Check for updates

AGU ADVANCING EARTH AND SPACE SCIENCES

Geophysical Research Letters®

RESEARCH LETTER
10.1029/2023GL106324

Key Points:

- 1D model of quasi-biennial oscillation (QBO) and gravity waves is used as a testbed for training neural network (NN)-based parameterizations
- Offline training NNs in small-data regimes yields unstable QBOs that are rectified by online re-training using only time-averaged statistics
- Fourier analysis of NNs reveals that they learn specific filters that are consistent with the dynamics of wave propagation and dissipation

Supporting Information:

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

Correspondence to:

H. A. Pahlavan,
hamid.a.pahlavan@gmail.com

Citation:

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. https://doi.org/10.1029/2023GL106324

Received 18 SEP 2023
Accepted 4 JAN 2024

© 2024 The Authors.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Explainable Offline-Online Training of Neural Networks for Parameterizations: A 1D Gravity Wave-QBO Testbed in the Small-Data Regime

Hamid A. Pahlavan[1], Pedram Hassanzadeh[1], and M. Joan Alexander[2]

[1]Rice University, Houston, TX, USA, [2]NorthWest Research Associates, Boulder, CO, USA

Abstract There are different strategies for training neural networks (NNs) as subgrid-scale parameterizations. Here, we use a 1D model of the quasi-biennial oscillation (QBO) and gravity wave (GW) parameterizations as testbeds. A 12-layer convolutional NN that predicts GW forcings for given wind profiles, when trained offline in a *big-data* regime (100-year), produces realistic QBOs once coupled to the 1D model. In contrast, offline training of this NN in a *small-data* regime (18-month) yields unrealistic QBOs. However, online re-training of just two layers of this NN using ensemble Kalman inversion and only time-averaged QBO statistics leads to parameterizations that yield realistic QBOs. Fourier analysis of these three NNs' kernels suggests why/how re-training works and reveals that these NNs primarily learn low-pass, high-pass, and a combination of band-pass filters, potentially related to the local and non-local dynamics in GW propagation and dissipation. These findings/strategies generally apply to data-driven parameterizations of other climate processes.

Plain Language Summary Due to computational limits, climate models estimate (i.e., parameterize) small-scale physical processes, such as atmospheric gravity waves (GWs), since they occur on scales smaller than the models' grid size. Recently, machine learning techniques, especially neural networks (NNs), have emerged as promising tools for learning these parameterizations from data. Offline and online learning are among the main strategies for training these NN-based parameterizations. Offline learning, while straightforward, requires extensive, high-quality data from small-scale processes, which are scarce. Alternatively, online learning only needs time or space-averaged data based on large-scale processes, which are more accessible. However, online learning can be computationally expensive. Here, we explore various learning strategies using an NN-based GW parameterization, within a simple model of the quasi-biennial oscillation (QBO), an important quasi-periodic wind pattern in the tropics. When supplied with a large 100-year data set, the offline-trained NN accurately replicates wind behaviors once coupled to the QBO model. Yet, when limited to an 18-month training data set (which is more realistic), its performance degrades. Interestingly, by online re-training specific parts of this NN using only time-averaged QBO statistics, its accuracy is restored. We term this approach an "offline-online" learning strategy. Our findings also benefit parameterization efforts for other climate processes.

1. Introduction

Due to the current resolution limitations of general circulation models (GCMs), many crucial subgrid-scale (SGS) physical processes remain unresolved and are instead represented through parameterization. The conventional physics-based parameterizations are based on simplified theories, which introduce significant uncertainties in climate modeling (e.g., Richter et al., 2022). Recently, machine learning (ML) techniques, particularly deep neural networks (NNs), have emerged as novel tools for developing parameterizations. Different strategies exist for training these ML-based parameterizations. In the common offline learning approach, the learnable parameters of NNs (i.e., weights and biases) are trained using stochastic gradient descent through backpropagation to find a nonlinear mapping between the resolved and SGS processes. However, this approach demands an extensive training data set that includes the true SGS terms obtained from high-fidelity sources such as high-resolution observations and/or simulations, which are typically scarce. To add to the challenge, the true SGS terms must be properly extracted from these sources, which can be sensitive to separation methods, as well as the filtering and coarse-graining operations (Grooms et al., 2021; Sun, Hassanzadeh, et al., 2023).

19448007, 2024, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL106324 by University Of Chicago Library, Wiley Online Library on [03/12/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Alternatively, learning SGS parameterization can be approached as an "online task," which allows learning from partial observations or statistics. However, learning from statistics requires performing long-term simulations during training, making the learning process challenging (Schneider et al., 2023). For online learning, various methods such as reinforcement learning (Mojgani et al., 2023; Novati et al., 2021), differentiable programming (Frezat et al., 2022; Gelbrecht et al., 2023), and ensemble Kalman inversion (EKI) (Iglesias et al., 2013; Lopez-Gomez et al., 2022) can be used. Here, we use EKI, a gradient-free algorithm (Kovachki & Stuart, 2019), which is ideal for GCMs where computing derivatives can be challenging.

Atmospheric gravity waves (GWs) are among the physical processes that are not fully resolved in the current GCMs as they span scales of $O(1)$ to $O(1{,}000)$ km. GWs play a crucial role in the transport of energy and momentum through the atmosphere (Fritts & Alexander, 2003). With decades of developments, GW parameterization (GWP) is now a critical component of GCMs for reproducing realistic atmospheric circulation mean and variability (Kruse et al., 2023). For instance, GCMs require skillful GWPs to naturally produce the quasi-biennial oscillation (QBO) (Richter et al., 2020), which is characterized by the downward propagation of successive westerly and easterly winds with an average period of ∼28 months (Baldwin et al., 2001). The QBO is the primary mode of interannual variability in the tropical stratosphere with links to subseasonal-to-seasonal forecast skills (Anstey et al., 2022). GWs are believed to contribute significantly to the forcing of the QBO (Ern et al., 2014; Kawatani et al., 2010; Kim & Chun, 2015; Pahlavan et al., 2021; Richter et al., 2014).

Recently, ML has been used to emulate or calibrate existing physics-based GWP schemes (Chantry et al., 2021; Espinosa et al., 2022; Hardiman et al., 2023; Mansfield & Sheshadri, 2022; Sun, Pahlavan, et al., 2023), or to estimate the GWs variability or structure from high-resolution reanalysis data (Amiramjadi et al., 2023; Matsuoka et al., 2020). In contrast to the emulation of an existing scheme, the task of developing new GWP schemes from high-resolution data sets and coupling such new schemes to GCMs is much more ambitious and challenging, and the data sets needed for such work have just started to emerge (Sun, Hassanzadeh, et al., 2023).

In this study, we use a conceptual 1D model of the QBO and GWP as testbeds to explore various learning strategies and challenges arising from the scarcity of high-resolution training data to inform future studies with GCMs. We show that a 12-layer convolutional neural network (CNN)-based GWP, when trained offline in a *big-data* regime spanning 100 years, generates accurate QBOs once coupled to the 1D model. However, offline training the CNN in a *small-data* regime covering only 18 consecutive months yields unrealistic QBOs. This *small-data* regime represents the common situations with limited availability of high-quality SGS data for training. Remarkably, by selectively online re-training only two layers of this CNN with EKI and using only time-averaged QBO statistics, we obtain GWPs that reproduce realistic QBOs. We refer to this approach as "offline-online" learning. We also use the Spectral Analysis of Regression Kernels and Activations (SpArK) framework (introduced in Subel et al. (2023)) to provide physically interpretable insights into what these three CNNs learn. While this study primarily addresses GWP, the findings are expected to be applicable broadly to data-driven parameterizations of other climate processes.

2. Methods

2.1. 1D-QBO Model

The 1D-QBO model represents a 1D model of the tropical stratosphere (Holton & Lindzen, 1972; Plumb, 1977). With a source of parameterized waves at its lower boundary, the model is a minimal configuration that represents the wave-mean flow interaction. In this study, the 1D-QBO is structured as a forced advection-diffusion model:

$$\frac{\partial u}{\partial t} + \omega\frac{\partial u}{\partial z} - \kappa\frac{\partial^2 u}{\partial z^2} = G(u) + \eta(t) \tag{1}$$

with zonal wind $u(t, z)$ as a function of time t and height z, upwelling $\omega = 0$, diffusivity $\kappa = 0.3\ \mathrm{m^2\ s^{-1}}$, and GW drag G. By setting $\omega = 0$, we exclude vertical advection for simplicity. η is a stochastic forcing, which represents the missing physics within the 1D model, and its importance is further detailed in Supporting Information S1.

Following Plumb (1977), the model is driven by two vertically propagating GWs with zonal phase speeds $(c_1, c_2) = (-30, +30)\ \mathrm{m\ s^{-1}}$. As these waves propagate upward, they dissipate and force the mean flow toward their phase speed. The vertical group velocity of these waves depends non-linearly on the difference between their phase speed and the mean flow, becoming smaller when the zonal wind is close to its phase speed, and reaching

19448007, 2024, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL106324 by University Of Chicago Library, Wiley Online Library on [03/12/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

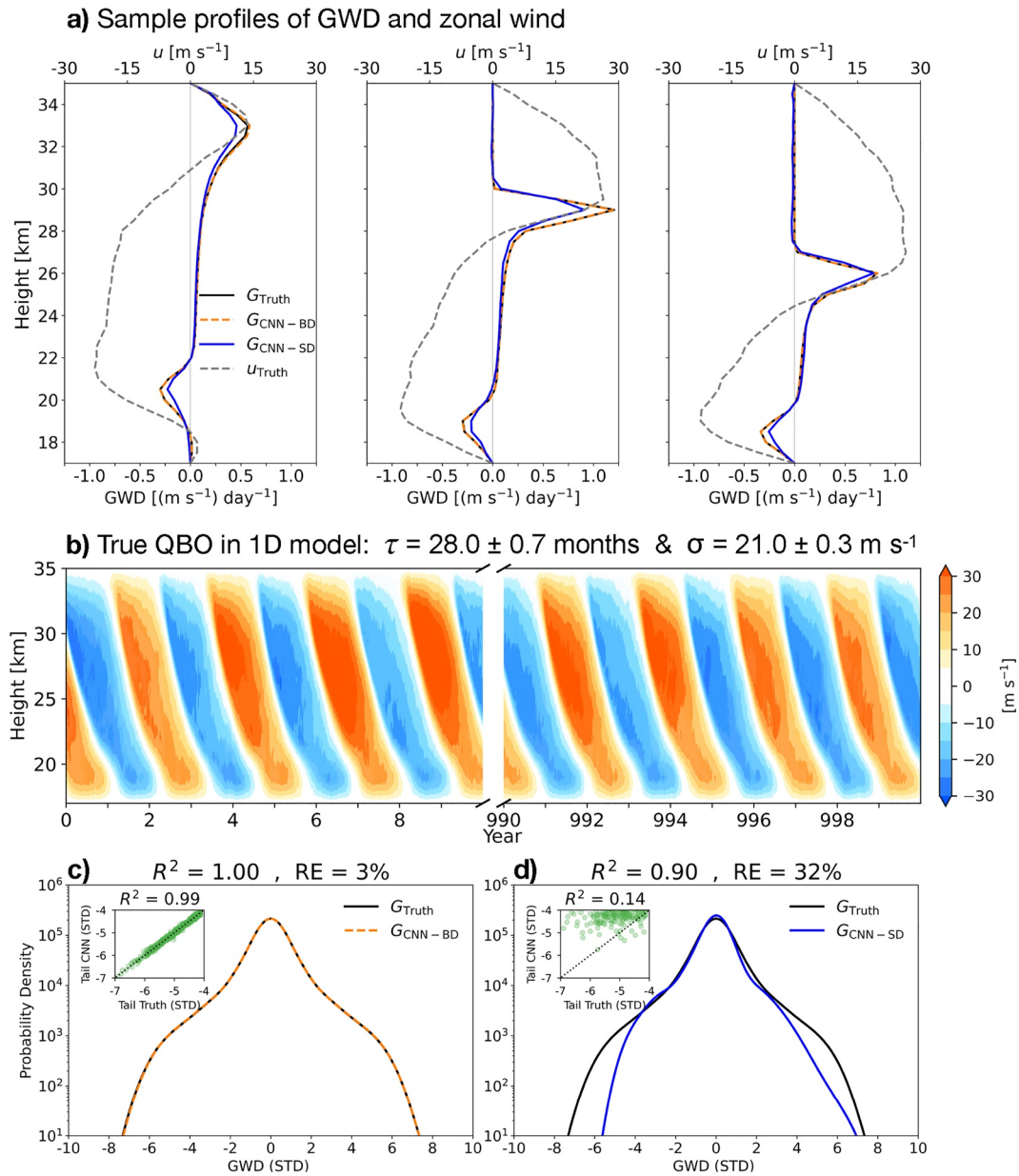

Figure 1. (a) Sample profiles of the true GWD (G) and zonal wind (u), spaced 150 days apart. Also shown is the a priori (offline) predicted GWD from convolutional neural networks (CNNs) that predict G as a function of u, and are trained either in the *big-data* regime (CNN-BD) or in the *small-data* regime (CNN-SD). (b) Time-height section of zonal wind of the true quasi-biennial oscillation in the 1D model (see Methods). Note the time axis break, such as only the first and last 10 years of the simulation are shown. The a priori performances of CNN-BD, and CNN-SD are shown in panels (c, d). (c) Probability density function (PDF) of the true and predicted G by CNN-BD. (d) As in (c), but for G predicted by CNN-SD. In panels (c, d), the insets show scatter plots representing the tails of the PDFs, identified as the top 1% of magnitudes. The x-axis is normalized by the standard deviation. In these panels, R^2 is the squared of the Pearson correlation coefficients between true and predicted GWD. Relative error is defined as $|G - G_{CNN}|/|G|$, where $|.|$ denotes the average of absolute values over all model levels.

zero at a critical level where $u = c$. With constant dissipation, slower ascent results in more dissipation per unit height.

Figure 1a shows sample profiles of GW drag (GWD) and zonal wind, illustrating the downward propagation of a QBO westerly phase from the upper boundary. Note the concentration of GWD over a shallow vertical extent

19448007, 2024, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL106324 by University Of Chicago Library, Wiley Online Library on [03/12/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where the eastward wave ($c = 30$ m s^{-1}) reaches its critical level. At this layer, the wave breaks and transfers its momentum to the mean flow, causing the GWD at higher levels to become zero.

More details on the model configuration are provided in Supporting Information S1. This setup yields an oscillation with a period (τ) of 28.0 ± 0.7 months, and an amplitude (σ) of 21 ± 0.3 m s^{-1} at the 25 km altitude (Figure 1b). The standard deviations of the period and amplitude are based on ∼430 QBO cycles in a 1000-year simulation. This simulation is our "truth" and is used to evaluate the performance of CNN-based GWP. Using the same setup, we produce an independent 100-year data set specifically for training and validation purposes.

2.2. CNN-Based GWP

We explore various learning strategies by emulating the GWD, $G(u)$ in Equation 1, using a CNN, denoted as $G_{\text{CNN}}(u, \theta)$, with θ being the learnable parameters of the CNN. This CNN consists of 12 sequential 1D convolutional layers. Each hidden layer has 15 channels, each with 15 kernels with a size of 5, resulting in ∼11,600 learnable parameters. The activation function is hyperbolic tangent (tanh). Training a CNN means learning the parameters θ, either offline or online, as detailed below.

2.3. Offline Learning

Offline learning seeks to find the optimal θ values by matching G_{CNN} and the true G profiles for a given profile of $u(t, z)$, which is achieved by minimizing the following loss:

$$\mathcal{L}_{offline} = \frac{1}{n}\sum_{i=1}^{n} \|G(u_i) - G_{\text{CNN}}(u_i, \theta)\|_2^2 \tag{2}$$

where n is the number of training samples and $\|.\|_2$ is the L_2 norm. We train the CNN offline under two distinct data regimes: (a) *big-data*, denoted as CNN-BD, which uses 100 years of sequential data, representing an ideal scenario with ample data, and (b) *small-data* (CNN-SD), which includes only 18 consecutive months of data, representing a more realistic scenario given the cost associated with, for instance, GW-resolving global simulations. Further insights on using a more strategically sampled 18 months, instead of a continuous span, are detailed in the Discussion section.

2.4. Online Learning

In online learning, we learn the parameters θ by using time-averaged statistics of the QBO, and minimizing the following loss:

$$\mathcal{L}_{online} = \left\| \mathcal{H}\Big(\Psi(u, G(u))\Big) - \mathcal{H}\Big(\Psi\big(u, G_{\text{CNN}}(u, \theta)\big)\Big) \right\|_{\Gamma}^2 \tag{3}$$

where $\|.\|_{\Gamma}$ is the Mahalanobis norm, Γ denotes the variance of the system's internal noise, and Ψ is the forward model, the numerical solver of the 1D-QBO model in this case, $\mathcal{H}$ is the observational map, which encapsulates all averaging and post-processing operations necessary to derive the desired statistics from an observable field, zonal wind u in this case. See Lopez-Gomez et al. (2022) for more details.

Various optimization methods can be used to minimize $\mathcal{L}_{online}$. As highlighted earlier, we employ EKI, which has been increasingly used for parameter estimation in recent climate studies (Cleary et al., 2021; Dunbar et al., 2022; Lopez-Gomez et al., 2022). Briefly, as an iterative method to solve inverse problems, EKI starts with an ensemble of model parameters θ drawn from a prior distribution. As the iterations proceed, these parameters are updated based on the discrepancies between statistics simulated with the model and the true statistics, usually obtained from observations, reanalysis, or high-resolution simulations. Once the algorithm converges, the optimal parameter values are then the ensemble mean of the last iteration.

The 1D model is very sensitive to the changes in GWD profiles, and for some sets of parameters, simulations become physically or numerically unstable, preventing the EKI algorithm from converging. We address these model failures following the methodology proposed in Lopez-Gomez et al. (2022).

For the online training of the CNN using EKI, denoted as CNN-EKI, our setup includes 200 ensemble members, 10 iterations, and 85 statistics, derived from a 10-year span of zonal wind u from our true QBO simulation, as defined earlier. We run each of the ensemble members for 15 years, then calculate the desired statistics from the last 10 years of those runs. The EKI's efficacy can be notably impacted by these choices, with poor selections

19448007, 2024, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL106324 by University Of Chicago Library, Wiley Online Library on [03/12/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

potentially causing instabilities, underscoring the challenges associated with online learning. Section 3.2 offers further details on the statistics and prior distributions used in this study.

3. Results

3.1. Offline Learning in the Big and Small-Data Regimes

We begin by evaluating the a priori (offline) performances of the CNN-BD and CNN-SD. Figure 1a shows sample GWD profiles predicted by these two CNNs, compared with the true GWD profiles. Both CNNs capture the general structure of the true GWD. However, $G_{\mathrm{CNN-BD}}$ aligns perfectly with the true GWD profiles, while $G_{\mathrm{CNN-SD}}$ exhibits some discrepancies, especially in representing the peaks.

Figure 1c compares the probability density function (PDF) of $G_{\mathrm{CNN-BD}}$ with that of the truth. The outstanding a priori performance is evident from their overlap, further highlighted by a mere 3% relative error (RE). In contrast, the a priori performance of the CNN-SD, shown in Figure 1d, clearly diminishes, evidenced by the increase in the RE to 32%. Furthermore, a pronounced decline in R^2 can be observed at the tails, decreasing from 0.99 to 0.14. These findings are in line with expectations. When abundant training data are provided, deep NNs such as our CNN can be effectively trained offline and demonstrate accurate a priori performance. Conversely, with limited data, the accuracy decreases, especially when predicting rare (but large) events at the tails. It is noteworthy that in the context of the *small-data* regime, our extensive experiments with smaller CNNs and other NN architectures with fewer parameters did not yield successful results (not shown).

In the a posteriori (online) evaluations, we replace $G(u)$ with $G_{\mathrm{CNN}}(u, \theta)$ and run the model for 1,000 years. Figure 2a shows the a posteriori performance of the CNN-BD, demonstrating a QBO whose structure, period, and amplitude closely match that of the true QBO. In contrast, Figure 2b reveals that the QBO simulated with the CNN-SD becomes unrealistic after the initial QBO cycles, with intensified westerly phases and diminished easterly phases. The PDFs of GWD and zonal wind are presented in Figures 2c and 2d. The indistinguishable overlap between the PDFs of truth and the CNN-BD underscores its outstanding a posteriori performance. In contrast, the PDFs for the CNN-SD demonstrate a significant deviation from those of the truth.

This specific unrealistic behavior of the CNN-SD is a result of the specific 18-month segment used for training. When we choose a different 18-month segment, the QBO exhibits other unrealistic deviations. The key takeaway, however, is that an 18-month sequential data set is not adequate to achieve accurate a posteriori QBO in the 1D model. Notably, Espinosa et al. (2022) achieved stable a posteriori QBO by training their ML-based emulator of the physics-based GWP using only 12 months of data, which were dominated by the westerly phase of the QBO. However, their use of global data suggests that the emulator might have learned from regions with easterly winds outside the tropical stratosphere. This is consistent with the findings of Chantry et al. (2021) and Hardiman et al. (2023), who similarly achieved stable a posteriori QBO using one and 2.5 years of global data, respectively. It should also be noted that in the more complex 3D climate models, the QBO is further constrained by other dynamics, such as meridional circulations and tropical upwelling, compared to its representation in the 1D model (Hardiman et al., 2023).

Our results so far indicate that the CNN-BD shows outstanding a priori and a posteriori performances. In contrast, the CNN-SD yields unstable a posteriori QBO and, given the appropriate metrics (e.g., PDF tails), a poor a priori performance too. This finding prompts one of the central questions of this study: Is it possible to use online learning to improve the CNN-SD and rectify this unrealistic QBO behavior? Note that the terms *small-* and *big-data* regimes in our context refer to the number of G snapshots available for offline learning. For online learning in the *small-data* regime, we assume that we have access to the time-averaged statistics of the true QBO but have limited snapshots of G.

In the context of online learning, one approach might be to train a CNN from scratch (i.e., from random initialization of θ) using methods like EKI, and the QBO statistics as the targets. However, as we discuss below, the priors (initialization of θ) are critical for the convergence of online learning methods, and poor priors, such as random ones, can lead to failed learning. Another approach, that we pursue here, is to use the parameters of the CNN-SD as priors. We will refer to this as the "offline-online" learning approach, as further elaborated below.

3.2. Offline-Online Learning in the Small-Data Regime

A critical question in using statistics for parameter estimation, as in the EKI method, is determining whether the targeted statistics are adequate to constrain the learnable parameters, which could be high-dimensional. Here, we

19448007, 2024, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL106324 by University Of Chicago Library, Wiley Online Library on [03/12/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

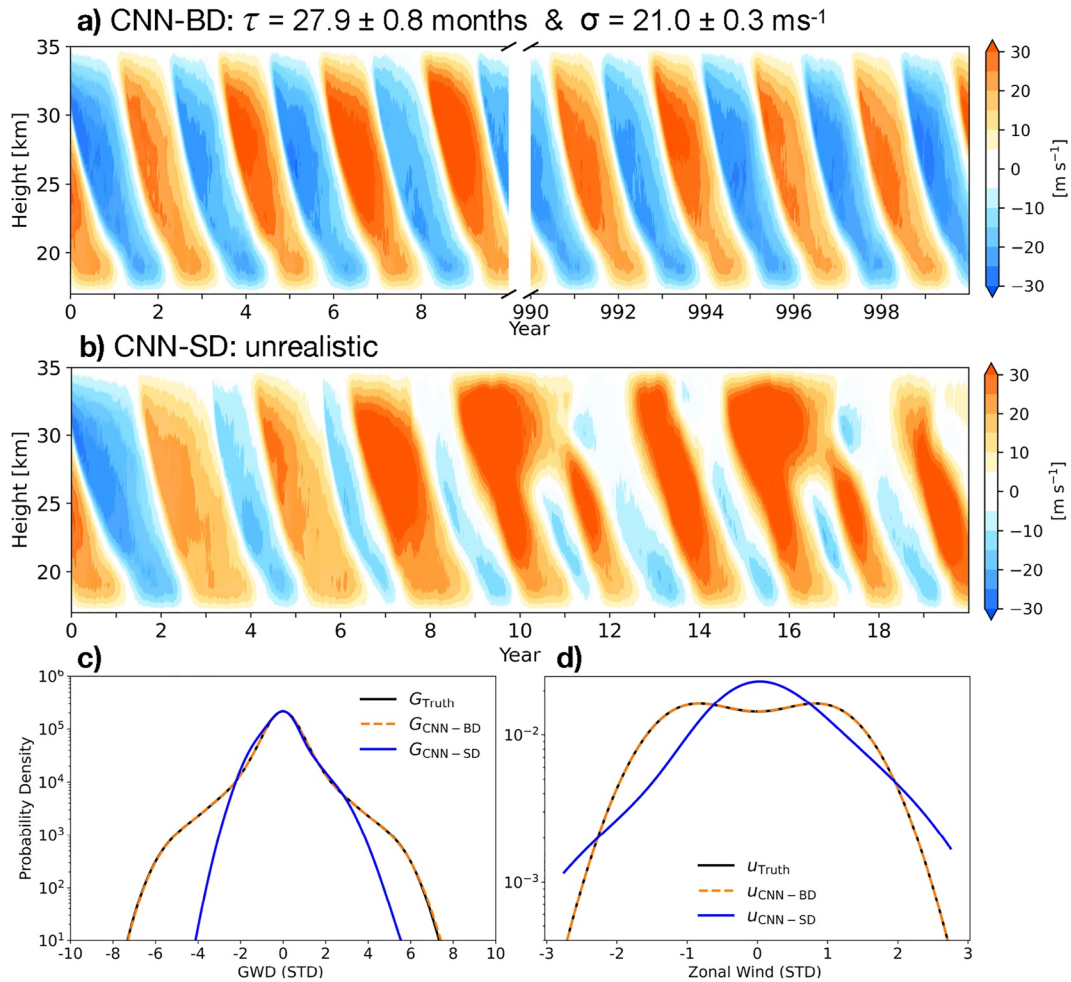

Figure 2. (a) The a posteriori (online) performance of the convolutional neural network (CNN)-BD. The quasi-biennial oscillation (QBO) remains stable for 1,000 years, with its period and amplitude closely matching the true QBO. (b) As in (a), but for the CNN-SD. The QBO is unrealistic. (c) Probability density functions of the true and a posteriori predicted GWD (G). (d) As in (c), but for zonal wind (u).

began by using the common targets for QBO, the period and amplitude, as our time-averaged statistics. However, we found that various unrealistic oscillations can misleadingly mimic the true QBO's period and amplitude, including an upward propagating QBO, suggesting non-uniqueness of the parameters and under-constrained optimization. To address this, we expanded our targeted statistics to include cross-covariances between various QBO levels, aiming to better capture its downward propagation. This led us to use an extensive list of 85 statistics (see Supporting Information S1).

As discussed above, another key element of EKI, and other online learning methods, is the prior distributions, with their choice having a significant influence on EKI's performance. Good priors can notably reduce the number of iterations and potentially the ensemble size. Conversely, poor priors can result in unsuccessful learning. Particularly, we are unable to online train a CNN from a random initialization of its weights and biases: over 95% of the ensemble members fail at each iteration, preventing the EKI algorithm from converging. Consequently, we use the weights and biases from CNN-SD as our priors. These serve as the mean values for unconstrained Gaussian priors with a standard deviation of 0.01, given that their magnitude is around $O(10^{-1})$.

Through further trial-and-error experiments, we discovered that it is unnecessary to online re-train every layer of the CNN-SD to achieve a stable QBO. By online re-training of only the shallowest and deepest hidden layers of the CNN-SD (i.e., layers 2 and 11), we obtain results that are on par with full re-training. Consequently, we confine our subsequent discussions to these findings. By re-training only two layers, we engage a significantly

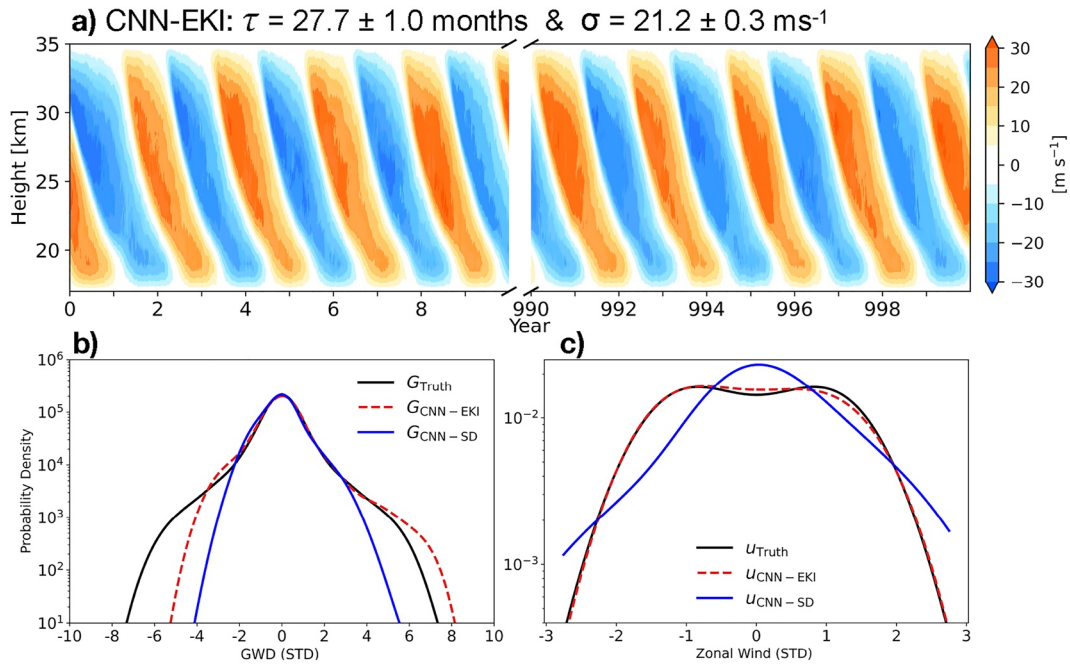

Figure 3. (a) The a posteriori performance of the convolutional neural network (CNN) after online re-training the CNN-SD, referred to as CNN-EKI. (b) Probability density functions of the true and a posteriori predicted GWD (G), before and after online re-training. (c) As in (b), but for zonal wind (u).

reduced parameter set, simplifying the analysis and enhancing interpretability, which we discuss in the next section. It is noteworthy that Subel et al. (2023) offers more structured strategies for determining the optimal layers for re-training in the context of transfer learning.

During the online re-training process, the EKI error decreases sharply in the first iteration (See also Figure S2 in Supporting Information S1). Subsequent iterations further reduce the mismatch between the statistics of the model (1D-QBO with CNN-based GWP) and the true QBO statistics. The ensemble mean of the parameters from the last iteration is then considered as the optimal parameter set for the CNN, referred to as CNN-EKI.

The a posteriori performance of the CNN-EKI is illustrated in Figure 3. Panel (a) shows an accurate QBO with period and amplitude closely agreeing with those of the true QBO. Figures 3b and 3c show the PDFs of GWD and zonal wind before and after the online re-training, with a comparison to the truth. The zonal wind's PDF demonstrates a significant improvement, closely aligning with the true QBO, albeit with minor deviations. This is despite the smaller improvement in the PDF of CNN-EKI's GWD, which better matches the PDF of the truth, but still deviates at the tails beyond four standard deviations. This could be expected, considering that we only use the statistics of zonal wind for online re-training, which does not necessarily constrain the PDF of GWD. In other words, the values of GWD beyond four standard deviations, which occur orders of magnitude less frequently than GWD values within two standard deviations, do not heavily influence the QBO period, amplitude, and its overall structure. Also, note that improvement in a posteriori performance without any improvement to the a priori one (or even its degradation) is a common feature of online learning, as reported in several past studies (Gelbrecht et al., 2023).

3.3. Explainable Learning Using SpArK

Next, we use the SpArK framework (Subel et al., 2023) to gain insights into the inner workings of the CNNs and connect them to the underlying physics of the GW propagation. Briefly, Subel et al. (2023) applied Fourier transformation and convolution theorem to the governing equations of CNNs. They showed that the kernels of CNNs used for SGS closure modeling of turbulent flows, while appearing meaningless in the physical space, are meaningful spectral filters in the Fourier domain, comprising low-, high-, band-pass, and Gabor filters. They further demonstrated that examining changes in the spectra of the kernels before and after re-training can explain the physics learned during transfer learning.

19448007, 2024, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL106324 by University Of Chicago Library, Wiley Online Library on [03/12/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

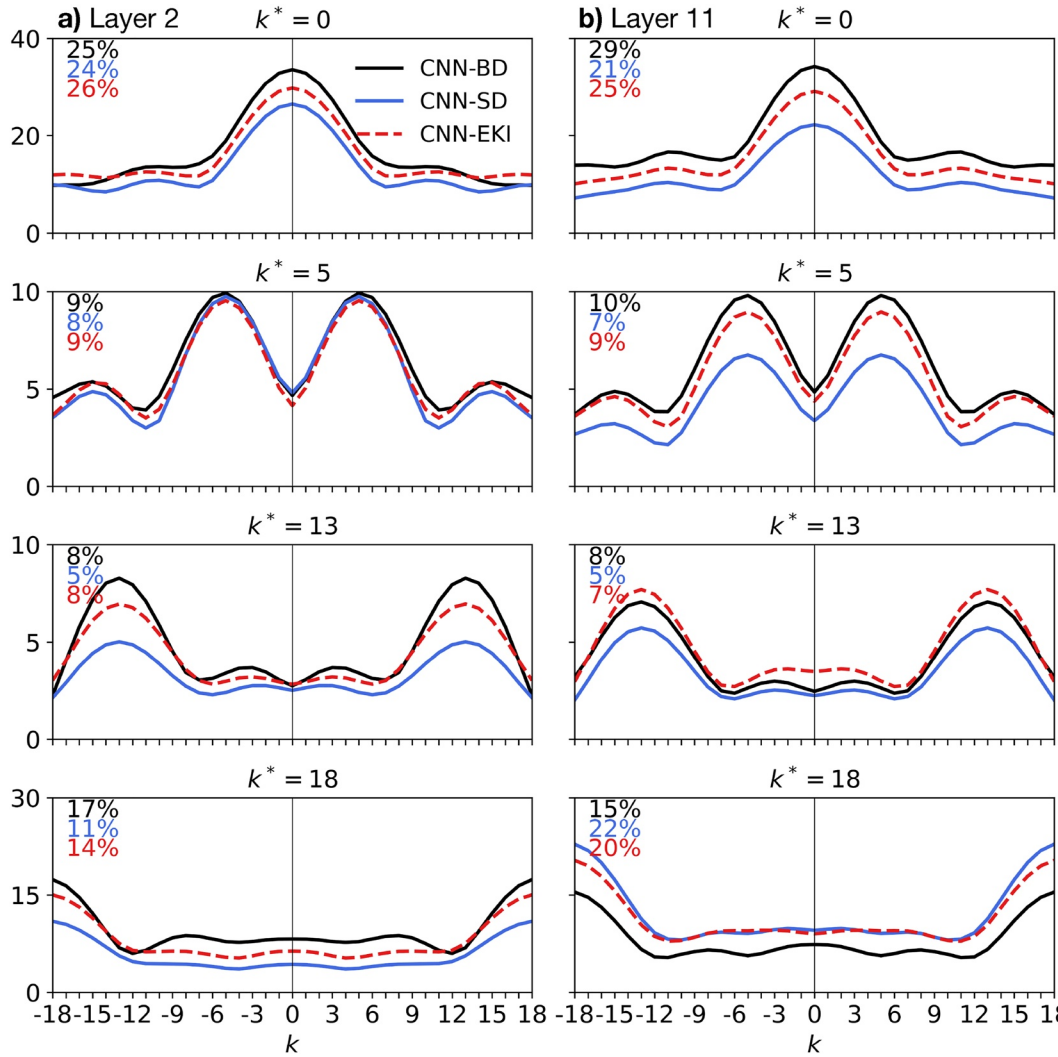

Figure 4. The sum of the Fourier spectra of kernels for the four most frequent wavenumber peaks k^* for (a) layer 2 and, (b) layer 11 for the three convolutional neural network (CNNs). The frequency of each kernel within its respective layer and for each of the three CNNs is indicated in the top left corner of each panel.

We start by extending each CNN kernel, originally of size 5 (doing convolution on activations of size 37), to match the size of the activations by zero-padding, resulting in kernels of size 37 (activation is the output of a layer after applying filters and non-linearity). We then simply transfer them into a spectral space using a Fourier transform. Upon close inspection of the Fourier spectra of kernels, it becomes evident that they are a combination of low-, high-, and band-pass filters. The similarity across the spectra of many kernels allows us to meaningfully categorize them by their dominant wavenumber (k^*), that is, the wavenumber where the spectrum peaks in magnitude.

The spectra of kernels for the four most frequent wavenumbers are shown in Figure 4. The dominant spectra have $k^* = 0$ (low-pass filters), followed by $k^* = 18$ (high-pass filters). $k^* = 5$ and $k^* = 13$ come next, each representing band-pass filters. While Figure 4 showcases the composited kernels from layers 2 and 11, similar patterns are observed across other layers. Collectively, these four wavenumbers account for ~65% of all the kernels, with $k^* = 0$ and $k^* = 18$ together constituting ~45% of the total. The frequent appearance of low- and high-pass filters, and to some degree, band-pass filters, might be connected to the dynamics of GW propagation and dissipation. On one hand, the GWD at a given level depends on the local zonal wind conditions. As discussed earlier, a wave propagates upward more slowly when the local wind is close to its phase speed, leading to increased dissipation. This dissipation is especially pronounced near the critical level, highlighting the essential role of local dynamics.

19448007, 2024, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL106324 by University Of Chicago Library, Wiley Online Library on [03/12/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19448007, 2024, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL106324 by University Of Chicago Library, Wiley Online Library on [03/12/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

On the other hand, the cumulative wind profile below a given level significantly impacts the GWD, underscoring the relevance of non-local dynamics. For any GWP scheme, capturing both local and non-local dynamics is necessary to be able to generate a spontaneous QBO (Campbell & Shepherd, 2005).

The prevalence of low-pass filters aligns well with the need to capture non-local dynamics, as these filters extract large scales and perform averaging. On the other hand, high-pass filters capture more local dynamics by extracting smaller scales. The band-pass filters, which resemble wavelets, extract specific scales, local in space. That said, the presence of layers and non-linearity further influence the output of each kernel, obscuring further understanding of the role of each kernel in connecting the profiles of the input (zonal winds) to the output (GWDs). Still, the understanding that emerges from SpArK can enable future work to exploit some of the novel mathematical tools from the deep learning community, in particular, those that leverage wavelet-analysis of NNs (Ha et al., 2021; Mallat, 2016). Finally, it should be highlighted that in this study, we focus on the Fourier spectra, and categorize the kernels based on k^*. A deeper analysis of both real and imaginary parts of the Fourier transformation of kernels and activations, as well as metrics beyond frequency, will be needed to gain further insight into how the NNs represent the GW dynamics.

While the full explainability of each NN remains a challenge, SpArK offers more insight into how an NN changes after online re-training. As shown in Figure 4, while these four wavenumbers retain their dominance in CNN-SD, their frequency deviates from those observed in CNN-BD. Yet, following online re-training, the frequency of these kernels aligns more closely with those in CNN-BD. This suggests a transformation of kernels from potentially ineffective wavenumbers to more efficient filters. This kind of analysis can provide insights into the calibration processes of ML-based parameterizations.

4. Discussion and Summary

The results presented in this study provide a proof-of-concept for the "offline-online" learning approach based on the parameterization of GWs in the computationally affordable 1D-QBO model. However, learning from time-averaged statistics necessitates long model simulations during training. Moreover, in complex and high-dimensional parameter spaces, EKI and similar optimization methods require large ensembles and more iterations to achieve an accurate estimate of the optimal parameters. Collectively, these factors can increase the computational cost, and when the forward model is expensive, as is the case for GCMs, the overall cost of EKI can become unfeasible (while here we focus on EKI, it should be noted that other online learning methods such as reinforcement learning suffer from the same challenges). Therefore, efficient strategies are needed to reduce the ensemble size and iterations. For EKI, techniques such as localization, inflation, and regularization are proposed in other studies for these situations (Huang et al., 2022; Iglesias, 2015, 2016; Iglesias & Yang, 2021; Lee, 2021; Tong & Morzfeld, 2022).

A common guideline for EKI suggests starting with ensemble numbers that are 10 times the count of parameters. However, a notable observation from this study is that the number of required ensemble members for EKI algorithm to converge does not directly correlate with the number of parameters of the CNN. In our experiments with CNNs containing approximately 1,000, 5,000, and 10,000 parameters, we consistently needed only 200 ensemble members for successful EKI convergence. A simple, yet untested, hypothesis is that increasing the number of parameters might not necessarily expand the dimensions of the parameter space, given the over-parameterized nature of NNs. This finding suggests the possibility of using a manageable ensemble size for online training of deep NNs using EKI, when coupled to GCMs.

We find that a consecutive 18-month span is inadequate for offline training of a CNN-based GWP in this 1D model. This is expected since this duration does not cover even a full QBO cycle. Alternatively, we can select 72 weeks, spaced a month apart, covering nearly three QBO cycles, while still being an 18-month-long data set. Offline training the CNN using this *strategically sampled small-data* regime, denoted as CNN-S3D, yields notably improved results compared to the CNN-SD (Figure S3 in Supporting Information S1). The a priori performance sees the R^2 value rise from 0.9 to 0.98, but more importantly, at the tails of the GWD PDF, a significant enhancement from 0.14 to 0.7. This results in an a posteriori accurate QBO.

This experiment highlights the importance of strategic sampling. Rather than continuous runs, commonly seen in current high-resolution modeling efforts (e.g., Satoh et al., 2019; Wedi et al., 2020), a more effective approach might be to create a library of shorter runs, sampling different regimes/phases important for a given physical

process. These runs would provide a diverse sampling of various climate and weather conditions without additional computational cost, an approach echoed in Shen et al. (2022), and Sun, Hassanzadeh, et al. (2023).

The process of calibration, upon which the "offline-online" learning strategy is suggested, is an essential element in the simulation of complex systems and is central to climate model development (Balaji et al., 2022). In this study, we primarily focus on the necessity of online re-training in the context of the *small-data* regime. However, it is essential to highlight that online re-training is probably indispensable when incorporating any data-driven parameterization scheme into the numerical models. This necessity arises from the fundamental differences (e.g., in numerics) between base models, like high-resolution simulations that supply the training data, and the target models, such as operational GCMs. Additionally, potential misalignments between a priori metrics and a posteriori performances further emphasize this need. While our focus in this study is on GWP, the findings can also be applied to other SGS modeling efforts.

Data Availability Statement

We use version 1.0.0 of open source software EnsembleKalmanProcesses.jl (Dunbar et al., 2022) for EKI analysis, accessible at Dunbar et al. (2023). The code for the 1D-QBO model and the specifically modified EKI software used in this study can be accessed at Pahlavan (2023b). The data are available at Pahlavan (2023a).

Acknowledgments
We are grateful to Oliver Dunbar, Tapio Schneider, Ofer Shamir, and Y. Qiang Sun for valuable discussions. We appreciate the CliMA project for providing public access to EKI. This work was supported by Grants from the NSF OAC CSSI program (2005123 and 2004512), and by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program (to P.H. and J.A.), by an Office of Naval Research (ONR) Young Investigator Award N00014-20-1-2722 (to P.H.), and by a Rice Academy Postdoctoral Fellowship (to H.P.). Computational resources were provided by NCAR's CISL (allocation URIC0009), and NSF XSEDE (allocation ATM170020).

References

Amiramjadi, M., Plougonven, R., Mohebalhojeh, A. R., & Mirzaei, M. (2023). Using machine learning to estimate nonorographic gravity wave characteristics at source levels. *Journal of the Atmospheric Sciences*, *80*(2), 419–440. https://doi.org/10.1175/jas-d-22-0021.1

Anstey, J. A., Osprey, S. M., Alexander, J., Baldwin, M. P., Butchart, N., Gray, L., et al. (2022). Impacts, processes and projections of the quasi-biennial oscillation. *Nature Reviews Earth & Environment*, *3*(9), 588–603. https://doi.org/10.1038/s43017-022-00323-7

Balaji, V., Couvreux, F., Deshayes, J., Gautrais, J., Hourdin, F., & Rio, C. (2022). Are general circulation models obsolete? *Proceedings of the National Academy of Sciences of the United States of America*, *119*(47), e2202075119. https://doi.org/10.1073/pnas.2202075119

Baldwin, M., Gray, L., Dunkerton, T., Hamilton, K., Haynes, P., Randel, W., et al. (2001). The quasi-biennial oscillation. *Reviews of Geophysics*, *39*(2), 179–229. https://doi.org/10.1029/1999rg000073

Campbell, L. J., & Shepherd, T. G. (2005). Constraints on wave drag parameterization schemes for simulating the quasi-biennial oscillation. part I: Gravity wave forcing. *Journal of the Atmospheric Sciences*, *62*(12), 4178–4195. https://doi.org/10.1175/jas3616.1

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I., & Palmer, T. (2021). Machine learning emulation of gravity wave drag in numerical weather forecasting. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002477. https://doi.org/10.1029/2021ms002477

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

Dunbar, O. R., Constantinou, N. C., Lopez-Gomez, I., Iñigo, A. G., Bolewski, J., Howland, M., et al. (2023). Clima/ensemblekalmanprocesses.jl: v1.0.0 [Software]. Zenodo. https://doi.org/10.5281/zenodo.7806813

Dunbar, O. R., Lopez-Gomez, I., Garbuno-Inigo, A., Huang, D. Z., Bach, E., & Wu, J.-L. (2022). EnsembleKalmanProcesses.jl: Derivative-free ensemble-based model calibration [Software]. Journal of Open Source Software, *7*(80), 4869. https://doi.org/10.21105/joss.04869

Ern, M., Ploeger, F., Preusse, P., Gille, J., Gray, L., Kalisch, S., et al. (2014). Interaction of gravity waves with the QBO: A satellite perspective. *Journal of Geophysical Research: Atmospheres*, *119*(5), 2329–2355. https://doi.org/10.1002/2013jd020731

Espinosa, Z. I., Sheshadri, A., Cain, G. R., Gerber, E. P., & DallaSanta, K. J. (2022). Machine learning gravity wave parameterization generalizes to capture the QBO and response to increased CO_2. *Geophysical Research Letters*, *49*(8), e2022GL098174. https://doi.org/10.1029/2022gl098174

Frezat, H., Le Sommer, J., Fablet, R., Balarac, G., & Lguensat, R. (2022). A posteriori learning for quasi-geostrophic turbulence parametrization. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2022MS003124. https://doi.org/10.1029/2022ms003124

Fritts, D. C., & Alexander, M. J. (2003). Gravity wave dynamics and effects in the middle atmosphere. *Reviews of Geophysics*, *41*(1), 1003. https://doi.org/10.1029/2001rg000106

Gelbrecht, M., White, A., Bathiany, S., & Boers, N. (2023). Differentiable programming for Earth system modeling. *Geoscientific Model Development*, *16*(11), 3123–3135. https://doi.org/10.5194/gmd-16-3123-2023

Grooms, I., Loose, N., Abernathey, R., Steinberg, J., Bachman, S. D., Marques, G., et al. (2021). Diffusion-based smoothers for spatial filtering of gridded geophysical data. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002552. https://doi.org/10.1029/2021ms002552

Ha, W., Singh, C., Lanusse, F., Upadhyayula, S., & Yu, B. (2021). Adaptive wavelet distillation from neural networks through interpretations. *Advances in Neural Information Processing Systems*, *34*, 20669–20682.

Hardiman, S. C., Scaife, A. A., Niekerk, A. V., Prudden, R., Owen, A., Adams, S. V., et al. (2023). Machine learning for non-orographic gravity waves in a climate model. In *Artificial intelligence for the Earth systems*.

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

Huang, D. Z., Huang, J., Reich, S., & Stuart, A. M. (2022). Efficient derivative-free Bayesian inference for large-scale inverse problems. *Inverse Problems*, *38*(12), 125006. https://doi.org/10.1088/1361-6420/ac99fa

Iglesias, M. A. (2015). Iterative regularization for ensemble data assimilation in reservoir models. *Computational Geosciences*, *19*(1), 177–212. https://doi.org/10.1007/s10596-014-9456-5

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

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

19448007, 2024, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL106324 by University Of Chicago Library, Wiley Online Library on [03/12/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Iglesias, M. A., & Yang, Y. (2021). Adaptive regularisation for ensemble Kalman inversion. *Inverse Problems*, *37*(2), 025008. https://doi.org/10.1088/1361-6420/abd29b

Kawatani, Y., Watanabe, S., Sato, K., Dunkerton, T. J., Miyahara, S., & Takahashi, M. (2010). The roles of equatorial trapped waves and internal inertia–gravity waves in driving the quasi-biennial oscillation. part I: Zonal mean wave forcing. *Journal of the Atmospheric Sciences*, *67*(4), 963–980. https://doi.org/10.1175/2009jas3222.1

Kim, Y.-H., & Chun, H.-Y. (2015). Contributions of equatorial wave modes and parameterized gravity waves to the tropical QBO in HadGEM2. *Journal of Geophysical Research: Atmospheres*, *120*(3), 1065–1090. https://doi.org/10.1002/2014jd022174

Kovachki, N. B., & Stuart, A. M. (2019). Ensemble Kalman inversion: A derivative-free technique for machine learning tasks. *Inverse Problems*, *35*(9), 095005. https://doi.org/10.1088/1361-6420/ab1c3a

Kruse, C. G., Richter, J. H., Alexander, M. J., Bacmeister, J. T., Heale, C., & Wei, J. (2023). Gravity wave drag parameterizations for Earth's atmosphere. *Fast Processes in Large-Scale Atmospheric Models: Progress, Challenges, and Opportunities*, *282*, 229.

Lee, Y. (2021). Sampling error correction in ensemble Kalman inversion. arXiv preprint arXiv:2105.11341.

Lopez-Gomez, I., Christopoulos, C., Langeland Ervik, H. L., Dunbar, O. R., Cohen, Y., & Schneider, T. (2022). Training physics-based machine-learning parameterizations with gradient-free ensemble Kalman methods. *Journal of Advances in Modeling Earth Systems*, *14*(8), e2022MS003105. https://doi.org/10.1029/2022ms003105

Mallat, S. (2016). Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *374*(2065), 20150203. https://doi.org/10.1098/rsta.2015.0203

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*(11), e2022MS003245. https://doi.org/10.1029/2022ms003245

Matsuoka, D., Watanabe, S., Sato, K., Kawazoe, S., Yu, W., & Easterbrook, S. (2020). Application of deep learning to estimate atmospheric gravity wave parameters in reanalysis data sets. *Geophysical Research Letters*, *47*(19), e2020GL089436. https://doi.org/10.1029/2020gl089436

Mojgani, R., Waelchli, D., Guan, Y., Koumoutsakos, P., & Hassanzadeh, P. (2023). Extreme event prediction with multi-agent reinforcement learning-based parametrization of atmospheric and oceanic turbulence. arXiv preprint arXiv:2312.00907.

Novati, G., de Laroussilhe, H. L., & Koumoutsakos, P. (2021). Automating turbulence modelling by multi-agent reinforcement learning. *Nature Machine Intelligence*, *3*(1), 87–96. https://doi.org/10.1038/s42256-020-00272-0

Pahlavan, H. A. (2023a). Dataset for "Explainable Offline-Online Training of Neural Networks for Parameterizations: A 1D Gravity Wave-QBO Testbed in the Small-data Regime" by Pahlavan et al. (2023) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.10278373

Pahlavan, H. A. (2023b). Software for "explainable offline-online training of neural networks for parameterizations: A 1d gravity wave-QBO testbed in the small-data regime" by Pahlavan et al. (2023) [Software]. Zenodo. https://doi.org/10.5281/zenodo.10278470

Pahlavan, H. A., Wallace, J. M., Fu, Q., & Kiladis, G. N. (2021). Revisiting the quasi-biennial oscillation as seen in ERA5. part II: Evaluation of waves and wave forcing. *Journal of the Atmospheric Sciences*, *78*(3), 693–707. https://doi.org/10.1175/jas-d-20-0249.1

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

Richter, J. H., Anstey, J. A., Butchart, N., Kawatani, Y., Meehl, G. A., Osprey, S., & Simpson, I. R. (2020). Progress in simulating the quasi-biennial oscillation in CMIP models. *Journal of Geophysical Research: Atmospheres*, *125*(8), e2019JD032362. https://doi.org/10.1029/2019jd032362

Richter, J. H., Butchart, N., Kawatani, Y., Bushell, A. C., Holt, L., Serva, F., et al. (2022). Response of the quasi-biennial oscillation to a warming climate in global climate models. *Quarterly Journal of the Royal Meteorological Society*, *148*(744), 1490–1518. https://doi.org/10.1002/qj.3749

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*(6), 3045–3062. https://doi.org/10.1002/2013jd021122

Satoh, M., Stevens, B., Judt, F., Khairoutdinov, M., Lin, S.-J., Putman, W. M., & Düben, P. (2019). Global cloud-resolving models. *Current Climate Change Reports*, *5*(3), 172–184. https://doi.org/10.1007/s40641-019-00131-0

Schneider, T., Behera, S., Boccaletti, G., Deser, C., Emanuel, K., Ferrari, R., et al. (2023). Harnessing AI and computing to advance climate modelling and prediction. *Nature Climate Change*, *13*(9), 887–889. https://doi.org/10.1038/s41558-023-01769-3

Shen, Z., Sridhar, A., Tan, Z., Jaruga, A., & Schneider, T. (2022). A library of large-eddy simulations forced by global climate models. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002631. https://doi.org/10.1029/2021ms002631

Subel, A., Guan, Y., Chattopadhyay, A., & Hassanzadeh, P. (2023). Explaining the physics of transfer learning in data-driven turbulence modeling. *PNAS nexus*, *2*(3), pgad015. https://doi.org/10.1093/pnasnexus/pgad015

Sun, Y. Q., Hassanzadeh, P., Alexander, M. J., & Kruse, C. G. (2023). 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*(5), e2022MS003585. https://doi.org/10.1029/2022ms003585

Sun, Y. Q., Pahlavan, H. A., Chattopadhyay, A., Hassanzadeh, P., Lubis, S. W., Alexander, M. J., et al. (2023). Data imbalance, uncertainty quantification, and generalization via transfer learning in data-driven parameterizations: Lessons from the emulation of gravity wave momentum transport in WACCM. arXiv preprint arXiv:2311.17078.

Tong, X. T., & Morzfeld, M. (2022). Localization in ensemble Kalman inversion. arXiv preprint arXiv:2201.10821.

Wedi, N. P., Polichtchouk, I., Dueben, P., Anantharaj, V. G., Bauer, P., Boussetta, S., et al. (2020). A baseline for global weather and climate simulations at 1 km resolution. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002192. https://doi.org/10.1029/2020ms002192

19448007, 2024, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL106324 by University Of Chicago Library, Wiley Online Library on [03/12/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License