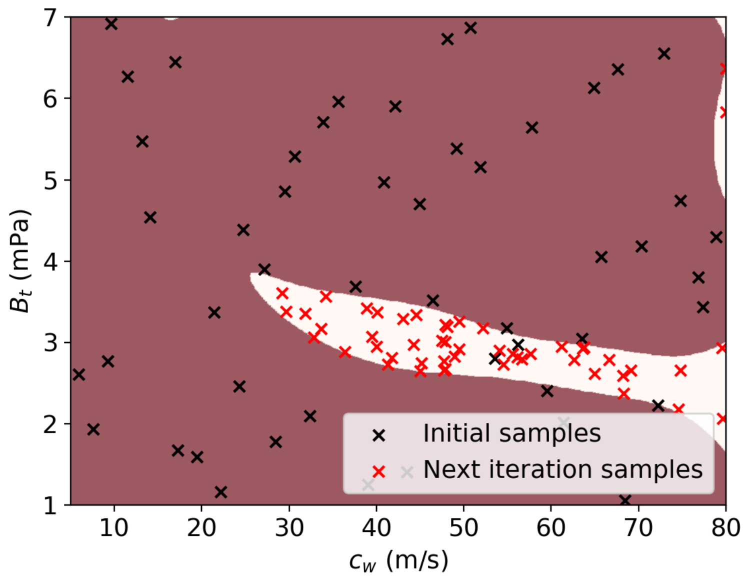


Figure 4: Demonstration of applying the χ^2 exclusion criterion and uniformly sampling the next iteration NROY space.

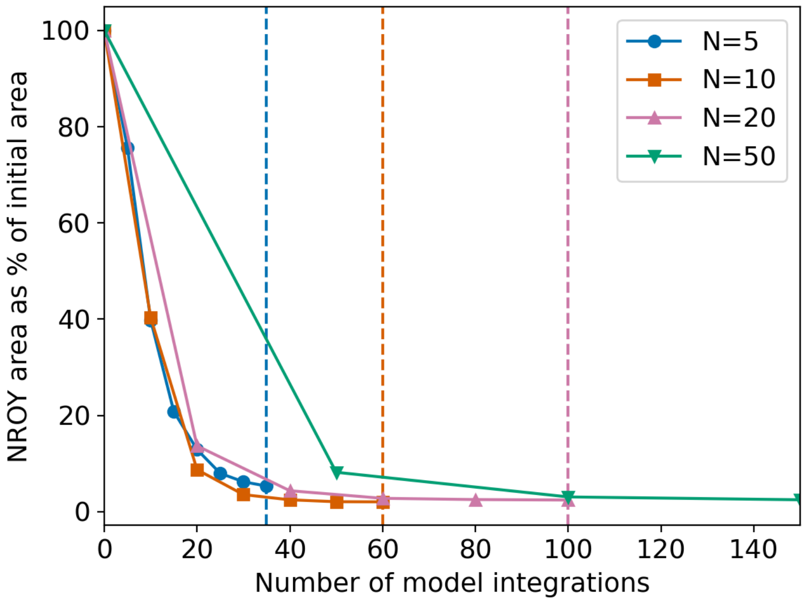


Figure 5: Comparison of the area of the NROY space as function of the number of 20 year forward integrations of MiMA that were required indicated for various number of sample points per iteration.

After applying the implausibility cutoff using a χ^2 distribution, we obtain the next iteration NROY space from which samples can be drawn uniformly. These are indicated in figure 4, which also shows the evolution of the NROY space and the samples taken for the next iteration of history matching. As seen, after this first iteration the area of the NROY space is reduced substantially, by a factor of 91.8%.

We define the NROY space as being converged once the relative change in the area of the NROY space from one iteration to the next is less than 5%. For measuring the speed of convergence, a convenient metric is the total number of forward integrations of MiMA that needed to be performed, as this represents by far the largest computational cost of the calibration. For the $N = 50$ case demonstrated above, convergence was obtained after 5 iterations, which required a total of 250 forward integrations of MiMA.

As mentioned in section 2, a range of N are investigated. A reduced number of sample points will result in a less accurate and confident emulator, however this has the benefit that the emulator will be updated more frequently. This can allow for obviously implausible regions of space to be ruled out without requiring that region of space to be directly sampled. Figure 5 displays the area of the NROY space against the cumulative number of MiMA forward integrations that were performed for each sample size. It can be seen that using fewer sample points per iteration attains convergence with the fewest model integrations, with $N = 5$ converging after 40 forward integrations of MiMA. However, as indicated in figure 5, this convergence is reached with a larger NROY space which was greater than 5% of the original space area. This was substantially higher than configurations with higher N, all of which achieved convergence with an NROY area of less than 2.5% of the original area.

As such, a configuration with $N = 10$ is seen to be the best performing, reaching convergence after 60 forward integrations of MiMA. The full chain of subsequent NROY

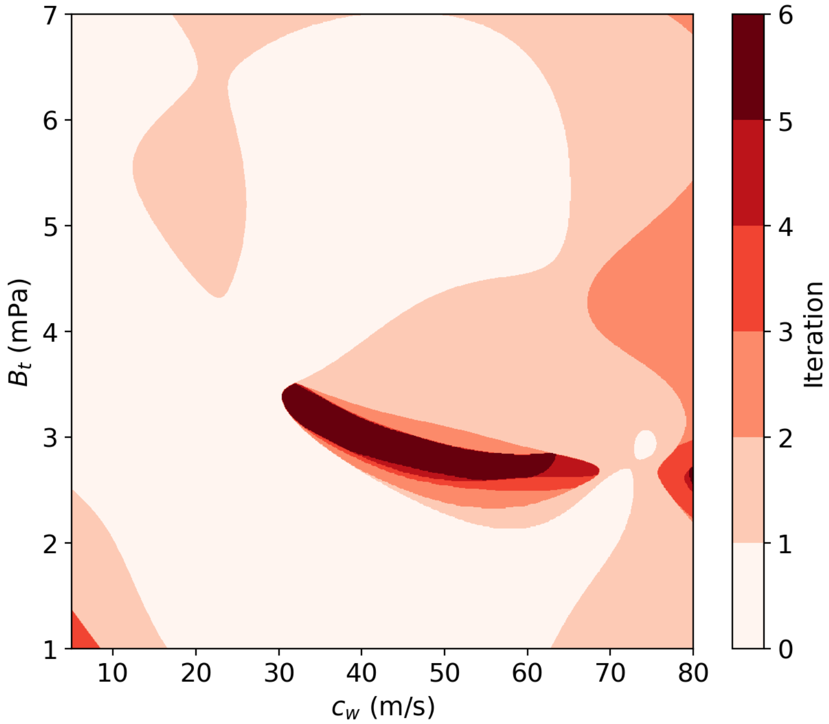


Figure 6: Demonstration of the convergence of the NROY space for the $N = 10$ case.

spaces can be seen in figure 6. The final converged area of this run was 2.06% the size of the original space area for this configuration and is centered on the point $\overline{c_w} = 45.62$ m/s & $\overline{B_t} = 2.94$ mPa. As mentioned by various authors, History Matching does not give preference to any one point located within the final NROY space, as all points are assumed to be equally "plausible" (Andrianakis et al., 2015; Williamson et al., 2015).

3.1 Comparison with Ensemble Kalman Inversion

As introduced in section 2.4, the Ensemble Kalman Inversion method is also used for the calibration of AD99 parameters utilizing the same QBO mean statistics as the ground truth observations. Analogously to History Matching, ensemble sizes of 5, 10, 20 & 50 were used. In the optimization framework of EKI, these ensemble sizes can be considered similar to batch sizes in mini-batch gradient descent, where smaller batches run quicker however take less accurate steps whilst larger batches are more computationally intensive with more accurate individual steps.

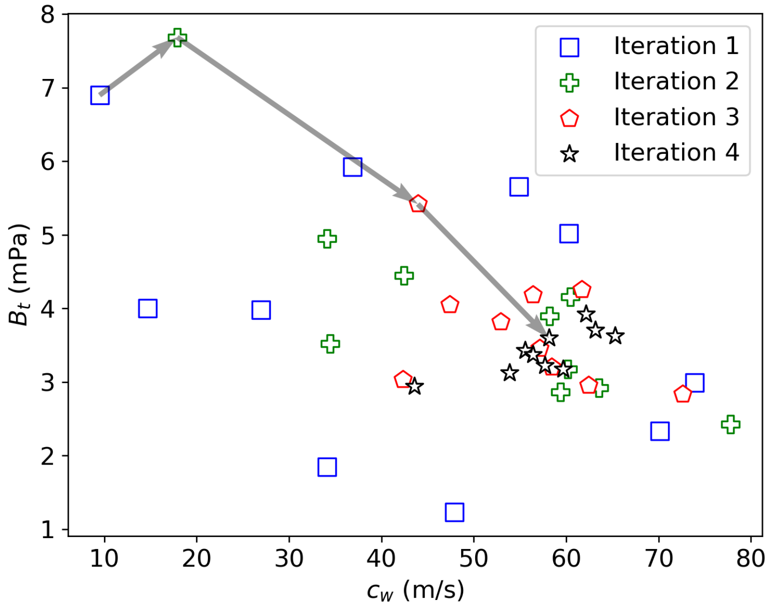


Figure 7: Positions of ensemble members during EKI for the first 4 iterations with an ensemble size of 10. The grey arrow indicates the trajectory taken by a single member under EKI

Figure 7 shows the ensemble members for the first 4 iterations of EKI using an ensemble size of 10 particles with the trajectory of a single member of EKI, indicated in grey. It is seen that the ensemble members gradually converge towards the bottom right of the figure, which is similar to the location seen with History Matching. The ensemble mean represents the current best estimate of the optimal parameter value. The evolution of the ensemble mean with increased iterations is seen in figure 8 for each ensemble size investigated. It is evident that the ensemble mean position converges to a single point with subsequent iterations. The exact location of this optimum along the c_w axis appeared to be substantially different in the $N = 20$ case than for the smaller ensemble sizes. This is likely due to the presence of multiple local optima in the parameter space, with the difference between their calibrated values indistinguishable from each other when accounting for the process level noise in the true QBO signal.

This behavior under EKI means that the centroid of the ensemble represents the best estimate of a calibrated parameter at any given iteration, in contrast to History Matching where there is no preference given to the centroid over any other point. Figure 8 demonstrates that regardless of choice of ensemble size, this centroid always converges on a singular point. This is in contrast to what is seen with history matching in figure 9 where the centroid often appears to move erratically. However it can be seen that for all N, approximately the same centroid point is obtained. For EKI, convergence about the final point can be seen to take approximately 5 to 6 iterations for all the ensemble sizes indicated above, implying the speed of convergence is not a strong function of ensemble size. This is further indicated in figure 10 which shows the root mean squared (RMS) magnitude of all the ensemble update vectors obtained via equation 11. In this we observe that this update vector magnitude decays at a similar rate for all ensemble sizes considered. This ensemble size invariance indicates that EKI is a strong algorithm if the objective of the calibration is to obtain a single best estimate of a parameter, and such a calibration can be performed rapidly with a small ensemble size, with $N = 5$ result-

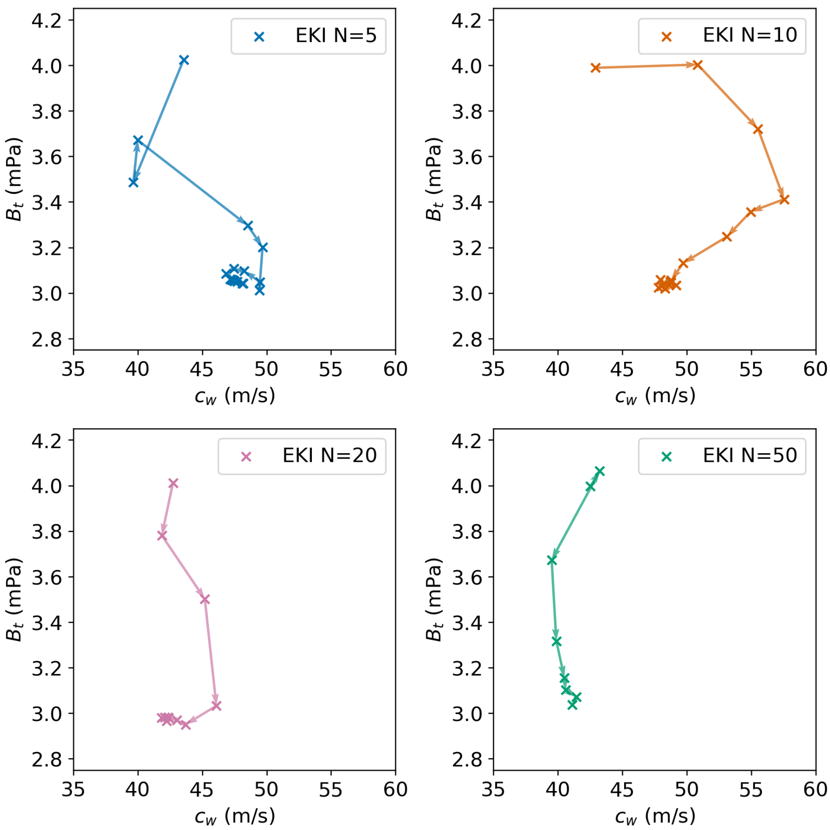


Figure 8: Evolution of the centroid for each ensemble size under EKI.

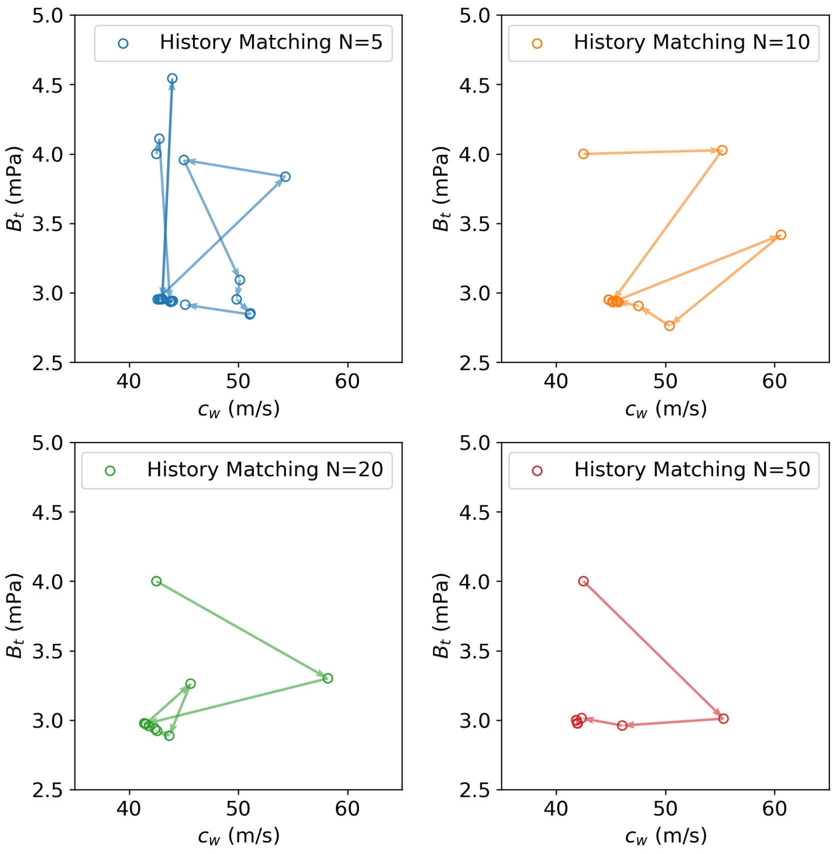


Figure 9: Evolution of the centroid for each ensemble size under History Matching.

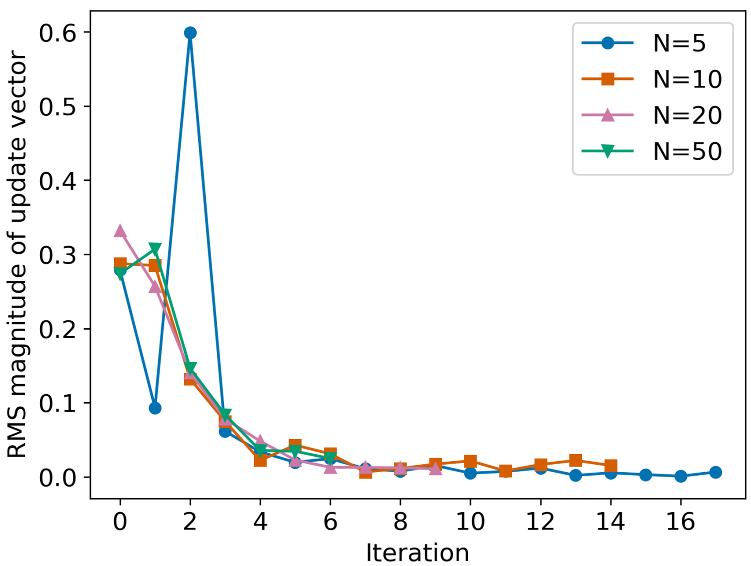
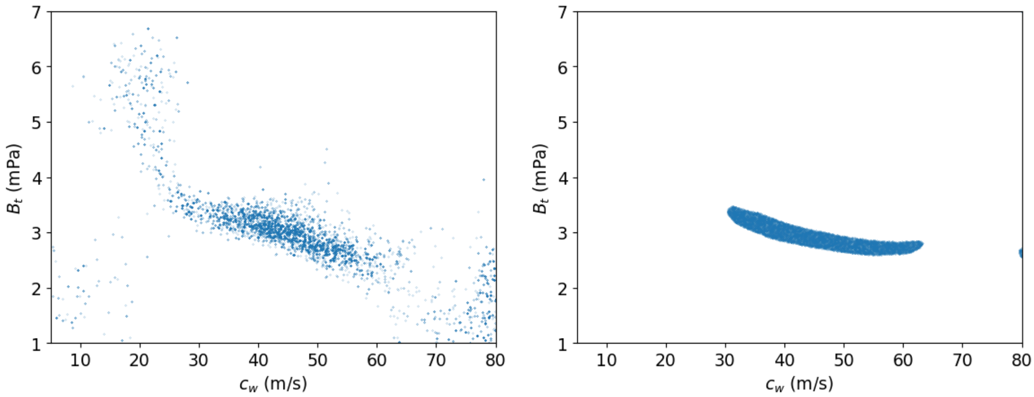


Figure 10: Normalized root mean squared magnitude of the update vectors under EKI for each ensemble size.



(a) EKI

(b) History Matching

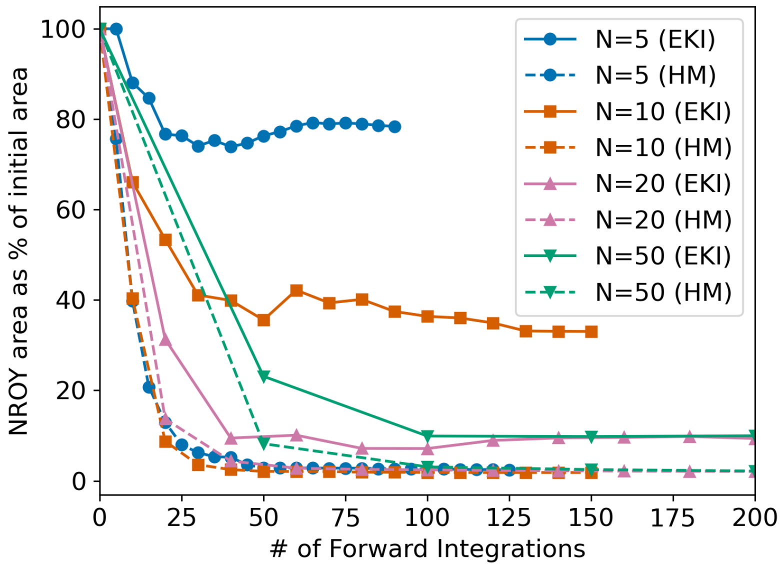
Figure 11: Comparison between estimates of the posterior distribution $p(\theta|y)$ between History Matching and EKI after 6 iterations, $N = 10$.

ing in an approximately converged centroid after 25 total GCM integrations, with figure 10 showing minimal updates to the ensemble members beyond this point.

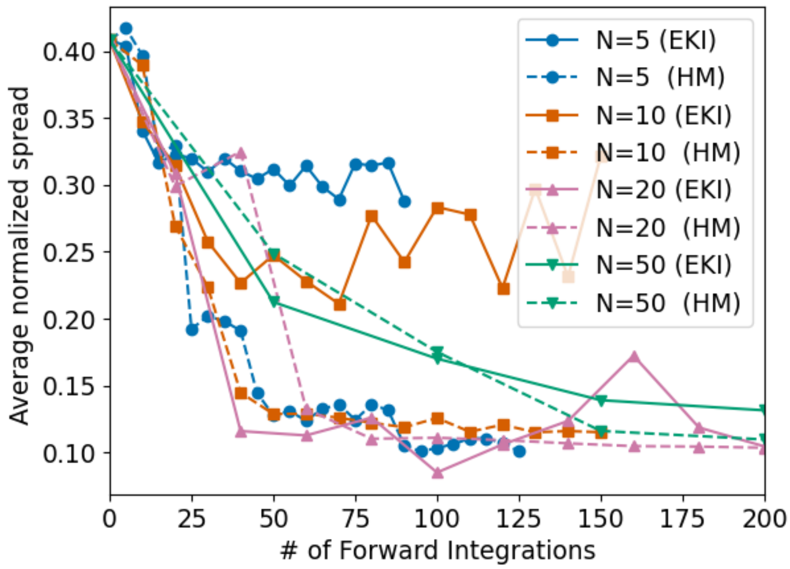
As mentioned above, History Matching does not produce an equivalent "best estimate" and thus to provide a comparison, the ability of both approaches to quantify the uncertainty in the calibrated parameters is estimated. In quantitative terms this translates to obtaining an estimate of the posterior distribution: $p(\theta|y)$. In an ideal case, this posterior distribution would be as "compact" as possible given the observation noise level, indicating that we have a narrow set of calibrated values that reproduce consistent observables. The NROY space from History Matching provides a rough heuristic for this posterior distribution subject to a uniform assumption whilst the CES methodology as described in section 2.4 can be used to obtain an estimate of the full posterior distribution (Cleary et al., 2021). In figure 11a we show 10,000 sample points drawn from the estimated posterior distribution sampled via the Metropolis Hasting algorithm for an EKI calibration at iteration 6 and $N = 10$. Meanwhile figure 11b shows an equivalent sample of 10,000 sample points drawn from the iteration 6, $N = 10$ History Matching NROY space, which was the first iteration to meet the $N = 10$ NROY convergence criterion.

To gain an estimate of the compactness of each sample space, we can calculate the normalized average spread of the posterior sampled points about the centroid for both EKI and History Matching, shown in figure 12b as a function of the number of model integrations. An equivalent estimate of the NROY space can also be calculated by use of equation 10 for the implausibility calculated using EKI via the CES emulator across the entire space and utilizing the same implausibility cutoff as for history matching. In other words, this represents where 95% of the probability mass lies. This is seen in figure 12a.

It is evident from both figures that for low ensemble sizes, history matching is able to obtain a substantially more compact calibrated space when considering both NROY and normalized spread. This is in contrast to the behavior in figures 8 and 9 where for small ensemble sizes EKI is able to obtain a converged centroid much more rapidly than history matching at small N, indicating the relative strengths of these approaches. For the larger ensemble size of $N = 20$ & $N = 50$, the differences between the two approaches



(a) NROY area



(b) Normalized Spread

Figure 12: Comparison of the NROY area and normalized spread method for quantifying the relative uncertainty in the EKI and History Matching calibrations.

become less apparent. However, the NROY comparison shows that the EKI equivalent NROY space is not able to collapse as compactly as seen under history matching. This comparison is limited however, as it neglects to take into account that the estimated NROY under EKI is not sampled from uniformly as it is under History Matching. As the comparison using the RMS posterior sample spread takes into account the non-uniform nature of the EKI posterior compared against the uniform history matching approach, the spread likely better reflects the true quantification of the compactness. The main downside of comparing RMS spread in comparison to the NROY area for history matching is that in cases where the centroid of the NROY space is not itself within the NROY space is that the normalized spread will not tend to zero even as the NROY area does tend to zero. Figure 12b, indicates that for N=5 & N=10 history matching draws a far more compact set of posterior samples compared to EKI and requires only approximately 50 GCM integrations to do so. This can be understood by the approach that history matching takes as it spends more time obtaining samples near the edges of the initial parameter space compared to EKI, allowing for more confident emulator performance in these regions yielding the more compact posterior. Thus in contrast to the centroid result described above, figure 12 indicates that History Matching can provide a converged posterior distribution of the plausible parameters with only 5 to 10 ensemble members, in concurrence with the results described in section 3.1.

4 Discussion

In this investigation we presented an implementation of the History Matching procedure for calibrating the AD99 gravity wave parameterization based on observations of the QBO. We showed that a chain of Gaussian Process regression emulators is capable of acting as a feasible emulator across the entire parameter space. The history matching procedure was successful at converging the initial NROY space by a factor of up to 98%, producing a compact region of plausible parameters. We showed that this result is robust across different choices of ensemble sizes, with a size of $N = 10$ converging the fastest.

We also compared history matching with an alternative calibration method, Ensemble Kalman Inversion. We found that this algorithm is capable of obtaining a single optimal calibrated value in best agreement with the observations, which it can do rapidly at small ensemble sizes. The Calibrate-Emulate-Sample (CES) method was used to obtain an estimate of the posterior distribution across the entire parameter space for comparison to the NROY space generated by history matching. It was found when considering the mean spread of the samples drawn from both methodologies that for large ensemble sizes of $N = 20$ & $N = 50$, both methods gave posteriors with a similar degree of compactness, however the smaller ensembles showed that history matching was able to obtain a stronger degree of compactness.

One key constraint that was imposed for simplicity in this work was the low dimensional space chosen for both the observables and the input parameters. Such a constraint was useful for reducing the number of iterations required to obtain convergence for both methodologies, in addition to making the outputs easy to visualize. Future work could look to increase the dimensionality of both the observable vector and the input parameters. In the case of the QBO, introducing a 3rd observable variable would allow the calibration to be based on the peak easterly and westerly velocities of the QBO instead of the amplitude. This could be significant given the acknowledged westerly bias present in GCMs (Bushell et al., 2022). Palmer demonstrated how this could be alleviated with an orographic gravity wave parameterization scheme (Palmer et al., 1986) and in ongoing work, we are considering the calibration of both orographic and non-orographic schemes in conjunction. Finally, we also restricted the calibration in this work to just the tropical parameters for AD99, however extra-tropical and polar parameters in principle also need to be calibrated.

As these additional considerations all increase the number of dimensions of the input and output spaces, both History Matching and EKI may require a dramatically increased number of iterations to converge. It is possible that History Matching may be a challenge, as for the ensemble sizes considered in this work, the density of members in the NROY space will decrease exponentially with the number of parameters, resulting in a reduced support for the emulator and thus a greatly less confident one. This is in contrast to EKI which has demonstrated efficiency even at large numbers of parameters (Kovachki & Stuart, 2019; Pahlavan et al., 2023). Another calibration algorithm that could be investigated in future work is Bayesian Optimization (BO) (Garnett, 2023; Shahriari et al., 2016) which has proven popular within the domain of hyper-parameter optimization for machine learning methods. This method works similarly to history matching as it involves using GPR to approximate the behavior of the model at different parameters. Unlike History Matching, an "acquisition function" is also obtained which is used to determine regions in the parameter space to be sampled for future iterations. Such functions often make a trade off between exploring unsampled regions of the space and exploitation of regions of the space where the error between the predictions and observations is minimized. This approach should in principle provide for a more optimal sampling in high dimensional spaces compared to the uniform approach of History Matching; however, this does come at the cost of more user-defined choices in the acquisition function.

Other calibration methods within the same family as EKI also exist. An example is Ensemble Kalman Sampling (EKS) (Garbuno-Inigo et al., 2020; Ding & Li, 2021) which includes an additional random walk component on top of the EKI update step. Such a random walk prevents the EKI ensemble members from falling into local minima during the loss function optimization and should lead to the final EKS ensemble members being distributed according to the posterior distribution without CES being explicitly required. EKS can be shown to produce exact results and converge in finite time in the case where the posterior distribution is Gaussian. However, no such assertion can be made for the more general non linear case. Unscented Kalman Inversion (UKI) is another recent method in the Kalman filter family of calibration methods that also aims to directly capture the posterior distribution (Huang et al., 2022) by allowing for nonlinear effects to be estimated during the update step in a Kalman filter.

Overall, our calibration of AD99 in MiMA using History Matching and EKI showed that both methods are able to competently reduce a large initial range of parameters and produce a compact space of plausible parameters that result in QBO statistics that resemble observations. Techniques such as BO as well as the above mentioned newly developed techniques have not yet been widely applied to aiding climate model development. We expect that future work probing the utility of these techniques for climate model calibration should prove useful in further constraining the plausible range of parameters, and thus potentially allow for more accurate model predictions with uncertainty quantification. These techniques also allows us to determine the future range of variability in observables such as the QBO period and amplitude under various CO_2 forcing scenarios using the current calibrated parameters.

5 Open Research

The "Quasi-Biennial-Oscillation (QBO) Data Series" developed by the Freie Universität Berlin(Kunze, 2007) was used as the source of zonal wind observations of the QBO. This dataset can be found at `https://www.geo.fu-berlin.de/en/met/ag/strat/produkte/qbo/index.html`. The Model of an idealized Moist Atmosphere GCM codebase can be found at `https://github.com/mjucker/MiMA`. The code developed during the course of this work is available in two repositories: one for the generic History Matching implementation & another for performing the analysis and model runs specific to the AD99 calibration. The History Matching code is made available at `https://github.com/`

`Eddy-Stanford/History-Matching-Core` and can also be installed via the `history-matching` package available on PyPy. The analysis and model run code is available at `https://github.com/Eddy-Stanford/QBO-History-Matching`.

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 also acknowledges support from the National Science Foundation through grant OAC-2004492. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results. Finally we would like to thank V. Balaji and Oliver Dunbar for very useful discussions during the development of this work.

References

Alexander, M. J., & Dunkerton, T. J. (1999, December). A Spectral Parameterization of Mean-Flow Forcing due to Breaking Gravity Waves. *Journal of the Atmospheric Sciences*, *56*(24), 4167–4182. Retrieved 2023-01-31, from `https://journals.ametsoc.org/view/journals/atsc/56/24/1520-0469_1999_056_4167_aspomf_2.0.co_2.xml` (Publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences) doi: 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., … Watanabe, S. (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. Retrieved 2023-02-08, from `https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.637` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.637) doi: 10.1002/qj.637

Andrianakis, I., Vernon, I. R., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., … White, R. G. (2015, January). Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda. *PLOS Computational Biology*, *11*(1), e1003968. Retrieved 2022-12-28, from `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003968` (Publisher: Public Library of Science) doi: 10.1371/journal.pcbi.1003968

Baldwin, M. P., Gray, L. J., Dunkerton, T. J., Hamilton, K., Haynes, P. H., Randel, W. J., … Takahashi, M. (2001). The quasi-biennial oscillation. *Reviews of Geophysics*, *39*(2), 179–229. Retrieved 2022-11-09, from `https://onlinelibrary.wiley.com/doi/abs/10.1029/1999RG000073` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/1999RG000073) doi: 10.1029/1999RG000073

Bushell, A. C., Anstey, J. A., Butchart, N., Kawatani, Y., Osprey, S. M., Richter, J. H., … Yukimoto, S. (2022). Evaluation of the Quasi-Biennial Oscillation in global climate models for the SPARC QBO-initiative. *Quarterly Journal of the Royal Meteorological Society*, *148*(744), 1459–1489. Retrieved 2023-01-16, from `https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3765` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3765) doi: 10.1002/qj.3765

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., & Stuart, A. M. (2021, January). Calibrate, emulate, sample. *Journal of Computational Physics*, *424*, 109716. Retrieved 2023-05-10, from `https://www.sciencedirect.com/science/article/pii/S0021999120304903` doi: 10.1016/j.jcp.2020.109716

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., ... Xu, W. (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. Retrieved 2023-01-26, from https://onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002217 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002217) doi: 10.1029/2020MS002217

Craig, P. S., Goldstein, M., Seheult, A. H., & Smith, J. A. (1997). Pressure Matching for Hydrocarbon Reservoirs: A Case Study in the Use of Bayes Linear Strategies for Large Computer Experiments. In C. Gatsonis, J. S. Hodges, R. E. Kass, R. McCulloch, P. Rossi, & N. D. Singpurwalla (Eds.), *Case Studies in Bayesian Statistics* (pp. 37–93). New York, NY: Springer. doi: 10.1007/978-1-4612-2290-3_2

Ding, Z., & Li, Q. (2021, January). Ensemble Kalman Sampler: Mean-field Limit and Convergence Analysis. *SIAM Journal on Mathematical Analysis*, *53*(2), 1546–1578. Retrieved 2023-11-28, from https://epubs.siam.org/doi/abs/10.1137/20M1339507 (Publisher: Society for Industrial and Applied Mathematics) doi: 10.1137/20M1339507

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. Retrieved 2023-07-03, from https://onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002454 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002454) doi: 10.1029/2020MS002454

Dunkerton, T. J. (1997). The role of gravity waves in the quasi-biennial oscillation. *Journal of Geophysical Research: Atmospheres*, *102*(D22), 26053–26076. Retrieved 2023-07-21, from https://onlinelibrary.wiley.com/doi/abs/10.1029/96JD02999 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/96JD02999) doi: 10.1029/96JD02999

Fritts, D. C., & Alexander, M. J. (2003). Gravity wave dynamics and effects in the middle atmosphere. *Reviews of Geophysics*, *41*(1). Retrieved 2022-11-09, from https://onlinelibrary.wiley.com/doi/abs/10.1029/2001RG000106 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2001RG000106) doi: 10.1029/2001RG000106

Garbuno-Inigo, A., Hoffmann, F., Li, W., & Stuart, A. M. (2020, January). Interacting Langevin Diffusions: Gradient Structure and Ensemble Kalman Sampler. *SIAM Journal on Applied Dynamical Systems*, *19*(1), 412–441. Retrieved 2023-07-19, from https://epubs.siam.org/doi/10.1137/19M1251655 (Publisher: Society for Industrial and Applied Mathematics) doi: 10.1137/19M1251655

Garfinkel, C. I., White, I., Gerber, E. P., Jucker, M., & Erez, M. (2020, July). The Building Blocks of Northern Hemisphere Wintertime Stationary Waves. *Journal of Climate*, *33*(13), 5611–5633. Retrieved 2023-06-06, from https://journals.ametsoc.org/view/journals/clim/33/13/jcli-d-19-0181.1.xml (Publisher: American Meteorological Society Section: Journal of Climate) doi: 10.1175/JCLI-D-19-0181.1

Garnett, R. (2023). *Bayesian Optimization*. Cambridge University Press. (Google-Books-ID: MBCrEAAAQBAJ)

Golomb, S., & Posner, E. (1964, July). Rook domains, Latin squares, affine planes, and error-distributing codes. *IEEE Transactions on Information Theory*, *10*(3), 196–208. (Conference Name: IEEE Transactions on Information Theory) doi: 10.1109/TIT.1964.1053680

Holton, J. R., & Lindzen, R. S. (1972, September). An Updated Theory for the Quasi-Biennial Cycle of the Tropical Stratosphere. *Journal of the Atmospheric Sciences*, *29*(6), 1076–1080. Retrieved 2023-07-21, from `https://journals.ametsoc.org/view/journals/atsc/29/6/1520-0469_1972_029_1076_autftq_2_0_co_2.xml` (Publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences) doi: 10.1175/1520-0469(1972)029⟨1076:AUTFTQ⟩2.0.CO;2

Huang, D. Z., Schneider, T., & Stuart, A. M. (2022, August). Iterated Kalman methodology for inverse problems. *Journal of Computational Physics*, *463*, 111262. Retrieved 2023-11-28, from `https://www.sciencedirect.com/science/article/pii/S0021999122003242` doi: 10.1016/j.jcp.2022.111262

Iglesias, M. A., Law, K. J. H., & Stuart, A. M. (2013, March). Ensemble Kalman methods for inverse problems. *Inverse Problems*, *29*(4), 045001. Retrieved 2023-08-19, from `https://dx.doi.org/10.1088/0266-5611/29/4/045001` (Publisher: IOP Publishing) doi: 10.1088/0266-5611/29/4/045001

Jucker, M., & Gerber, E. P. (2017, September). Untangling the Annual Cycle of the Tropical Tropopause Layer with an Idealized Moist Model. *Journal of Climate*, *30*(18), 7339–7358. Retrieved 2023-06-17, from `https://journals.ametsoc.org/view/journals/clim/30/18/jcli-d-17-0127.1.xml` (Publisher: American Meteorological Society Section: Journal of Climate) doi: 10.1175/JCLI-D-17-0127.1

Kovachki, N. B., & Stuart, A. M. (2019, August). Ensemble Kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Problems*, *35*(9), 095005. Retrieved 2023-03-21, from `https://dx.doi.org/10.1088/1361-6420/ab1c3a` (Publisher: IOP Publishing) doi: 10.1088/1361-6420/ab1c3a

Kunze, M. (2007, October). *The Quasi-Biennial-Oscillation (QBO) Data Serie.* Retrieved 2023-01-31, from `https://www.geo.fu-berlin.de/en/met/ag/strat/produkte/qbo/index.html` (Published via the Freie Universität Berlin)

Lguensat, R., Deshayes, J., Durand, H., & Balaji, V. (2023). Semi-Automatic Tuning of Coupled Climate Models With Multiple Intrinsic Timescales: Lessons Learned From the Lorenz96 Model. *Journal of Advances in Modeling Earth Systems*, *15*(5), e2022MS003367. Retrieved 2023-05-10, from `https://onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003367` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003367) doi: 10.1029/2022MS003367

Lindzen, R., & Holton, J. (1968, November). A Theory of the Quasi-Biennial Oscillation. *J. Atmos. Sci.*, *25*, 1095–1107. doi: 10.1175/1520-0469(1968)025⟨1095:ATOTQB⟩2.0.CO;2

Lott, F., & Miller, M. J. (1997). A new subgrid-scale orographic drag parametrization: Its formulation and testing. *Quarterly Journal of the Royal Meteorological Society*, *123*(537), 101–127. Retrieved 2023-06-17, from `https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712353704` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49712353704) doi: 10.1002/qj.49712353704

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*(11), e2022MS003245. Retrieved 2022-11-14, from `https://onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003245` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003245) doi: 10.1029/2022MS003245

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, *21*(2), 239–245. Retrieved 2023-06-26, from `https://www.jstor.org/stable/1268522` (Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality]) doi:

10.2307/1268522

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953, March). *Equation of state calculations by fast computing machines* (Tech. Rep. No. AECU-2435; LADC-1359). Los Alamos Scientific Lab., Los Alamos, NM (United States); Univ. of Chicago, IL (United States). Retrieved 2023-12-07, from `https://www.osti.gov/biblio/4390578` doi: 10.2172/4390578

Pahlavan, H. A., Hassanzadeh, P., & Alexander, M. J. (2023, September). *Explainable Offline-Online Training of Neural Networks for Parameterizations: A 1D Gravity Wave-QBO Testbed in the Small-data Regime.* arXiv. Retrieved 2023-12-04, from `http://arxiv.org/abs/2309.09024` (arXiv:2309.09024 [physics]) doi: 10.48550/arXiv.2309.09024

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*(474), 1001–1039. Retrieved 2022-11-09, from `https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.49711247406` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49711247406) doi: 10.1002/qj.49711247406

Pukelsheim, F. (1994). The Three Sigma Rule. *The American Statistician*, *48*(2), 88–91. Retrieved 2023-01-31, from `https://www.jstor.org/stable/2684253` (Publisher: [American Statistical Association, Taylor & Francis, Ltd.]) doi: 10.2307/2684253

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning.* The MIT Press. Retrieved 2023-11-15, from `https://direct.mit.edu/books/book/2320/Gaussian-Processes-for-Machine-Learning` doi: 10.7551/mitpress/3206.001.0001

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. Retrieved 2023-05-05, from `https://onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032362` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019JD032362) doi: 10.1029/2019JD032362

Salter, J. M., & Williamson, D. (2016). A comparison of statistical emulation methodologies for multi-wave calibration of environmental models. *Environmetrics*, *27*(8), 507–523. Retrieved 2023-10-05, from `https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2405` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.2405) doi: 10.1002/env.2405

Schenzinger, V., Osprey, S., Gray, L., & Butchart, N. (2017, June). Defining metrics of the Quasi-Biennial Oscillation in global climate models. *Geoscientific Model Development*, *10*(6), 2157–2168. Retrieved 2023-07-05, from `https://gmd.copernicus.org/articles/10/2157/2017/` doi: 10.5194/gmd-10-2157-2017

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016, January). Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, *104*(1), 148–175. Retrieved 2023-11-27, from `https://ieeexplore.ieee.org/abstract/document/7352306` (Conference Name: Proceedings of the IEEE) doi: 10.1109/JPROC.2015.2494218

Vernon, I., Goldstein, M., & Bower, R. G. (2010, December). Galaxy formation : a Bayesian uncertainty analysis. *Bayesian analysis.*, *05*(04), 619–670. Retrieved 2023-06-17, from `http://ba.stat.cmu.edu/vol05is04.php` (Number: 04 Publisher: International Society for Bayesian Analysis) doi: 10.1214/10

-BA524

Williamson, D., Blaker, A. T., Hampton, C., & Salter, J. (2015, September). Identifying and removing structural biases in climate models with history matching. *Climate Dynamics*, *45*(5), 1299–1324. Retrieved 2023-11-21, from https://doi.org/10.1007/s00382-014-2378-z doi: 10.1007/s00382-014-2378-z

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., & Yamazaki, K. (2013, October). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dynamics*, *41*(7), 1703–1729. Retrieved 2022-11-09, from https://doi.org/10.1007/s00382-013-1896-4 doi: 10.1007/s00382-013-1896-4