AGU Word Manuscript Template

Uncertainty Quantification of a Machine Learning Subgrid-Scale Parameterization for Atmospheric Gravity Waves

L. A. Mansfield[1], and A. Sheshadri[1]

[1]Earth System Science, Doerr School of Sustainability, Stanford University, California

Corresponding author: Laura A. Mansfield (lauraman@stanford.edu)

Key Points:

- Using ensembles of neural networks, we learn parametric uncertainties associated with an emulator of a gravity wave parameterization.
- When coupled to the climate model, the ensemble of neural networks reveals increased climate variability.
- Parametric uncertainty dominates the Quasi-Biennial Oscillation statistics, although polar vortex properties remain robust to parameters.

Abstract

Subgrid-scale processes, such as atmospheric gravity waves, play a pivotal role in shaping the Earth's climate but cannot be explicitly resolved in climate models due to limitations on resolution. Instead, subgrid-scale parameterizations are used to capture their effects. Recently, machine learning has emerged as a promising approach to learn parameterizations. In this study, we explore uncertainties associated with a machine learning parameterization for atmospheric gravity waves. Focusing on the uncertainties in the training process (parametric uncertainty), we use an ensemble of neural networks to emulate an existing gravity wave parameterization. We estimate both offline uncertainties in raw neural network output and online uncertainties in climate model output, after the neural networks are coupled. We find that online parametric uncertainty contributes a significant source of uncertainty in climate model output that must be considered when introducing neural network parameterizations. This uncertainty quantification provides valuable insights into the reliability and robustness of machine learning-based gravity wave parameterizations, thus advancing our understanding of their potential applications in climate modeling.

Plain Language Summary

Climate models are unable to resolve processes that vary on length and time scales smaller than the model resolution and timestep. For example, atmospheric gravity waves, which are waves created when winds encounter disturbances to the flow, such as mountains, convection and fronts, can have wavelengths smaller than the spacing between grid cells. Climate models use "parameterizations" to capture the effect of these processes. Machine learning based parameterizations are becoming popular because they can learn relationships purely from data. However, we do not have a good understanding of the uncertainties introduced through machine learning parameterizations. This study estimates uncertainties associated with training a neural network gravity wave parameterization. We explore uncertainties in the neural network output, as well as the uncertainties in the climate model output, when the neural network is used for the gravity wave parameterization.

1 Introduction

1.1. Subgrid-scale parameterizations

Global climate models (GCMs) simulate the entire Earth system by coupling a dynamical core, which numerically solves the primitive equations for atmospheric flow, with other physical components called "subgrid-scale parameterizations". The latter includes dynamical processes occurring on scales smaller than the grid-scale (generally $O(100\ \text{km})$ for a typical GCM; Chen et al., 2021), such as convection and short wavelength gravity waves, and non-dynamical processes, such as radiation, atmospheric chemistry, and cloud and aerosol microphysics. Subgrid-scale parameterizations make up a large portion of the computational cost associated with GCM simulations and sometimes make drastic assumptions for the sake of computational cost, which can introduce additional sources of model uncertainty. This has motivated the demand for faster and/or higher accuracy schemes that use machine learning (ML)/artificial intelligence (AI), which hold out the potential for training on large volumes of training data and performing fast inferences when invoked.

ML-based subgrid-scale parameterizations have demonstrated skill across a wide range of atmospheric processes including convection, clouds, aerosols, radiation and gravity waves (e.g., Brenowitz et al., 2020; Brenowitz & Bretherton, 2019; Chantry et al., 2021; Chevallier et al., 2000; Espinosa et al., 2022; Gentine et al., 2018; Harder et al., 2022; Krasnopolsky & Fox-Rabinovitz, 2006; O'Gorman & Dwyer, 2018; Perkins et al., 2023; Rasp et al., 2018; Ukkonen, 2022; Yu et al., 2023; Yuval et al., 2021; Yuval & O'Gorman, 2020). However, few studies have explored the uncertainties associated with these. Stochastic subgrid-scale parameterizations have been developed by sampling from parametric distributions, learned through neural networks (Guillaumin & Zanna, 2021) and generative adversarial networks (GANs) (Gagne II et al., 2020; Nadiga et al., 2022; Perezhogin et al., 2023). These studies focus on stochastic representations to improve model accuracy since they may better represent scaling properties (Palmer, 2019). Including uncertainty estimates can also be beneficial in assessing the trustworthiness of model predictions (Haynes et al., 2023; McGovern et al., 2022), and has gained some attention in weather and climate prediction studies (e.g., Delaunay & Christensen, 2022; Gagne et al., 2014, 2017; Gordon & Barnes, 2022; Weyn et al., 2021). Here, we explore uncertainty quantification in a machine learning subgrid-scale parameterization (a type of *model uncertainty;* Hawkins & Sutton, 2009; Palmer, 2019), focusing on gravity wave parameterizations.

1.2 Atmospheric Gravity Waves

Atmospheric gravity waves (GWs) are important drivers of middle atmosphere circulation as they transport momentum upwards and away from their sources in the lower troposphere (Fritts & Alexander, 2003). They are forced by perturbations to a stable stratified flow, for instance, orography, convection, and frontogenesis. They propagate primarily in the vertical and, due to the decreasing density in the upper atmosphere, grow in amplitude until reaching a critical level, at which point they break and deposit momentum. This provides a forcing on the mean flow in the middle and upper atmosphere and has a substantial impact on atmospheric circulation, including in driving the Quasi-Biennial Oscillation (QBO) in the equatorial stratosphere (Baldwin et al., 2001) and affecting the occurrence of Sudden Stratospheric Warmings in the polar vortex during winter (Wang & Alexander, 2009), described further in Section 1.3.

GW wavelengths can range from $\mathcal{O}(1\text{ km})$ to $\mathcal{O}(1000\text{ km})$, which presents a challenge for accurate representation in global climate models (GCMs). While the primitive equations do capture GW dynamics, typical GCM resolutions are $\mathcal{O}(100\text{ km})$, resulting in a large portion of the GW spectrum being un- or under-resolved. Parameterizations must be employed to model the impacts of subgrid-scale GWs on the mean flow and are critical for obtaining realistic circulation, for example, to induce a spontaneous QBO (Bushell et al., 2020). Some studies find GW parameterizations to be necessary even in kilometer-scale resolution simulations (Achatz et al., 2023; Polichtchouk et al., 2023), suggesting that the need for accurate parameterizations will persist even as modeling centers move towards high resolution GCMs (or "digital twins"; e.g., Bauer et al., 2021).

1.2.2 Gravity wave parameterizations

GCMs usually make use of both an orographic and a non-orographic GW parameterization to capture their effects. Machine learning alternatives to GW parameterizations have recently gained attention in several forms. Chantry et al. (2021), Espinosa et al. (2022) and Hardiman et al. (2023) present machine learning emulators of existing non-orographic gravity wave schemes, while Dong et al. (2023) and Sun et al. (2023) use machine learning to learn gravity wave momentum fluxes from high resolution simulations.

This study can be viewed as a continuation of the work by Espinosa et al. (2022), which develops an emulator of a non-orographic GW parameterization designed primarily for convectively forced GWs (Alexander & Dunkerton, 1999). Note that this machine learning parameterization is, at best, as accurate as the scheme it aims to emulate and is not significantly faster than the original physics-based scheme, which could be due to coupling of the neural network within a Fortran-based GCM (Cambridge-ICCS, 2023). Rather, this neural network emulator is used as a first step towards probing uncertainties introduced when replacing a gravity wave parameterization with an emulator, when we have a "ground truth" parameterization for reference.

1.3 Gravity wave effects

1.3.1 Quasi-Biennial Oscillation

Gravity waves strongly influence the stratospheric circulation. In the tropical stratosphere, the dominant mode of variability is the Quasi-Biennial Oscillation (QBO), in which the equatorial stratospheric zonal winds alternate between easterly and westerly and descend downwards with time (Gray, 2010). The change in direction is driven by breaking waves across a range of scales (Baldwin et al., 2001; Lindzen & Holton, 1968), with modeling studies suggesting that non-orographic gravity wave parameterizations contribute around half of the forcing required for a simulated QBO (Holt et al., 2020).

In this study, we measure the performance of gravity wave parameterizations through the simulated QBO period and amplitudes at 10 hPa, where the QBO amplitude is generally a maximum (Bushell et al., 2020; Richter et al., 2020). We consider the QBO winds to be defined by the zonal mean zonal winds between $5°S$ and $5°N$. Following Schenzinger et al., (2017), we estimate the period of a QBO cycle by the length between transition times from westward and eastward flow, after applying a 5-month binomial filter to remove high frequency variability. The amplitude is estimated as the absolute maximum of the QBO winds during each cycle.

1.3.2 Stratospheric Polar Vortex

As well as driving the equatorial stratospheric circulation, gravity waves are also influential at high latitudes. Gravity waves affect the stratospheric polar vortex in both hemispheres, as they contribute to the breakdown of the polar vortices, influencing the frequency and properties of Sudden Stratospheric Warmings (SSWs) (Siskind et al., 2007, 2010; Wang & Alexander, 2009; Whiteway et al., 1997; Wright et al., 2010) and the timing of the Spring final warming (Gupta et al., 2021). SSWs are defined as a reversal of the zonal mean zonal winds at 60°N at 10 hPa (Butler et al., 2015) which is followed by large and rapid temperature increases

(>30-40 K) in the polar stratosphere. They occur around 6 times per decade in the Northern hemisphere, but are not common in the Southern hemisphere. In this study, we consider gravity wave parameterization effects on the number of Northern hemisphere SSWs per decade and the timing of the final warming of the Southern hemisphere polar vortex.

2. Uncertainty Quantification

Uncertainties can be categorized into two types: *aleatoric uncertainty* and *epistemic uncertainty* (Hüllermeier & Waegeman, 2021). Aleatoric uncertainty is used to describe the variability in a system that is due to inherently random effects (Haynes et al., 2023; Hüllermeier & Waegeman, 2021). It represents the statistical or stochastic nature of a system, such as flipping a coin or rolling a dice and in ML literature, refers to uncertainty in the data. It includes *internal variability* of the system and *observational uncertainties* in the data. In contrast, epistemic uncertainty is caused by a lack of knowledge about the best model for a system and refers to uncertainty in the model. It includes *structural uncertainties* from the choice of ML architecture, *parametric uncertainties* in estimating of model parameters, and *out-of-sample uncertainties* which arise when predicting outside of the range of the training data.

In this study, we aim to quantify parametric uncertainty, a type of epistemic uncertainty, in an ML-based parameterization for gravity waves. We expect this to also capture out-of-sample uncertainties, i.e., increased uncertainty when generalizing to a situation that lies outside of the training data distribution. For simplicity, we do not estimate aleatoric uncertainty in the training data, and we also do not consider structural uncertainty. Future studies may wish to account for these additional types of uncertainty for a more complete picture. There are several methods that could be used to estimate parametric uncertainty (Abdar et al., 2021). Here, we use an ensemble of deep neural networks or “deep ensembles”, which involves training multiple identical neural networks, each with a different initialization (Lakshminarayanan et al., 2017). Each neural network converges upon slightly different parameters which are then used to predict an ensemble, from which statistics can be obtained. This is a relatively simple approach to implement, although can be costly as it requires repetition during training and evaluation. Deep ensembles have been used in climate model applications for prediction (Weyn et al., 2021), but have not been used for subgrid-scale parameterizations. In this context, deep ensembles could be viewed as a machine learning complement to “perturbed parameter ensembles” (PPE), which involve perturbing physics-based parameters for uncertainty quantification (e.g., Murphy et al., 2007; Sengupta et al., 2021; Sexton et al., 2021).

3 Methods

3.1 Gravity Wave Parameterization Setup

Alexander & Dunkerton (1999; hereafter AD99) present a simple non-orographic, gravity wave parameterization that has been used in various GCMs, including GFDL’s Atmospheric Model 3 (Donner et al., 2011), Isca (Vallis et al., 2018), and MiMA (Jucker & Gerber, 2017). AD99 estimates *gravity wave drag* (GWD) in both the zonal and meridional directions for each level in a column, at each grid-cell and timestep. When coupled into a climate model, gravity wave drag or forcing acts to accelerate or decelerate winds (i.e., it is a wind tendency). As a spectral parameterization, AD99 defines a spectrum of gravity waves at a source level with momentum flux distributed by phase speeds, assumed to follow a Gaussian distribution centered at 0 m/s with half-width 35 m/s. This spectrum of gravity waves propagates upwards until the waves reach the critical level (when the wind speed equals the phase speed of the waves), when breaking occurs and drag is deposited.

3.2 Atmospheric Model Setup

We use an intermediate complexity GCM, a Model of an idealized Moist Atmosphere (MiMA) (Jucker & Gerber, 2017). It is run at spectral resolution T42, corresponding to 64 latitudes by 128 longitudes (approximately 2.8 degrees or 300 km grid spacing at the equator), with 40 model levels. The level top is 0.18 hPa, with a strong dissipating sponge layer in the upper three levels (0.85-0.18 hPa). AD99 is coupled into MiMA with the parameters described above and with a fixed source level defined to be 315 hPa in the tropics and decreasing in height with latitude, roughly in line with the tropopause. The model is run with an advection timestep of 10 minutes and a physics timestep, which includes calling the gravity wave parameterization, of 3 hours.

3.2 Machine Learning Setup

We use the neural network (NN) gravity wave parameterization developed by Espinosa et al. (2022). This is trained on MiMA simulations using the AD99 gravity wave parameterization, described above (Alexander & Dunkerton, 1999). Espinosa et al. (2022) show that the NN emulator, trained on one year of data, achieves an accurate representation of the AD99 scheme both offline and online. For the online tests, Espinosa et al. (2022) replace the original AD99 scheme in MiMA with the NN emulator within MiMA and show that these coupled NN simulations produce a Quasi-Biennial Oscillation consistent with original AD99 simulation. Furthermore, when tested on an out-of-sample climate under 4xCO2 forcing, the NN simulations remained stable and reproduced similar changes to the QBO as the AD99 simulations.

Espinosa et al. (2022) emulate the zonal and meridional GW drag with two independently trained but almost identical fully connected NNs. The inputs to the zonal GW drag network are zonal winds at all levels, u, temperature at all levels, T, surface pressure, p_s, and latitude, λ, and similarly for the meridional GW drag the inputs are meridional winds at all levels, v, T, p_s, and λ. MiMA uses 40 pressure levels, giving a total of 82 inputs into the NN. The architecture consists of four shared hidden layers followed by another four pressure level specific layers (see Supporting Information of Espinosa et al., 2022). The network outputs the zonal/meridional GW drag for all 40 pressure levels. Note that the pressure levels closest the surface always predict zero, where there is no GW drag below the source of the GWs. Although these layers are redundant, we include them because the AD99 gravity wave source level changes with latitude to follow the approximate level of the tropopause. Following Espinosa et al. (2022), we normalize the input and output data to have a zero mean and standard deviation of 1. For the pressure levels below the source level, where all GW drag values are exactly zero and standard deviation is undefined, we fix the outputs to zero. Although we follow the same architecture as Espinosa et al. (2022), there are some software differences in our implementation. Firstly, we opt for PyTorch (Paszke et al., 2019) rather than Keras and TensorFlow (Abadi et al., 2015; Chollet &

others, 2015) for the machine learning library. Secondly, Espinosa et al. (2022) use the forpy software (Rabel, 2019) to call python code in the fortran-based climate model. This resulted in a slow-down of roughly 2.5x when replacing AD99 with the NN emulator. Instead, we use FTorch (Cambridge-ICCS, 2023), a software package that directly calls the existing Torch C++ interface from Fortran resulting in faster inference. We find a 20% slow-down in the NN simulations relative to the AD99 simulations, although we have not explored if this could be optimized further.

In this study, we capture parametric uncertainty of the NN emulator presented in Espinosa et al. (2022) using deep ensembles (Lakshminarayanan et al., 2017). We repeatedly train an ensemble of size 30 independent NNs, each with the same architecture and trained on the same data but with different random seed initializations. The random seed affects the initialization of the NN parameters and the shuffling order of data during training, leading to slightly different parameters when converged. Following Espinosa et al. (2022), we train the NNs with one year of data, selected so that it contains a typical QBO cycle with a period and amplitude similar to the long-term mean period and amplitude. We use the following one year of data for the validation dataset, and the following 20 years are used for the test dataset, requiring 22 years of simulation data in total. Figure 1 shows (a) the QBO zonal winds and (b) the QBO zonal gravity wave drag over this dataset up to year 12.

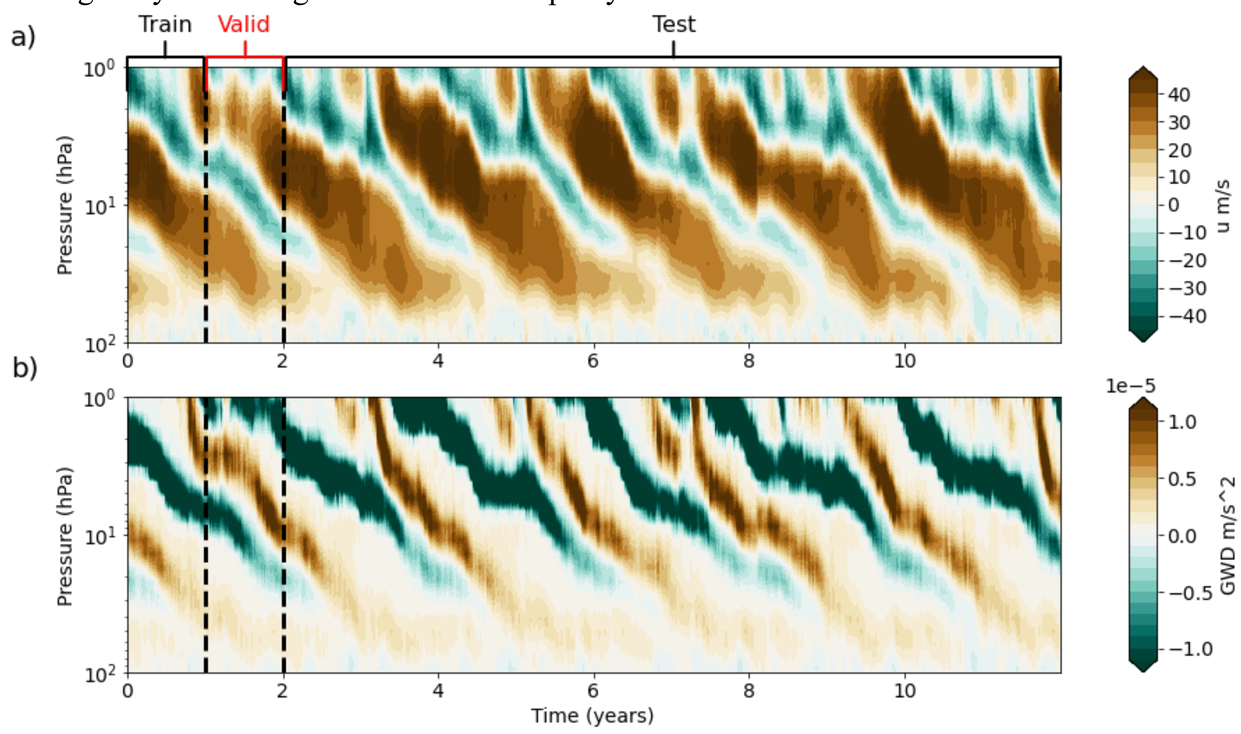

Figure 1 The QBO (a) zonal winds and (b) zonal gravity wave drag for the training, validation, and test dataset.

4 Results

4.1 Offline predictions

Figure 2 shows an example of gravity wave drag (GWD) profiles for a single grid cell close to the equator for a) the zonal component and b) the meridional component, with the black line indicating the ground truth from the AD99 parameterization and the red line indicating the mean prediction across all NN ensemble members. The orange shading represents 1 standard deviation across all ensemble members. Animations showing the evolution of this GWD profile can be found in the Supporting Materials. The NNs agree well on the gravity wave profiles and the ground truth falls within the 1 standard deviation range for across most model levels for the zonal component. The meridional component generally captures the patterns within the profile but is found to be less accurate, even when considering the uncertainty estimates.

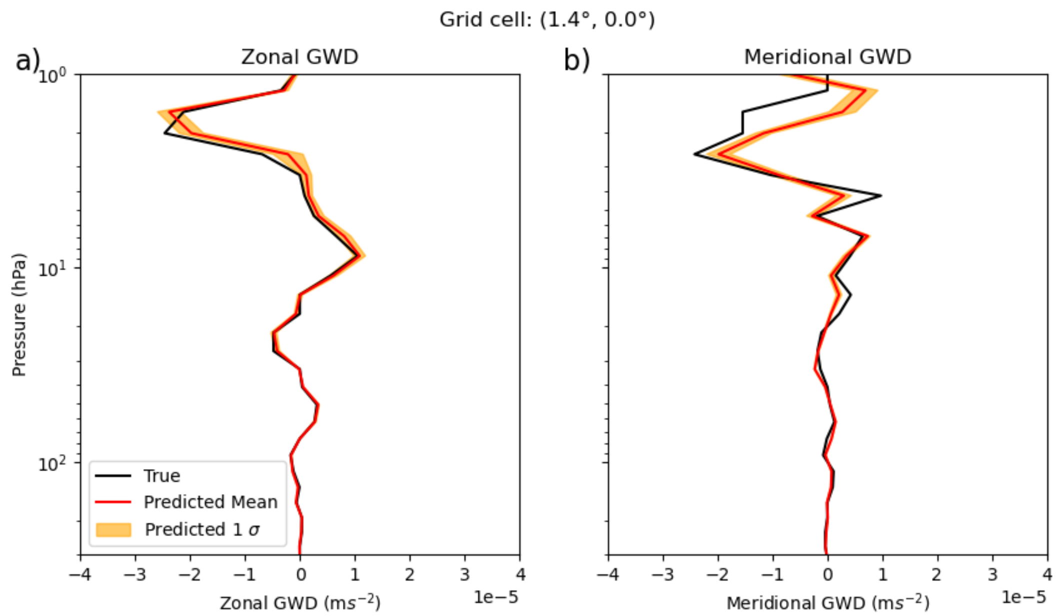

Figure 2 Example profiles of a) zonal and b) meridional gravity wave drag at one grid-cell and one timestep in the tropics where the black line indicates the ground truth from the AD99 parameterization, the red line indicates the mean prediction across all neural network ensembles and the orange shading indicates 1 standard deviation across these ensembles.

To measure the errors, we calculate the continuous ranked probability score (CRPS), a generalization of mean absolute error that allows for comparison of probability distributions. The use of CRPS to measure error between a predicted probability distribution and a single ground truth has long been used for verification of ensemble weather forecasts (Hersbach, 2000), and has recently been adopted for probabilistic machine learning (Gneiting & Raftery, 2007). Figure 3 shows CRPS for a) zonal and b) meridional gravity wave drag predictions over a range of latitudes. Note the scale of the axis is reduced by 10x relative to the gravity wave drag magnitudes in Figure 2. We find lower errors in the lower and mid-stratosphere that increase with height, where gravity wave drag magnitudes also increase. We see good performance across all latitudes.

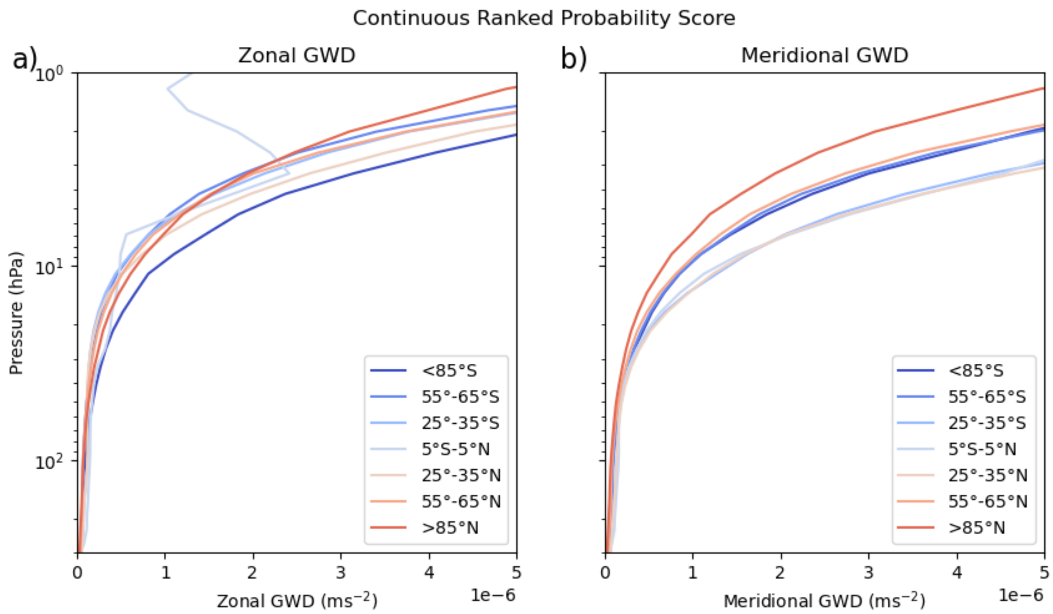

Figure 3 Continuous Ranked Probability Score for a) zonal and b) meridional gravity wave drag for different latitudes over the test dataset.

4.2 Offline uncertainty estimates

One common problem in uncertainty quantification of deep learning algorithms is in ensuring that uncertainty estimates are reasonable, often known as calibration of uncertainty (Lakshminarayanan et al., 2017). A well-calibrated machine learning model should predict low uncertainties when errors are small and high uncertainties when errors are large (for instance, when the data is out-of-sample). Figure 4 shows the 1 standard deviation uncertainty estimates against the ensemble mean absolute errors estimated for the test dataset, with the colors representing the density of points. Ideally, these should be correlated and lie approximately along the $y = x$ line shown in the dashed line. Points above the $y = x$ line are underconfident and points below are overconfident. Although the errors and predicted uncertainties are correlated, we see that the NNs suffer from overconfidence and frequently underestimate the uncertainty relative to the error. This is typical behavior for machine learning uncertainty estimates, including those based on deep ensembles (Abdar et al., 2021), and may be not be surprising given we only consider one type of uncertainty (parametric uncertainty) and do not consider structural uncertainty or data uncertainty in these estimates. This overconfidence is systematic across all levels of the stratosphere and occurs for both zonal and meridional NNs, but especially for the meridional predictions.

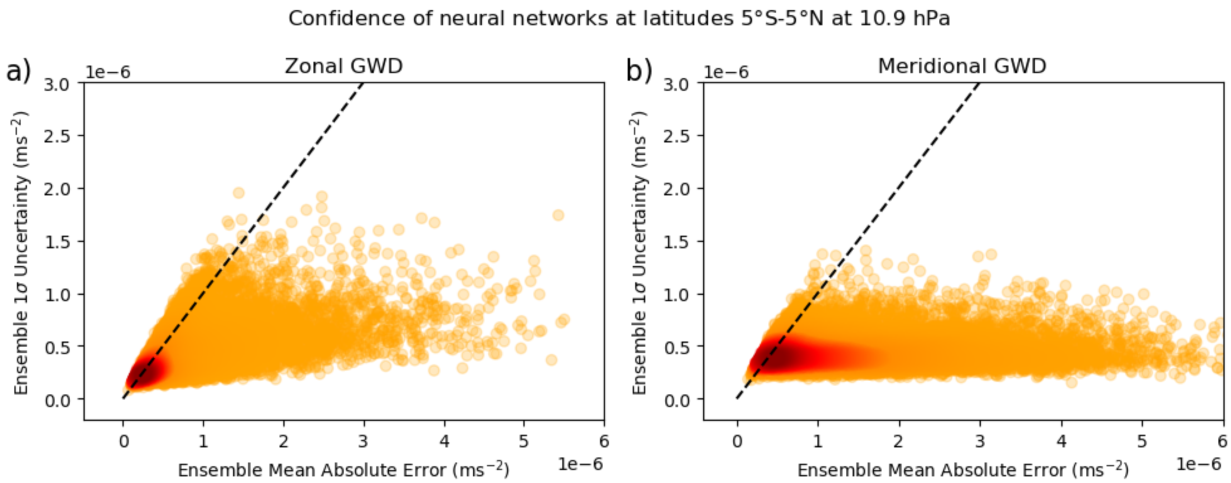

Figure 4 Ensemble uncertainty (measured as 1 standard deviation amongst the ensemble predictions) against ensemble error (measured as the mean absolute error across all ensemble predictions) for a) zonal and b) meridional gravity wave drag for test dataset between 5°S-5°N at 10 hPa. Each individual point represents a single prediction at one timestep and grid-cell and they are shaded according to density. The black dashed line shows the $y = x$ *line.*

4.3 Offline and Online Probability Distributions

Once coupled online into MiMA, the ensembles begin to diverge from each other even though they are initialized from the same state. This is partly due to the chaotic nature of the atmosphere where minute differences in one atmospheric variable can lead to very different atmospheric states after some time. Even introducing relatively minor differences in the GWD profiles, such as those in Figure 2, can lead to very different atmospheric states. Here, we aim to quantify how uncertainties in Figure 2 propagate into the GCM. We examine long-term statistics in order to separate out the NN parametric uncertainty from the internal variability.

We consider GWD in the tropics, due to its influence on the QBO. Figure 5 shows distributions of gravity wave drag in the upper stratosphere at 10 hPa for (a) zonal and (b) meridional components, where the black line indicates ground truth from the AD99 MiMA simulations, the blue line indicates the offline NN predicted gravity wave drag and the red line indicates the online NN predicted GWD. Both offline and online distributions are centered over the same location as AD99, indicating that the NN does not introduce a bias. In the lower stratosphere, the distributions are virtually indistinguishable (not shown). However, in the upper stratosphere at 10 hPa, the NN distributions take a different shape than AD99. This is particularly notable around the low negative zonal gravity wave drag values, where AD99 predicts an asymmetric gravity wave drag distribution with a positive skew. The NN distributions are more symmetric between positive and negative values. This may because machine learning optimizes for RMSE which may overly smooth gravity wave drag profiles, reducing asymmetry between positive and negative drag. The online NN distributions are slightly smoother than the offline NN distributions. We suggest that this must be caused by the interaction between the predicted gravity wave drag and the winds when coupled online. This is verified by Figure 6a, which shows distributions of zonal winds near the equator at 10 hPa, where online distributions tend to be smoother and weaker than the AD99 distributions.

Figure 5b shows that the online and offline meridional distributions are highly similar, even though they are smoothed out at low magnitudes. Even though the meridional NN is generally less accurate (e.g., Figure 2b), the meridional component of gravity wave drag does not appear to diverge when coupled online. Similarly, Figure 6b shows the distribution of the meridional winds to be unchanged when the NNs are coupled. This indicates that the meridional circulation is not highly sensitive to the effects of subgrid-scale gravity wave drag, possibly due to lower magnitude of the meridional winds.

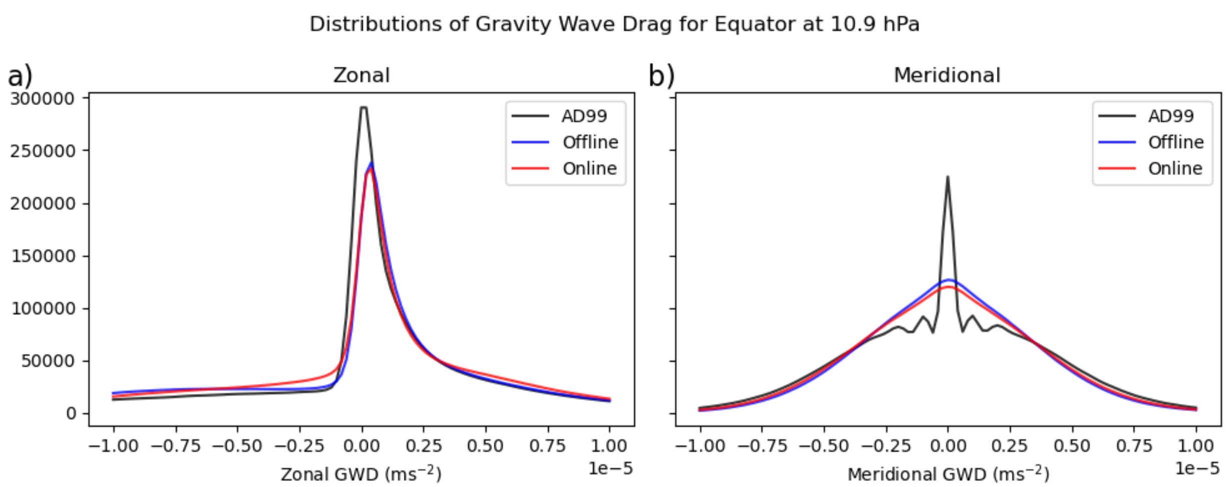

Figure 5 a) zonal and b) meridional gravity wave drag distributions for AD99 simulations (black), offline NN predictions (blue) and online NN simulations (red) at 10 hPa between 5°S-5°N.

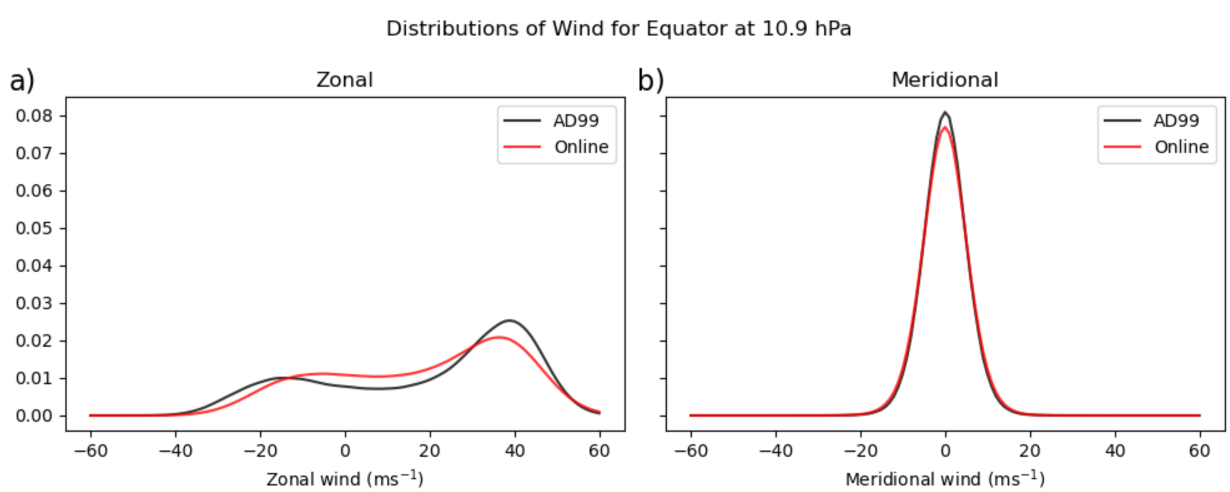

Figure 6 a) zonal and b) meridional wind distributions for AD99 (black) and online NN simulations (red) at 10 hPa between 5°S-5°N.

4.4 QBO uncertainties

Ultimately, we are interested in how the NN estimations for GWD influence the climatology and its variability when coupled into a GCM. We examine statistics of the QBO in MiMA by calculating the QBO period and amplitude at 10 hPa for each QBO cycle within 400 years of AD99 simulations and the 600 years of NN simulations (from 30 simulations each of 20 year simulations), shown in Figure 7. While the mean period of the QBO across all simulation years are similar, the NN ensembles show increased variability that can be attributed to the parametric uncertainty. The NNs also appear to introduce a bias that reduces the QBO amplitude, consistent with the reduction in QBO zonal winds (Figure 6). These increases in QBO variability originate from differences between NN ensemble members (and therefore from the learned NN parameters), each of which tend to maintain fairly consistent QBO periods and amplitudes within the 20 year simulation.

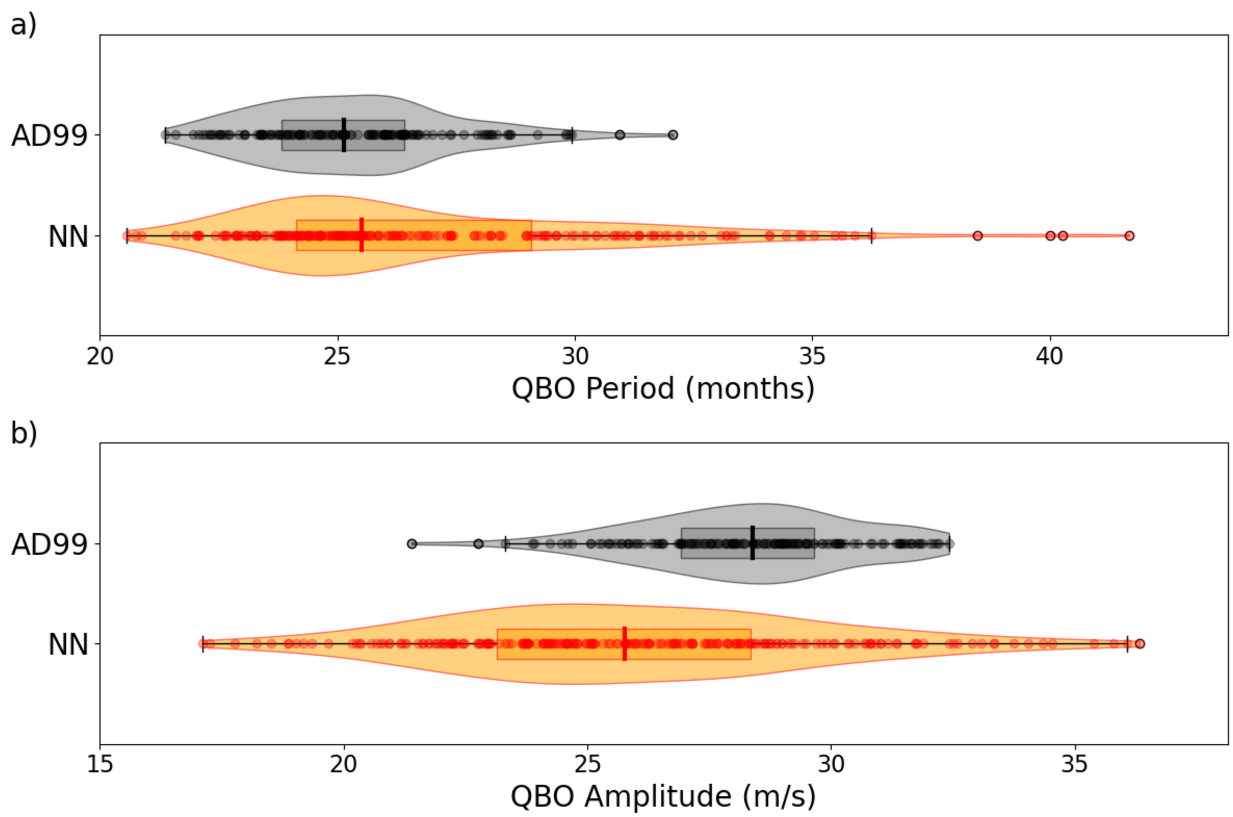

Figure 7 Violin plots showing distributions of QBO a) period and b) amplitude for the AD99 simulations in grey and for NN simulations in orange. The boxplots also show the median, upper and lower quartiles and each point represents a single QBO cycle.

We estimate parametric uncertainty by considering the increase in variability that arises due to the NNs. Assuming QBO cycles are normally distributed in both AD99 and in the ensemble of NNs, the additional variability from the uncertainty in parameters, σ_{param}, can be calculated as

$$\sigma_{param}^2 = \sigma_{AD99}^2 + \sigma_{NNs}^2$$

Equation 1

where σ_{AD99}^2 is the variance in the AD99 simulations and σ_{NNs}^2 is the total variance across all NN ensemble members. These results are shown in Table 1. Notably, the parametric uncertainty is significantly larger than the internal variability in the AD99 simulations, for both the QBO period and amplitude. It is possible that these uncertainties are underestimates of the true parametric uncertainty, given the overconfidence noted in offline tests (Figure 4). Still, the uncertainties in NN parameters are much greater than uncertainties in the parameters in the physics-based scheme AD99, estimated to be 1.53 months and 2.14 m/s for the period and amplitude respectively, in Mansfield & Sheshadri (2022) under the same model set-up. This highlights the importance of uncertainty quantification, regardless of whether the parameterization is physics-based or machine learning based.

Table 1 Mean and variability of QBO calculated across MiMA simulations using AD99 vs. the ensemble of NNs. Means are estimated across all QBO cycles in a 400 year long MiMA simulation using AD99 and in 600 years of simulations from the 30-member, 20 year long simulations from the ensemble of NNs. Variability is measured as 1 standard deviation between all QBO cycles. Parametric uncertainty is calculated assuming QBO cycles are normally distributed (Equation 1).

	Mean		Variability (measured as 1 standard deviation)		
	AD99	Ensemble of NNs	Internal variability in AD99 simulations	Total variability in ensemble of NNs	Parametric uncertainty
Period (months)	25.32	26.78	2.03	3.82	**3.25**
Amplitude (m/s)	28.29	25.91	2.17	3.86	**3.18**

4.5 Polar vortex uncertainties

The QBO is just one phenomenon that is strongly influenced by gravity wave dynamics. The stratospheric polar vortices in both hemispheres also depend upon gravity wave activity. In particular, the breakdown of the polar vortices during sudden stratospheric warmings (SSWs) and in the springtime final warming is driven by both planetary-scale and subgrid-scale gravity waves, and the variability of these events could also be impacted by changes to the gravity wave parameterization. For the northern hemisphere polar vortex, we consider the frequency of SSWs and for the southern hemisphere, we consider polar vortex lifetime. Figure 8 shows there is no obvious distinction between the variability of these properties between the AD99 and NN simulations, thus making the attribution of extratropical changes (and therefore, the calibration of extratropical parameters in AD99 and other schemes; Mansfield & Sheshadri, 2022) rather challenging.This may be because the breakdown of the polar vortices is driven by both planetary-scale waves and subgrid-scale gravity waves, thereby reducing the impact of any changes to the parameterization. Furthermore, some studies find there may be a compensation effect between resolved Rossby waves and unresolved gravity waves during SSW events (e.g., Cohen & Gerber, 2013), while some studies suggest that small scale gravity waves influence polar vortex recovery after a SSW more strongly than the breakdown itself (Wicker et al., 2023).

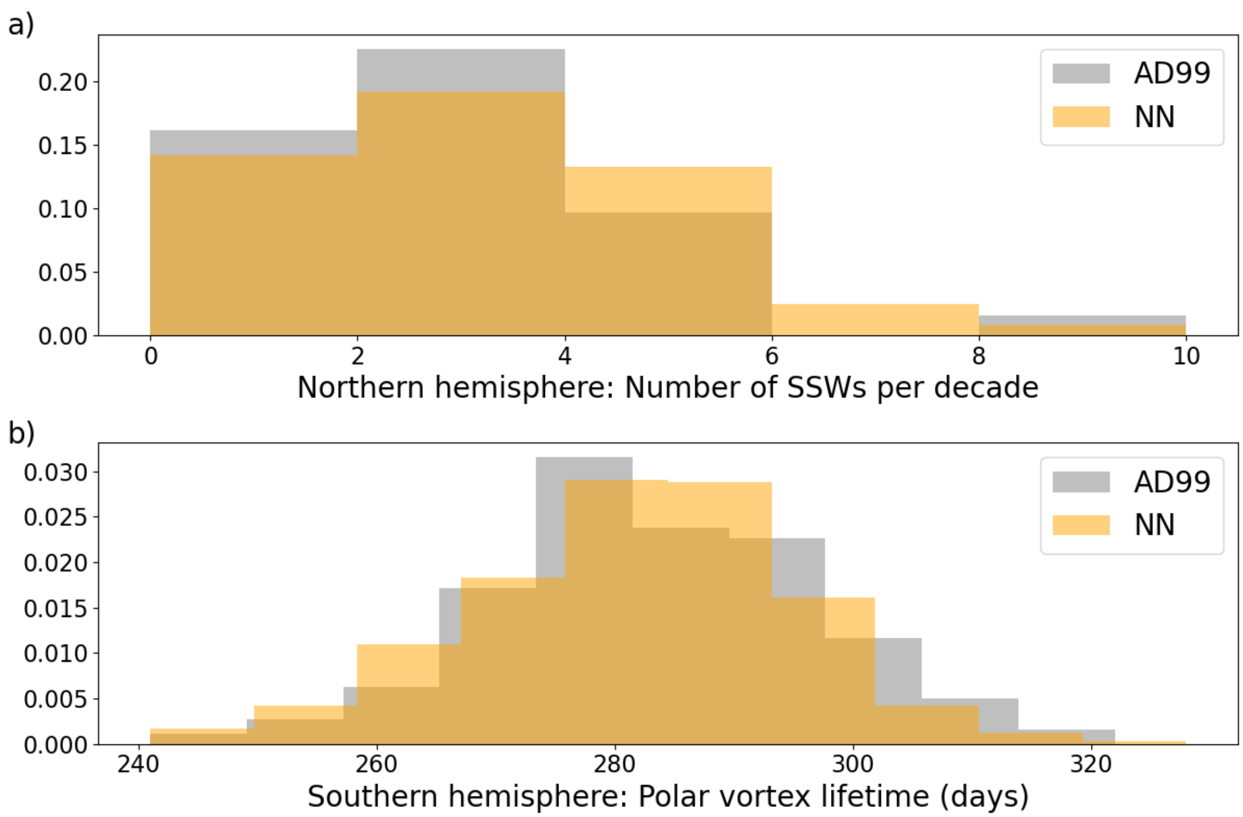

Figure 8 Histograms showing a) the Northern hemisphere number of SSWs per decade and b) the Southern hemisphere polar vortex lifetime for AD99 simulations in grey and the NN simulations in orange.

5 Conclusions

This study uses deep neural network ensembles to quantify parametric uncertainties in a machine learning parameterization of gravity wave drag. We use the neural network architecture of Espinosa et al. (2022) trained on one year of data simulated by the intermediate complexity GCM, MiMA, which uses AD99 gravity wave parameterization (Alexander & Dunkerton, 1999;

Jucker & Gerber, 2017). An ensemble of 30 identical neural networks are trained, each initialized with a different random seed. This ensemble allows us to estimate parametric uncertainties in neural network weights and biases. First, we assessed uncertainties in raw GWD output, which we refer to as *offline uncertainties*. We find fairly consistent results across all neural networks. Then, we used the FTorch library to couple the neural network into MiMA, allowing for GCM simulations that use the machine learning parameterization in place of the traditional physics-based scheme (Cambridge-ICCS, 2023). We assess uncertainties in GCM output for gravity wave drag and wind, refering to these as *online uncertainties*. We find increased online uncertainty, particularly for zonal winds.

Comparing long-term statistics of the climate within MiMA using the physics-based scheme AD99 and the ensemble of neural networks, showed that the use of NN emulators can alter the circulation significantly. We found that the NNs from the ensemble produce a bias in the QBO towards reduced amplitudes and dramatically increase the variability of the QBO, with uncertainty from NN parameters increasing the variability between QBO cycles by over 50%. Uncertainty quantification of parameterizations should therefore not be overlooked when developing ML-based schemes for future climate models. Our findings reiterate results from previous studies that find that, even when offline tests indicate "good" NN performance with relatively low uncertainties, the coupling of machine learning schemes into climate models can still introduce a significant source of uncertainty (Brenowitz et al., 2020; Lin et al., 2023). Learning distributions on the model parameters could provide a basis for further parameter refinement, for example, acting as a Bayesian prior distribution that could be constrained through *online calibration*, such as derivative-free optimization Ensemble Kalman methods (Pahlavan et al., 2023). As with traditional parameterization calibration, this could lead to improved QBO statistics and reduced parametric uncertainty. Interestingly, we find that the behavior and breakdown of the polar vortex is not strongly dependent on the parameterization, which may be partially due to influences from planetary-scale waves. This suggests that it may not be possible to further calibrate neural network parameters to polar vortex properties, and is comparable to the difficulties in calibration of extratropical parameters of AD99 (Mansfield & Sheshadri, 2022).

We only scratch the surface of uncertainty quantification for machine learning parameterizations. Firstly, we describe only one type of uncertainty: parametric uncertainty, a type of epistemic (model) uncertainty. There exist a wide range of machine learning approaches that could be used for this task, including Bayesian Neural Networks, Monte Carlo dropout generative models and deep ensembles (Abdar et al., 2021). We used deep ensemble methods for this task (Lakshminarayanan et al., 2017), due to their simplicity to implement. However, this approach is computationally costly during both training and evaluation, requiring the use of ensembles which is not feasible for long climate model integrations. A more complete picture would be given by also assessing aleatoric (data) uncertainties. We note that our parametric uncertainty estimates would change given a different training dataset, which makes detangling the effects of epistemic and aleatoric uncertainty a challenge (Haynes et al., 2023; Hüllermeier & Waegeman, 2021). Still, learning the relative contributions between model and data uncertainties would be insightful when designing machine learning parameterizations. Aleatoric uncertainties could be estimated through the use of Bayesian neural networks or Monte Carlo dropout (Abdar et al., 2021), by parameterizing gravity wave outputs as a distribution (Guillaumin & Zanna, 2021; Haynes et al., 2023), or through generative models such as GANs (Gagne II et al., 2020; Nadiga et al., 2022; Perezhogin et al., 2023).

Secondly, the machine learning parameterization used here is an emulator of an existing scheme, allowing us to compare against a ground truth simulation. Future studies may wish to extend this to train ML models on gravity-wave resolving simulations e.g., with kilometer-scale resolution models such as IFS (Anantharaj et al., 2022), WRF (Sun et al., 2023) or ICON (Hohenegger et al., 2023). When using novel training datasets from high resolution simulations, we do not have online "true" distributions to compare against, which could present challenges when disentangling the various sources of variability. Furthermore, it also raises the issue of understanding the role of aleatoric uncertainty, e.g., in the choice of training data and method for estimating gravity wave drag (Sun et al., 2023).

Thirdly, MiMA is an intermediate complexity atmospheric circulation model. One may expect that coupling this atmospheric model to other Earth system components, such as the ocean, land, and sea-ice, would introduce further uncertainties. Therefore, we might consider the results presented here as a lower bound on the uncertainties we could expect to see in fully operational Earth system models that employ ML parameterizations. Extending this study to higher complexity Earth system models would be significantly more costly, however, this could be worthwhile towards better informing the design of ML parameterizations, which ultimately could lead to efficient but accurate hybrid GCMs that combine traditional dynamical solvers with novel machine learning parameterizations.

Acknowledgments

This research was made possible by Schmidt Sciences, a philanthropic initiative founded by Eric and Wendy Schmidt, as part of the Virtual Earth System Research Institute (VESRI). AS acknowledges support from the National Science Foundation through grant OAC-2004492.

We would also like to thank our Datawave colleagues, in particular L. Minah Yang and Dave Connelly for their work on the PyTorch implementation of the machine learning model, and Simon Clifford, Jack Atkinson, Dominic Orchard and others at ICCS, for their help in setting up the FTorch coupler with the Fortran-based climate model. We also appreciate the Stanford high performance computing resources that made this work possible.

Open Research

The code to run simulations, train neural networks and replicate plots presented in this paper is available at https://github.com/lm2612/WaveNet_UQ. The data generated will be made available on the Stanford Digital Repository on publication. The FTorch library for coupling PyTorch to Fortrain is maintained by ICCS and can be found https://github.com/Cambridge-ICCS/FTorch. The Model of an idealized Moist Atmosphere (MiMA) is maintained by Martin Jucker and is available at https://github.com/mjucker/MiMA. The version of MiMA that uses FTorch for coupling to the PyTorch emulator used in this study can be found at https://github.com/lm2612/MiMA/tree/ML-laura.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. https://www.tensorflow.org/

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Achatz, U., Alexander, M. J., Becker, E., Chun, H.-Y., Dörnbrack, A., Holt, L., Plougonven, R., Polichtchouk, I., Sato, K., Sheshadri, A., Stephan, C. C., Niekerk, A. van, & Wright, C. J. (2023). Atmospheric Gravity Waves: Processes and Parameterization. *Journal of the Atmospheric Sciences*, *1*(aop). https://doi.org/10.1175/JAS-D-23-0210.1

Alexander, M. J., & Dunkerton, T. J. (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

Anantharaj, V., Hatfield, S., Polichtchouk, I., Wedi, N., O'Neill, M. E., Papatheodore, T., & Dueben, P. (2022). An open science exploration of global 1-km simulations of the earth's atmosphere. *2022 IEEE 18th International Conference on E-Science (e-Science)*, 427–428. https://doi.org/10.1109/eScience55777.2022.00071

Baldwin, M. P., Gray, L. J., Dunkerton, T. J., Hamilton, K., Haynes, P. H., Randel, W. J., Holton, J. R., Alexander, M. J., Hirota, I., Horinouchi, T., Jones, D. B. A., Kinnersley, J. S., Marquardt, C., Sato, K., & Takahashi, M. (2001). The quasi-biennial oscillation. *Reviews of Geophysics*, *39*(2), 179–229. https://doi.org/10.1029/1999RG000073

Bauer, P., Stevens, B., & Hazeleger, W. (2021). A digital twin of Earth for the green transition. *Nature Climate Change*, *11*(2), Article 2. https://doi.org/10.1038/s41558-021-00986-y

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and Stabilizing Machine-Learning Parametrizations of Convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357–4375. https://doi.org/10.1175/JAS-D-20-0082.1

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially Extended Tests of a Neural Network Parametrization Trained by Coarse-Graining. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2728–2744. https://doi.org/10.1029/2019MS001711

Bushell, A. C., Anstey, J. A., Butchart, N., Kawatani, Y., Osprey, S. M., Richter, J. H., Serva, F., Braesicke, P., Cagnazzo, C., Chen, C.-C., Chun, H.-Y., Garcia, R. R., Gray, L. J., Hamilton, K., Kerzenmacher, T., Kim, Y.-H., Lott, F., McLandress, C., Naoe, H., … Yukimoto, S. (2020). Evaluation of the Quasi-Biennial Oscillation in global climate models for the SPARC QBO-initiative. *Quarterly Journal of the Royal Meteorological Society*, *n/a*(n/a). https://doi.org/10.1002/qj.3765

Butler, A. H., Seidel, D. J., Hardiman, S. C., Butchart, N., Birner, T., & Match, A. (2015). Defining Sudden Stratospheric Warmings. *Bulletin of the American Meteorological Society*, *96*(11), 1913–1928. https://doi.org/10.1175/BAMS-D-13-00173.1

Cambridge-ICCS. (2023). *FTorch: A library for coupling (Py)Torch machine learning models to Fortran* [Computer software]. https://github.com/Cambridge-ICCS/FTorch

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

Chen, D., Rojas, M., Samset, B. H., Cobb, K., Diongue Niang, A., Edwards, P., Emori, S., Faria, S. H., Hawkins, E., Hope, P., Huybrechts, P., Meinshausen, M., Mustafa, S. K., Plattner, G.-K., & Tréguier, A.-M. (2021). Framing, Context, and Methods. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (Eds.), *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 147–286). Cambridge University Press. https://doi.org/10.1017/9781009157896.003

Chevallier, F., Morcrette, J.-J., Chéruy, F., & Scott, N. A. (2000). Use of a neural-network-based long-wave radiative-transfer scheme in the ECMWF atmospheric model. *Quarterly Journal of the Royal Meteorological Society*, *126*(563), 761–776. https://doi.org/10.1002/qj.49712656318

Chollet, F. & others. (2015). *Keras*. GitHub. https://github.com/fchollet/keras

Cohen, N. Y., & Edwin P. Gerber, and O. B. (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

Delaunay, A., & Christensen, H. M. (2022). Interpretable Deep Learning for Probabilistic MJO Prediction. *Geophysical Research Letters*, *49*(16), e2022GL098566. https://doi.org/10.1029/2022GL098566

Dong, W., Fritts, D. C., Liu, A. Z., Lund, T. S., Liu, H.-L., & Snively, J. (2023). Accelerating Atmospheric Gravity Wave Simulations Using Machine Learning: Kelvin-Helmholtz Instability and Mountain Wave Sources Driving Gravity Wave Breaking and Secondary Gravity Wave Generation. *Geophysical Research Letters*, *50*(15), e2023GL104668. https://doi.org/10.1029/2023GL104668

Donner, L. J., Wyman, B. L., Hemler, R. S., Horowitz, L. W., Ming, Y., Zhao, M., Golaz, J.-C., Ginoux, P., Lin, S.-J., Schwarzkopf, M. D., Austin, J., Alaka, G., Cooke, W. F., Delworth, T. L., Freidenreich, S. M., Gordon,

C. T., Griffies, S. M., Held, I. M., Hurlin, W. J., … Zeng, F. (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., 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 CO2. *Geophysical Research Letters*, *49*(8), e2022GL098174. https://doi.org/10.1029/2022GL098174

Fritts, D. C., & Alexander, M. J. (2003). Gravity wave dynamics and effects in the middle atmosphere. *Reviews of Geophysics*, *41*(1). https://doi.org/10.1029/2001RG000106

Gagne, D. J., McGovern, A., Haupt, S. E., Sobash, R. A., Williams, J. K., & Xue, M. (2017). Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Weather and Forecasting*, *32*(5), 1819–1840. https://doi.org/10.1175/WAF-D-17-0010.1

Gagne, D. J., McGovern, A., & Xue, M. (2014). Machine Learning Enhancement of Storm-Scale Ensemble Probabilistic Quantitative Precipitation Forecasts. *Weather and Forecasting*, *29*(4), 1024–1043. https://doi.org/10.1175/WAF-D-13-00108.1

Gagne II, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine Learning for Stochastic Parameterization: Generative Adversarial Networks in the Lorenz ’96 Model. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001896. https://doi.org/10.1029/2019MS001896

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

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Gordon, E. M., & Barnes, E. A. (2022). Incorporating Uncertainty Into a Regression Neural Network Enables Identification of Decadal State-Dependent Predictability in CESM2. *Geophysical Research Letters*, *49*(15), e2022GL098635. https://doi.org/10.1029/2022GL098635

Gray, L. J. (2010). Stratospheric Equatorial Dynamics. In *The Stratosphere: Dynamics, Transport, and Chemistry* (pp. 93–107). American Geophysical Union (AGU). https://doi.org/10.1002/9781118666630.ch5

Guillaumin, A. P., & Zanna, L. (2021). Stochastic-Deep Learning Parameterization of Ocean Momentum Forcing. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002534. https://doi.org/10.1029/2021MS002534

Gupta, A., Birner, T., Dörnbrack, A., & Polichtchouk, I. (2021). Importance of Gravity Wave Forcing for Springtime Southern Polar Vortex Breakdown as Revealed by ERA5. *Geophysical Research Letters*, *48*(10), e2021GL092762. https://doi.org/10.1029/2021GL092762

Harder, P., Watson-Parris, D., Stier, P., Strassel, D., Gauger, N. R., & Keuper, J. (2022). Physics-informed learning of aerosol microphysics. *Environmental Data Science*, *1*, e20. https://doi.org/10.1017/eds.2022.22

Hardiman, S. C., Scaife, A. A., Niekerk, A. van, Prudden, R., Owen, A., Adams, S. V., Dunstan, T., Dunstone, N. J., & Madge, S. (2023). Machine learning for non-orographic gravity waves in a climate model. *Artificial Intelligence for the Earth Systems*, *1*(aop). https://doi.org/10.1175/AIES-D-22-0081.1

Hawkins, E., & Sutton, R. (2009). The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108. https://doi.org/10.1175/2009BAMS2607.1

Haynes, K., Lagerquist, R., McGraw, M., Musgrave, K., & Ebert-Uphoff, I. (2023). Creating and Evaluating Uncertainty Estimates with Neural Networks for Environmental-Science Applications. *Artificial Intelligence for the Earth Systems*, *2*(2). https://doi.org/10.1175/AIES-D-22-0061.1

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Hohenegger, C., Korn, P., Linardakis, L., Redler, R., Schnur, R., Adamidis, P., Bao, J., Bastin, S., Behravesh, M., Bergemann, M., Biercamp, J., Bockelmann, H., Brokopf, R., Brüggemann, N., Casaroli, L., Chegini, F., Datseris, G., Esch, M., George, G., … Stevens, B. (2023). ICON-Sapphire: Simulating the components of the Earth system and their interactions at kilometer and subkilometer scales. *Geoscientific Model Development*, *16*(2), 779–811. https://doi.org/10.5194/gmd-16-779-2023

Holt, L. A., Lott, F., Garcia, R. R., Kiladis, G. N., Cheng, Y.-M., Anstey, J. A., Braesicke, P., Bushell, A. C., Butchart, N., Cagnazzo, C., Chen, C.-C., Chun, H.-Y., Kawatani, Y., Kerzenmacher, T., Kim, Y.-H., McLandress, C., Naoe, H., Osprey, S., Richter, J. H., … Yukimoto, S. (2020). An evaluation of tropical waves and wave forcing of the QBO in the QBOi models. *Quarterly Journal of the Royal Meteorological Society*, *n/a*(n/a). https://doi.org/10.1002/qj.3827

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Jucker, M., & Gerber, E. P. (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

Krasnopolsky, V. M., & Fox-Rabinovitz, M. S. (2006). Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural Networks*, *19*(2), 122–134. https://doi.org/10.1016/j.neunet.2006.01.002

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles* (arXiv:1612.01474). arXiv. http://arxiv.org/abs/1612.01474

Lin, J., Yu, S., Beucler, T., Gentine, P., Walling, D., & Pritchard, M. (2023). *Systematic Sampling and Validation of Machine Learning-Parameterizations in Climate Models* (arXiv:2309.16177). arXiv. http://arxiv.org/abs/2309.16177

Lindzen, R. S., & Holton, J. R. (1968). A Theory of the Quasi-Biennial Oscillation. *Journal of the Atmospheric Sciences*, *25*(6), 1095–1107. https://doi.org/10.1175/1520-0469(1968)025<1095:ATOTQB>2.0.CO;2

Mansfield, L. A., & Sheshadri, A. (2022). Calibration and Uncertainty Quantification of a Gravity Wave Parameterization: A Case Study of the Quasi-Biennial Oscillation in an Intermediate Complexity Climate Model. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2022MS003245. https://doi.org/10.1029/2022MS003245

McGovern, A., Bostrom, A., Davis, P., Demuth, J. L., Ebert-Uphoff, I., He, R., Hickey, J., Ii, D. J. G., Snook, N., Stewart, J. Q., Thorncroft, C., Tissot, P., & Williams, J. K. (2022). NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES). *Bulletin of the American Meteorological Society*, *103*(7), E1658–E1668. https://doi.org/10.1175/BAMS-D-21-0020.1

Murphy, J. M., Booth, B. B. B., Collins, M., Harris, G. R., Sexton, D. M. H., & Webb, M. J. (2007). A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philosophical*

Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, *365*(1857), 1993–2028. https://doi.org/10.1098/rsta.2007.2077

Nadiga, B. T., Sun, X., & Nash, C. (2022). Stochastic parameterization of column physics using generative adversarial networks. *Environmental Data Science*, *1*, e22. https://doi.org/10.1017/eds.2022.32

O'Gorman, P. A., & Dwyer, J. G. (2018). Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563. https://doi.org/10.1029/2018MS001351

Pahlavan, H. A., Hassanzadeh, P., & Alexander, M. J. (2023). *Explainable Offline-Online Training of Neural Networks for Parameterizations: A 1D Gravity Wave-QBO Testbed in the Small-data Regime* (arXiv:2309.09024). arXiv. https://doi.org/10.48550/arXiv.2309.09024

Palmer, T. N. (2019). Stochastic weather and climate models. *Nature Reviews Physics*, *1*(7), Article 7. https://doi.org/10.1038/s42254-019-0062-2

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Perezhogin, P., Zanna, L., & Fernandez-Granda, C. (2023). *Generative data-driven approaches for stochastic subgrid parameterizations in an idealized ocean model* (arXiv:2302.07984). arXiv. http://arxiv.org/abs/2302.07984

Perkins, W. A., Brenowitz, N. D., Bretherton, C. S., & Nugent, J. M. (2023). *Emulation of cloud microphysics in a climate model* [Preprint]. Preprints. https://doi.org/10.22541/essoar.168614667.71811888/v1

Polichtchouk, I., Niekerk, A. van, & Wedi, N. (2023). Resolved Gravity Waves in the Extratropical Stratosphere: Effect of Horizontal Resolution Increase from O(10) to O(1) km. *Journal of the Atmospheric Sciences*, *80*(2), 473–486. https://doi.org/10.1175/JAS-D-22-0138.1

Rabel, E. (2019). *forpy: A library for Fortran-Python interoperability* [Computer software]. 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

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

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

Sengupta, K., Pringle, K., Johnson, J. S., Reddington, C., Browse, J., Scott, C. E., & Carslaw, K. (2021). A global model perturbed parameter ensemble study of secondary organic aerosol formation. *Atmospheric Chemistry and Physics*, *21*(4), 2693–2723. https://doi.org/10.5194/acp-21-2693-2021

Sexton, D. M. H., McSweeney, C. F., Rostron, J. W., Yamazaki, K., Booth, B. B. B., Murphy, J. M., Regayre, L., Johnson, J. S., & Karmalkar, A. V. (2021). A perturbed parameter ensemble of HadGEM3-GC3.05 coupled model projections: Part 1: selecting the parameter combinations. *Climate Dynamics*, *56*(11), 3395–3436. https://doi.org/10.1007/s00382-021-05709-9

Siskind, D., Eckermann, S. D., Coy, L., McCormack, J. P., & Randall, C. E. (2007). On recent interannual variability of the Arctic winter mesosphere: Implications for tracer descent: MESOSPHERIC INTERANNUAL VARIABILITY. *Geophysical Research Letters*, *34*(9). https://doi.org/10.1029/2007GL029293

Siskind, D., Eckermann, S., McCormack, J., Coy, L., Hoppel, K., & Baker, N. (2010). Case studies of the mesospheric response to recent minor, major, and extended stratospheric warmings. *J. Geophys. Res*, *115*, 0–3. https://doi.org/10.1029/2010JD014114

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

Ukkonen, P. (2022). Exploring Pathways to More Accurate Machine Learning Emulation of Atmospheric Radiative Transfer. *Journal of Advances in Modeling Earth Systems*, *14*(4), e2021MS002875. https://doi.org/10.1029/2021MS002875

Vallis, G. K., Colyer, G., Geen, R., Gerber, E., Jucker, M., Maher, P., Paterson, A., Pietschnig, M., Penn, J., & Thomson, S. I. (2018). Isca, v1.0: A framework for the global modelling of the atmospheres of Earth and other planets at varying levels of complexity. *Geoscientific Model Development*, *11*(3), 843–859. https://doi.org/10.5194/gmd-11-843-2018

Wang, L., & Alexander, M. J. (2009). Gravity wave activity during stratospheric sudden warmings in the 2007–2008 Northern Hemisphere winter. *Journal of Geophysical Research: Atmospheres*, *114*(D18). https://doi.org/10.1029/2009JD011867

Weyn, J. A., Durran, D. R., Caruana, R., & Cresswell-Clay, N. (2021). Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002502. https://doi.org/10.1029/2021MS002502

Whiteway, J. A., Duck, T. J., Donovan, D. P., Bird, J. C., Pal, S. R., & Carswell, A. I. (1997). Measurements of gravity wave activity within and around the Arctic stratospheric vortex. *Geophysical Research Letters*, *24*(11), 1387–1390. https://doi.org/10.1029/97GL01322

Wicker, W., Polichtchouk, I., & Domeisen, D. I. V. (2023). Increased vertical resolution in the stratosphere reveals role of gravity waves after sudden stratospheric warmings. *Weather and Climate Dynamics*, *4*(1), 81–93. https://doi.org/10.5194/wcd-4-81-2023

Wright, C. J., Osprey, S. M., Barnett, J. J., Gray, L. J., & Gille, J. C. (2010). High Resolution Dynamics Limb Sounder measurements of gravity wave activity in the 2006 Arctic stratosphere. *Journal of Geophysical Research: Atmospheres*, *115*(D2). https://doi.org/10.1029/2009JD011858

Yu, S., Hannah, W. M., Peng, L., Lin, J., Bhouri, M. A., Gupta, R., Lütjens, B., Will, J. C., Behrens, G., Busecke, J. J. M., Loose, N., Stern, C., Beucler, T., Harrop, B. E., Hilman, B. R., Jenney, A. M., Ferretti, S. L., Liu, N., Anandkumar, A., … Pritchard, M. S. (2023). *ClimSim: An open large-scale dataset for training high-resolution physics emulators in hybrid multi-scale climate simulators* (arXiv:2306.08754). arXiv. https://doi.org/10.48550/arXiv.2306.08754

Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-17142-3

Yuval, J., O'Gorman, P. A., & Hill, C. N. (2021). Use of Neural Networks for Stable, Accurate and Physically Consistent Parameterization of Subgrid Atmospheric Processes With Good Performance at Reduced Precision. *Geophysical Research Letters*, *48*(6), e2020GL091363. https://doi.org/10.1029/2020GL091363