

## RESEARCH ARTICLE

# The graft-versus-host problem for data-driven gravity-wave parameterizations in a one-dimensional quasibiennial oscillation model

Ofer Shamir<sup>1</sup>  | David S. Connelly<sup>1</sup> | Steven C. Hardiman<sup>2</sup> | Zihan Shao<sup>1</sup> | L. Minah Yang<sup>1</sup> | Edwin P. Gerber<sup>1</sup> 

<sup>1</sup>Courant Institute of Mathematical Sciences, New York University, New York, New York

<sup>2</sup>Met Office Hadley Centre, Exeter, UK

## Correspondence

Ofer Shamir, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA.  
Email: [ofer.shamir@courant.nyu.edu](mailto:ofer.shamir@courant.nyu.edu)

## Present address

Zihan Shao, Department of Mathematics, University of California, San Diego, California

## Funding information

Schmidt Sciences: Virtual Earth System Research Institute; Division of Advanced Cyberinfrastructure, Grant/Award Number: OAC-2004572

## Abstract

Two key challenges in the development of data-driven gravity-wave parameterizations are generalization, how to ensure that a data-driven scheme trained on the present-day climate will continue to work in a new climate regime, and calibration, how to account for biases in the “host” climate model. Both problems depend fundamentally on the response to out-of-sample inputs compared with the training dataset, and are often conflicting. The ability to generalize to new climate regimes often goes hand in hand with sensitivity to model biases. To probe these challenges, we employ a one-dimensional (1D) quasibiennial oscillation (QBO) model with a stochastic source term that represents convectively generated gravity waves in the Tropics with randomly varying strengths and spectra. We employ an array of machine-learning models consisting of a fully connected feed-forward neural network, a dilated convolutional neural network, an encoder–decoder, a boosted forest, and a support-vector regression model. Our results demonstrate that data-driven schemes trained on “observations” can be critically sensitive to model biases in the wave sources. While able to emulate accurately the stochastic source term on which they were trained, all of our schemes fail to simulate fully the expected QBO period or amplitude, even with the slightest perturbation to the wave sources. The main takeaway is that some measures will always be required to ensure the proper response to climate change and to account for model biases. We examine one approach based on the ideas of optimal transport, where the wave sources in the model are first remapped to the observed one before applying the data-driven scheme. This approach is agnostic to the data-driven method and guarantees that the model adheres to the observational constraints, making sure the model yields the right results for the right reasons.

## KEYWORDS

data-driven, gravity waves, machine learning, quasibiennial oscillation, subgrid-scale parameterizations

# 1 | INTRODUCTION

A practical problem concerning the development of data-driven parameterizations of subgrid-scale processes is the “graft-versus-host” problem, where a data-driven scheme may be incompatible with its “host”, the large-scale model, due to model biases and nonlinear feedbacks between the resolved and parameterized scales. In the absence of sufficient observational constraints, traditional (physics-based) schemes are tuned on a per-host basis to overcome model biases and yield desirable results. However, with data-driven schemes one does not have this luxury. Aside from technical difficulties associated with the fact that the tunable parameters in such schemes are only latently related to the physical parameters, tuning an observationally constrained scheme works against its purpose, to represent the missing process faithfully. To the extent that the training dataset does, indeed, represent the observed conditions, the resulting data-driven scheme ought to be changed as little as possible.

In the present work, we probe the graft-versus-host problem in the context of data-driven gravity-wave (GW) parameterizations. The scenario we envisage is one where a data-driven model is trained on observations (or high-resolution model simulations) to “learn” a parameterization of the form “GW drag = GW drag(flow, GW sources)”. This parameterization is then transplanted into an operational climate model, which will almost certainly exhibit different (biased) wave sources. For convective GW in the Tropics, this is partly due to the representation of convection in the model and partly due to the fact that the GW sources are themselves dependent on the resolved flow, making them susceptible to model biases. For the procedure to succeed, the wave sources in the host model must be within the set of observations used for training, or the data-driven scheme must generalize out-of-set. Otherwise, even a “healthy” but incompatible parameterization could lead to “unhealthy” simulations, for example, an unrealistic quasibiennial oscillation (QBO).

The QBO is the dominant mode of variability in the tropical stratosphere, consisting of downwelling shear zones of alternating easterly and westerly winds with a period of about 28 months (Baldwin et al., 2001). It was first observed in the mid-1950s and early 1960s (Ebdon, 1960; Ebdon & Veryard, 1961; Reed et al., 1961) and was explained theoretically soon after, in the late 1960s and early 1970s (Holton & Lindzen, 1972; Lindzen & Holton, 1968), by means of a wave–mean-flow interaction driven by upward-propagating waves. As such, the QBO in general circulation models is particularly sensitive to the spectrum of the simulated waves and, ultimately, their momentum deposition. Due to limited (vertical)

resolution, insufficient for resolving wave generation, upward propagation, and the ensuing wave–mean-flow interactions, simulations of the QBO as an emergent phenomenon remained a challenging task for decades, until the late 1990s and early 2000s. Still, despite constant improvements in the representation of the (resolved) tropical wave spectrum, and in lieu of infeasible vertical resolutions, present-day global climate models (GCMs) generally rely on the addition of parameterized waves to obtain realistic QBOs (Geller et al., 2016; Holt et al., 2022; Richter et al., 2014, 2020). All but one of the models participating in the QBO initiative (QBOi) required parameterized GW to exhibit a QBO, and the majority of the wave forcing above the QBO base in those models was attributed to the parameterized waves (Bushell et al., 2022). Moreover, in practice, given limited observational constraints, the GW schemes in those models were likely tuned to yield realistic QBOs.

In addition to traditional GW schemes, the QBO has recently been used as a key metric for assessing data-driven schemes. Espinosa et al. (2022) and Connelly and Gerber (2023) used neural networks and random forests to emulate the Alexander and Dunkerton (1999, henceforth AD99) GW scheme in a Model of an idealized Moist Atmosphere (MiMA, Jucker & Gerber, 2017; Garfinkel et al., 2020). Mansfield and Shezhadri (2022) have also used Gaussian processes to emulate the AD99 scheme in MiMA en route to quantifying the uncertainties associated with the GW sources. Yang and Gerber (2024) used encoder–decoders to emulate AD99 in MiMA en route to developing optimal (re)sampling strategies. Finally, Hardiman et al. (2023) used convolutional neural network for emulating the Warner and McIntyre GW scheme (Warner & McIntyre, 1999, 2001) in the Met Office HadGEM3-GA8.0 climate model (in an atmosphere-only configuration), while comparing different inputs. These studies have all considered the fidelity of the QBO (among other criteria) to assess the “online” performance of their respective emulators, that is, when coupled to the climate model in place of the original (physics-based) scheme. While demonstrating the feasibility of emulating physics-based GW schemes, they also raise questions about the implementation of purely data-driven ones.

To tackle the graft-versus-host challenge in a controlled environment, remove climate model complexities, and facilitate the development of data-driven GW parameterizations, we employ a one-dimensional (1D) QBO model based on the classic model in Lindzen and Holton (1968), Holton and Lindzen (1972), and Plumb (1977). Aside from explaining the governing mechanism of the QBO itself, this model has proven to be a useful abstraction for explaining other properties of the QBO: for example, the

formation of the buffer zone below the QBO base (Match & Fueglistaler, 2020) and the QBO disruption-triggering mechanism (Match & Fueglistaler, 2021). In the present work, we add a stochastic source term to the model, mimicking convectively generated gravity waves in the Tropics with randomly varying strengths and phase speeds.

In addition to representing the relevant physical scenario better, this setup enables us to examine the sensitivity of the QBO to the source spectrum parameters, namely the source flux and spectral width. This allows us to explore two related questions. First, in a climate-change context, how well can a data-driven scheme trained on today's climate generalize to a perturbed climate (i.e., a warmer world)? Second, can a data-driven scheme trained on observations be calibrated to yield the correct macroscopic behavior, that is, the "right" QBO, when grafted into a host climate model with biased GW sources?

We implement an array of machine learning (ML) models consisting of a neural network (NN), a convolutional neural network (CNN), an encoder-decoder (ED), a boosted random forest (BF), and a support-vector regression (SVR) model. While able to "learn" the GW drags corresponding to the GW source distribution on which they were trained, they all fail to capture the sensitivity of the QBO to perturbations in the source distribution fully, that is, fail to generalize to new climate conditions. In addition, a data-driven scheme trained on observations leads to a biased simulation of the QBO when fed perturbed GW sources, that is, when grafted into the host. A key conclusion is the fact that some remedy to this problem will always be required. In the present work, we suggest a pre-conditioning step based on the ideas of optimal transport, where the biased source distribution is first remapped back to the "observed" one before being fed into the data-driven model. Aside from guaranteeing that the graft and host are compatible, the advantages of this approach are that it is agnostic to the data-driven method, and that it guarantees that the model adheres to the observational constraints, and so the model yields the right results for the right reasons.

We start with a short description of the physical model, including our modifications and the control experiment, in Sections 2 and 2.1, respectively. In preparation for studying our envisaged scenario with data-driven methods, we first examine, in Section 2.2, the sensitivity of the QBO to changes in the GW sources in the physical model. The data-driven models are presented in Section 3, their results on the control experiment in Section 4.1, and their sensitivity to changes in GW sources in Section 4.2. Our suggested remedy for a model with biased forcing parameters is presented in Section 4.3.

## 2 | THE PHYSICAL MODEL: A STOCHASTICALLY DRIVEN 1D QBO

The 1D QBO model of the present work is a hybrid of the models introduced in Holton and Lindzen (1972) and Plumb (1977), coupled with a stochastic source term to mimic randomly generated GW. The model equation consists of an advection-diffusion equation for the zonal mean zonal wind ( $u$ ) with a source term ( $S$ ) due to GW momentum deposition:

$$\frac{\partial u}{\partial t} + w \frac{\partial u}{\partial z} - \kappa \frac{\partial^2 u}{\partial z^2} = S(z, u), \quad (1)$$

where  $t$  is time,  $z$  is the vertical coordinate,  $w = w(t, z)$  is the (potentially) time- and height-dependent residual vertical wind, and  $\kappa$  is a constant diffusion coefficient. The source term on the right-hand side (RHS) originates from the divergence of upward zonal momentum fluxes, and, as such, needs to be parameterized further in terms of the zonal wind. For upward-propagating Kelvin(-like) waves in a slowly varying zonal flow, the resulting forcing due to a sum of monochromatic waves is (Lindzen, 1971)

$$S(z, u) = -\frac{1}{\rho} \frac{\partial}{\partial z} \sum_n B_n \exp \left\{ -\int_{z_1}^z \frac{\alpha(z') N}{k_n(u - c_n)^2} dz' \right\}, \quad (2)$$

where  $\rho(z)$  is the density,  $B_n$  are the wave amplitudes,  $k_n$  are the wavenumbers,  $c_n$  are the phase speeds,  $N$  is the Brunt-Väisälä frequency, and  $\alpha(z)$  is the wave dissipation.

In general, the wave amplitudes can be chosen independently of the wave phase speeds, provided only that  $\text{sgn}(B_n) = \text{sgn}(c_n)$ , so as to guarantee that westerly (easterly) waves carry westerly (easterly) momentum upward. A more physically plausible assumption in the presence of many waves is that of a continuous spectrum. In the present work, we assume a Gaussian wave spectrum similar to one used in Alexander and Dunkerton (1999), namely

$$B(c) \propto \text{sgn}(c) \exp \left[ -\ln 2 \left( \frac{c}{c_w} \right)^2 \right], \quad (3)$$

where  $c_w$  is the spectral half width.

Aside from the introduction of a continuous wave spectrum, the key distinction between our 1D QBO model and its predecessors in Holton and Lindzen (1972) and Plumb (1977) is the addition of stochasticity to the wave forcing. At each time step, the total (absolute) source flux  $F_{S0} = \rho_0 \sum_n |B_n|$  and spectral width  $c_w$  are drawn from a bivariate log-normal distribution, with the proportionality coefficient in Equation 3 determined by  $F_{S0}$ . Physically,

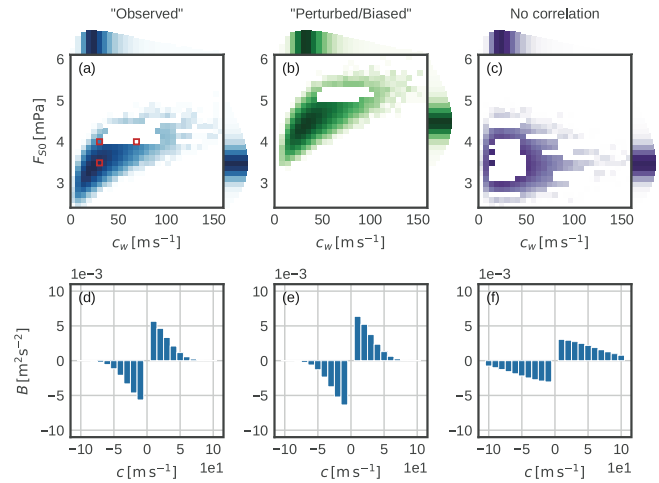
one can think of convectively generated gravity waves in the Tropics having randomly varying strengths and spectra, with more intense convection causing stronger fluxes and deeper convection exciting faster waves, and hence broader spectra (Alexander et al., 2021). The bivariate log-normal distribution is a minimal distribution, having just five parameters, capable of describing the above physical picture, while also guaranteeing that  $F_{S0}$  and  $c_w$  are strictly positive.

Figure 1 shows the GW sources in the stochastic model, including three wave source distributions (a–c) and the spectra of three randomly sampled wave packets (e, f) from the observed distribution in Figure 1a. The control distribution Figure 1a was chosen to produce the “observed” amplitude and period of the QBO, as described in the next section. As  $F_{S0}$  relates to the square of the total latent heating (or total precipitation) and  $c_w$  to the depth of convection, we chose them to be positively correlated. The perturbed distribution (Figure 1b) can be viewed as the forcing under a climate perturbation (here, stronger and slightly deeper convection), or as a host model with a biased source distribution. Finally, Figure 1c illustrates a hypothetical distribution with no correlation between  $F_{S0}$  and  $c_w$ , used to test the sensitivity of data-driven models to the correlation between the two.

The resulting QBO in response to the control and perturbed wave forcing is shown in Figure 2a,h. The control simulation nearly matches the observed QBO by construction, while our “warmer world” exhibits a slower, but more intense QBO. (To be clear, we have not modified any other model parameter in the perturbed run, e.g., a change in the vertical advection  $w$ , which would also impact the QBO.) Aside from the internal variability, the main notable difference from the classic model is the replacement of the critical level mechanism by a filtering mechanism, where the low phase speed waves break first as the wind amplifies. This is the result of using more than two waves in the present model, not the added stochasticity. Thus, our model maintains the essential physics of the classic model but allows us to link the above forcing to variability in the intensity and depth of convection, as in more advanced GW parameterizations (e.g., Beres et al., 2004).

## 2.1 | Control experiment: The “observed” QBO in our 1D model

Traditional, physics-based, GW schemes are often tuned to yield the observed/realistic QBO. Among the first parameters tuned are those associated with the GW sources. For convective GW in the Tropics, this is partly due to the representation of convection in the model, and partly due to

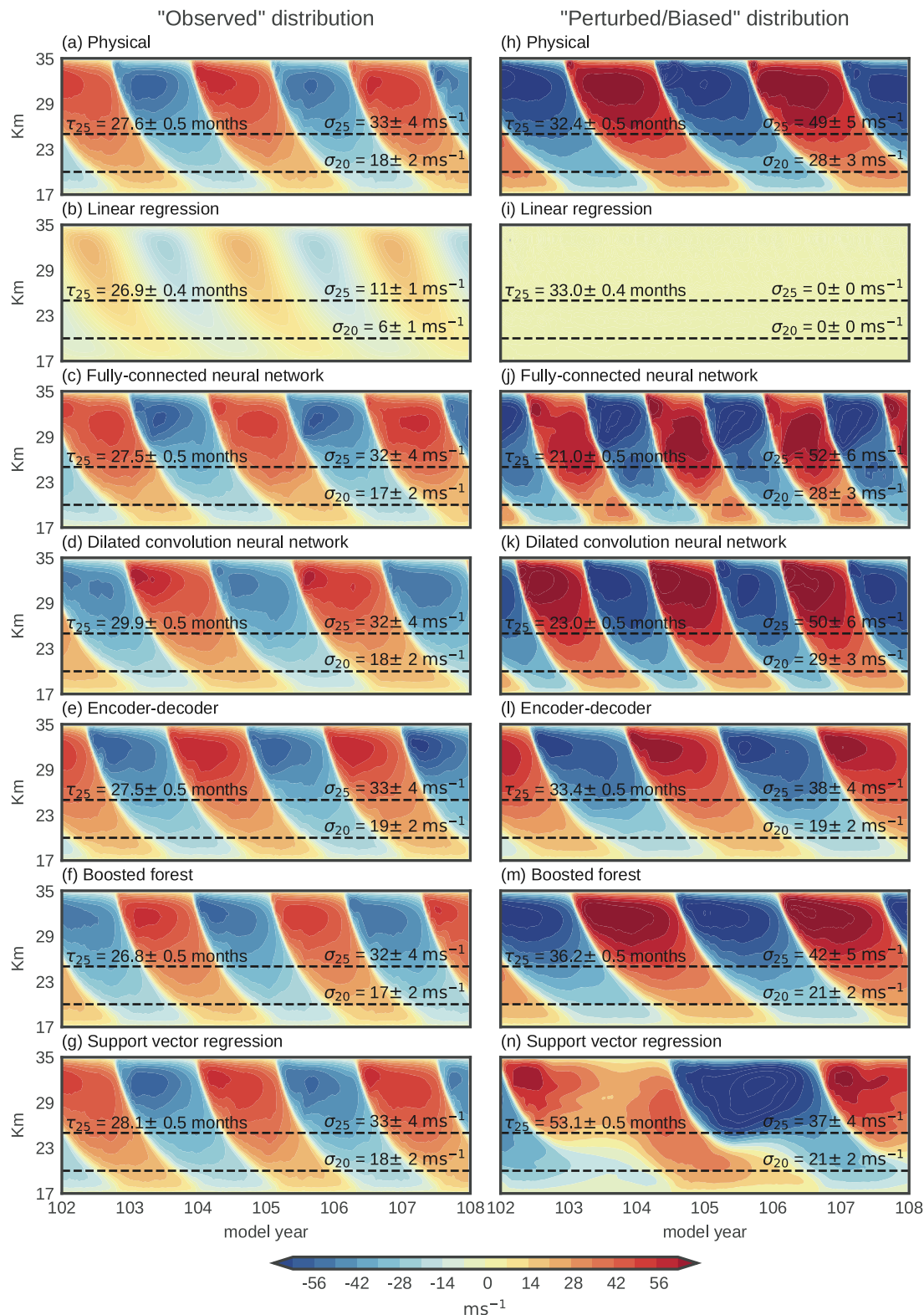


**FIGURE 1** The gravity-wave sources. At each time step the total source flux  $F_{S0}$  and spectral width  $c_w$  are drawn from one of three bivariate log-normal distributions: (a) the control distribution, obtained as described in Section 2.1 and representing the “observed” distribution, (b) the effective distribution in the “host” model, which represents model biases or climate change, and (c) a hypothetical distribution used to test the learning sensitivity to the correlation (same as the control distribution, but with no correlation). The distribution parameters are given in Table B1 of Appendix B. (d–f) Three sample wave packets drawn from the control distribution in (a), with (d)  $F_{S0} = 3.5$  mPa and  $c_w = 32$  m·s<sup>−1</sup>, (e)  $F_{S0} = 4.0$  mPa and  $c_w = 32$  m·s<sup>−1</sup>, and (f)  $F_{S0} = 4.0$  mPa and  $c_w = 70$  m·s<sup>−1</sup>, also indicated by red squares in (a). The resulting gravity-wave drags are shown in Figure 5. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

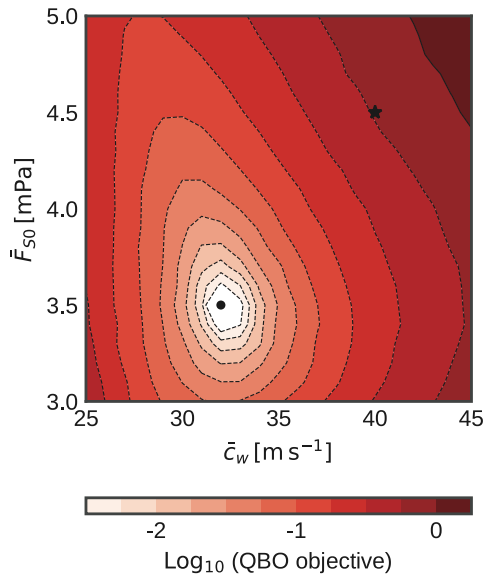
the fact that the GW sources are themselves dependent on the resolved flow, making them susceptible to model biases. Accordingly, our experimental parameters consist of the mean source flux  $\bar{F}_{S0}$  and spectral width  $\bar{c}_w$ , with the control experiment defined by the combination of the two that yields the amplitude and period of the “observed” QBO, defined here by minimizing

$$\begin{aligned} & \frac{[\sigma(25 \text{ km}) - 33 \text{ m} \cdot \text{s}^{-1}]^2}{[33 \text{ m} \cdot \text{s}^{-1}]^2} \\ & + \frac{[\sigma(20 \text{ km}) - 18 \text{ m} \cdot \text{s}^{-1}]^2}{[18 \text{ m} \cdot \text{s}^{-1}]^2} \\ & + \frac{[\tau(25 \text{ km}) - 28 \text{ months}]^2}{[28 \text{ months}]^2}, \end{aligned} \quad (4)$$

where  $\sigma$  denotes the QBO amplitude in m·s<sup>−1</sup> and  $\tau$  the QBO period in months. Following Garfinkel et al. (2022), the QBO amplitude is evaluated in the mid ( $z = 25$  km) and low ( $z = 20$  km) stratosphere by means of the zonal wind standard deviation, and the QBO period is evaluated in the mid stratosphere by means of the dominant Fourier mode. While this choice of vertical levels is arbitrary, the results are insensitive to variations, provided



**FIGURE 2** The simulated QBO in the data-driven models. Left column: in response to the “true” GW sources. Right column: in response to the biased GW sources. From top to bottom, (a,h) the physical model for comparison, (b,i) the linear regression model, (c,j) the fully connected neural network, (d,k) the dilated convolutional neural network, (e,l) the encoder–decoder, (f,m) the boosted forest, (g,n) the support-vector regression model. The color scale is determined by the global absolute maximum of the zonal wind in the control experiment of the physical model, with 21 equally spaced contours between  $\pm \max |u|$ , and is uniform across all panels. The high-level amplitude ( $\sigma_{25}$ ), low-level amplitude ( $\sigma_{20}$ ), and period ( $\tau_{25}$ ) of the simulated QBO in each model, estimated as detailed in Section 2.1, are indicated in the panels. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]



**FIGURE 3** The QBO objective. The (log-scaled) objective in Equation 4 as a function of the mean source flux (ordinate) and spectral width (abscissa). The optimum at  $\bar{F}_{S0} = 3.5$  mPa,  $\bar{c}_w = 32$  m·s<sup>-1</sup> (precise to 0.1 mPa in  $\bar{F}_{S0}$  and 1 m·s<sup>-1</sup> in  $\bar{c}_w$ ), indicated by a black dot, corresponds to the control source distribution and represents the “observed” distribution. The black star at  $\bar{F}_{S0} = 4.5$  mPa,  $\bar{c}_w = 40$  m·s<sup>-1</sup> indicates the biased distribution in the “host” model. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

one avoids getting too close to the lower boundary at  $z = 17$  km, where the winds are held fixed. The incorporation of the lower-level amplitude helps narrow down the optimum but is not essential. The existence of a well-defined dominant Fourier mode in our simulations is confirmed in Figure S1 of the supplementary material.

Figure 3 shows the resulting (log-scaled) objective in Equation 4 as a function of the mean source flux and spectral width. The experimental range ( $3 \leq \bar{F}_{S0} \leq 5$  mPa,  $25 \leq \bar{c}_w \leq 45$  m·s<sup>-1</sup>) was chosen by trial and error to be as large as possible while maintaining numerical stability, a meaningful QBO (exhibiting descending shear zones), and a “not too long” QBO period (up to 56 months). Reassuringly, this range covers the main portion of the corresponding parameter space ( $1.3 \leq \bar{F}_{S0} \leq 6.3$  mPa,  $5 \leq \bar{c}_w \leq 45$  m·s<sup>-1</sup>) in Garfinkel et al. (2022), who studied the QBO’s sensitivity to these parameters in MiMA. Within the sampled resolution, the optimum is located around  $\bar{F}_{S0} = 3.5$  mPa and  $\bar{c}_w = 32$  m·s<sup>-1</sup>, denoted by a black dot. The simulated QBO at this point, which serves as our control experiment, is shown in Figure 2a for the last 6 years of a 108-year-long simulation. The remaining physical model parameters used throughout this work are listed in Table B1 of Appendix B.

## 2.2 | Perturbation experiments: Sensitivity to the source spectrum

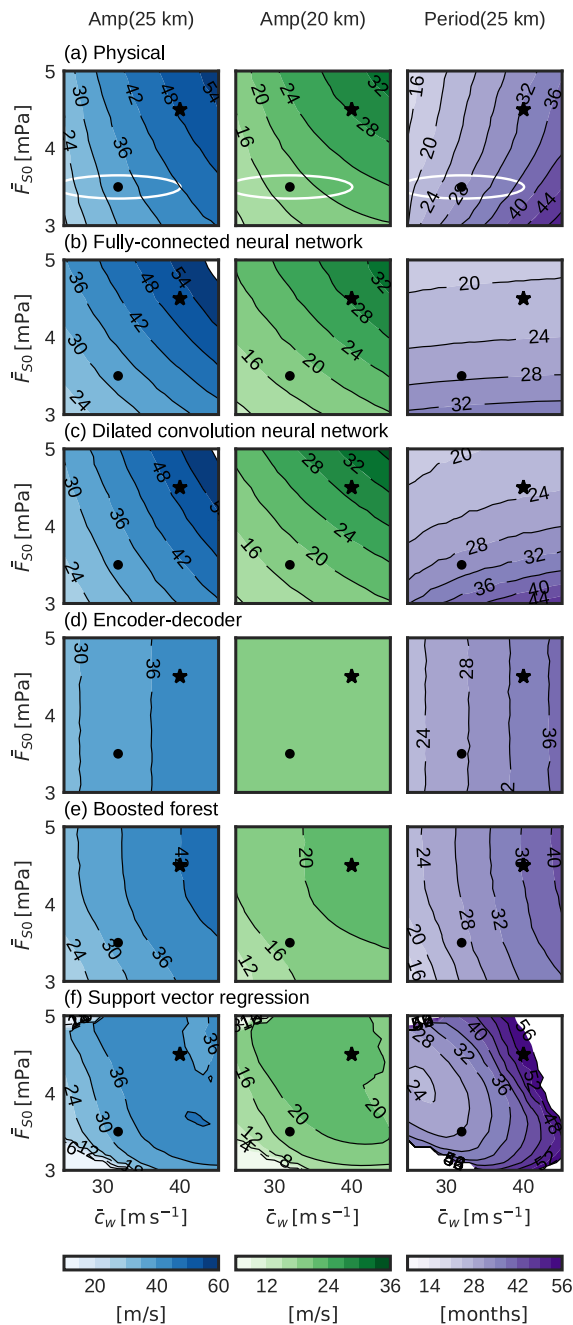
Having identified the optimal source distribution (the one leading to the observed QBO amplitude and period), we now examine the sensitivity of the solutions to changes in the mean source flux  $\bar{F}_{S0}$  and spectral width  $\bar{c}_w$ . These two parameters correspond to primary sources of uncertainty in GCMs, total precipitation and depth of convection, respectively, and constitute our key experimental parameters.

First, compare the perturbed QBO in Figure 2h, corresponding to a source distribution with ( $\bar{F}_{S0} = 4.5$  mPa,  $\bar{c}_w = 40$  m·s<sup>-1</sup>), with the “observed” one in Figure 2a, corresponding to a source distribution with ( $\bar{F}_{S0} = 3.5$  mPa,  $\bar{c}_w = 32$  m·s<sup>-1</sup>). Increasing the mean source flux and spectral width results, in this case, in a slower and stronger QBO. The period increases from  $27.6 \pm 0.5$  months to  $32.4 \pm 0.5$  months. The high-level amplitude, for example, increases from  $33 \pm 4$  m·s<sup>-1</sup> to  $49 \pm 5$  m·s<sup>-1</sup>.

Next, in order to examine the sensitivity to changes in the mean source flux and spectral width in more detail, we integrate the model for each combination of  $\bar{F}_{S0}$  and  $\bar{c}_w$  in our experimental range for 108 years and calculate the amplitude and period after 12 years of “spinup”. Figure 4a shows the resulting high-level amplitude (left, blue shading), low-level amplitude (center, green shading), and period (right, purple shading) in the physical model, as functions of  $\bar{c}_w$  (abscissa) and  $\bar{F}_{S0}$  (ordinate). The control distribution is indicated by a black dot and the perturbed one by a black star. Indeed, increasing both  $\bar{F}_{S0}$  and  $\bar{c}_w$  results in a slower and stronger QBO. More accurately, however, increasing  $\bar{c}_w$  while holding  $\bar{F}_{S0}$  fixed leads to stronger QBO amplitudes and longer QBO periods, whereas increasing the source flux  $\bar{F}_{S0}$  while holding  $\bar{c}_w$  fixed leads to stronger amplitudes but shorter periods.

The sensitivities of the QBO amplitude and period in our stochastic model are consistent with those predicted by the classic model in Holton and Lindzen (1972) and Plumb (1977). As in the classic model, distancing the critical levels (increasing  $\bar{c}_w$ ) increases the wind amplitudes and lengthens the time taken for the winds above the shear zone to dissipate, while increasing the wave amplitude (increasing  $\bar{F}_{S0}$ ) shortens the time taken for the shear zones to descend and form in the first place. Unlike Holton and Lindzen (1972) and Plumb (1977), the presence of more than two waves results in a wave filtering mechanism, instead of the critical levels mechanism, so the effect of increasing  $\bar{F}_{S0}$  does not saturate.

The QBO amplitude sensitivity in our stochastic 1D model is also qualitatively consistent with that found in Garfinkel et al. (2022) using MiMA, but is quantitatively



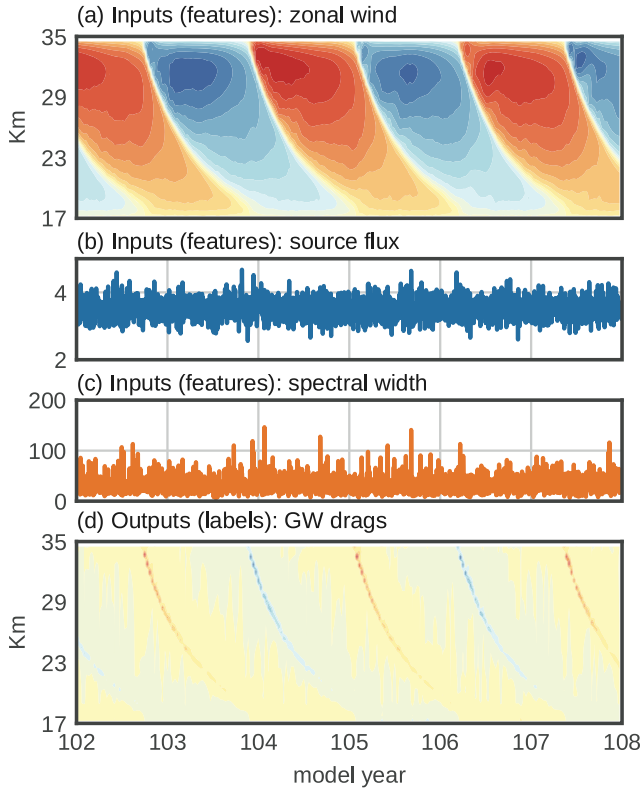
**FIGURE 4** The QBO sensitivity in the data-driven models. The sensitivities of the high-level amplitude (left, blue shading), low-level amplitude (center, green shading), and period (right, purple shading) to changes in the mean spectral width  $\bar{c}_w$  (abscissa) and source flux  $\bar{F}_{S0}$  (ordinate). From top to bottom: (a) the physical model for comparison, (b) the fully connected neural network, (c) the dilated convolutional neural network, (d) the encoder–decoder, (e) the boosted forest, (f) the support-vector regression model. The black dot at  $(c_w = 32 \text{ m s}^{-1}, F_{S0} = 3.5 \text{ mPa})$  in each panel indicates the control experiment used for training. The black star at  $(\bar{F}_{S0} = 4.5 \text{ mPa}, \bar{c}_w = 40 \text{ m s}^{-1})$  indicates the biased distribution. The white ellipse in panel (a) indicates the standard deviation of  $F_{S0}$  and  $c_w$  samples in the training dataset. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

more sensitive. For the same range of source fluxes and spectral widths, the total amplitude variation found in Garfinkel et al. (2022) was about 30%–50%, compared with 100%–250% in the present work. This is to be expected, at the very least considering the fact that the resolved and parameterized waves in QBOi models have comparable contributions (Bushell et al., 2022). The 1D model is also more sensitive in terms of the QBO period. Garfinkel et al. (2022) found little to no change in the QBO period over the range of source fluxes and spectral widths considered here. While we cannot expect a more quantitative agreement, it is worthwhile mentioning that the control values ( $F_{S0} = 3.5 \text{ mPa}$ ,  $c_w = 32 \text{ m s}^{-1}$ ) are remarkably close to those used in Garfinkel et al. (2022) ( $F_{S0} = 4.3 \text{ mPa}$ ,  $c_w = 35 \text{ m s}^{-1}$ ).

### 3 | DATA-DRIVEN MODELS

Recall the scenario envisaged: a data-driven GW parameterization is trained on observed GW drags, as well as some proxies of the GW sources, to yield a relation of the form “GW drag = GW drag(flow, GW sources)”, and this parameterization is then implemented in an operational GCM having perturbed (biased) sources. For convective GW in the Tropics, this is partly due to the representation of convection in the model, and partly due to the fact that the GW sources are themselves dependent on the resolved flow, making them susceptible to model biases. The question is then how will a data-driven model trained on a certain distribution fare when fed a perturbed one, and how will it respond to changes in the sources under climate perturbations?

The learning task at hand is a supervised regression task consisting of finding a function,  $f : [u, F_{S0}, c_w] \rightarrow S$ , that best fits the given data samples  $\{[u, F_{S0}, c_w]_i, S_i\}_{i=1}^{N_{\text{samp}}}$ . In other words, our inputs, or features, consist of the zonal wind, source flux, and spectral width, and the outputs, or labels, consist of the resulting wave drag. Our training dataset, shown in Figure 5, consists of 96 years of daily samples, after 12 years of spinup, simulated using the control parameters described in Section 2.1, and representing the “observed” record. The samples are not necessarily sequentially ordered during training. Due to the temporal correlation, consecutive samples add less information to the training data. Indeed, for the neural networks, we found that randomly shuffling the training data improves the convergence. This is also consistent with the results of Pahlavan et al. (2024), who found that 72 weeks of data are sufficient for training a CNN on the 1D QBO model, provided they are spaced 1 month apart to cover different phases of the QBO.



**FIGURE 5** The training dataset. A total of 6 years of daily time samples, out of the 96 years available in the training dataset, are shown. The samples need not be ordered sequentially during training. For each sample, the inputs consist of (a) the zonal wind profile (excluding the top and bottom boundaries), (b) the source flux, and (c) the spectral width of the GW packet. The output consists of the GW drags (excluding the boundaries). During training, the data are subject to standardization, and hence the units are arbitrary. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

In order to account for the zero wind (Dirichlet) boundary conditions imposed in the physical model, we found it easiest to exclude the boundaries during training. After removing the top and bottom boundaries,  $[u, F_{S0}, c_w] \in \mathbb{R}^{(N_{\text{samp}}) \times (N_{\text{lev}} - 2)}$  and  $S \in \mathbb{R}^{(N_{\text{samp}}) \times (N_{\text{lev}} - 2)}$ .

We consider the following six models.

- (i) **A linear regression model:** While  $S$  is a nonlinear and nonlocal function of  $u$ , empirically it is similar to the zonal wind shear  $\partial u / \partial z$ . This is a manifestation of the idealized case of constant wave flux density studied in Lindzen and Holton (1968), where the forcing is exactly proportional to the zonal wind shear. Since  $\partial u / \partial z$  can be linearly approximated by  $u$  (to any desired accuracy), it is conceivable that linear regression will approximate the forcing to some degree. The linear model used here includes a bias term, that is, we seek the least-squares fit to  $S = [u, F_{S0}, c_w]W + b$ , where  $W \in \mathbb{R}^{(N_{\text{lev}}) \times (N_{\text{lev}} - 2)}$ , and  $b \in \mathbb{R}^{(N_{\text{lev}} - 2)}$ . Hence,

the total number of “trainable” parameters is 5254.

- (ii) **A fully connected feed-forward neural network (NN):** The theoretical basis for this type of model is the universal approximation theorem(s), which, generally speaking, establishes their ability to approximate nonlinear functions to any desired accuracy provided there are sufficient degrees of freedom. See, for example, Goodfellow et al. (2016) for an exposition and Espinosa et al. (2022) for an application to GW parameterizations. We start with a fully connected, feed-forward, network, where each neuron in one layer is connected to all neurons in the following layer and the information flows sequentially from the input layer, through the hidden layers, to the output layer. The network’s architecture and optimization parameters are given in Table B2 of Appendix B. The training dataset was first randomly shuffled and then split in half, for a total of 48 years of training samples and 48 years of validation samples. The data were propagated through the network in batches of 360 days for 100 epochs. The output  $S$  was scaled by  $\bar{F}_{S0}$ . No scaling was applied on the inputs  $u$ ,  $F_{S0}$ , or  $c_w$ .

- (iii) **A dilated convolutional neural network (CNN):** CNNs are a specialized form of NNs particularly suitable for data made up of distinct and repeatable “building blocks”, for example, headlights and bumpers of motor vehicles, or the shear zones of the QBO. See, for example, Goodfellow et al. (2016) for an exposition and Chattopadhyay et al. (2020) for an application to climate data. An important feature of CNNs is that they are less prone to overfitting than fully connected networks. The CNN used here has the same architecture as the dilated CNN used in Hardiman et al. (2023), consisting of one-dimensional filters with fixed kernel size and increasing dilation (increasing the filters’ receptive fields). The network architecture is given in Table B3 of Appendix B. The optimization parameters and training procedure were identical to those used for the NN.

- (iv) **An encoder–decoder (ED):** This encoder–decoder structure is inspired by CNN variational autoencoders. ED is not an autoencoder, since the input and output are not the same, but the same structure is used to encode and decode information. Convolutional layers are used to encode the input information locally onto a reduced dimension latent space, where global interactions are processed with dense layers. The resulting latent-space variable is then decoded with transposed convolutional layers to yield the output. See Kingma and Welling (2013) for a general exposition and Yang and Gerber (2024) for

application to GW parameterizations. The parameters of the ED used here are given in Table B4 of Appendix B. The training dataset was split in half, for a total of 48 years of training samples and 48 years of validation samples.

- (v) **A boosted forest (BF):** Regression trees make predictions by traversing a binary tree according to the components of the input vector. At each level, the traversal moves to the left or right by comparing a particular component of the input against a pre-determined threshold. The returned value is the mean of the training samples that reached the same leaf as the input. Boosted forests are ensembles of trees where each new tree is trained on the residuals of those trained before it, so that the ensemble prediction zeros in on the correct answer. See Breiman et al. (1984) and Friedman (2001) for expositions of regression trees and boosted forests, respectively, and Connelly and Gerber (2023) for an application to GW parameterizations. The parameters of the BF used here are given in Table B5 of Appendix B.
- (vi) **A support-vector regression (SVR) model:** SVR is a variation of support-vector machines, a classification algorithm that attempts to make the data linearly separable by mapping them to higher dimensions. Similarly, by mapping the data with the so-called “kernel trick”, SVR aims to restrict data points within an  $\epsilon$ -tube of a hyperplane. Intuitively, the kernel allows one to narrow the space of comparison for an input sample, allowing for nonlinear regression. Also, like support-vector machines, one then finds only a subset of input data points, called supported vectors, that have contributed to determining the SVR model. See, for example, Drucker et al. (1996) and Smola and Schölkopf (2004) for a general exposition. Since SVR, by its nature, is designed for one-dimensional output regression, we used a column of independent one-dimensional SVR models in our task. This model architecture harms the efficiency, as each of these one-dimensional SVR models uses a different subset of support vectors. In practice, we minimize the size of the training dataset to control the number of support vectors. We found that an SVR model can emulate the source term satisfactorily even with only 1% of the dataset used for training (less than 1 year of data). The parameters of the SVR model used here are given in Table B6 of Appendix B.

These models were chosen to provide a “representative” sample of frequently used data-driven methods and highlight potential strengths and weakness of different ML strategies, but not to make definitive statements that approach A is always better than approach B. Importantly,

they were not purposely designed for the present work. For example, the CNN was optimized to emulate the Warner and McIntyre scheme in the Met Office HadGEM3-GA8.0 climate model (Hardiman et al., 2023); we use the same architecture, but trained on our control experiment. Likewise, the ED and BF were designed to emulate the AD99 (Alexander & Dunkerton, 1999) GW parameterization in MiMA. In general, any ML strategy can be further optimized, and the “best” approach depends on the circumstances. For a climate model, for instance, the constraint is ultimately the best skill for the least amount of computational time, but even “skill” can be subjective: should we require the best climatology, or the best representation of extreme events?

## 4 | RESULTS

We are interested in the coupled problem, where Equation 1 is integrated numerically with the RHS replaced by the corresponding model, often referred to as an “online” simulation. When doing so, it is imperative to assess the skill of the models in their intended modus operandi. Attempting to optimize the models based on offline metrics can lead to online instability, which is perhaps associated with overfitting. We therefore focus on our main goal: how do these models perform online when grafted into the host?

### 4.1 | Control experiment: Simulating the present day climate

Figure 2 shows the QBO generated by the various ML models, compared with the physical model in panels (a,h), for the last 6 years of a 108-year-long simulation. The left column shows the QBO in response to the control GW source distribution, and the right column the response to the perturbed wave distribution. That is, the left column shows the response to the wave distribution on which the ML models were trained, while the right column shows the QBO in response to an out-of-set GW source distribution. Aside from the generation and maintenance of the QBO, performance can be gauged by the amplitude and period of the resulting QBO, indicated in each panel.

Starting with the control experiment, all models (except linear regression) are able to capture both the QBO period and amplitude quite well. To be more precise, the estimated uncertainty in the QBO period for these centennial length integrations is 0.5 months (0.4 for the linear model), based on the width of the dominant Fourier mode (Figure S1 in the supplementary material). All models agree with the “observed” QBO period (the physical

model) to within this uncertainty, except for the CNN, where the period is biased long by 2 months. The QBO amplitude is also within the estimated uncertainty for all cases (except linear regression), based on the standard error of the standard deviation.

The linear model is able to capture the QBO period surprisingly well, but not the amplitude. Examining the zonal wind as a function of time (Figure S2 of the supplementary material), it is evident that the linear model is unstable. There is a slow but steady trend in the amplitude, where changes in the wind strength feedback on the forcing. In this particular case, the trend happens to be negative (the wind diminishes with time), but in other cases the linear model showed a positive trend (e.g., with an annually varying vertical wind). With  $L_2$  regularization (i.e., ridge regression) the results are nearly identical. In particular, the model is still unstable. With  $L_1$  regularization (i.e., Lasso regression) the model does not converge, which is perhaps another manifestation of the nonlinearity of the source term.

The fact that all of the data-driven schemes perform well makes it easy to gloss over the key result in Figure 2b–g: all ML schemes produce a stable and accurate simulation of the QBO when forced using the control GW sources. Their stability is confirmed in Figure S2 of the supplementary material, which verifies that there is no trend in the zonal wind at  $z = 25$  km for up to 108 years. The stability of the simulated QBO is not a trivial result. It is an “open secret” in the community that high accuracy during training does not guarantee online stability (e.g., Brenowitz et al., 2020), and a stable QBO is, after all, the *raison d'être* of a QBO model.

## 4.2 | Source spectrum sensitivity: Capturing the response to a climate perturbation

We now examine the ability of our data-driven models to capture the sensitivities of the QBO amplitude and period to changes in  $\bar{F}_{S0}$  and  $\bar{c}_w$ . The context of this experiment is climate change. We use data-driven models trained only on the control GW source distribution ( $\bar{F}_{S0} = 3.5$  mPa,  $\bar{c}_w = 32$  m·s<sup>−1</sup>) to simulate the QBO in a perturbed climate where the source parameters have changed. This is a challenging test. While neural networks are capable of extrapolation, the BF and SVR methods can only predict some combination of the data they have seen during training. Still, due to the variability in  $F_{S0}$  and  $c_w$  within the control integration (which represents the natural variability in observations), even these two methods have a chance. The question is whether it is possible to learn enough from variability in the “observations” to capture systematic changes

introduced by the climate perturbation, at least to some extent.

First, consider the response to the perturbed source distribution ( $\bar{F}_{S0} = 4.5$  mPa,  $\bar{c}_w = 40$  m·s<sup>−1</sup>) in Figure 2i–n. This systematic increase in both the source intensity and spectral width could reflect a warmer climate with more intense and deeper convection. The QBO amplitude in the fully connected NN and dilated CNN models increases in response to this change in the wave sources, in agreement with the physical model, the differences being well within the sampling uncertainty. However, the QBO period in both simulations decreases relative to the control, in contrast to the physical model where the period increased! That is, these models fail to capture the sensitivity of the QBO period to the change in source distribution, even qualitatively. In contrast, the encoder–decoder model does capture the increase in the QBO period, but not the amplitude. In fact, the amplitude of the QBO in the perturbed simulation of the ED model is almost the same as that of the “observed” QBO, suggesting the ED model has not “learned” the QBO amplitude sensitivity at all. The SVR model fails to capture the changes in both the amplitude and period, even qualitatively. The BF is the only one that captures the qualitative changes in both the QBO amplitude and period, namely a slower and stronger QBO, but it fails to capture the changes quantitatively.

The perturbation experiment considered above represents just one, perhaps extreme, scenario. We now consider the sensitivities of the different models to changes in the mean source flux and spectral width across the ( $\bar{F}_{S0}$ ,  $\bar{c}_w$ ) plane. We integrate the models for each combination of  $\bar{F}_{S0}$  and  $\bar{c}_w$  in our experimental range ( $3 \leq \bar{F}_{S0} \leq 5$  mPa,  $25 \leq \bar{c}_w \leq 45$  m·s<sup>−1</sup>). For each integration, we compute the amplitude of the QBO at 25 and 20 km and its period, summarizing the results in Figure 4b–f. The black dot in all panels indicates the control experiment on which the models were trained. The white ellipse in Figure 4a indicates the standard deviation of  $F_{S0}$  and  $c_w$  samples in the training dataset. The perturbation experiment shown in Figure 2 is marked by the black star, highlighting its distance from the control experiment.

In terms of the QBO amplitude, the different models succeed to varying extent. The NN and CNN capture the amplitude sensitivity “quite well”, perhaps even quantitatively considering the estimated uncertainty. The BF and SVR model capture the amplitude sensitivity qualitatively, in the sense that the amplitude increases with increasing  $\bar{F}_{S0}$  and  $\bar{c}_w$ , while the ED struggles to capture the amplitude sensitivity even qualitatively. In terms of the QBO period, all five models fail to fully capture the period sensitivity. The CNN and SVR model capture the period sensitivity qualitatively, in the sense that the period increases with decreasing  $\bar{F}_{S0}$  and increasing  $\bar{c}_w$ .

The NN captures the qualitative increase in the period with decreasing  $\bar{F}_{S0}$ , but struggles to capture the sensitivity to  $\bar{c}_w$  altogether. The ED and BF capture the qualitative increase in the period with increasing  $\bar{c}_w$ , but struggle to capture the sensitivity to  $\bar{F}_{S0}$ .

Recall that  $F_{S0}$  and  $c_w$  are positively correlated, representing the positive correlation between the total precipitation and depth of convection. Is it possible that this correlation is the reason why the NN, ED, and BF are only able to learn the period sensitivity to one of them? Unlikely, considering that the NN and BF do capture the sensitivity of the amplitude to both of these parameters and considering the CNN does capture the period sensitivity to both. However, in order to rule out this hypothesis we have repeated the above calculation using the neural network with zero correlation between  $F_{S0}$  and  $c_w$  and the results are nearly identical (Figures S3 and S4 in the supplementary material).

The observed QBO in the atmosphere is more irregular than that in our simple model, due to the annual cycle in the vertical advection and GW sources and random fluctuations from synoptic and planetary-scale waves. Can we improve the ML scheme's ability to generalize by training on less regular data, allowing the models to "see" a wider range of wind profiles? To test this hypothesis we have repeated the above calculations for all five models with an annual cycle added to the vertical wind. Instead of the constant Brewer–Dobson upwelling  $w = 3 \times 10^{-4} \text{ m s}^{-1}$  used above, we repeated the calculations for

$$w(t) = \left[ 3 + 2 \sin\left(\frac{2\pi t}{360 \text{ days}}\right) + \varepsilon_w \right] \times 10^{-4} \text{ m s}^{-1}, \quad (5)$$

where  $\varepsilon_w \sim U(-0.5, 0.5)$  is white noise. Figures S5 and S6 in the supplementary material show that our models learn the new, less regular, control QBO. Despite this, we observed no improvements in terms of their ability to capture the sensitivity to changes in  $\bar{F}_{S0}$  and  $\bar{c}_w$ .

Having examined the global sensitivity of the solutions over our wide experimental range, we now take a closer look at the local sensitivity in the vicinity of the control source distribution, which would represent a gradual climate drift. Figure 6 shows the partial derivatives of the high-level amplitude (left column, blue shading), low-level amplitude (center column, green shading), and period (right column, purple shading) with respect to  $\bar{F}_{S0}$  (top row) and  $\bar{c}_w$  (bottom row). The partial derivatives in this figure are normalized on the corresponding derivatives in the physical model. Thus a value of 1 corresponds to the correct response, and values greater (less) than 1 indicate an exaggerated (muted) response relative to the physical model.

No one method perfectly captures the partial derivatives at the control distribution. The NN and CNN are overly sensitive to changes in  $\bar{F}_{S0}$  for both the QBO amplitude and period, but underpredict the response to changes in  $\bar{c}_w$ . As observed above, the ED struggles to capture any response to changes in  $\bar{F}_{S0}$  and responds only weakly to changes in  $\bar{c}_w$ . The BF performs well across most metrics, but captures the wrong sign of the period response to changes in  $\bar{F}_{S0}$ . Finally, considering the estimated uncertainty, the SVR model is perhaps the most accurate locally.

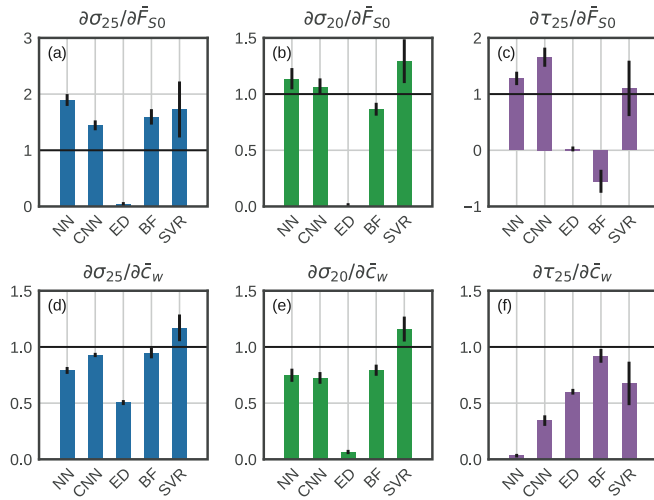
### 4.3 | Calibration: Preconditioning the source distribution

In the climate-change context of the previous section, we wanted the data-driven schemes to capture the response to changes in the source distribution. However, this skill does not assist with the calibration problem and can work against it. A scheme perfectly capable of generalizing will react to model biases and cement them when grafted into a numerical model with biased sources. Thus, a different measure is needed to account for model biases, and, to the extent that the training dataset does, indeed, represent the observed conditions, the scheme ought to be changed as little as possible.

A simple way to overcome a model bias in the sources, while also adhering to the observational constraints, is a preconditioning step where the wave sources are first remapped to the observed distribution before being fed to the GW scheme. For example, if the precipitation in a model is systematically too large relative to the observations, one would always need to reduce the value of  $F_{S0}$  provided by the model before passing it to the data-driven scheme; otherwise, the GW momentum forcing would be systematically larger, biasing the QBO. The mapping is done by means of the cumulative distribution function (CDF) of the sources, such that the amplitude of convection at the 95th percentile level in the model is rescaled to that of the 95th percentile in the observations, and so forth. The advantage of this approach is that it is agnostic to the chosen data-driven method.

The procedure is greatly simplified by the fact that our data-driven schemes are only weakly sensitive to the correlation between the source flux and spectral width (Section 4.2 and Figures S3 and S4 in the supplementary material), and can be treated as independent random variables. Using (informally) the solution of the one-dimension optimal transport problem, they are remapped as follows:

$$X_{\text{remapped}} = CDF_{\text{observed}}^{-1} \circ CDF_{\text{biased}}(X_{\text{biased}}), \quad (6)$$



**FIGURE 6** The QBO local sensitivity. The gradient of the high-level amplitude (left, blue shading), low-level amplitude (center, green shading), and period (right, purple shading) for the control distribution. Top: the partial derivative with respect to the mean source flux  $\bar{F}_{S0}$ . Bottom: the partial derivative with respect to the mean spectral width  $\bar{c}_w$ . The derivatives for each model (each bar) are normalized by the corresponding derivative of the physical model. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

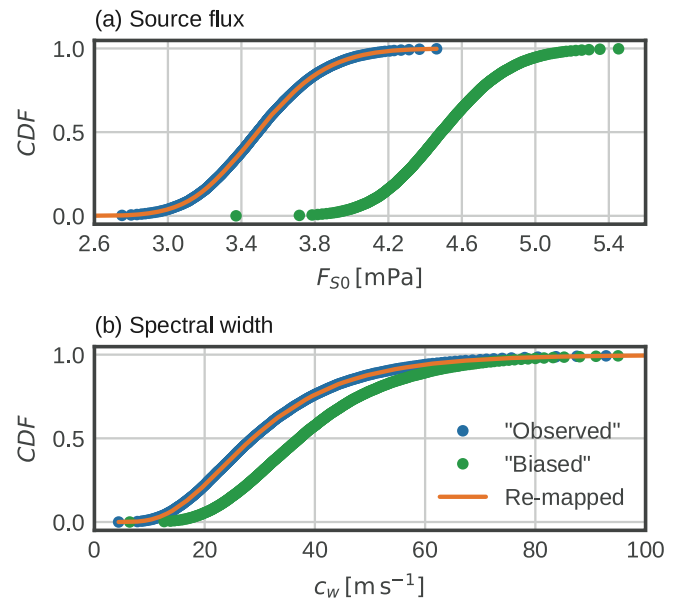
where  $X \in \{F_{S0}, c_w\}$ . The CDFs are evaluated empirically as

$$CDF(x) = \frac{1}{N_{\text{samples}}} \sum_{i=1}^{N_{\text{samples}}} \mathbb{1}_{X_i \leq x}, \quad (7)$$

where  $\mathbb{1}_{X \leq x}$  is the indicator function (i.e.,  $\mathbb{1}_{X \leq x} = 1$  for  $X \leq x$  and 0 otherwise). The observed sources enter Equation 6 implicitly via their estimated distribution. In practice, Equation 7 is evaluated on the sampled wave sources. In order to apply  $CDF_{\text{observed}}^{-1}$  to arbitrary images of  $CDF_{\text{biased}}$ , the CDFs were linearly interpolated.

Figure 7 shows the CDFs of the “observed” (blue dots), biased (green dots), and (a) the remapped source flux and (b) spectral width, confirming that the remapped sources are distributed according to the “observed” distribution. Figure 8 shows the simulated QBO, using the neural network, (a) before and (b) after remapping the sources. Upon remapping the sources, the neural network yields the correct QBO amplitude and period and is stable for at least 108 years, confirming, a posteriori, our assumption that  $\bar{F}_{S0}$  and  $\bar{c}_w$  can be treated independently for the purpose of modeling the wave drags.

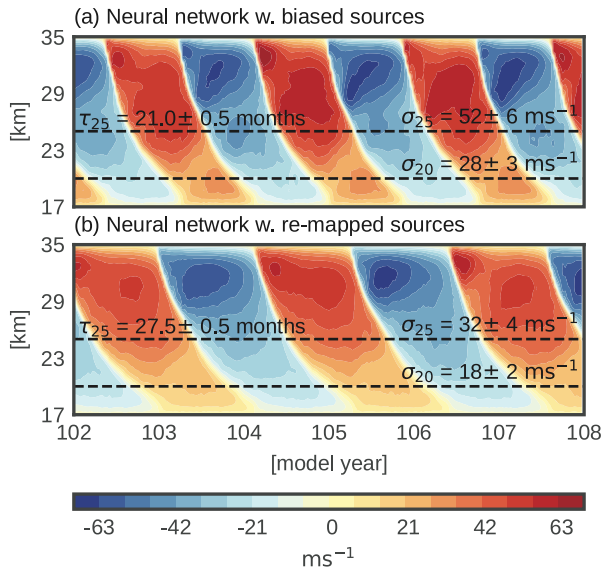
While providing a simple solution in our idealized 1D model, the application of this preconditioning approach to parameterized GWs in GCMs is more involved. Most notably, if the source flux and spectral width in observations cannot be remapped independently, as in the present case, one has to solve a two-dimensional optimal transport problem. Moreover, this approach relies on having faithful estimations of the GW source distribution in both the observations and the host. Specifically, a continuous approximation of the CDFs is required. The first can be difficult to obtain from observations with limited spatio-temporal coverage, and would likely have to be estimated from high-resolution global climate models. The second can be computationally taxing. In the present



**FIGURE 7** Source distribution preconditioning. The “observed” (blue dots), biased (green), and remapped (orange) CDFs of the (a) source flux and (b) spectral width. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

work, the wave sources can be drawn from the bivariate log-normal distribution upfront, so the empirical CDFs can be evaluated a priori, and the interpolation has to be applied only once. In GCMs, the wave sources are generated online, so the CDFs have to be evaluated and interpolated repeatedly, every time the wave sources in the model are regenerated.

In addition, while this preconditioning approach works well for calibrating the scheme to work in the current climate, it is unclear how much it can be trusted in a climate-change context. It would provide the correct response if the relative change in the source distribution in the host matches the relative change in the “true” source



**FIGURE 8** Source remapping in the neural network model. The simulated QBO using the neural network (a) with biased sources and (b) after remapping the biased sources to the observed ones. The color scale is determined by the global absolute maximum of the zonal wind in the unbiased physical model (i.e., as in Figure 2), with 21 equally spaced contours between  $\pm \max |u|$ . The high-level amplitude ( $\sigma_{25}$ ), low-level amplitude ( $\sigma_{20}$ ), and period ( $\tau_{25}$ ) of the simulated QBO in each model, estimated as detailed in Section 2.1, are indicated in the panels. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

distribution. This is admittedly a tall order if the source distribution in the host is different from that in the control climate to begin with.

## 5 | DISCUSSION

A primary concern with the advent of machine learning for climate modeling is making sure that the models yield the right results for the right reasons. The particular example studied here is the graft-versus-host problem, where a data-driven scheme might be incompatible with its host climate model, leading to erroneous results. This problem emerges in two related physical contexts, a climate-change context (the generalization problem) and model biases context (the calibration problem). In the former, we want to capture the response to a physically induced change in the sources. In the latter, we have to correct for a nonphysical bias in the model's sources. These can be conflicting aims. A scheme can be well calibrated to provide the right results for the control climate, but have no ability to generalize to different conditions. The best we can hope for is a scheme that generalizes well and can be adjusted sensibly to work in the control climate.

We considered the graft-versus-host problem for data-driven gravity wave (GW) parameterizations in a stochastically driven 1D quasibiennial oscillation (QBO) model, where both climate change and model biases are represented by perturbations in the GW source distribution. The 1D model provides a control environment where the two issues can be examined independently, as opposed to a climate model where the drifting climate is not easily disentangled from model biases. We first examined the expected response of the physical model to perturbations in the GW source distribution. Having trained various data-driven schemes on the “control” distribution, we examined the ability of these schemes to capture the expected perturbations, compared with the physical model, that is, their ability to generalize. We then proposed a preconditioning approach to calibrate a perturbed model.

Our results demonstrate that data-driven schemes trained on “observations” are sensitive to perturbations in the wave sources. While all methods considered here were able to emulate accurately the stochastic source term on which they were trained, no one method was able to generalize fully to perturbations in the wave sources, in terms of the amplitude and period of the resulting QBO. Some methods were able to capture the sensitivity of the QBO amplitude to changes in the wave sources (even quantitatively), others captured the sensitivity of the QBO period (mostly qualitatively), but no method captured the full response.

We showed that a scheme can be calibrated by preconditioning the sources to account for differences in the source distribution between the observed climate and host model. For a relatively low-dimensional problem like this, optimal transport allows us to remap the source distributions, so that a data-driven scheme sees the same wave distribution when grafted into the host as it did from the observations. This approach, however, will only generalize to new climate conditions to the extent that a data-driven scheme can learn the climate sensitivity from the observed, present-day, variability.

We have focused on the sensitivity to perturbations in the wave sources, which is a primary source of uncertainty in climate models. Other model biases can trigger the graft-versus-host problem as well, for example, biases in the Brewer–Dobson circulation (represented here by the upwelling velocity  $w$ ), or differences in the resolved wave forcing. This raises the difficult issue of making a data-driven scheme scale-aware. In a realistic context, one must make assumptions about what is “resolved” versus “unresolved” in the construction of the training dataset. Ideally, one could custom-build the training dataset for a given model, but when this

is not practical, transfer learning may be an option. Transfer learning is also an option for preparing a scheme to work in a climate-change context if one can obtain limited data from the future climate, for example, from a high-resolution model with modified boundary conditions taken from a climate-change scenario integration Sun et al. (2023b).

Another issue concerning the development of data-driven parameterizations is the length of the training set. In the present work, we considered a plentiful data limit in order to test our schemes at their best. In practice, however, the high-frequency, high-resolution, outputs required to resolve gravity waves limit the length of the records. A typical training set is expected to cover short periods, of the order of weeks (e.g., Sun et al., 2023a) to months. Pahlavan et al. (2024) studied the small versus large data regimes in a 1D QBO model in more detail. They found that 18 months of data were insufficient for emulating a (physically and numerically) stable QBO using a 12-layer CNN (with about 11,000 trainable parameters), even for a simpler configuration of the model with only two waves, no vertical advection, and a white noise forcing. They were able to make their scheme stable by retraining the second and last layers of their CNN iteratively (with additional data), a form of transfer learning termed “offline–online learning”.

The limited data problem, however, is exacerbated in the 1D model. A data-driven method trained on higher complexity climate model outputs benefits from additional variables and geographical regions, such as outside the Tropics. Indeed, Espinosa et al. (2022) were able to learn the AD99 scheme in MiMA using only 12 months of global data, when the QBO was in its westerly phase: the key was training data from the midlatitudes, which provided a wider range of wind and momentum deposition profiles.

Transfer learning seems to be a useful approach for handling both the scale-awareness and small-data problems, and perhaps other problems in the context of data-driven parameterizations. However, it also raises another problem, the fidelity problem. The challenge is that, more so than traditional physics-based schemes, a data-driven scheme must adhere to the observational constraints imposed during training. Retraining parts of the scheme using data from the host model, for example, runs the risk of overriding the observational constraints.

The strength of the 1D model is that it allows us to explore all the difficult issues on the machine-learning side with a very simple atmospheric model, here simply the left-hand side of Equation 1. It allowed us to explore

a wide range of data-driven approaches in the coupled context, highlighting strengths and weaknesses of each approach. In future work, we plan to use it to explore these thornier questions of calibration in the context of more varied model biases, and to extend the offline–online learning approach of Pahlavan et al. (2024) to the climate-change context.

## AUTHOR CONTRIBUTIONS

**Ofer Shamir:** conceptualization; formal analysis; investigation; methodology; writing – original draft. **David S. Connelly:** formal analysis; investigation; methodology. **Steven C. Hardiman:** software; writing – review and editing. **Zihan Shao:** formal analysis; investigation; methodology. **L. Minah Yang:** formal analysis; investigation; methodology. **Edwin P. Gerber:** conceptualization; funding acquisition; supervision; writing – original draft.

## ACKNOWLEDGEMENTS

This work was supported by Schmidt Sciences, as part of the Virtual Earth System Research Institute (VESRI). We also acknowledge support from the US National Science Foundation through award OAC-2004572.

## CONFLICT OF INTEREST STATEMENT

We declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## ORCID

Ofer Shamir  <https://orcid.org/0000-0003-3079-5812>

Edwin P. Gerber  <https://orcid.org/0000-0002-6010-6638>

## REFERENCES

- Alexander, M. & Dunkerton, T. (1999) A spectral parameterization of mean-flow forcing due to breaking gravity waves. *Journal of the Atmospheric Sciences*, 56, 4167–4182.
- Alexander, M.J., Liu, C., Bacmeister, J., Bramberger, M., Hertzog, A. & Richter, J.H. (2021) Observational validation of parameterized gravity waves from tropical convection in the whole atmosphere community climate model. *Journal of Geophysical Research: Atmospheres*, 126, e2020JD033954.
- Baldwin, M., Gray, L., Dunkerton, T., Hamilton, K., Haynes, P., Randel, W.J. et al. (2001) The quasi-biennial oscillation. *Reviews of Geophysics*, 39, 179–229.
- Beres, J.H., Alexander, M.J. & Holton, J.R. (2004) A method of specifying the gravity wave spectrum above convection based on latent heating properties and background wind. *Journal of the Atmospheric Sciences*, 61, 324–337.
- Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984) *Classification and regression trees*. New York: Chapman and Hall/CRC.

- Brenowitz, N.D., Beucler, T., Pritchard, M. & Bretherton, C.S. (2020) Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, 77, 4357–4375.
- Bushell, A., Anstey, J., Butchart, N., Kawatani, Y., Osprey, S., Richter, J. et al. (2022) Evaluation of the quasi-biennial oscillation in global climate models for the sparc qbo-initiative. *Quarterly Journal of the Royal Meteorological Society*, 148, 1459–1489.
- Butchart, N. (2014) The brewer-dobson circulation. *Reviews of Geophysics*, 52, 157–184.
- Chattopadhyay, A., Hassanzadeh, P. & Pasha, S. (2020) Predicting clustered weather patterns: a test case for applications of convolutional neural networks to spatio-temporal climate data. *Scientific Reports*, 10, 1317.
- Connelly, D.S. & Gerber, E.P. (2023) *Regression forest approaches to gravity wave parameterization for climate projection*. ESS Open Archive.
- Drucker, H., Burges, C.J., Kaufman, L., Smola, A. & Vapnik, V. (1996) Support vector regression machines. *Advances in Neural Information Processing Systems*, 9.
- Ebdon, R. (1960) Notes on the wind flow at 50 mb in tropical and sub-tropical regions in january 1957 and january 1958. *Quarterly Journal of the Royal Meteorological Society*, 86, 540–542.
- Ebdon, R. & Veryard, R. (1961) Fluctuations in equatorial stratospheric winds. *Nature*, 189, 791–793.
- Espinosa, Z.I., Sheshadri, A., Cain, G.R., Gerber, E.P. & DallaSantana, K.J. (2022) Machine learning gravity wave parameterization generalizes to capture the qbo and response to increased CO<sub>2</sub>. *Geophysical Research Letters*, 49, e2022GL098174.
- Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29, 1181–1232.
- Garfinkel, C.I., Gerber, E.P., Shamir, O., Rao, J., Jucker, M., White, I. et al. (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, e2021MS002568.
- 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, 5611–5633.
- Geller, M.A., Zhou, T., Shindell, D., Ruedy, R., Aleinov, I., Nazarenko, L. et al. (2016) Modeling the QBO—improvements resulting from higher-model vertical resolution. *Journal of Advances in Modeling Earth Systems*, 8, 1092–1105.
- Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep learning*. Cambridge, MA: MIT Press. <http://www.deeplearningbook.org>
- Hardiman, S.C., Scaife, A.A., van Niekerk, A., Prudden, R., Owen, A., Adams, S.V. et al. (2023) Machine learning for non-orographic gravity waves in a climate model. *Artificial Intelligence for the Earth Systems*, 2(4), e220081.
- Holt, L.A., Lott, F., Garcia, R.R., Kiladis, G.N., Cheng, Y.-M., Anstey, J.A. et al. (2022) An evaluation of tropical waves and wave forcing of the qbo in the qboi models. *Quarterly Journal of the Royal Meteorological Society*, 148, 1541–1567.
- Holton, J.R. & Lindzen, R.S. (1972) An updated theory for the quasi-biennial cycle of the tropical stratosphere. *Journal of Atmospheric Sciences*, 29, 1076–1080.
- Jucker, M. & Gerber, E. (2017) Untangling the annual cycle of the tropical tropopause layer with an idealized moist model. *Journal of Climate*, 30, 7339–7358.
- Kingma, D.P. & Ba, J. (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- Kingma, D.P. & Welling, M. (2013) Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
- Lindzen, R.S. (1971) Equatorial planetary waves in shear. Part i. *Journal of Atmospheric Sciences*, 28, 609–622.
- Lindzen, R.S. & Holton, J.R. (1968) A theory of the quasi-biennial oscillation. *Journal of Atmospheric Sciences*, 25, 1095–1107.
- Mansfield, L. & 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.
- Match, A. & Fueglistaler, S. (2020) Mean-flow damping forms the buffer zone of the quasi-biennial oscillation: 1d theory. *Journal of the Atmospheric Sciences*, 77, 1955–1967.
- Match, A. & Fueglistaler, S. (2021) Anomalous dynamics of qbo disruptions explained by 1d theory with external triggering. *Journal of the Atmospheric Sciences*, 78, 373–383.
- Pahlavan, H.A., Hassanzadeh, P. & Alexander, M.J. (2024) Explainable offline-online training of neural networks for parameterizations: a 1d gravity wave-qbo testbed in the small-data regime. *Geophysical Research Letters*, 51, e2023GL106324.
- Plumb, R. (1977) The interaction of two internal waves with the mean flow: implications for the theory of the quasi-biennial oscillation. *Journal of Atmospheric Sciences*, 34, 1847–1858.
- Reed, R.J., Campbell, W.J., Rasmussen, L.A. & Rogers, D.G. (1961) Evidence of a downward-propagating, annual wind reversal in the equatorial stratosphere. *Journal of Geophysical Research*, 66, 813–818.
- Richter, J.H., Anstey, J.A., Butchart, N., Kawatani, Y., Meehl, G.A., Osprey, S. et al. (2020) Progress in simulating the quasi-biennial oscillation in CMIP models. *Journal of Geophysical Research: Atmospheres*, 125, e2019JD032362.
- Richter, J.H., Solomon, A. & Bacmeister, J.T. (2014) On the simulation of the quasi-biennial oscillation in the community atmosphere model, version 5. *Journal of Geophysical Research: Atmospheres*, 119, 3045–3062.
- Smola, A.J. & Schölkopf, B. (2004) A tutorial on support vector regression. *Statistics and Computing*, 14, 199–222.
- Sun, Y.Q., Hassanzadeh, P., Alexander, M.J. & Kruse, C.G. (2023a) Quantifying 3d gravity wave drag in a library of tropical convection-permitting simulations for data-driven parameterizations. *Journal of Advances in Modeling Earth Systems*, 15, e2022MS003585.
- Sun, Y.Q., Pahlavan, H.A., Chattopadhyay, A., Hassanzadeh, P., Lubis, S.W., Alexander, M.J. et al. (2023b) Data imbalance, uncertainty quantification, and generalization via transfer learning in data-driven parameterizations: Lessons from the emulation of gravity wave momentum transport in waccm.
- Warner, C. & McIntyre, M. (2001) An ultrasimple spectral parameterization for nonorographic gravity waves. *Journal of the Atmospheric Sciences*, 58, 1837–1857.
- Warner, C.D. & McIntyre, M.E. (1999) Toward an ultra-simple spectral gravity wave parameterization for general circulation models. *Earth, Planets and Space*, 51, 475–484.
- Yang, L.M. & Gerber, E.P. (2024) Overcoming set imbalance in data driven parameterization: A case study of gravity wave momentum transport.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Shamir, O., Connelly, D.S., Hardiman, S.C., Shao, Z., Yang, L.M. & Gerber, E.P. (2024) The graft-versus-host problem for data-driven gravity-wave parameterizations in a one-dimensional quasibiennial oscillation model. *Quarterly Journal of the Royal Meteorological Society*, 1–18. Available from: <https://doi.org/10.1002/qj.4707>

## APPENDIX A. NUMERICAL SCHEME

Equation 1 is discretized using a semi-implicit scheme, combining an implicit scheme for the advection/diffusion terms (centered in both time and space) and a (explicit) leapfrog scheme for the source term. The discretized model equation on the numerical grid  $(i\Delta t, j\Delta z)$ , for fixed  $\Delta t, \Delta z, i = 1, 2, 3, \dots, N$ , and  $j = 1, 2, 3, \dots, N$ , is

$$\begin{aligned} & [\mathbf{I} + \Delta t(\text{diag}(\mathbf{w}^{\tau+1})\mathbf{D1} - \kappa\mathbf{D2})]\mathbf{u}^{\tau+1} \\ & = [\mathbf{I} - \Delta t(\text{diag}(\mathbf{w}^{\tau-1})\mathbf{D1} - \kappa\mathbf{D2})]\mathbf{u}^{\tau-1} - 2\Delta t\mathbf{S}^{\tau}, \quad (\text{A1}) \end{aligned}$$

where the superscripts denote the time step,  $\mathbf{u}^{\tau} = (u_1^{\tau}, \dots, u_N^{\tau})^T$  is the vector of discretized unknowns,  $\mathbf{S}^{\tau} = (S_1^{\tau}, \dots, S_N^{\tau})^T$  is the vector of discretized source terms,

$\mathbf{w}^{\tau} = (w_1^{\tau}, \dots, w_N^{\tau})^T$  is the vector of discretized vertical wind,  $\mathbf{I}$  is the  $N \times N$  identity, and  $\mathbf{D1}$  and  $\mathbf{D2}$  are the differentiation matrices for the first- and second-order derivatives, defined here as

$$D1_{ij} = \begin{cases} 0 & \text{for } i = 1, 1 \leq j \leq N, \\ (\delta_{i,j+1} - \delta_{i,j-1})/2\Delta z & \text{for } 1 < i < N, 1 \leq j \leq N, \\ 0 & \text{for } i = N, 1 \leq j \leq N, \end{cases} \quad (\text{A2})$$

and

$$D2_{ij} = \begin{cases} 0 & \text{for } i = 1, 1 \leq j \leq N, \\ (\delta_{i,j+1} - 2\delta_{i,j} + \delta_{i,j-1})/\Delta z^2 & \text{for } 1 < i < N, 1 \leq j \leq N, \\ 0 & \text{for } i = N, 1 \leq j \leq N, \end{cases} \quad (\text{A3})$$

where  $\delta_{i,j}$  is the Kronecker delta.

Note, zeroing-out the first and last rows of  $\mathbf{D1}$  and  $\mathbf{D2}$  implies that the tendency at the boundaries is determined by the source term, that is,

$$u_{\{1,N\}}^{\tau+1} = u_{\{1,N\}}^{\tau-1} - 2\Delta t S_{\{1,N\}}^{\tau}, \quad (\text{A4})$$

where the subscripts denote the vertical levels. Numerically, we compute the source term by applying  $\mathbf{D1}$  to the flux (the sum in Equation 2), which zeros-out the source term at the boundaries. Thus, if the initial wind and wind tendency at the boundaries are zero,  $u_{\{1,N\}}^{\tau}$  remain zero for all  $\tau = 0, 1, 2, \dots$ .

Model parameters and various neural network parameters are given in Tables B1–B4, while Table B5 lists boosted forest parameters and Table B6 SVR model parameters.

## APPENDIX B. PARAMETER TABLES

TABLE B1 Physical model parameters.

| Domain                                          |                                                                                                                                                                                                                                                                        |                                    |
|-------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------|
| Final time ( $t_f$ )                            | 108 years <sup>a</sup>                                                                                                                                                                                                                                                 |                                    |
| Temporal spacing ( $\Delta t$ )                 | 86,400 s                                                                                                                                                                                                                                                               |                                    |
| Bottom boundary ( $z_1$ )                       | $17 \times 10^3$ m                                                                                                                                                                                                                                                     |                                    |
| Top boundary ( $z_N$ )                          | $35 \times 10^3$ m                                                                                                                                                                                                                                                     |                                    |
| Vertical spacing ( $\Delta z$ )                 | 250 m                                                                                                                                                                                                                                                                  |                                    |
| Background state                                |                                                                                                                                                                                                                                                                        |                                    |
| Density profile <sup>b,c</sup> ( $\rho$ )       | $(P_0/R_d T_0) \exp[-(g/R_d T_0)z]$                                                                                                                                                                                                                                    |                                    |
| Reference pressure ( $P_0$ )                    | 101,325 Pa                                                                                                                                                                                                                                                             |                                    |
| Gas constant for dry air ( $R_d$ )              | $287.04 \text{ J} \cdot \text{kg}^{-1} \cdot \text{K}^{-1}$                                                                                                                                                                                                            |                                    |
| Reference temperature ( $T_0$ )                 | 204 K                                                                                                                                                                                                                                                                  |                                    |
| Gravitational acceleration ( $g$ )              | $9.8 \text{ m} \cdot \text{s}^{-2}$                                                                                                                                                                                                                                    |                                    |
| Brunt–Väisälä frequency <sup>d</sup> ( $N$ )    | $2.16 \times 10^{-2} \text{ s}^{-1}$                                                                                                                                                                                                                                   |                                    |
| Model                                           |                                                                                                                                                                                                                                                                        |                                    |
| Diffusion coefficient <sup>b</sup> ( $\kappa$ ) | $0.3 \text{ m}^2 \cdot \text{s}^{-1}$                                                                                                                                                                                                                                  |                                    |
| Vertical wind <sup>e,f</sup> ( $w$ )            | $3 \times 10^{-4} \text{ m} \cdot \text{s}^{-1}$                                                                                                                                                                                                                       |                                    |
| GW forcing                                      |                                                                                                                                                                                                                                                                        |                                    |
| Number of waves ( $N_{\text{waves}}$ )          | 20                                                                                                                                                                                                                                                                     |                                    |
| Zonal wavenumbers ( $k_n$ )                     | $2 \times 2\pi/(4 \times 10^7) \text{ m}^{-1}$ for $n = 1, \dots, 20$                                                                                                                                                                                                  |                                    |
| Phase speeds ( $c_n$ )                          | $\begin{cases} -100 \text{ m} \cdot \text{s}^{-1} + 10(n-1) \text{ m} \cdot \text{s}^{-1} & \text{for } n = 1, \dots, 10 \\ 10(n-10) \text{ m} \cdot \text{s}^{-1} & \text{for } n = 11, \dots, 20 \end{cases}$                                                        |                                    |
| Wave dissipation <sup>b</sup> ( $\alpha$ )      | $\begin{cases} \frac{1}{21} \text{ days}^{-1} + \left(\frac{z-17}{13}\right) \frac{2}{21} \text{ days}^{-1} & \text{for } 17 \text{ km} \leq z \leq 30 \text{ km} \\ \frac{3}{21} \text{ days}^{-1} & \text{for } 30 \text{ km} \leq z \leq 35 \text{ km} \end{cases}$ |                                    |
| GW source distribution                          |                                                                                                                                                                                                                                                                        | “Perturbed/biased”                 |
| Mean total source flux ( $\overline{F}_{S0}$ )  | 3.5 mPa                                                                                                                                                                                                                                                                | 4.5 mPa                            |
| STD total source flux                           | 0.3 mPa                                                                                                                                                                                                                                                                | 0.3 mPa                            |
| Mean spectral width ( $\overline{c}_w$ )        | $32 \text{ m} \cdot \text{s}^{-1}$                                                                                                                                                                                                                                     | $40 \text{ m} \cdot \text{s}^{-1}$ |
| STD spectral width ( $\text{STD}c_w$ )          | $16 \text{ m} \cdot \text{s}^{-1}$                                                                                                                                                                                                                                     | $16 \text{ m} \cdot \text{s}^{-1}$ |
| Correlation                                     | 0.75                                                                                                                                                                                                                                                                   | 0.75                               |

<sup>a</sup> Using a 30-day month calendar (i.e., 1 year = 360 days).<sup>b</sup> Following Holton and Lindzen (1972).<sup>c</sup> Assuming an isothermal atmosphere.<sup>d</sup> Corresponding to the chosen values of  $P_0$  and  $T_0$  for an isothermal atmosphere (i.e., not an additional free parameter).<sup>e</sup> Corresponding to the Brewer–Dobson circulation in the Tropics at 70 hPa (Butchart, 2014).<sup>f</sup> In Section 4.2 we also examine an annually varying vertical wind with noise as detailed in Equation 5.

**TABLE B2** Fully connected feed-forward neural network parameters.

| Architecture         | [Linear(# inputs, # outputs), activation]                                                |
|----------------------|------------------------------------------------------------------------------------------|
| Input layer          | [Linear(nlev-2 + 2, nlev), ReLU] <sup>a,b</sup>                                          |
| (Hidden) layers 2–9  | [Linear(nlev, nlev), ReLU]                                                               |
| Output layer         | [Linear(nlev, nlev-2), None]                                                             |
| Trainable parameters | 53,872                                                                                   |
| Optimization         |                                                                                          |
| Loss function        | Relative MSE = $\frac{\sum (\text{prediction} - \text{target})^2}{\sum \text{target}^2}$ |
| Optimizer            | Adam <sup>c,d</sup>                                                                      |
| Learning rate        | 10 <sup>-3</sup>                                                                         |

<sup>a</sup>The linear layers are written in PyTorch syntax.<sup>b</sup>Similar results were obtained using a tanh activation instead.<sup>c</sup>Kingma and Ba (2014).<sup>d</sup>Similar results were obtained using stochastic gradient descent instead.**TABLE B3** Dilated convolutional neural network parameters.

| Architecture         | [Conv1D (in channels, out channels, kernel size, stride, padding, dilation), activation] <sup>a</sup> |
|----------------------|-------------------------------------------------------------------------------------------------------|
| Layer 1              | [Conv1D(1, 20, 5, 1, 2, 0), ReLU]                                                                     |
| Layer 2              | [Conv1D(20, 40, 5, 1, 6, 3), ReLU]                                                                    |
| Layer 3              | [Conv1D(40, 60, 5, 1, 10, 5), ReLU]                                                                   |
| Layer 4              | [Conv1D(60, 80, 5, 1, 22, 11), ReLU]                                                                  |
| Layer 5              | [Conv1D(80, 60, 5, 1, 10, 5), ReLU]                                                                   |
| Layer 6              | [Conv1D(60, 40, 5, 1, 6, 3), ReLU]                                                                    |
| Layer 7              | [Conv1D(40, 20, 5, 1, 2, 0), None]                                                                    |
| Layer 8              | [Conv1D(20, 1, 5, 1, 2, 0), None]                                                                     |
| Trainable parameters | 80,521                                                                                                |

Note: The optimization parameters are the same as for the NN in Table B2.

<sup>a</sup>The 1D convolution layers are written in PyTorch syntax.**TABLE B4** Encoder–decoder network parameters.

| Architecture         | Encoder–dense–decoder                       |
|----------------------|---------------------------------------------|
| Trainable parameters | 13,261                                      |
| Activation function  | Exponential linear unit function            |
| Optimization         |                                             |
| Loss function        | Mean-squared error                          |
| Optimizer            | Adam <sup>a</sup>                           |
| Learning rate        | Start at 1e–3, and reduce on plateau by 0.5 |

<sup>a</sup>Kingma and Ba (2014).**TABLE B5** Boosted forest parameters.

| Architecture                              |                      |
|-------------------------------------------|----------------------|
| Tree maximum depth <sup>a</sup>           | 15                   |
| Number of trees                           | 72                   |
| Fraction of samples per tree <sup>a</sup> | 0.5                  |
| Fraction of features per node             | 0.5                  |
| Trainable parameters                      | 1,111,866            |
| Optimization                              |                      |
| Impurity                                  | Gini                 |
| Learning rate <sup>a</sup>                | 0.05                 |
| Validation set <sup>b</sup>               | 20% of training data |

<sup>a</sup>Selected with threefold cross-validation.<sup>b</sup>Used to determine when to stop adding trees to the ensemble.**TABLE B6** Support-vector regression model parameters.

| Architecture             |                                                                      |
|--------------------------|----------------------------------------------------------------------|
| # support vectors        | 20% (train/test split) * 34,560<br>(size of dataset) = 6912          |
| Kernel: RBF kernel       | $K(x, y) = \exp(-\gamma \ x - y\ _2^2)$ ,<br>where $\gamma = 0.05^a$ |
| Trainable parameters     | 490,823                                                              |
| Optimization             |                                                                      |
| Loss function            | Hinge loss with $\epsilon = 1e - 4$                                  |
| Regularization (penalty) | $C = 16$                                                             |

<sup>a</sup>The notation follows Sklearn syntax.