Geophysical Research Letters

RESEARCH LETTER
10.1029/2022GL098174

Key Points:

- Neural networks trained on one annual cycle accurately emulate a physics-based gravity wave parameterization (GWP) when coupled to a climate model
- Although trained on only one phase of the Quasi-Biennial Oscillation, the emulator generates the entire cycle of the oscillation
- The emulator captures key qualitative features of the response of the original GWP to enhanced CO_2

Supporting Information:

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

Correspondence to:

Z. I. Espinosa,
zespinos@stanford.edu

Citation:

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*, e2022GL098174. https://doi.org/10.1029/2022GL098174

Received 1 FEB 2022
Accepted 27 MAR 2022

Author Contributions:

Conceptualization: Aditi Sheshadri, Edwin P. Gerber
Data curation: Zachary I. Espinosa, Kevin J. DallaSanta
Formal analysis: Zachary I. Espinosa, Aditi Sheshadri, Edwin P. Gerber
Funding acquisition: Aditi Sheshadri
Investigation: Zachary I. Espinosa, Aditi Sheshadri
Methodology: Zachary I. Espinosa, Aditi Sheshadri, Gerald R. Cain, Edwin P. Gerber, Kevin J. DallaSanta
Project Administration: Aditi Sheshadri
Software: Zachary I. Espinosa, Gerald R. Cain
Supervision: Aditi Sheshadri, Edwin P. Gerber
Visualization: Zachary I. Espinosa

Machine Learning Gravity Wave Parameterization Generalizes to Capture the QBO and Response to Increased CO_2

Zachary I. Espinosa[1], Aditi Sheshadri[1], Gerald R. Cain[2], Edwin P. Gerber[3], and Kevin J. DallaSanta[4,5]

[1]Department of Earth System Science, Stanford University, Stanford, CA, USA, [2]Department of Computer Science, Stanford University, Stanford, CA, USA, [3]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA, [4]NASA Goddard Institute for Space Studies, New York, NY, USA, [5]Universities Space Research Association, Columbia, MD, USA

Abstract We present single-column gravity wave parameterizations (GWPs) that use machine learning to emulate non-orographic gravity wave (GW) drag and demonstrate their ability to generalize out-of-sample. A set of artificial neural networks (ANNs) are trained to emulate the momentum forcing from a conventional GWP in an idealized climate model, given only one view of the annual cycle and one phase of the Quasi-Biennial Oscillation (QBO). We investigate the sensitivity of offline and online performance to the choice of input variables and complexity of the ANN. When coupled with the model, moderately complex ANNs accurately generate full cycles of the QBO. When the model is forced with enhanced CO_2, its climate response with the ANN matches that generated with the physics-based GWP. That ANNs can accurately emulate an existing scheme and generalize to new regimes given limited data suggests the potential for developing GWPs from observational estimates of GW momentum transport.

Plain Language Summary Atmospheric gravity waves (GWs) or "buoyancy waves" are generated by perturbations in a stably-stratified environment. They mediate momentum transport between the lower and middle atmospheres and play a leading-order role in driving middle atmospheric circulation. Due to computational constraints, global climate models "parameterize" or estimate the effect of GWs on the large-scale flow. Current climate predictions are sensitive to uncertainties in these representations. Here, we examine whether machine learning, given limited data, can be used for gravity wave parameterization (GWP) in climate prediction. This approach represents an appealing technique to build data-driven GWPs that can reduce existing uncertainties by incorporating observations.

1. Introduction

Atmospheric gravity waves (GWs) play an important role in surface climate (Palmer et al., 1986) and a leading-order role in driving the middle atmospheric circulation and its variability (Fritts & Alexander, 2003). By transporting momentum, GWs impact the mean climatology (Bretherton, 1969; Sato & Hirano, 2019) and atmospheric variability, affecting storm tracks (Fritts & Nastrom, 1992) and stratospheric dynamics (Antonita et al., 2007; Kang et al., 2018; Limpasuvan et al., 2012). Since much of the GW spectrum is too fine to be captured at current model resolutions, models typically mimic its effects on the resolved circulation via explicit gravity wave parameterizations (GWPs) (Fritts & Alexander, 2003; Richter et al., 2010).

Incorporating a realistic representation of these effects in numerical models, however, remains challenging for several reasons. The true GW momentum flux is not well-constrained by observational or intermodel studies (Geller et al., 2013) and GWs are generated by a variety of sources, including orography, convection, and frontogenesis (Fritts & Alexander, 2003), many of which must be parameterized themselves, that vary greatly in their representation across models. Hence the details of GWPs differ greatly between models (Butchart et al., 2018), and even small changes within the same model can lead to diverging regional climate projections (Schirber, 2015). Given these limitations, GWPs are optimized toward maximizing global forecast skill scores (Alexander et al., 2019) or "tuned" to reduce climatological biases (Garcia et al., 2017). Projections of the tropospheric and stratospheric circulations' response to anthropogenic forcing are sensitive to uncertainties in GWPs (Polichtchouk et al., 2018; Sigmond & Scinocca, 2010), requiring continued development of GWP and the need for more observational constraints on their behavior.

Writing – original draft: Zachary I. Espinosa
Writing – review & editing: Zachary I. Espinosa, Aditi Sheshadri, Edwin P. Gerber

An alternative approach to physics-based parameterization has emerged in the form of machine learning (ML). ML has been employed to parameterize processes such as convection (Gentine et al., 2018; Rasp et al., 2018) and radiation (Brenowitz & Bretherton, 2018; Roh & Song, 2020). Matsuoka et al. (2020) demonstrated that a convolutional neural network can estimate the GW structure over Hokkaido, Japan when trained on high-resolution reanalysis data and Chantry et al. (2021) used an artificial neural network (ANN) to emulate a non-orographic GWP in a weather forecasting system for seasonal prediction.

In this study, we investigate the efficacy of ML as a non-orographic GWP for climate projection. We demonstrate that the GW momentum forcing and resultant circulation can be stably and accurately generated using ANNs that are trained to emulate a physics-based GWP in a global atmospheric model. Most critically, we ask whether the ANN, which we call WaveNet, can be coupled with the atmospheric model under out-of-sample conditions, subjecting it to two tests. First, when trained on only one phase of the Quasi-Biennial Oscillation (QBO), can it reproduce the entire oscillation? And second, trained on only output from a climate that resembles that of the present day, can the scheme faithfully reproduce the impact of enhanced CO_2 forcing? With this second test, we can only assess whether the ANN emulator is able to reproduce the response of the model with the physics-based GWP; the response of true GWs remains unknown. The success of our emulator to meet both challenges given only a single annual cycle demonstrates the potential for developing ANNs using observationally constrained estimates of GW momentum forcing.

2. Data

We train the ANN to emulate the Alexander and Dunkerton (1999) GWP as incorporated into an atmospheric model of intermediate complexity, the Model of an Idealized Moist Atmosphere (MiMA) (DallaSanta et al., 2019; Jucker & Gerber, 2017). MiMA captures key dynamical features of the stratosphere-troposphere system that depend critically on GWs at a resolution comparable to state-of-the-art stratosphere resolving global climate models (GCMs). For additional details, refer to Jucker and Gerber (2017) and DallaSanta et al. (2019). MiMA is integrated with T42 spectral resolution (triangular truncation at wavenumber 42, roughly equivalent to a 2.8-degree or 310 km grid at the equator) in the horizontal and 40 vertical levels, with a model lid at 0.18 hPa and 23 levels above 100 hPa.

The implementation of the Alexander and Dunkerton (1999) GWP, hereafter referred to as AD99, is based on its formulation within the GFDL Flexible Modeling System. The AD99 scheme aims to capture the effect of non-orographic GW drag. It is the same scheme employed by GFDL Atmospheric Model 3 (Donner et al., 2011), except modified by Cohen et al. (2013) to ensure that all momentum that reaches 0.85 hPa is uniformly distributed to layers above the stratopause. AD99 assumes a Gaussian spectrum of GWs, discretized as a function of phase speed and launched from the upper troposphere (315 hPa). A broad spectrum of phase speeds is chosen, with a half width of 35 m/s, and centered around the speed of the zonal wind at the launch level. The total amplitude of the momentum stress is 0.0043 Pa. The broad spectrum ensures that there is a rich source of GWs at MiMA's lower boundary. The momentum associated with each wave is deposited at its linear breaking level. The intermittency of observed GWs is taken into account via a scaling parameter. The breaking level is based on the behavior of a large wave, indicative of the median amplitude of the distribution, but the momentum flux is rescaled to provide the momentum deposition associated with an average wave. This better captures the true breaking level of GWs, but smooths out the momentum deposition in time.

The chief idealization of the scheme is in our choice of source spectrum. We assume a uniform spectrum that varies only with respect to the zonal winds at the source level. This crudely accounts for filtering of the wave spectrum by the troposphere and was critical for the simulation of the QBO.

We utilize five years of six-hourly output from one integration of MiMA, yielding 11,796,480 training samples per year. The second year of output is used for training and validation, while all others are reserved for testing. Not all runs utilize the full year of training data, and differences are specified per experiment. All training data are standardized by removing the mean and scaling to unit variance and calculated for each variable across each pressure level. Test data is similarly standardized using the mean and variance calculated from training data. Output variables are inverse transformed before presentation.

3. Neural Network Architecture

An ANN is a computing system of interconnected layers of computational nodes. For a detailed review of ANNs, we recommend Brenowitz and Bretherton (2018). We trained two ANNs to generate zonal and meridional drag separately. The ANN architecture is described in Figure S1 in Supporting Information S1.

The input layer of both ANNs accepts a single vertical column of resolved flow variables from MiMA (Table S1 in Supporting Information S1). The output layer produces a single vertical column of 33 gravity wave drag (GWD) values, corresponding to the upper 33 model levels (approximately 0.18–436 hPa). This includes the upper three sponge layers and three layers below AD99's average launch level. During online simulations with the ANN, no drag is specified below level 33; that is, zeros are appended to WaveNet's output to extend to the full length of the vertical column. The result of each node in the output layer does not pass through an activation function. For all other nodes, we use the Rectified Linear Unit. We consider a variety of ANN configurations, varying the number and width of layers. Total trainable parameters vary from 3,848,481 to 104,097, associated with a depth of 9–4 layers, respectively. Offline results presented in Section 4 use the largest network unless otherwise stated, and an analysis of performance sensitivity to ANN size is presented in Section 4.2.

The loss function - in our case, the logcosh error - is computed for a minibatch of 1,024 training samples that are drawn from a shuffled training data set. The logcosh error is defined as

$$L\left(y, y^{p}\right) = \sum_{i=1}^{n} \log\left(\cosh\left(y_i^{p} - y_i\right)\right)$$

where n is the size of the data set, and y_i and y_i^p are the i^{th} truth and prediction, respectively. The logcosh error is found to consistently outperform mean-squared and mean-absolute loss functions when comparing R^2 performance. Parameter updates are performed according to Adam optimization (Kingma & Ba, 2014). We started each training session with a learning rate of 10^{-3} and reduced it when improvement plateaued for more than 5 epochs, with an epoch defined as 1,500 batches. We stopped training when performance plateaued for more than 10 epochs, which occurred at 200 epochs on average. We did not perform any regularization. All weights are initialized using Xavier initialization (Glorot & Bengio, 2010).

4. Offline Performance

We start by training WaveNet on 1 year of MiMA output and testing it offline over the 4 years consisting of the year previous to and the 3 years following the training year. We analyze the vertical profile of momentum tendencies at 5°S - 5°N and 60°N, which show performance in the QBO region and boreal polar vortex, respectively (Figure 1). WaveNet is trained on 1 year of global data (year 2), dominated by the westerly phase of the QBO in the mid stratosphere and a full seasonal cycle of the polar vortex. In the tropics, the ANN captures the changes in GW driving associated with the QBO phases preceding and succeeding the training period. This suggests that WaveNet can generalize learning from regions (horizontally or vertically) containing easterly winds to a region where it has not experienced similar samples during the training period. At 60°N, WaveNet captures the response to variability in the vortex unobserved during the training year, notably the disturbance near the end of year 1, which appears as a spike in easterly forcing at upper levels and westerly forcing at lower levels associated with the breakdown of the vortex.

While comparison between Figures 1a and 1d and Figures 1c and 1f shows that WaveNet can capture the key features of the "out-of-sample" periods, differences between AD99 and WaveNet are larger at these times, as seen in the middle panels. Notably, in the tropics the errors decrease around month 48, when the winds are more similar in structure to the training period. This suggests more varied training data or regularization techniques may improve WaveNet's performance. Movies S1–S4 further document the consistency between the horizontal GWD generated by WaveNet and AD99 at 10 and 100 hPa for one year time series of test data. These results suggest WaveNet's potential to generalize from regional observations.

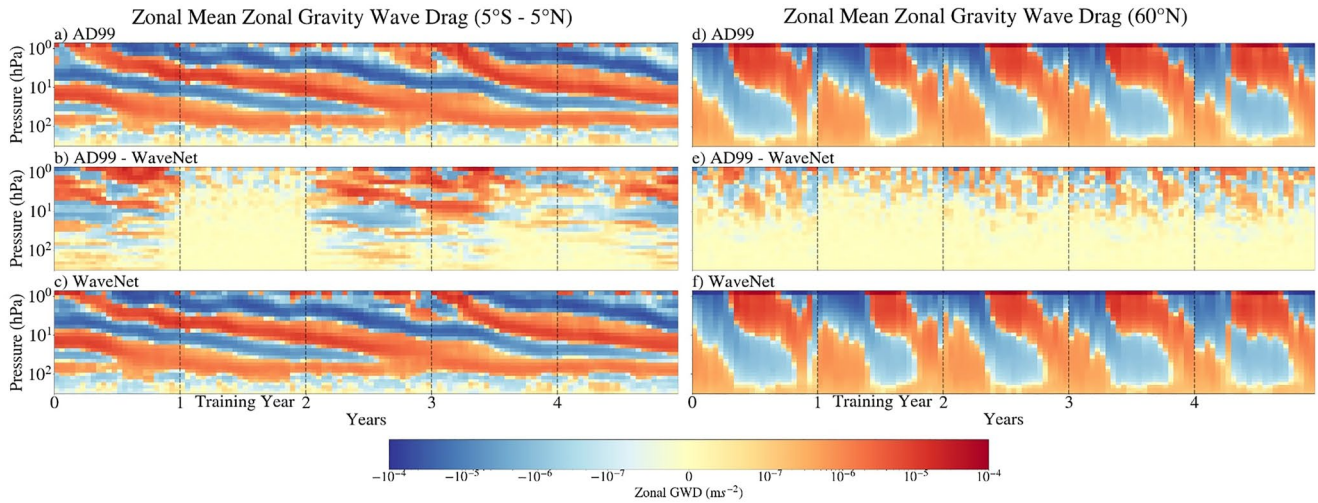

Figure 1. Pressure-time profiles of 15-day averaged zonal wind tendencies at 5°S - 5°N and 60°N for (a and d) AD99, (c and f) WaveNet, and their difference (b and e) during offline, uncoupled testing. Vertical dashed lines separate years. The artificial neural network is trained on the second year and tested on all other years. The easterly (blue) and westerly (red) bands in the left column correspond to drag associated with the Quasi-Biennial Oscillation, and in the right column correspond to drag associated with the seasonal cycle of the polar vortex.

4.1. Interpretability of the ANN

To further evaluate the quality of predictions, we calculate the R^2 coefficient of determination for a 1 year time series of test data at each spatial location and average for each pressure level. R^2 is defined as one minus the proportion of the sum of squares of residuals to the total sum of squares. We also train WaveNet using 10 distinct subsets of the resolved flow variables to determine which training features are most critical (Figures 2a and 2b). When trained with the full feature set (light blue, circle inscribed lines), the average pressure weighted R^2 value above the source level is 0.91 for zonal and 0.88 for meridional GWD predictions. Below this level, performance significantly degrades as AD99 overwhelmingly outputs trivial, nonphysical GWD. The zonal (meridional) wind component is the only flow variable necessary to retain 94% (96%) of the ANN's performance for zonal (meridional) GWD predictions.

To further investigate the role of horizontal wind features, we calculate Shapley Additive Explanations (SHAP) (Lundberg & Lee, 2017a). The SHAP value for a feature is the weighted sum of the conditional expectation of the marginal contribution of including that feature in the prediction. The SHAP values for this study are calculated using Deep SHAP, an approximation technique that combines DeepLIFT with Shapley values from collinear cooperative game theory (Shrikumar et al., 2017; Lundberg & Lee, 2017b). Figure S2 in Supporting Information S1 shows that on average the horizontal wind components on levels directly above and below the level of prediction have the largest contribution to the model's prediction. This is consistent with our physical understanding of GWs, where dissipation is linked to critical levels (i.e., where the phase speed of a wave is equal to the speed of the background flow). This spatially local dependence suggests that our approach may generalize well to observational datasets that contain measurements at a small range of pressure levels, for example, observations from superpressure balloons (Lindgren et al., 2020; Podglajen et al., 2016).

4.2. Sensitivity to Amount of Training Data and Complexity

To understand how performance degrades as less training data is made available to the ANN, we incrementally decrease the number of training samples from 1 year to 1 day (Figures 2c and 2d). To account for the effects of seasonality, we sample uniformly in space and time from 1 year of training data to generate a subset with coverage of the entire annual cycle. We find that a quarter of a year of data (equivalent to 3 months or 2.9 million samples) is sufficient to retain 98% of WaveNet's performance compared to training with 1 year of data. This suggests the promise of a data-driven GWP to be trained and validated on observational datasets of similar resolution and size.

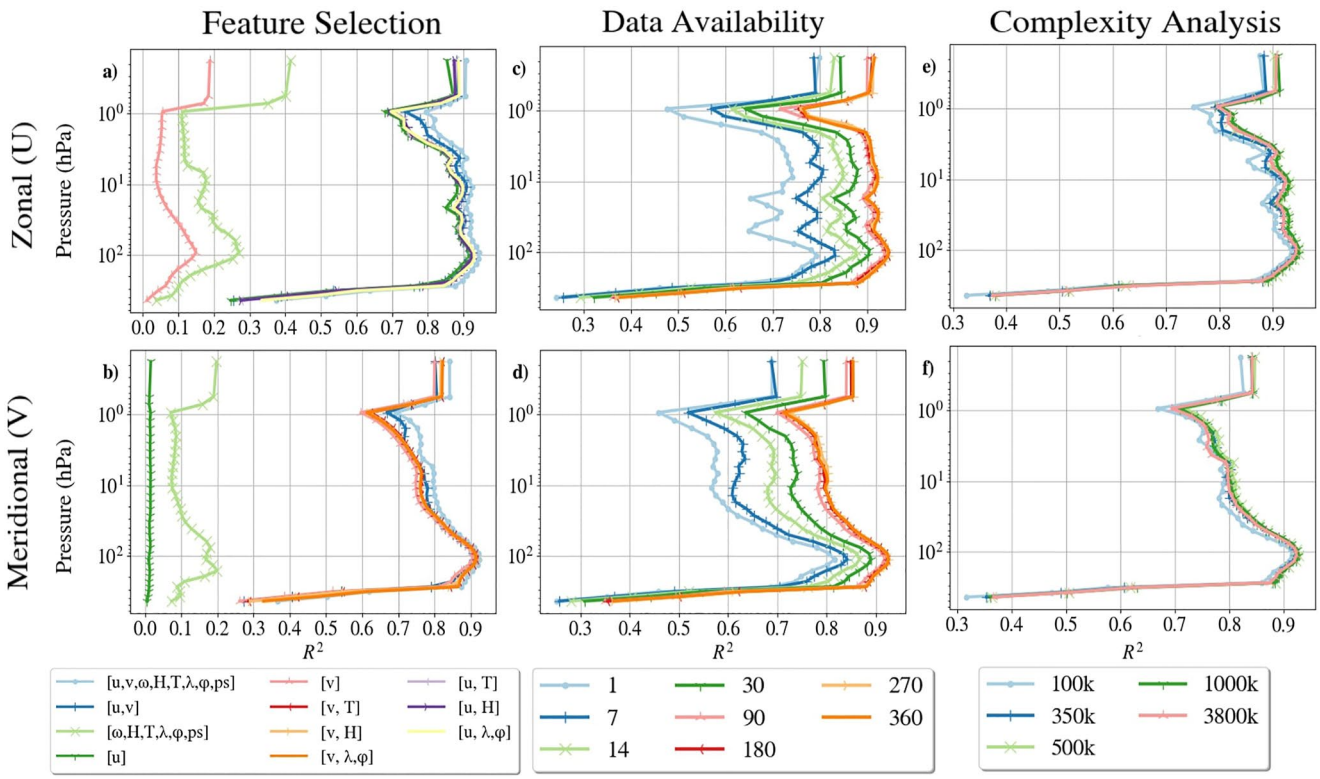

Figure 2. R^2 values are computed for 1 year of test data at each spatial location and averaged for each model pressure level. Presented are the R^2 values for three sets of experiments: (1) (a and b) feature experiments, where 10 artificial neural networks (ANNs) with 3800K parameters are trained with different subsets of the full input variable on 360 days of data; (2) (c and d) data availability experiments, where ANNs with 3800K parameters are trained with the number of days specified in the legend using the full feature set; (3) and (e and f) complexity experiments, where the number of parameters and layers is incrementally decreased for ANNs trained using 360 days of data and the full feature set.

To assess how complexity impacts performance, we train ANNs with incrementally fewer trainable parameters. WaveNet's offline performance exhibits a modest response to changes in trainable parameters (Figures 1e and 1f), hardly perceptible in the R^2 metric until the number of parameters is reduced by a factor of 10 or more. Its online performance, however, is found to be more sensitive to complexity.

5. Online Performance

Accurate offline emulation does not necessarily engender good online performance (Brenowitz & Bretherton, 2018). We replace the AD99 scheme with WaveNet, coupling the ANN, written in Python, with MiMA, written in Fortran, using the interoperability package *forpy*. Different approaches can be taken to couple ML algorithms with Fortran (Chantry et al., 2021; Ott et al., 2020); the main advantage of *forpy* is that it supports complex network structures, allowing one to easily switch between ML schemes. It is thus ideal for research and development, but not for long climate simulation, as it is slow compared with alternate approaches. When coupled, WaveNet slows the GCM by roughly 2.5x, a result of the *forpy* interface and WaveNet's size. The runtime can be substantially optimized by using a faster interface (i.e., directly from Fortran to C), performing quantization and pruning, reducing the number of trainable weights, and migrating to a GPU-compatible GCM. Optimizing and analyzing WaveNet's run-time is a subject of ongoing study and beyond the scope of this work.

For all subsequent online analysis, we use the version of WaveNet containing approximately 350K parameters trained with 1 year of data that accept as input u,T (zonal ANN) and v,T (meridional ANN). While versions of WaveNet with fewer parameters score well in offline tests, we required an ANN with roughly 350K parameters to accurately capture the QBO (Figure S3 in Supporting Information S1).

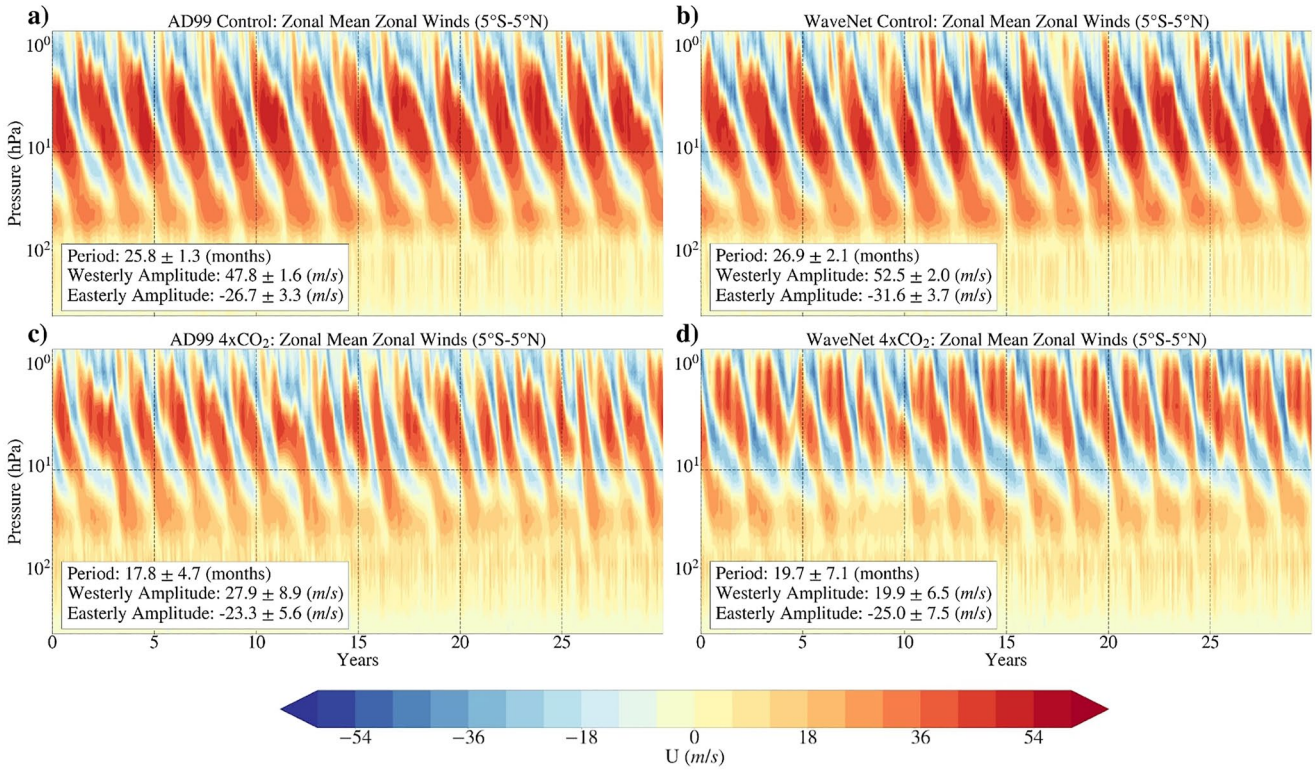

Figure 3. Pressure-time profiles of the zonal mean zonal wind, averaged between 5°S and 5°N and smoothed with a 15-day low pass filter, show the behavior of the Quasi-Biennial Oscillation (QBO) in integrations of (a) the control version of Model of an Idealized Moist Atmosphere with the AD99 parameterization, (b) the model coupled with WaveNet, (c) a 4xCO_2 integration with the AD99 parameterization, and (d) a 4xCO_2 integration coupled with WaveNet. Vertical dashed lines separate 5 years segments. The westerly (red) and easterly (blue) bands correspond to winds associated with opposite phases of the QBO. The QBO period and amplitudes are calculated using the transition time (TT) method. The dashed-horizontal line in each panel delineates the model level (≈10 hPa) where the TT method is used.

To demonstrate the ability of WaveNet to stably and accurately capture the QBO and CO_2 response, we complete 60-year integrations of MiMA coupled to WaveNet and AD99, respectively. The first 30 are discarded as spin-up, and results presented here are for the final 30 years of each integration. Figure 3 shows pressure-time profiles of 15-day averaged zonal mean zonal winds between 5°S and 5°N for (a and c) AD99 and (b and d) WaveNet. Following the TT method described in Richter, Anstey, et al. (2020), we calculate the period of each QBO cycle as the difference in time between every other phase change for the 5°S–5°N averaged zonal mean zonal wind near 10 hPa. The westerly (easterly) amplitude is taken as the maximum (minimum) value of the time series for each QBO cycle. These statistics are shown in the lower left corner of each panel. From the same experiments, we plot the average zonal winds as a function of pressure and latitude in Figure 4.

5.1. Capturing the QBO and Climatology

The first experiment utilizes the same model configuration as was used to generate training data. This experiment serves to evaluate WaveNet's ability to stably reproduce the QBO on climate timescales, having trained on only one annual cycle. For this scenario, WaveNet produces a QBO with an average period of 26.9 ± 2.1 months, comparable to the 25.8 ± 1.3 months period generated by AD99. WaveNet produces remarkable consistency in the overall height and fine-scale features of the QBO. The asymmetry between easterly and westerly amplitudes, which is well appreciated in models and observations (DallaSanta et al., 2021), is fully captured by WaveNet. This suggests that training an ANN on limited observations (less than one QBO cycle) may provide spectral insight even if the entire amplitude cycle is not well-sampled. WaveNet produces a climatology similar to AD99 (Figure 4). WaveNet generates a stronger polar vortex (Figure S4 in Supporting Information S1) and cooler temperatures in the upper atmosphere in both poles. The general agreement between WaveNet and AD99 observed in these results is a strong indicator that WaveNet can act as a faithful emulator of AD99 when given limited training data.

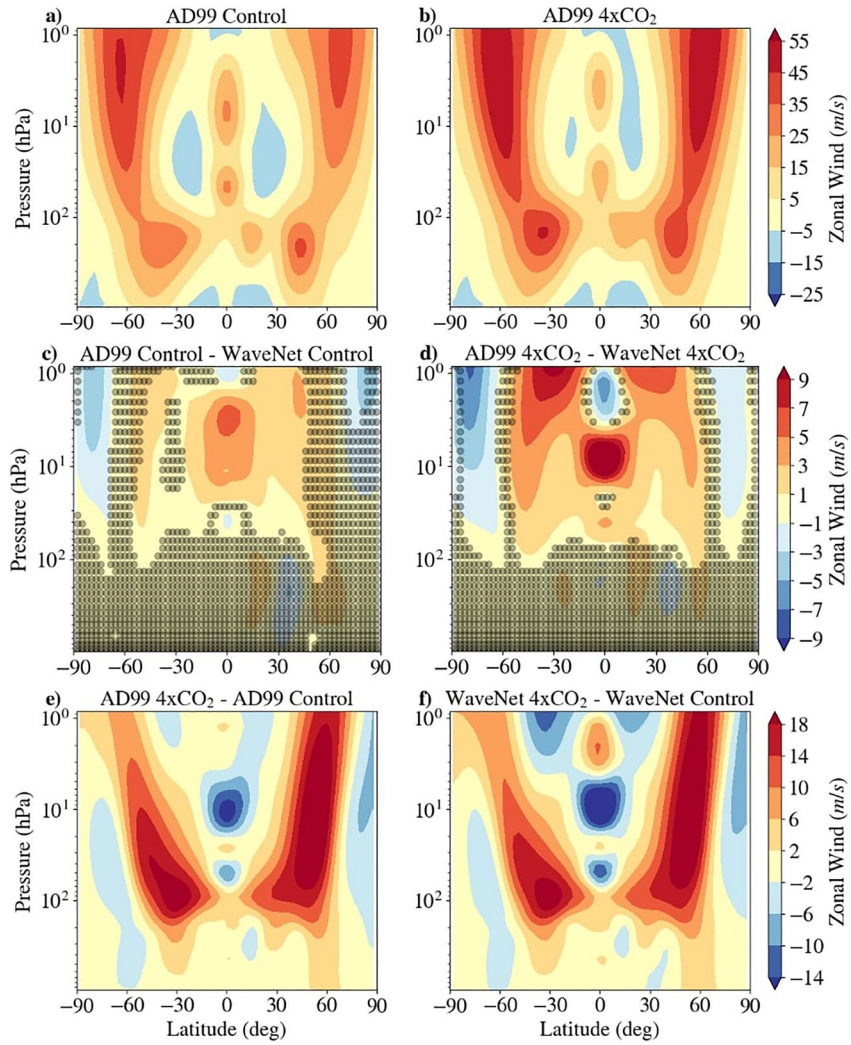

Figure 4. Pressure-latitude profiles of average zonal winds in the control (a) and 4xCO_2 (b) simulations for AD99. (c and d) present the climatological difference between AD99 and WaveNet for average zonal winds for the control and 4xCO_2 simulations, respectively. Regions that are not statistically distinguishable ($p > 0.05$) via the Student's *t*-test are dotted. The climate change signal for average zonal winds for AD99 and WaveNet are presented in panels (e and f). Note that colorbar magnitudes vary between rows: (e and f) the climate response is nearly double the (c and d) magnitude of performance error.

5.2. Capturing the CO_2 Response

To be used for climate modeling, a data-driven GWP must accurately emulate GWD across a range of model configurations and scenarios not fully captured in the training data. The second experiment presented here serves as a first step toward understanding WaveNet's fitness for climate projection. Here, we completed a coupled integration using four times pre-industrial concentrations of CO_2 (1200 ppm) or roughly triple that of the control run (390 ppm), which was optimized for early 21st century scenarios. Figures 3c and 3d show that when using either AD99 or WaveNet the period and amplitude of the QBO decrease. The QBO generated by WaveNet has a similar response in magnitude and period, however, the westerly phase lasts longer than in the QBO generated by AD99 and peaks at roughly 4 hPa, compared to a peak at 8 hPa with AD99. An amplitude decrease is found in state-of-the-art models (Richter, Butchart, et al., 2020) and is attributed to the expansion of the troposphere (Match & Fueglistaler, 2021).

Comprehensive climate models do not agree on the response of the QBO period to CO_2 and a mechanism is not yet clear (Richter, Butchart, et al., 2020). A reduction in the period could be due to an increase in GW momentum flux, but the intermodel correlation between period and flux vanishes when only fixed-source GWPs are

considered. Thus, while it is encouraging that WaveNet captures the same reduction of the QBO period as that observed with the AD99 parameterization, we do not know if this is the correct climate response. Our idealized model results point to further investigation with more complex models, for example, within the GFDL Flexible Modeling System.

Figures 4 and S4 in Supporting Information S1 show that the model with WaveNet and AD99 generates a similar but weaker climatological response to enhanced CO_2. They both project an increase in the strength of the stratospheric polar jets, which is more consistent through the depth of the stratosphere in the Northern Hemisphere, and a poleward shift in the tropospheric jets. The strength of the zonal jet response to $4xCO_2$ is weaker in WaveNet compared to AD99. The fraction of total distinguishable points for the $4xCO_2$ and control scenarios are 48% and 33%, respectively.

6. Discussion and Conclusion

We have demonstrated that ANNs of moderate complexity can skillfully learn the salient features of GW momentum transport by a conventional GWP directly from resolved flow variables. Our emulator stably couples with an idealized atmospheric model under out-of-sample conditions capturing the full cycle of the QBO when only trained on one phase of the oscillation, and key qualitative features of the response of the model with the original GWP to $4xCO_2$ when only trained on the present climate. The most important input features for WaveNet's predictions are the horizontal wind components local to the vertical level of prediction, consistent with our physical understanding of the importance of critical levels for GW momentum deposition. The success of these experiments may open a new avenue to incorporate the advantages of high-resolution GW simulations (Remmler et al., 2015) and observational datasets (Lindgren et al., 2020) into current GCMs.

There remain challenges, however, before the advantages of this approach can be fully realized for climate projection. First, are there sufficient observationally constrained estimates of GW momentum transport available for training? The AD99 parameterization is idealized relative to true GWs, particularly in its treatment of sources. While observationally constrained estimates of GW momentum transport could likely provide a reasonable estimate of an annual cycle of GW activity—on par with the training input used by WaveNet—more data may be required to capture the complexity of true GWs. A second critical question is computational efficiency. Radical improvements may come from optimizing ANN complexity, employing alternate coupling schemes and/or utilizing GPU hardware.

Nevertheless, our results suggest that ML may represent a powerful alternative to existing GWPs. The approach presented here constitutes an important step toward obtaining such non-orographic GWPs for global climate modeling.

Data Availability Statement

The artificial neural network (ANN) is built using Keras, a deep learning framework that wraps Tensorflow. All training source code is available at https://doi.org/10.5281/zenodo.4428931 (Espinosa, 2021). Training took on the order of 12 hr on a graphical processing unit and varied according to data size and trainable parameters. The implementation of Shapley Additive Explanations values is available at https://github.com/slundberg/shap and is not maintained or owned by this project group (Lundberg & Lee, 2017a). Model of an Idealized Moist Atmosphere (MiMA) is documented by Jucker and Gerber (2017) and Garfinkel et al. (2020), maintained at https://github.com/mjucker/MiMA and available at https://doi.org/10.5281/zenodo.3984605. The model code, forpy coupling code, trained ANNs, run parameters, and modified configuration for MiMA are available at https://doi.org/10.5281/zenodo.5533166. The coupling library, forpy, developed and maintained by Elias Rabel is well documented and available at https://github.com/ylikx/forpy (Rabel et al., 2018).

Acknowledgments
We thank Joan Alexander and Pedram Hassanzadeh for useful discussions. Z. I. Espinosa and A. Sheshadri acknowledge support from the National Science Foundation (NSF) through grant OAC-2004492. E. P. Gerber acknowledges support from the NSF through grant OAC-2004572. K. J. DallaSanta was funded by the NASA Postdoctoral Program at the Goddard Institute for Space Studies. This research received support by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program. We thank two anonymous reviewers for their constructive feedback, which greatly improved the manuscript.

References

Alexander, M. J., Bacmeister, J., Ern, M., Gisinger, S., Hoffmann, L., Holt, L., et al. (2019). *Seeking new quantitative constraints on orographic gravity wave stress and drag to satisfy emerging needs in seasonal-to-subseasonal and climate prediction*. Retrieved from http://www.issibern.ch/teams/consonorogravity/#

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

Antonita, T. M., Ramkumar, G., Kumar, K. K., Appu, K., & Nambhoodiri, K. (2007). A quantitative study on the role of gravity waves in driving the tropical Stratospheric Semiannual Oscillation. *Journal of Geophysical Research*, *112*(D12), D12115. https://doi.org/10.1029/2006jd008250

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. https://doi.org/10.1029/2018gl078510

Bretherton, F. P. (1969). Momentum transport by gravity waves. *Quarterly Journal of the Royal Meteorological Society*, *95*(404), 213–243. https://doi.org/10.1002/qj.49709540402

Butchart, N., Anstey, J., Hamilton, K., Osprey, S., McLandress, C., Bushell, A., et al. (2018). Overview of experiment design and comparison of models participating in phase 1 of the SPARC Quasi-Biennial Oscillation initiative (QBOi). *Geoscientific Model Development*, *11*(3), 1009–1032. https://doi.org/10.5194/gmd-11-1009-2018

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

Cohen, N. Y., Gerber, E. P., & Oliver Bühler, E. P. (2013). Compensation between resolved and unresolved wave driving in the stratosphere: Implications for downward control. *Journal of the Atmospheric Sciences*, *70*(12), 3780–3798. https://doi.org/10.1175/jas-d-12-0346.1

DallaSanta, K., Gerber, E. P., & Toohey, M. (2019). The circulation response to volcanic eruptions: The key roles of stratospheric warming and eddy interactions. *Journal of Climate*, *32*(4), 1101–1120. https://doi.org/10.1175/jcli-d-18-0099.1

DallaSanta, K., Orbe, C., Rind, D., Nazarenko, L., & Jonas, J. (2021). Dynamical and trace gas responses of the Quasi-Biennial Oscillation to increased CO_2. *Journal of Geophysical Research: Atmospheres*, *126*(6), e2020JD034151. https://doi.org/10.1029/2020jd034151

Donner, L. J., Wyman, B. L., Hemler, R. S., Horowitz, L. W., Ming, Y., Zhao, M., et al. (2011). The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component AM3 of the GFDL global coupled model CM3. *Journal of Climate*, *24*(13), 3484–3519. https://doi.org/10.1175/2011jcli3955.1

Espinosa, Z. I. (2021). Learning-GWD-with-MiMA [Computer Software]. Zenodo. https://doi.org/10.5281/zenodo.4428931

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

Fritts, D. C., & Nastrom, G. D. (1992). Sources of mesoscale variability of gravity waves. Part II: Frontal, convective, and jet stream excitation. *Journal of the Atmospheric Sciences*, *49*(2), 111–127. https://doi.org/10.1175/1520-0469(1992)049<0111:somvog>2.0.co;2

Garcia, R. R., Smith, A. K., Kinnison, D. E., Cámara, Á. d. l., & Murphy, D. J. (2017). Modification of the gravity wave parameterization in the Whole Atmosphere Community Climate Model: Motivation and results. *Journal of the Atmospheric Sciences*, *74*(1), 275–291. https://doi.org/10.1175/jas-d-16-0104.1

Garfinkel, C. I., White, I., Gerber, E. P., Jucker, M., & Erez, M. (2020). The building blocks of Northern Hemisphere wintertime stationary waves. *Journal of Climate*, *33*(13), 5611–5633. https://doi.org/10.1175/jcli-d-19-0181.1

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

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*(11), 5742–5751. https://doi.org/10.1029/2018gl078202

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256).

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

Kang, M.-J., Chun, H.-Y., Kim, Y.-H., Preusse, P., & Ern, M. (2018). Momentum flux of convective gravity waves derived from an offline gravity wave parameterization. Part II: Impacts on the quasi-biennial oscillation. *Journal of the Atmospheric Sciences*, *75*(11), 3753–3775. https://doi.org/10.1175/jas-d-18-0094.1

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.

Limpasuvan, V., Richter, J. H., Orsolini, Y. J., Stordal, F., & Kvissel, O.-K. (2012). The roles of planetary and gravity waves during a major stratospheric sudden warming as characterized in WACCM. *Journal of Atmospheric and Solar-Terrestrial Physics*, *78*, 84–98. https://doi.org/10.1016/j.jastp.2011.03.004

Lindgren, E. A., Sheshadri, A., Podglajen, A., & Carver, R. W. (2020). Seasonal and latitudinal variability of the gravity wave spectrum in the lower stratosphere. *Journal of Geophysical Research: Atmospheres*, *125*(18), e2020JD032850. https://doi.org/10.1029/2020jd032850

Lundberg, S. M., & Lee, S.-I. (2017a). A unified approach to interpreting model predictions. In I. Guyon, et al. (eds.). In *Advances in neural information processing systems*, (Vol. 30, pp. 4765–4774). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Lundberg, S. M., & Lee, S.-I. (2017b). A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, (pp. 4765–4774).

Match, A., & Fueglistaler, S. (2021). Large internal variability dominates over global warming signal in observed lower stratospheric QBO amplitude. *Journal of Climate*, *34*(24), 9823–9836. https://doi.org/10.1175/jcli-d-21-0270.1

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

Ott, J., Pritchard, M., Best, N., Linstead, E., Curcic, M., & Baldi, P. (2020). A Fortran-Keras deep learning bridge for scientific computing. *Scientific Programming*, *2020*.

Palmer, T., Shutts, G., & Swinbank, R. (1986). Alleviation of a systematic westerly bias in general circulation and numerical weather prediction models through an orographic gravity wave drag parametrization. *Quarterly Journal of the Royal Meteorological Society*, *112*(474), 1001–1039. https://doi.org/10.1002/qj.49711247406

Podglajen, A., Hertzog, A., Plougonven, R., & Legras, B. (2016). Lagrangian temperature and vertical velocity fluctuations due to gravity waves in the lower stratosphere. *Geophysical Research Letters*, *43*(7), 3543–3553. https://doi.org/10.1002/2016gl068148

Polichtchouk, I., Shepherd, T. G., & Byrne, N. J. (2018). Impact of parametrized nonorographic gravity wave drag on stratosphere-troposphere coupling in the northern and southern hemispheres. *Geophysical Research Letters*, *45*(16), 8612–8618. https://doi.org/10.1029/2018gl078981

Rabel, E., Ruger, R., Govoni, M., & Ehlert, S. (2018). *Forpy: A library for fortran-python interoperability*. Retrieved from https://github.com/ylikx/forpy

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Remmler, S., Hickel, S., Fruman, M. D., & Achatz, U. (2015). Validation of large-eddy simulation methods for gravity wave breaking. *Journal of the Atmospheric Sciences*, *72*(9), 3537–3562. https://doi.org/10.1175/jas-d-14-0321.1

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. (2020). Response of the Quasi-Biennial Oscillation to a warming climate in global climate models. *Quarterly Journal of the Royal Meteorological Society*, 1–29.

Richter, J. H., Sassi, F., & Garcia, R. R. (2010). Toward a physically based gravity wave source parameterization in a general circulation model. *Journal of the Atmospheric Sciences*, *67*(1), 136–156. https://doi.org/10.1175/2009jas3112.1

Roh, S., & Song, H.-J. (2020). Evaluation of neural network emulations for radiation parameterization in cloud resolving model. *Geophysical Research Letters*, *47*(21), e2020GL089444. https://doi.org/10.1029/2020gl089444

Sato, K., & Hirano, S. (2019). The climatology of the Brewer–Dobson circulation and the contribution of gravity waves. *Atmospheric Chemistry and Physics*, *19*(7), 4517–4539. https://doi.org/10.5194/acp-19-4517-2019

Schirber, S. (2015). Influence of ENSO on the QBO: Results from an ensemble of idealized simulations. *Journal of Geophysical Research: Atmospheres*, *120*(3), 1109–1122. https://doi.org/10.1002/2014jd022460

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). *Learning important features through propagating activation differences*. arXiv preprint arXiv:1704.02685.

Sigmond, M., & Scinocca, J. F. (2010). The influence of the basic state on the Northern Hemisphere circulation response to climate change. *Journal of Climate*, *23*(6), 1434–1446. https://doi.org/10.1175/2009jcli3167.1