Machine Learning Emulation of Parameterized Gravity Wave Momentum Fluxes in an Atmospheric Global Climate Model

Zachary I. Espinosa[1], Aditi Sheshadri[1], Gerald R. Cain [2], Edwin P. Gerber[3], 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

Key Points:

- We train a machine learning model to emulate a parameterization of gravity wave momentum transport
- The model reproduces the key features of the physics-based gravity wave parameterization
- The horizontal winds are the primary input features used by the model to generate gravity wave drag

Corresponding author: Zachary Espinosa, `zespinos@stanford.edu`

Abstract

We present a novel, single-column gravity wave parameterization (GWP) that uses machine learning to emulate a physics-based GWP. An artificial neural network (ANN) is trained with output from an idealized atmospheric model and tested in an offline environment, illustrating that an ANN can learn the salient features of gravity wave momentum transport directly from resolved flow variables. We demonstrate that when trained on the westward phase of the Quasi-Biennial Oscillation, the ANN can skillfully generate the momentum fluxes associated with the eastward phase. We also show that the meridional and zonal wind components are the only flow variables necessary to predict horizontal momentum fluxes with a globally and temporally averaged R^2 value over 0.8. State-of-the-art GWPs are severely limited by computational constraints and a scarcity of observations for validation. This work constitutes a significant step towards obtaining observationally validated, computationally efficient GWPs in global climate models.

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 and a lack of observations, global climate models “parameterize” or crudely estimate the effect of GWs on the large-scale flow. Current climate predictions are sensitive to uncertainties in these representations, particularly at the regional scale. Here, we present a novel approach to parameterizing GWs by training a neural network to emulate an existing gravity wave parameterization in a global climate model. This approach represents an appealing technique to build data-driven gravity wave schemes that can reduce existing uncertainties.

1 Introduction

Atmospheric gravity waves (GWs) play a leading-order role in driving middle atmospheric circulation, structure, and variability (Fritts & Alexander, 2003). By transporting momentum, GWs impact the mean climatology (Bretherton, 1969; Sato & Hirano, 2019). They are also critical for variability, impacting the jet stream and storm tracks (Fritts & Nastrom, 1992), stratospheric dynamics (Antonita et al., 2007; Kang et al., 2018; Limpasuvan et al., 2012), and processes such as stratospheric cloud formation (Hoffmann et al., 2017; S. Alexander et al., 2013). Since much of the wave spectrum (10^1 to 10^5 kilometers) 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).

However, realistic representation of these effects in numerical models remains challenging for several reasons: i) The absolute magnitude of GW momentum flux is not particularly well-constrained by observational or intermodel studies (Geller et al., 2013). ii) GWs are generated by a variety of sources, including orography, convection, and frontogenesis (Fritts & Alexander, 2003), but the representation of their sources is not uniform among models. iii) For a given source, the details of the GWP can vary greatly between models (Butchart et al., 2018), and even small changes within the same model can lead to diverging regional climate projections (Schirber, 2015). iv) The horizontal propagation of GWs is usually neglected in parameterizations, which is nonphysical and has an impact on the middle atmosphere (Xu et al., 2017). v) There tends to be a compensation between resolved and unresolved waves (Cohen et al., 2013), complicating observational and intermodel comparisons.

Despite these limitations, GWPs can be optimized towards maximizing global forecast skill scores (Alexander et al., 2019) or "tuned" to reduce climatological biases (Garcia et al., 2017). However, these limitations become apparent in simulations of future climate. Projections of the tropospheric and stratospheric circulations' response to anthropogenic forcing are sensitive to uncertainties in GWPs (Sigmond & Scinocca, 2010; Polichtchouk et al., 2018). Such limitations indicate the need for the continued development of GWPs, preferably with observational validation, computational efficiency, and minimal bias stemming from underlying physical assumptions.

An alternative approach to physics-based parameterization has emerged in the form of machine learning. For atmospheric sciences, machine learning has been employed to parameterize processes such as convection (Rasp et al., 2018; Gentine et al., 2018) and radiation (Brenowitz & Bretherton, 2018; Roh & Song, 2020), among other examples. These applications are particularly valuable when algorithms derived from first principles (e.g., advection on the sphere) cannot be defined with great success. GWPs are thus ripe for investigation with machine learning techniques, and relatively little work has been done in this area. Matsuoka et al. (2020) made the important demonstration that a convolutional neural network can estimate the GW structure over Hokkaido, Japan when trained on high-resolution reanalysis data.

Here, we perform a novel investigation into the efficacy of machine learning as a GWP. We demonstrate that the drag due to breaking GWs can be faithfully represented using an artificial neural network (ANN) that is trained using data from a global atmospheric model embedded with an existing GWP. The machine learning model, which we call WaveNet, differs from the Matsuoka et al. (2020) approach in that it is trained using output from an existing GWP, generates GW momentum tendencies rather than wind anomalies, and provides global coverage. The value of a data-driven GWP lies in its computational efficiency after training and its ability to be trained and evaluated using an arbitrary amount of input data, which may include observations, reanalysis, and targeted integrations (e.g., at ultra-high resolutions). The results presented here are a natural first step towards developing a computationally efficient, three-dimensional, observationally validated GWP.

2 Data

We train the ANN to emulate the M. Alexander and Dunkerton (1999) GWP as incorporated into an atmospheric model of intermediate complexity, the Model of an Idealized Moist Atmosphere (MiMA; Jucker and Gerber (2017)). This choice of GWP and model were based on two factors. First, it is desirable to use a model and GWP that produce somewhat realistic GW behavior. 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). It also produces a realistic representation of stratospheric variability (i.e., the frequency and intensity of sudden stratospheric warmings and a self-generated Quasi-Biennial Oscillation (QBO)). Second, it is advantageous to limit additional degrees of freedom that could cause underfitting in the ANN. MiMA's key simplification relative to a comprehensive atmospheric model lies in its idealized treatment of the hydrological cycle, its lower boundary (a purely thermodynamic, or slab ocean), and the absence of cloud and aerosol processes.

MiMA is integrated with T42 spectral resolution (triangular truncation at wavenumber 42, roughly equivalent to a 2.8-degree grid) in the horizontal and 40 vertical levels, with a model lid at 0.18 hPa and 23 levels above 100 hPa. Following Garfinkel et al. (2020), its simple thermodynamic ocean includes a crude parameterization of oceanic heat transport specified by steady heat flux within the oceanic layer, often referred to as a "Q-flux".

The implementation of the M. Alexander and Dunkerton (1999) GWP, hereafter referred to as the AD99 scheme, is based on its formulation within the GFDL Flexible Modeling System. 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 no momentum flux escapes the model lid. Here, all momentum that reaches 0.85 hPa is uniformly distributed to levels layers above the stratopause. The scheme 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 width and momentum stress were optimized to simulate the QBO, but still provide a reasonable representation of waves in the extratropics. The broad spectrum ensures that there is a rich source of GWs at MiMA's lower boundary, and the amplitude and variability of the extratropical polar vortices are well captured in the simulation. The momentum associated with each wave is deposited at its linear breaking level. The intermittency (i.e., highly skewed distribution) 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. Such fixed source schemes are widely employed in atmospheric models (Butchart et al., 2018), but are highly simplified relative to real GW sources.

We utilized five years of one integration of MiMA, yielding around 12 million training samples per year, with one year of six-hourly data representing approximately 20 Gb. The second year of output is used for training, while all others are reserved for testing. Not all runs utilize the full year of training data, and discrepancies are specified per experiment. All training data are standardized by removing the mean and scaling to unit variance, calculated for each variable across each pressure level. Test data was similarly standardized using the mean and variance calculated for training data. Output variables were inverse transformed before presentation.

3 Neural Network Architecture

An ANN is a computing system of interconnected layers of computational nodes. Each node is comprised of a linear component, which adds a bias parameter to the inner product of a feature vector and a learnable weight vector. A nonlinear component then maps the linear output to an activation function. The resulting scalar from each node in a layer is passed as input to each node in the subsequent layer. With incremental weight and bias adjustments, an ANN attempts to approximate nonlinear relationships between input and output features. The first layer, or input layer, of WaveNet accepts a stacked vector representing a single vertical column of resolved flow variable output from MiMA. The last layer, or output layer, produces a stacked vector of gravity wave drag (GWD) generated by the ANN for a single vertical column (Table S1). The layers between the first and last layers are called hidden layers. We trained two ANNs, one to generate zonal drag and another to generate meridional drag. Both ANNs contain four hidden layers, each with 256 nodes. The fourth hidden layer splits into 33 branches, one for each nontrivial vertical level (the first seven layers are below the source level). We allow the network to use resolved flow variables from below the source level since this may relate to the drag above. Each branch contains four pressure-level specific hidden layers containing 256, 128, 64, and 32 nodes. The final pressure-level specific layer feeds into an output layer. The result of each node in the output layer does not pass through an activation function and produces a GWD value for the vertical column (Figure S1). For all other nodes, we use the Rectified Linear Unit (ReLU). In total, this ANN architecture contains 3,848,481 trainable parameters when using the full set of resolved flow variables and shifts slightly, with a lower limit of 3,806,753, when training with a subset of resolved flow variables. We did not perform an analysis of performance sensitivity to the number of learnable parameters. Rather, we followed standard literature suggesting that with an abundance of data available, deeper neural networks generally produce better scores than shallow networks (Liang & Srikant, 2016).

During training, the ANN attempts to minimize the loss by incrementally nudging each trainable parameter by a scaled version of the gradient of the loss with respect to that parameter. The loss function - in our case, the logcosh error - is computed for a minibatch of 1,024 training samples that are drawn from a pseudo-shuffled training dataset. The logcosh error is defined as

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

where n is the size of the dataset and y_i and y_i^p are the i^{th} truth and prediction, respectively. 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 a single pass through the training data. 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); however, initialization between runs proved to have no impact on the final results between training events.

4 Results

4.1 Evaluation of ANN Predictions

We start by training WaveNet on one year of MiMA output and testing it on the three years proceeding and one year preceding the training period. All subsequent analyses are completed using our best performing networks and the full vertical column of resolved flow variables, unless otherwise noted. Figure 1 shows strong similarities on a global scale between the zonal and meridional GWD generated by WaveNet and AD99 at 10 hPa (panels a through f) and 100 hPa (panels g through l) for a single time step. Similar similarities are seen at all vertical layers and across time (Movie S1 and S2). All reported tests are conducted “offline”, such that WaveNet is not directly coupled to MiMA (i.e. WaveNet’s output is not used by MiMA to generate data at subsequent time steps).

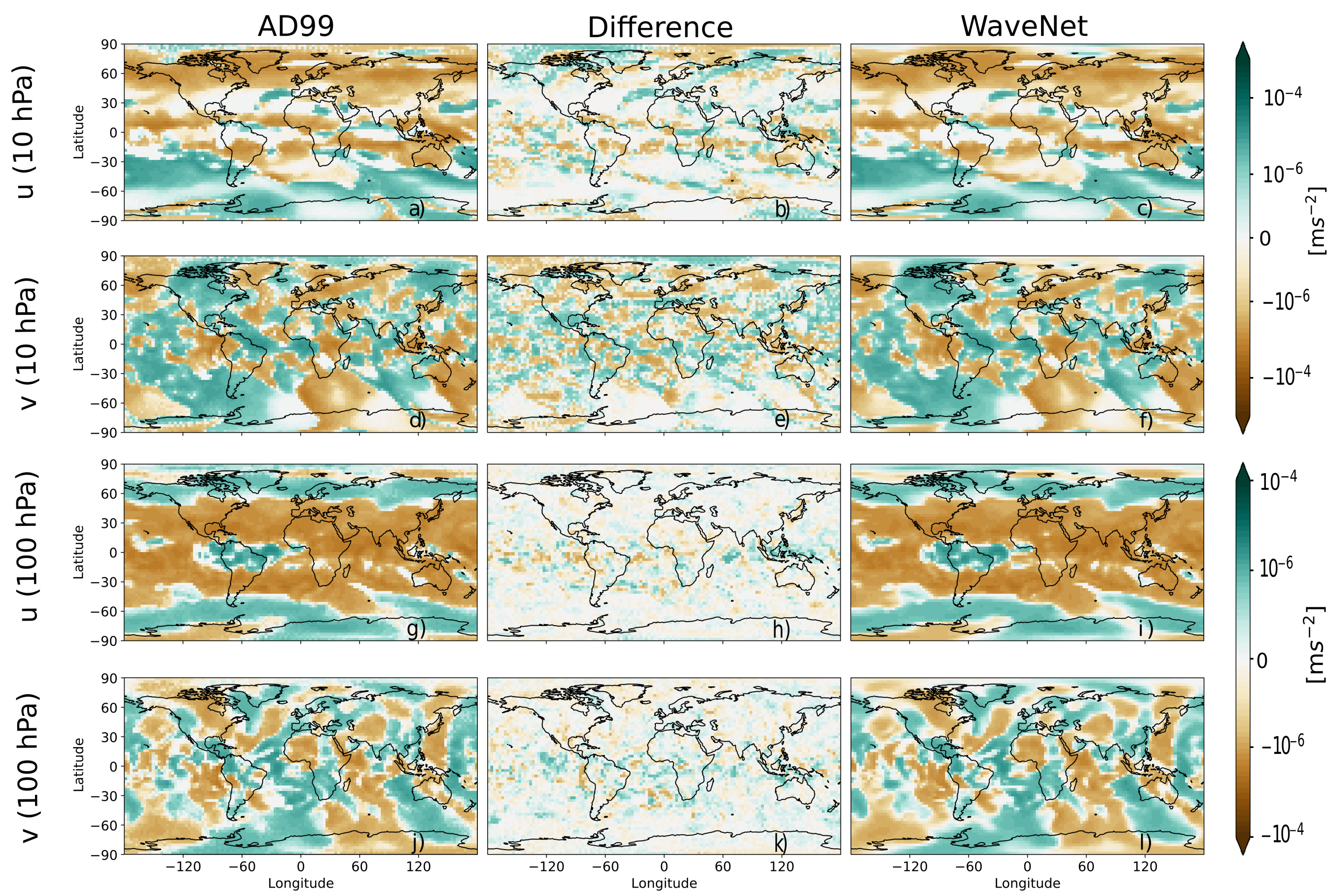

Figure 1. Latitude-longitude snapshot of zonal and meridional AD99 generated drag (panels a, d, g, and j), the corresponding ANN predictions in an offline test (panels c, f, i, and l), and their difference (panels b, e, h, and k) at model level 13 (10 hPa) and 23 (100 hPa) for one time step in the test set.

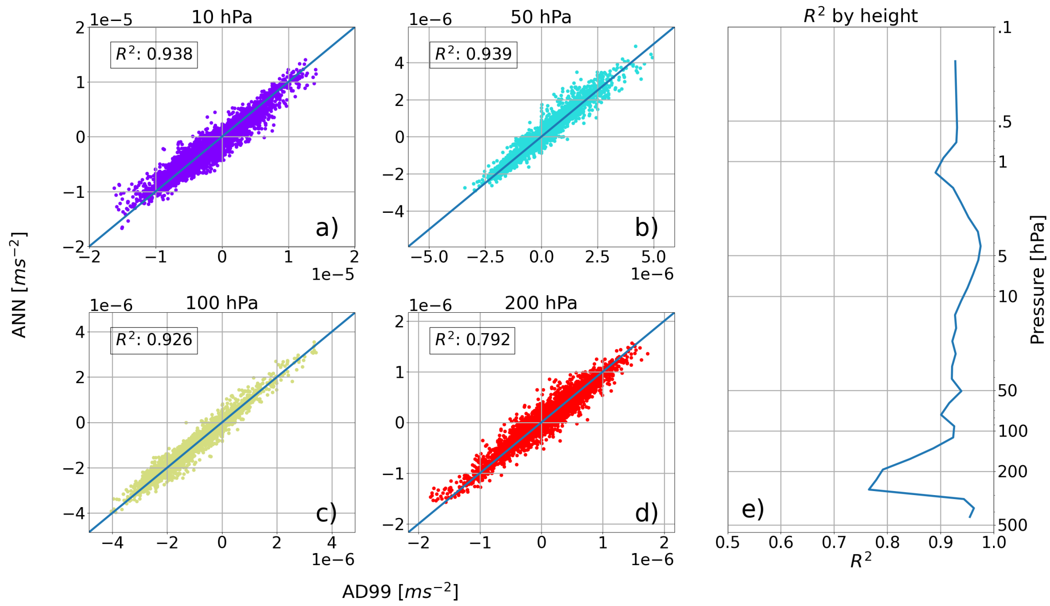

Figure 2. Panels a) through d) show ANN zonal predictions versus AD99 truth at 10 hPa, 50 hPa, 100 hPa and 200 hPa, respectively, for 10k samples in the test set. Panel e) shows pressure versus horizontal and time averaged R^2 values generated from one year of test data for zonal predictions.

To evaluate the quality of predictions, we calculate the R^2 coefficient of determination averaged over time and horizontal dimensions for each vertical level (Figure 2). R^2 is defined as one minus the proportion of the sum of squares of residuals to the total sum of squares (this is also equal to the square of the correlation coefficient):

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where y_i and f_i are the AD99 and WaveNet generated drag for the i^{th} sample, respectively, and $\bar{y}$ is the globally and temporally averaged AD99 generated drag. Figure 2 shows that the ANN can skillfully generate GWD similar to that generated by AD99. WaveNet produces R^2 values for its zonal and meridional predictions averaged across space and time of .92 and .85, respectively. For all metrics and across experiments, WaveNet performs better on zonal tendencies than meridional tendencies (Figure S2). The variance of meridional GWD generated by AD99 is smaller in absolute magnitude than the variance of zonal GWD but greater relative to its mean at all pressure levels. As a result, meridional drag is likely more difficult to learn.

In order to assess how well WaveNet captures GWD associated with large-scale circulation, we analyze the vertical profile of equatorial drag tendencies (Figure 3). WaveNet is trained on one year of global data (months 12-24), containing the westward phase of the QBO. The ANN captures the changes in GW driving associated with the eastward phases preceding and proceeding the training period. While this is an offline test, this

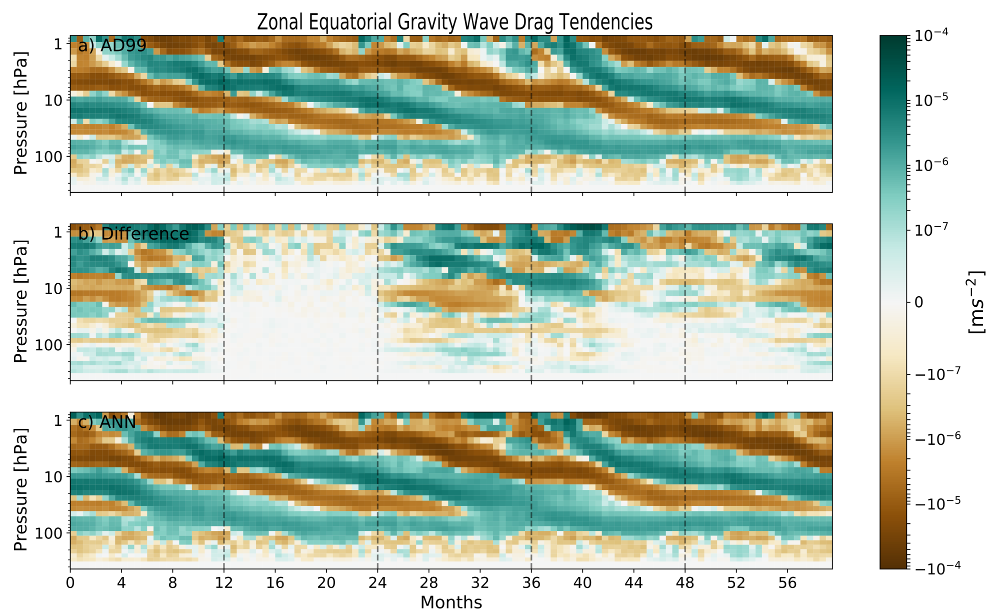

Figure 3. Pressure-Time profile of zonal mean drag 15-day tendencies at 0° latitude for zonal AD99 generated GW drag (a), the corresponding ANN predictions (c), and their difference (b). Vertical dashed lines separate years. The ANN is trained on months 12-24 and tested on all other months. The westward (brown) and eastward (green) bands correspond to drag associated with breaking GWs in opposite phases of the QBO.

result suggests that WaveNet can generalize outside of its training sample. That is, it can transfer learning from other regions (horizontally or vertically) containing eastward wind samples to a region where it has not experienced similar samples during the training period. While comparison between the top and bottom panels of Figure 3 shows that WaveNet can capture the gross features of the "out-of-sample" periods, differences between AD99 and WaveNet are larger at these times, as seen in the middle panel. Notably, the errors decrease around month 48, when the winds are more similar in structure to the training period. This suggests more varied training data may improve WaveNet's performance. Nonetheless, this result is reassuring with respect to WaveNet's potential to globally generalize from regional observations or datasets generated from high-resolution simulations, a subject of future studies. Furthermore, it suggests that the ANN does not trivially depend on the source function's artificial behavior. A similar analysis performed at 60N reveals that WaveNet captures the seasonal cycle of GWD associated with the polar vortex, further supporting the claim that WaveNet can capture large-scale circulation patterns (Figure S3).

4.2 Interpretability of ANN

Although ANNs have achieved great success in a range of applications, their lack of interpretability has become a significant obstacle to their widespread acceptance and made them generally unsuitable for conceptual model building. Here, WaveNet is trained using ten distinct subsets of the full resolved flow variable set (Figure 4ab) to determine which features are most critical. Figure 4a and 4b show horizontally and temporally averaged R^2 values by height per feature set for zonal and meridional drag predictions. We conclude that the zonal and meridional wind components are the only flow variables necessary to predict horizontal drag with a globally and temporally averaged R^2 value over 0.8. Additional training features mildly improve performance, with varied effects at different pressure levels. Note that with the exclusion of latitude and longitude features, the general performance of WaveNet does not significantly drop. These results are in agreement with those in subsection 4.1, in that the ANN is learning to emulate GW dissipation rather than climatological GWD properties or a function of latitude. For these experiments, the total number of trainable parameters varies by roughly 1.0%, with fewer input features corresponding to fewer parameters. This variance did not impact the relative performance of each experiment.

As a preliminary analysis of the role of horizontal wind components in predictions, we calculate Shapley Additive Explanations (SHAP; Lundberg and Lee (2017)). SHAP values represent the effect a feature has on the model's prediction if that feature is included in the input. To compute SHAP values, the model is retrained on all feature subsets S $\subseteq$ F, where F is the set of all features. The SHAP value for a feature is then 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, 2017). This approximation technique avoids retraining the network N times, where N is the power set of F. Unsurprisingly, the results show 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 (Figure S4). 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, e.g., observations from superpressure balloons (Podglajen et al., 2016; Lindgren et al., 2020).

4.3 Sensitivity to Amount of Training Data

To understand how performance degrades as less training data is made available to the ANN, we incrementally decrease the number of training samples from one year to one day (Figure 4cd). To account for the effects of seasonality, for each test, we sample uniformly in space and time from one year of training data to generate a subset with coverage of the entire seasonal cycle. While more data generally leads to better performance, we find that one fourth of one year of data (equivalent to three months or roughly 2.9 million training samples) is sufficient to learn most of the salient features of AD99.

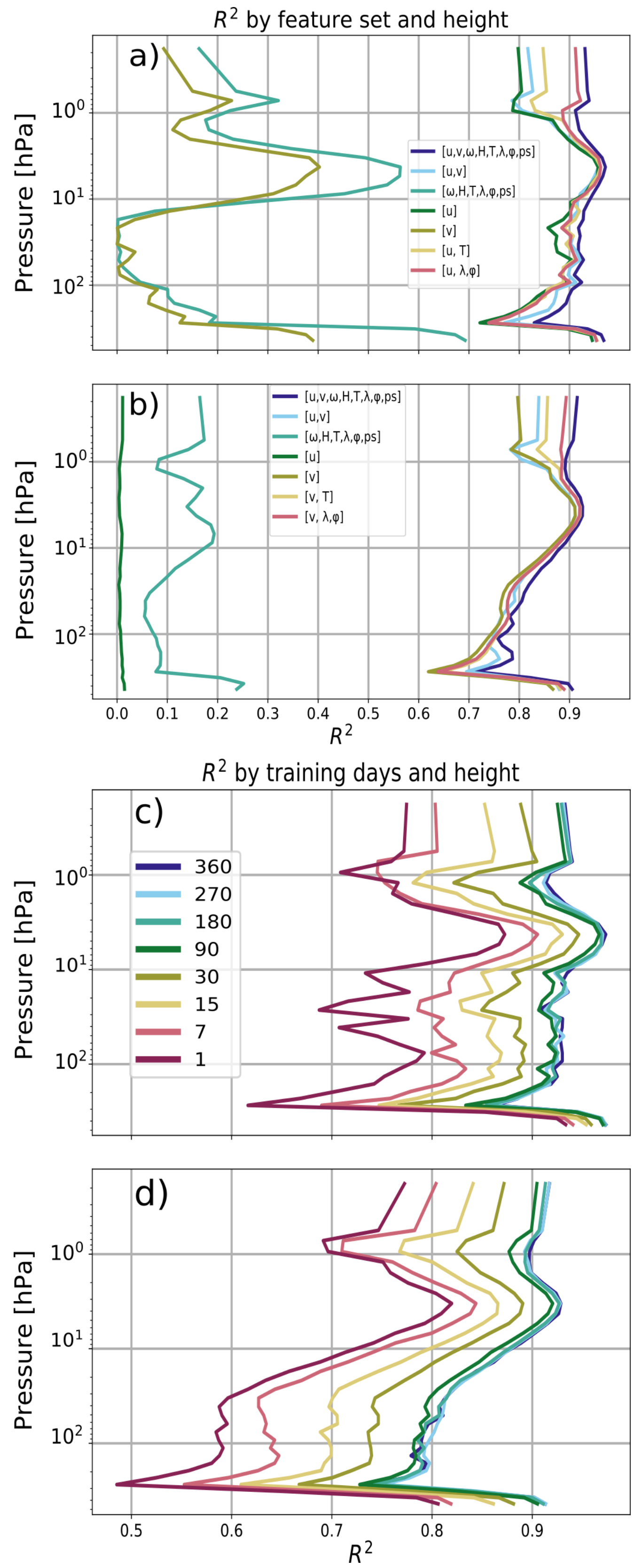

Figure 4. Panels a) through d) show horizontally and temporally averaged R^2 values computed for each model pressure level for either each feature experiment (a and b) or for each data availability experiment (c and d).

This suggests the promise of a data-driven GWP to be trained and validated on observational datasets of similar resolution and size, as well as to generalize to simulations that more realistically model the physics of GW generation, propagation, and dissipation. Additionally, given the deep architecture of WaveNet, that no effort was made to improve performance at each duration, and that it is known that deep networks require large training sets, three months can be regarded as a threshold beyond which optimal performance is likely for an ANN similar to WaveNet.

5 Discussion and Conclusion

We have demonstrated that an ANN can skillfully learn the salient features of GW momentum transport directly from resolved flow variables. The concept was demonstrated in an idealized setting using the M. Alexander and Dunkerton (1999) GWP in a simplified atmospheric model, MiMA. We have shown that the most important input features for WaveNet's predictions are the horizontal wind components local to the vertical level of prediction. Moreover, WaveNet is skillfully able to reproduce drag associated with the eastward phase of the QBO after being trained on data representing the westward phase. In doing so, we have demonstrated that WaveNet can spatially and temporally generalize. The success seen in this context implies that an approach like WaveNet may open a new avenue by which the advantages of high-resolution GW simulations (Remmler et al., 2015) or observational datasets (Lindgren et al., 2020) can be incorporated into current GCMs.

There are, however, a number of challenges that may emerge before the advantages of an approach like WaveNet can be fully realized in a GCM. First, many studies (e.g., Brenowitz and Bretherton (2018)) have shown that machine learning schemes which perform very well offline, i.e., reproducing the correct tendencies, given the correct model state, do not work as well (or at all) when the scheme is coupled with a GCM in an "online" integration. Second, ANNs do not inherently conserve energy or momentum, and additional assumptions may be made to conserve these quantities: for example, artificially scaling the positive and negative fluxes, or depositing the remaining momentum at pre-determined levels. Third, the lack of interpretability of ANNs may serve as a substantial barrier to their widespread adoption. Additional effort is necessary to consider how WaveNet's behavior may relate to the GW dispersion relations. Fourth, WaveNet is an extremely large network. In order to make coupling WaveNet with a GCM computationally feasible, a cost-performance analysis should be performed to reduce WaveNet's complexity. Finally, a next test is to examine how WaveNet generalizes when trained on regional datasets and orographic and nonorographic GWPs.

Nevertheless, our results suggest that machine learning may represent a powerful alternative to existing GWPs. An approach like WaveNet is naturally suited for data assimilation, and WaveNet may be completely or partially trained and validated using observational datasets. Moreover, existing GWPs ignore horizontal GW propagation due to computational limitations. A machine learning approach such as WaveNet may afford the computational efficiency needed to develop a three-dimensional GWP. Projections of the large-scale climate response to anthropogenic warming are sensitive to uncertainties in existing GWPs. Developing an observationally validated, three-dimensional GWP may more accurately capture the physics of GWs. The approach presented here constitutes a first step toward obtaining such GWPs for global climate prediction.

Acknowledgments

We thank Joan Alexander and Pedram Hassanzadeh for useful discussions. ZIE and AS acknowledge support from the NSF through grant OAC-2004492. EPG acknowledges support from the NSF through grant OAC-2004572. KD was funded by the NASA Postdoctoral Program at the Goddard Institute for Space Studies.

Data Availability

The ANN is built using Keras, a deep learning framework that wraps Tensorflow. All source code is available at `https://doi.org/10.5281/zenodo.4428931`. Training took on the order of 48 hr on a graphical processing unit and varied according to data size and trainable parameters. The implementation of SHAP values is available at `https://github.com/slundberg/shap` and is not maintained or owned by this project group. MiMA is documented by Jucker and Gerber (2017) and Garfinkel et al. (2020). The source code, run parameters, and modified configuration for MiMA are available at: `https://doi.org/10.5281/zenodo.1401407`

References

Alexander, Bacmeister, J., Ern, M., Gisinger, S., Hoffmann, L., Holt, L., … Wright, C. (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., & Dunkerton, T. (1999). A spectral parameterization of mean-flow forcing due to breaking gravity waves. *Journal of the Atmospheric Sciences*, *56*(24), 4167–4182.

Alexander, S., Klekociuk, A., McDonald, A., & Pitts, M. (2013). Quantifying the role of orographic gravity waves on polar stratospheric cloud occurrence in the antarctic and the arctic. *Journal of Geophysical Research: Atmospheres*, *118*(20), 11–493.

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: Atmospheres*, *112*(D12).

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

Bretherton, F. P. (1969). Momentum transport by gravity waves. *Quarterly Journal of the Royal Meteorological Society*, *95*(404), 213–243.

Butchart, N., Anstey, J., Hamilton, K., Osprey, S., McLandress, C., Bushell, A., ... others (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).

Cohen, N. Y., Gerber, & 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.

Donner, L. J., Wyman, B. L., Hemler, R. S., Horowitz, L. W., Ming, Y., Zhao, M., ... others (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.

Fritts, D. C., & Alexander, M. J. (2003). Gravity wave dynamics and effects in the middle atmosphere. *Reviews of geophysics*, *41*(1).

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.

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.

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.

Geller, M. A., Alexander, M. J., Love, P. T., Bacmeister, J., Ern, M., Hertzog, A., ... others (2013). A comparison between gravity wave momentum fluxes in observations and climate models. *Journal of Climate*, *26*(17), 6383–6405.

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.

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).

Hoffmann, L., Spang, R., Orr, A., Alexander, M. J., Holt, L. A., & Stein, O. (2017).

A decadal satellite record of gravity wave activity in the lower stratosphere to study polar stratospheric cloud formation. *Atmospheric Chemistry and Physics*, *17*(4), 2901–2920.

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.

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.

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

Liang, S., & Srikant, R. (2016). Why deep neural networks for function approximation? *arXiv preprint arXiv:1610.04161*.

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.

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.

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

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.

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.

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.

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.

Remmler, S., Hickel, S., Fruman, M. D., & Achatz, U. (2015). Validation of largeeddy simulation methods for gravity wave breaking. *Journal of the Atmospheric Sciences*, *72*(9), 3537–3562.

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.

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

Sato, K., & Hirano, S. (2019). The climatology of the brewer–dobson circulation and the contribution of gravity waves. *Atmos. Chem. Phys*, *19*, 4517–4539.

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.

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.

Xu, X., Wang, Y., Xue, M., & Zhu, K. (2017). Impacts of horizontal propagation of orographic gravity waves on the wave drag in the stratosphere and lower mesosphere. *Journal of Geophysical Research: Atmospheres*, *122*(21), 11–301.

Supporting Information for "Machine Learning Emulation of Parameterized Gravity Wave Momentum Fluxes in an Atmospheric Global Climate Model"

Zachary I. Espinosa[1], Aditi Sheshadri[1], Gerald R. Cain [2], Edwin P. Gerber[3],

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

Contents of this file

Additional Supporting Information (Files uploaded separately)

1. Captions for Movies S1 and S2

Corresponding author: Zachary I. Espinosa, (zespinos@stanford.edu)

Introduction The supporting information includes one table, four figures and two movies. Table S1 shows a list of input variables accepted by the ANN and output variables generated by the ANN. Figure S1 shows a schematic of the ANN architecture. Figure S2 shows ANN meridional predictions versus AD99 truth at four model levels and a globally and temporally averaged R^2-Pressure plot for one year of test data. Figure S3 is a pressure-time profile of zonal mean drag 15-day tendencies at 60N for zonal ANN predictions, the corresponding AD99 truths, and their difference. Figure S4 shows SHAP bar plots of the ten most important meridional and zonal wind features used by WaveNet to generate GWD at 10 hPa and 100 hPa. Movies S1 and S2 are a time series animation of Figure 2 at 10 hPa, for zonal and meridional predictions, respectively.

Movie S1. A latitude-longitude time series of zonal ANN predictions, AD99 truth, and their difference at model level 13 (10 hPa) for half a year of test data. Panels a through c in Figure 1 are a single snapshot of this animation.

Movie S2. A latitude-longitude time series of meridional ANN predictions, AD99 truth, and their difference at model level 13 (10 hPa) for half a year of test data. Panels d through f in Figure 1 are a single snapshot of this animation.

Table S1. List of input variables accepted by the ANN and output variables generated by the ANN. The total input feature vector contains 203 elements, and the output is 33 GWD values. Two networks are trained, one for zonal drag and one for meridional drag.

List of Input and Output Variables Used for ANN			
Input Variables	Vertical Levels	Output Variables	Vertical Levels
Zonal Wind ($\frac{m}{s}$)	40	GW zonal drag ($\frac{m}{s^2}$)	33
Meridional Wind ($\frac{m}{s}$)	40	GW meridional drag ($\frac{m}{s^2}$)	33
Vertical Wind ($\frac{m}{s}$)	40		
Temperature (K)	40		
Height (m)	40		
Latitude (λ)	1		
Longitude (ϕ)	1		
Surface Pressure (hPa)	1		
Size of Stacked Array	203		33

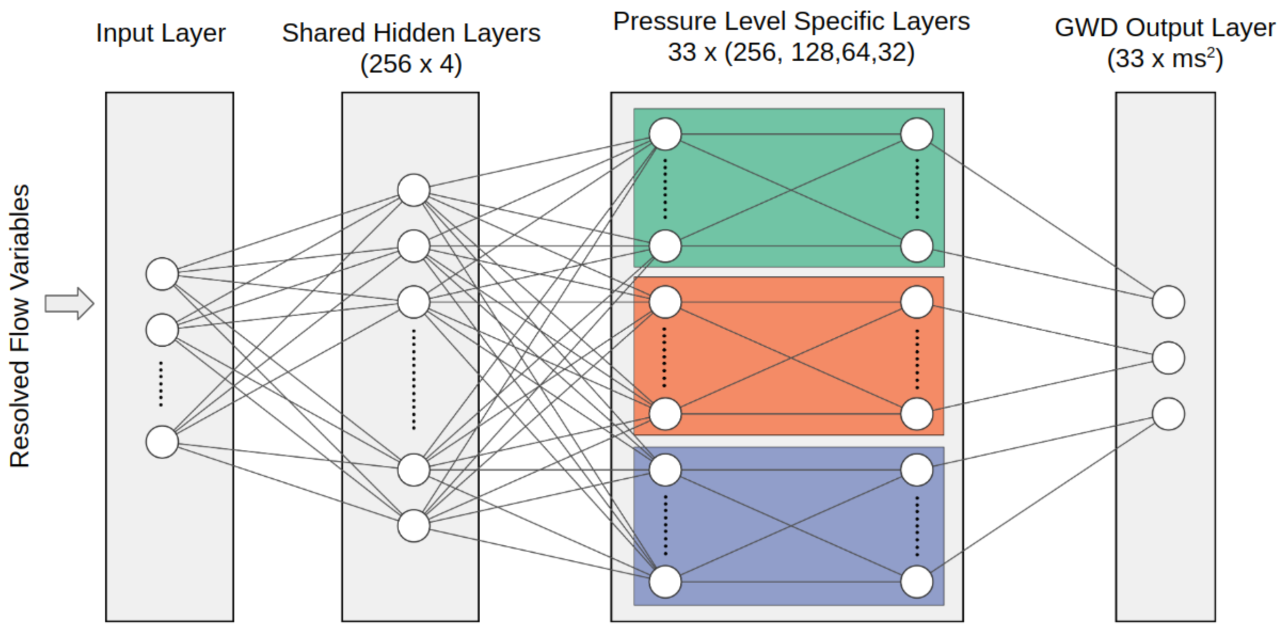

Figure S1. WaveNet contains 4 shared hidden layers, each with 256 neurons. WaveNet then splits into 33 branches (one branch per nontrivial vertical layer) each containing 4 pressure level specific layers with 256, 128, 64, and 32 neurons, respectively. Each branch then outputs a single value corresponding to the gravity wave drag at that vertical layer.

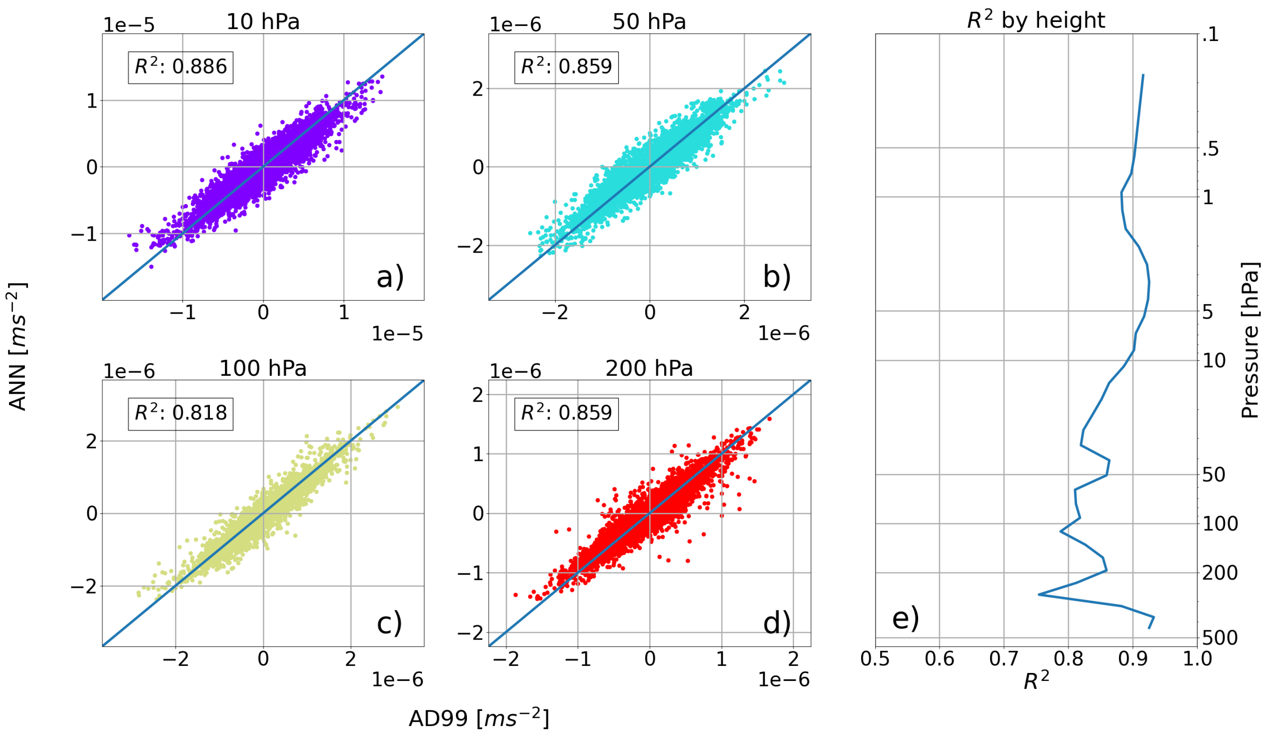

Figure S2. As in Figure 2, but for meridional GWD, panels a) through d) show ANN meridional GWD versus AD99 truth at 10 hPa, 50 hPa, 100 hPa and 200 hPa, respectively, for 10k samples in the test set. Panel e) shows pressure versus horizontally and temporally averaged R^2 values generated from one year of test data for zonal predictions.

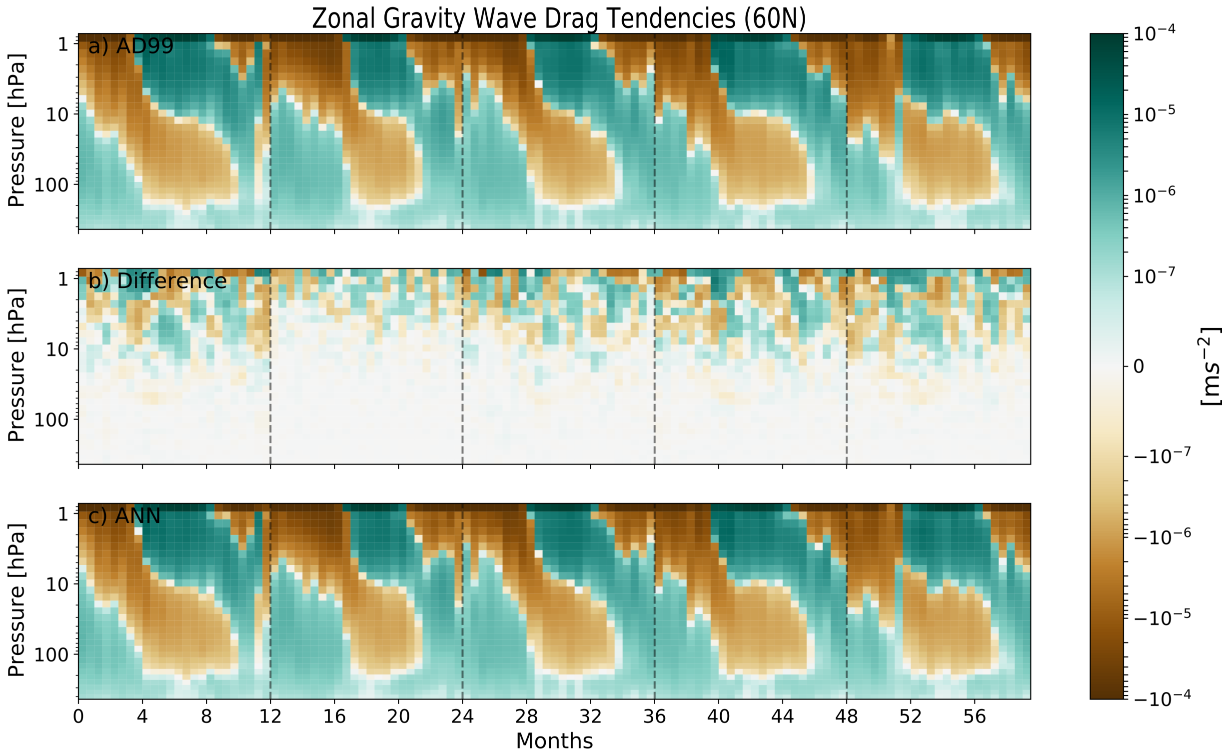

Figure S3. Pressure-Time profile of zonal mean drag 15-day tendencies at $60N$ latitude for zonal AD99 generated GW drag (a), the corresponding ANN predictions (c), and their difference (b). Vertical dashed lines separate years. The ANN is trained on months 12-24, and tested on all other months. The westward (brown) and eastward (green) bands correspond to drag associated with the seasonal cycle of the Polar Vortex.

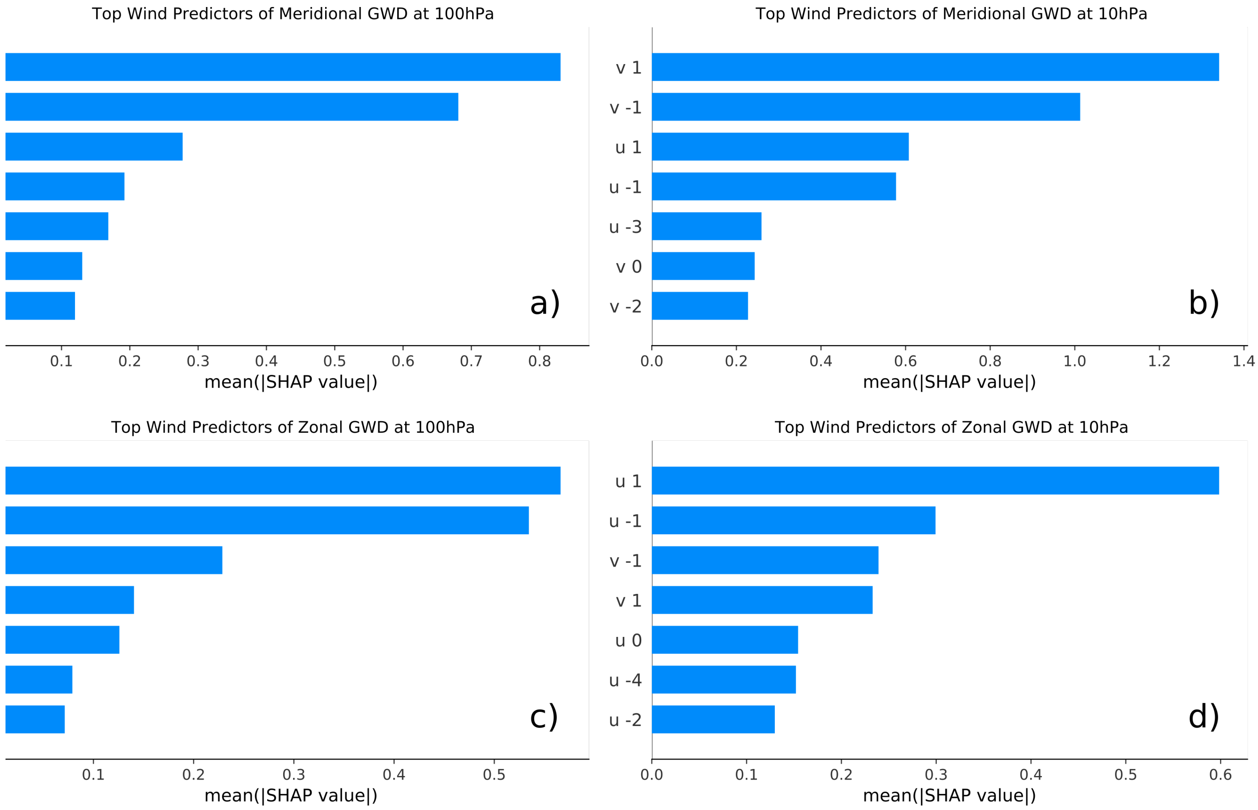

Figure S4. Panels a) through d) are SHAP bar plots of the ten most important meridional (v; panels a and b) and zonal (u; panels c and d) wind features used by WaveNet to generate meridional or zonal GWD at 10 hPa and 100 hPa. The vertical axes' values indicate displacements in vertical pressure levels, with positive and negative values being above and below the level of prediction, respectively (e.g., u -1 indicates zonal wind at the vertical level directly below the level of prediction). The results suggest that vertically local wind fields are the dominant features used by WaveNet to generate GWD.