Data Imbalance, Uncertainty Quantification, and Generalization via Transfer Learning in Data-driven Parameterizations: Lessons from the Emulation of Gravity Wave Momentum Transport in WACCM

Y. Qiang Sun[1], Hamid Pahlavan[2], Ashesh Chattopadhyay[3], Pedram Hassanzadeh[1], Sandro W. Lubis[4], M. Joan Alexander[5], Edwin P Gerber[6], Aditi Sheshadri[7], and Yifei Guan[1]

[1]Rice University
[2]Unknown
[3]UC Santa Cruz
[4]Pacific Northwest National Laboratory (DOE)
[5]NorthWest Research Associates, CoRA Office
[6]New York University
[7]Stanford University

December 27, 2023

Abstract

Neural networks (NNs) are increasingly used for data-driven subgrid-scale parameterization in weather and climate models. While NNs are powerful tools for learning complex nonlinear relationships from data, there are several challenges in using them for parameterizations. Three of these challenges are 1) data imbalance related to learning rare (often large-amplitude) samples; 2) uncertainty quantification (UQ) of the predictions to provide an accuracy indicator; and 3) generalization to other climates, e.g., those with higher radiative forcing. Here, we examine the performance of methods for addressing these challenges using NN-based emulators of the Whole Atmosphere Community Climate Model (WACCM) physics-based gravity wave (GW) parameterizations as the test case. WACCM has complex, state-of-the-art parameterizations for orography-, convection- and frontal-driven GWs. Convection- and orography-driven GWs have significant data imbalance due to the absence of convection or orography in many grid points. We address data imbalance using resampling and/or weighted loss functions, enabling the successful emulation of parameterizations for all three sources. We demonstrate that three UQ methods (Bayesian NNs, variational auto-encoders, and dropouts) provide ensemble spreads that correspond to accuracy during testing, offering criteria on when a NN gives inaccurate predictions. Finally, we show that the accuracy of these NNs decreases for a warmer climate (4XCO2). However, the generalization accuracy is significantly improved by applying transfer learning, e.g., re-training only one layer using ˜1% new data from the warmer climate. The findings of this study offer insights for developing reliable and generalizable data-driven parameterizations for various processes, including (but not limited) to GWs.

Data Imbalance, Uncertainty Quantification, and Generalization via Transfer Learning in Data-driven Parameterizations: Lessons from the Emulation of Gravity Wave Momentum Transport in WACCM

Y. Qiang Sun[1], Hamid A. Pahlavan[1], Ashesh Chattopadhyay[1], Pedram Hassanzadeh[1], Sandro W. Lubis[1,3], M. Joan Alexander[2], Edwin Gerber[4], Aditi Sheshadri[5], Yifei Guan[1]

[1]Rice University, Houston, TX, 77005
[2]NorthWest Research Associates, Boulder, CO, 80301, USA
[3]Pacific Northwest National Laboratory, Richland, WA, 99354, USA
[4]New York University, New York, NY 10012, USA
[5]Stanford University, Stanford, CA, 94305, USA

Key Points:

- WACCM's orographic, convective, and frontal gravity wave parameterizations are emulated using neural nets to inform future modeling efforts
- Out-of-distribution generalization (extrapolation) of the neural nets under $4\times CO_2$ forcing is enabled via transfer learning with 1% new data
- Data imbalance is addressed via resampling and weighted loss; uncertainty quantification via Bayesian, dropout, and variational methods

Corresponding author: Y. Qiang Sun and Pedram Hassanzadeh, ys91@rice.edu and pedramh@uchicago.edu

Abstract

Neural networks (NNs) are increasingly used for data-driven subgrid-scale parameterization in weather and climate models. While NNs are powerful tools for learning complex nonlinear relationships from data, there are several challenges in using them for parameterizations. Three of these challenges are 1) data imbalance related to learning rare (often large-amplitude) samples; 2) uncertainty quantification (UQ) of the predictions to provide an accuracy indicator; and 3) generalization to other climates, e.g., those with higher radiative forcing. Here, we examine performance of methods for addressing these challenges using NN-based emulators of the Whole Atmosphere Community Climate Model (WACCM) physics-based gravity wave (GW) parameterizations as the test case. WACCM has complex, state-of-the-art parameterizations for orography-, convection- and frontal-driven GWs. Convection- and orography-driven GWs have significant data imbalance due to the absence of convection or orography in many grid points. We address data imbalance using resampling and/or weighted loss functions, enabling the successful emulation of parameterizations for all three sources. We demonstrate that three UQ methods (Bayesian NNs, variational auto-encoders, and dropouts) provide ensemble spreads that correspond to accuracy during testing, offering criteria on when a NN gives inaccurate predictions. Finally, we show that accuracy of these NNs decreases for a warmer climate ($4\times CO_2$). However, the generalization accuracy is significantly improved by applying transfer learning, e.g., re-training only one layer using $\sim 1\%$ new data from the warmer climate. The findings of this study offer insights for developing reliable and generalizable data-driven parameterizations for various processes, including (but not limited) to GWs.

Plain Language Summary

Scientists are increasingly using machine learning methods, especially neural networks (NNs), to improve weather and climate models. However, it can be challenging for a NN to learn rare, large-amplitude events, because they are infrequent in training data. Also, NNs need to express their confidence (certainty) about a prediction and work effectively across different climates, e.g., warmer climates due to increased CO_2. Traditional NNs often struggle with these challenges. Here, we share insights gained from emulating the complex physics-based parameterization schemes for gravity waves in a state-of-the-art climate model. We propose specific strategies for addressing imbalanced data, uncertainty quantification (UQ), and making accurate predictions across various climates. For instance, to manage data balance, one such strategy involves amplifying the impact of infrequent events in the training data. We also demonstrate that several UQ methods could be useful in determining the accuracy of predictions. Furthermore, we show that NNs trained on simulations of the historical period do not perform as well in warmer climates. However, we improve the NNs' performance by employing transfer learning using limited data from warmer climates. This study provides lessons for developing robust and generalizable approaches for using NNs to improve models in the future.

1 Introduction

Small-scale processes such as moist convection, gravity waves, and turbulence are key players in the variability of the climate system and its response to increased greenhouse gases. However, as these processes cannot be resolved, entirely or partially, by the coarse-resolution general circulation models (GCMs), they need to be represented as functions of the resolved dynamics via subgrid-scale (SGS) parameterization schemes (e.g., Kim et al., 2003; Stensrud, 2007; Prein et al., 2015). Many of these parameterization schemes are based on heuristic approximations and simplifications, introducing large parametric and epistemic uncertainties in GCMs (Schneider et al., 2017; Hourdin et al., 2017; Palmer, 2019).

Recently, there has been a growing interest in developing data-driven SGS parameterizations for different complex processes in the Earth system using machine learning (ML) techniques, particularly deep neural networks (NNs). Promising results have been demonstrated in a wide range of idealized applications, including prototype systems (Maulik et al., 2019; Gagne et al., 2020; Rasp, 2020; Chattopadhyay, Subel, & Hassanzadeh, 2020; Frezat et al., 2022; Guan et al., 2022; Pahlavan et al., 2023), ocean turbulent processes (Bolton & Zanna, 2019; C. Zhang et al., 2023), moist convection in the atmosphere (O'Gorman & Dwyer, 2018; Brenowitz & Bretherton, 2019; Yuval & O'Gorman, 2020; Beucler et al., 2021; Iglesias-Suarez et al., 2023), radiation (Krasnopolsky et al., 2005; Belochitski & Krasnopolsky, 2021; Song & Roh, 2021), and microphysics (Seifert & Rasp, 2020; Gettelman et al., 2021). The ultimate promise of data-driven parameterizations, learned from observation-derived data and/or high-fidelity high-resolution simulations, is that they might have smaller parametric/structural errors, thus reducing the biases of GCMs and producing more reliable climate change projections (e.g., Schneider et al., 2017; Reichstein et al., 2019; Schneider et al., 2021).

However, there are major challenges in developing trustworthy, interpretable, stable, and generalizable data-driven parameterizations that can be used for such climate change projection efforts. Discussing and even listing all of these challenges is well beyond the scope of this paper. Well-known challenges such as interpretability and stability have been extensively discussed in a number of recent studies (e.g., McGovern et al., 2019; Beck et al., 2019; Brenowitz et al., 2020; Balaji, 2021; Clare et al., 2022; Mamalakis et al., 2022; Guan et al., 2022; Subel et al., 2023; Pahlavan et al., 2023). Here, we focus on three other key issues:

1. Data imbalance, and related to that, learning rare/extreme events,
2. Uncertainty quantification (UQ) of the NN-based SGS parameterization outputs,
3. Out-of-distribution (OOD) generalization (e.g., extrapolation to climates with higher radiative forcings).

Below we briefly discuss the importance of 1-3 and the current state-of-the-art methods in addressing them in the climate and ML literature. Data imbalance is a well-known problem in the ML literature, especially in the context of classification tasks (e.g., Japkowicz &

Stephen, 2002; G. Wu & Chang, 2003; Chawla et al., 2004; Sun et al., 2009; Huang et al., 2016; Ando & Huang, 2017; Buda et al., 2018; Johnson & Khoshgoftaar, 2019). The problem becomes particularly significant when one aims to learn rare/extreme events (Maalouf & Trafalis, 2011; Maalouf & Siddiqi, 2014; Baldi et al., 2014; Liu et al., 2016; O'Gorman & Dwyer, 2018; Qi & Majda, 2020; Chattopadhyay, Nabizadeh, & Hassanzadeh, 2020; Miloshevich et al., 2023; Finkel et al., 2023; Shamekh et al., 2023). For example, suppose we aim to learn the binary classification of the 99 percentile of temperature anomalies using a NN. In this case, label 0 (no extreme) will constitute 99% of the training (or testing) set while label 1 (extreme) will be just 1%. With many common loss functions such as mean squared error (MSE) or root-mean squared-error (RMSE), training a NN will result in one that predicts 0 for any sample (extreme or no extreme) while having a seemingly high accuracy of 99% (of course, other metrics such as precision/recall will show the shortcoming, see Chattopadhyay, Nabizadeh, & Hassanzadeh (2020)). The most common remedy to this problem for classification tasks is resampling. An example is down-sampling non-extreme cases by a factor of 100, which effectively balances the dataset.

In addition to *classification* tasks, Data imbalance also presents a significant challenge in *regression* tasks required for parameterization schemes in climate models. As highlighted by Chantry et al. (2021), such imbalances contributed to the unsuccessful emulation of their orographic gravity wave parameterization (GWP) scheme, largely because orography affects the gravity wave (GW) drag in only a fraction of the grid columns. This challenge also persists in emulating GWP for non-orographic GWs, especially when GWs are intricately linked to their sources. For instance, the presence of zero convective GW drag at numerous grid points due to the absence of convection creates a notably imbalanced dataset. This issue will be explored further in the results section. In regression tasks, data imbalance may also manifest in the form of difficulty in learning large-amplitude (extreme) outputs, which are rare and constitute only a small fraction of the training set. In the case of GWs, Observations have shown that gravity wave amplitudes are highly intermittent such that the largest 10% events alone can contribute more than 50% of the total momentum flux (Hertzog et al., 2012), so the extreme events will contribute an outsized fraction of the total drag. Nonetheless, poorly learning these large-amplitude outputs, like drag forces, can result in instabilities (e.g., Guan et al., 2022). Addressing data imbalance in climate applications has received relatively limited attention. In this study, we propose several remedies based on resampling techniques and weighted loss functions, demonstrating their advantages in enabling successful emulations of all GWP schemes and improving the learning of rare/extreme events.

Quantifying the uncertainties in outputs from NN-based parameterization schemes is essential when employing these schemes, particularly for high-stakes decision-making tasks such as climate change projections. Crucially, during testing when we are unable to directly determine a prediction's accuracy, we need a UQ method that can provide a credible *confidence level* for each prediction, serving as a reliable indicator of its accuracy. During inference, the output of an NN can be inaccurate for various reasons, including poor approximation (e.g., due to poor NN architecture), poor within-distribution generalization (e.g.,

for inputs that are rare events), or poor optimization (collectively referred to as *epistemic uncertainty*), as well as because of OOD generalization errors due to input samples from a distribution different from that of the training set (Abdar et al., 2021; Lu et al., 2021; Krueger et al., 2021; Miller et al., 2021; Shen et al., 2021; D. Wu et al., 2021; Ye et al., 2021; D. Zhang et al., 2021; Subel et al., 2023). Quantifying the level of uncertainty would then allow us to avoid using a data-driven parameterization scheme when it is inaccurate due to one of the aforementioned reasons (Maddox et al., 2019; Zhu et al., 2019; Li et al., 2022; Psaros et al., 2023). In the context of data-driven parameterization in climate modeling, the two most challenging sources of uncertainty are rare/extreme events and OOD generalization errors. The latter is a concern, particularly when the GCM is used for climate change studies (see below for more discussions).

Developing UQ methods for NNs is an active area of research in the ML community, and there is not a generally applicable rigorous method yet. For instance, techniques like Markov-Chain Monte Carlo can be prohibitively expensive, especially when dealing with high-dimensional systems (Oh et al., 2005; Ballnus et al., 2017; Chen & Majda, 2019). For a comprehensive review in the context of scientific ML, refer to Psaros et al. (2023). The topic has also started to increasingly gain attention in the climate literature (Guillaumin & Zanna, 2021; Gordon & Barnes, 2022; Haynes et al., 2023; Barnes et al., 2023). In this study, we will assess the performance of three common UQ methods (Bayesian, dropout, and variational NNs) by analyzing the relationships between uncertainty and accuracy during inference testing. We will also consider scenarios involving OOD generalization errors resulting from global warming.

As already mentioned above, OOD generalization (extrapolation to a test data distribution different from that of the training set) is a major challenge for applications involving non-stationarity, like a changing climate. Studies have already shown that the lack of OOD generalization in data-driven parameterizations leads to inaccurate and unstable simulation (Rasp et al., 2018; O'Gorman & Dwyer, 2018; Chattopadhyay, Subel, & Hassanzadeh, 2020; Guan et al., 2022; Nagarajan et al., 2020). A general and powerful method for improving the OOD generalization capability of NNs is transfer learning (TL), which involves re-training a few or all of the layers of a NN using a small amount of data from the new system (Yosinski et al., 2014). This approach has already shown remarkable success in enabling data-driven parameterization schemes to extrapolate across the parameter space (e.g., to 100$\times$ higher Reynolds number) in canonical test cases (Chattopadhyay, Subel, & Hassanzadeh, 2020; Subel et al., 2021; Guan et al., 2023; Subel et al., 2023; C. Zhang et al., 2023). In particular, Subel et al. (2023) introduced SpArK (Spectral Analysis of Regression Kernels and Activations) showing that re-training even one layer can lead to successful OOD generalization, although this optimal layer, unlike the rule of thumb in the ML literature, may not be the deepest but the shallowest hidden layer. Here, we further leverage these studies and show how TL can enable OOD generalization of data-driven parameterization schemes in state-of-the-art GCMs.

The methods used in this study and the learned lessons apply to a broad range of processes and applications in climate modeling. However, the results are presented for a single test case, that is based on the emulation of complex physics-based GWP schemes in version 6 of the Whole Atmosphere Community Climate Model (WACCM), a state-of-the-art GCM (Gettelman et al., 2019). Here, we use the emulations of current physics-based parameterization schemes as a stepping stone towards learning data-driven parameterizations from observations and high-fidelity simulations by testing ideas for addressing items 1-3 listed earlier. Furthermore, developing better representations of un- and under-resolved GWs in GCMs is an important problem on its own (Kim et al., 2003; Alexander et al., 2010; Achatz, 2022). A number of recent studies have taken the first steps in learning data-driven GWP from observations and high-resolution simulations (Matsuoka et al., 2020; Amiramjadi et al., 2022; Sun et al., 2023; Dong et al., 2023), though careful and time-consuming steps are needed in producing, analyzing, and using such data. Furthermore, two recent studies focused on emulators of simpler GWP schemes in a forecast model and idealized GCM have readily shown the usefulness of lessons learned from emulators (Chantry et al., 2021; Espinosa et al., 2022; Hardiman et al., 2023). This further motivates the focus on using emulators for testing ideas for addressing data imbalance, UQ, and OOD generalization.

This paper is structured as follows. Section 2 introduces the WACCM simulations and the NN architectures used in this study. The findings, detailed in Section 3, emphasize the insights gained in addressing data imbalance and UQ, alongside OOD generalization of the emulators under warmer climate conditions. Consistent with Chantry et al. (2021), we find that using an NN to emulate the parameterization of orographic GWs is significantly more challenging than non-orographic GWs. This necessitated additional steps to achieve reasonable offline performance, as detailed in Section 4. To the best of our knowledge, this stands as the first NN-based emulation of orographic GWs to address the challenges in Chantry et al. (2021). Finally, we provide a concluding summary in Section 5.

2 Data and Methods

2.1 The Whole Atmosphere Community Climate Model (WACCM)

The NCAR's WACCM version 6 introduced in Gettelman et al. (2019) is used in this study. WACCM has state-of-the-art GWP schemes for GWs from three different sources: orography (OGWs), convection (CGWs), and fronts (FGWs). These complex sources make the emulation of the GWP schemes in WACCM a challenging task. This is, therefore, a suitable test case to investigate ideas for learning rare events, UQ, and OOD generalization to benefit the future efforts for the much more complex task, that is learning data-driven GWP schemes from observations and/or high-resolution GW-resolving simulations (Amiramjadi et al., 2022; Sun et al., 2023).

The configuration of the WACCM used in this study is identical to the public version in Gettelman et al. (2019), with a horizontal resolution of $0.95^\circ \times 1.25^\circ$ and 70 vertical levels. The two non-orographic GWP schemes in WACCM both follow Richter et al. (2010), yet

allow separate specifications of FGW and CGW sources. For OGWs, WACCM uses an updated planetary boundary layer form drag scheme from Beljaars et al. (2004), near-surface nonlinear drag processes following Scinocca & McFarlane (2000), and a ridge-finding algorithm to define orographic sources based on Bacmeister et al. (1994). A full documentation of WACCM OGWs can also be found in Kruse et al. (2022).

We conduct two sets of simulations: A 10-year pre-industrial "control" run, and a 10-year pseudo-global-warming "future" run with $4\times CO_2$ and uniform +4 K sea-surface temperature increases. In each run, we save, on the native grid, all the inputs and outputs for each of the three GWPs every 3 hours to capture the diurnal cycle. A complete list of these inputs/outputs, which are used in the training of the NN-based emulators, is presented in Appendix A.

We train separate NNs for emulating the three GWP schemes that have different sources. We use the first 6 years of the control run for training and the last 4 years for validation (years 7 and 8) and testing (years 9 and 10). With a grid resolution of $\sim 1^\circ$, there are 55,296 columns for each time snapshot, resulting in approximately 960 million input/output columns during the 6-year training period. Given the strong temporal correlation between the 3-hourly samples, we perform sub-sampling on both the training and validation data to reduce the dataset size. To accomplish this, we begin by shuffling all the input/output column pairs in time at each latitude/longitude grid point. Then, we randomly select 2,000 input/output pairs at each location for training and 500 pairs for validation.

To give the readers a general idea of the parameterized GWs and large-scale circulation in WACCM, Figure 1 shows the zonal-mean climatology for zonal GW drag/forcing, hereinafter referred to as GWD, arises from the divergence of gravity wave momentum transport (fluxes), from all three sources, computed from the 6-year training period in the control run. The zonal-mean zonal wind climatology is also shown. Seasonal dependency for both the GWD and the circulation is observed in the simulations. At levels below 100 hPa, the tendencies of non-orographic GW are relatively small compared to those from OGWs; however, their amplitudes increase significantly at higher altitudes. While the parameterized effect of GWs is generally to decelerate the zonal flow, there are exceptions, notably in regions like the equatorward flanks of the stratospheric polar night jets, where FGWs can accelerate the flow. For more information on the GWP schemes and circulations in WACCM, see Garcia et al. (2017) and Gettelman et al. (2019).

2.2 The NNs and UQ

2.2.1 The Deterministic Fully Connected NN

Here we briefly describe the general structure of the NN-based regression models trained as emulators for GWP schemes. For the deterministic artificial NN, denoted as ANN in this study, we use multilayer perceptrons (MLP). MLPs, which are feedforward fully connected NNs, take inputs through successive layers of linear transformation and non-linear activation

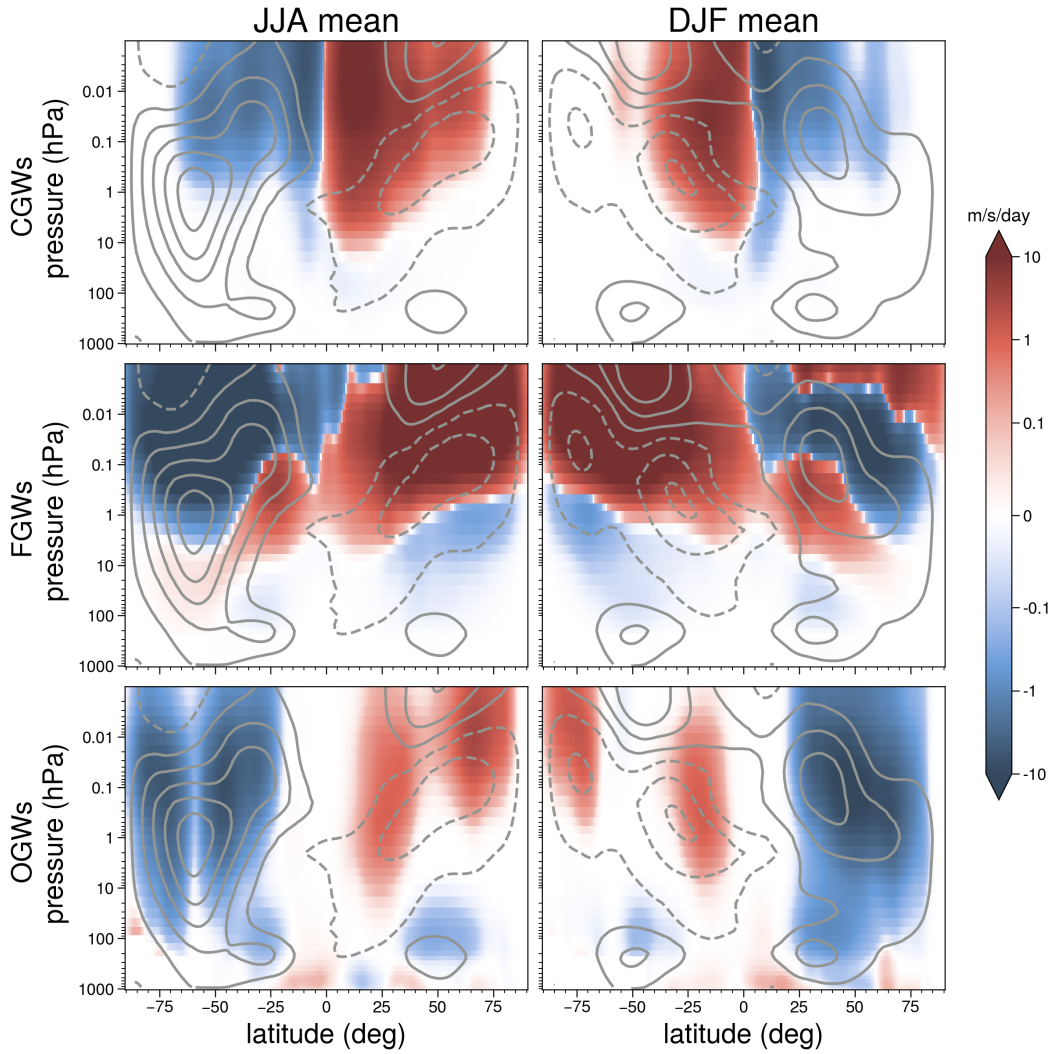

Figure 1. Climatology of zonal-mean GWD during summers (JJA) and winters (DJF) from all 3 sources in the control (pre-industrial) WACCM simulations. Top: CGWs; middle: FGWs; bottom: OGWs. The climatology of the zonal-mean zonal wind is also shown (grey lines), with an interval of 20 m/s. Dashed lines indicate negative values. Zero lines are omitted.

functions to produce an output, so as to learn a functional relationship between the input and output (Figure 2a). Deep MLPs have multiple layers of weights, which are optimized over many samples of input-output data pairs. Such MLPs are thus very powerful in terms of learning complicated functional relationships. Generally, we can write the governing equations of an MLP as

$$z^{\ell} = \sigma\left(W^{\ell} z^{\ell-1} + b^{\ell}\right), \tag{1}$$

where z^{ℓ} is the activation (output) of layer ℓ, W^{ℓ} is the weight matrix connecting layers ℓ and $\ell - 1$, and b^{ℓ} is the bias at layer ℓ, which allows the network to fit the data even when all input features are equal to 0. σ is the non-linear activation function.

In this study, we employ the same NN structure while training three distinct NNs, each for GWP originating from one of the three unique GW sources. The input layer contains the same input variables (see Appendix A) used by the WACCM GWPs across all vertical levels. There are 10 hidden layers in total (Figure 2a), and there are 500 neurons in each hidden layer. In the output layer, both zonal and meridional GWD are predicted. The activation function in each layer, σ, is chosen to be swish (Ramachandran et al., 2017), except for the output layer, where it is linear. During training, W^ℓ and b^ℓ are randomly initialized and learned by minimizing a loss function using an ADAM optimizer, with a fixed learning rate of $\alpha = 0.0001$. One of the loss functions used here is the common MSE, i.e.,

$$\mathcal{L}(\Theta) = \frac{1}{n}\sum_{i=1}^{n} \left\| \mathbf{NN}\left(x_i, \Theta\right) - y_i \right\|_2^2 \tag{2}$$

Here, n is the number of training samples and $\|.\|_2$ is the L_2 norm. For training sample i, vector x_i contains all the inputs to the NN (Appendix A), vector y_i contains the true zonal and meridional GWD at each vertical level, and $\Theta = \{\theta_j\}_{j=1\cdots p}$ denotes the trainable parameters, i.e., the weights ($p \approx 3 \times 10^6$).

2.2.2 The UQ Methods and Metrics

Although deterministic NNs are powerfully expressive and can exhibit high out-of-sample predictive skills, they do not provide estimates of the uncertainty associated with their predictions. As mentioned earlier, currently there is no rigorous method to estimate the uncertainty of an NN prediction. That said, a variety of techniques have been developed for UQ in NNs, though the validity and usefulness of the estimated uncertainty for scientific applications remain subjects of ongoing investigations (e.g., Psaros et al., 2023; Haynes et al., 2023). In this paper, we use three different and widely used approaches to perform UQ from the ML literature: Bayesian neural network (BNN), dropout neural network (DNN), and variational auto-encoder (VAE). A brief overview of these approaches is provided below.

Bayesian neural network (BNN): A BNN combines the deterministic NN described earlier and in Figure 2a with Bayesian inference (Blundell et al., 2015). Simply speaking, a BNN estimates distributions of the weights, rather than point values (as in a deterministic NN). The posterior distributions in the BNN (i.e., the distributions of the weights and biases) are calculated using the Bayes rule. In this study, we follow the standard practice and assume that all variational forms of the posterior are normal distributions. Furthermore, to accelerate the training process, we use the normal distribution $\mathcal{N}(\mu, 1)$ for all the priors in the BNN (where μ is obtained from parameters of the trained deterministic NN). Note that while we are assuming normal distributions for the trainable parameters, the predictions generated by BNN can fit different distributions due to the use of nonlinear activation functions. The resulting distribution of the predictions during inference gives an estimate of their uncertainty.

Dropout neural network (DNN): A DNN is developed by randomly eliminating all outgoing connections from some of the nodes (Figure 2a) in each hidden layer of a deterministic

NN during the training and the inference (Srivastava et al., 2014). The fraction of nodes "dropped" on average in each layer is called the dropout ratio. Mathematically, Equation (1) can be reformulated for a DNN as:

$$z^{\ell} = \sigma\left(D^{\ell} W^{\ell} z^{\ell-1} + b^{\ell}\right), \tag{3}$$

where the dropout matrix D^{ℓ} is a square diagonal binary matrix of integers 0 or 1. The diagonal elements of D^{ℓ} follow a Bernoulli distribution where the probability of zero is the dropout ratio.

Dropout was initially developed as a regularization technique to prevent over-fitting in NNs. However, Gal & Ghahramani (2016) showed that training a NN with the dropout technique approximates a Bayesian NN. In this study, we use a dropout rate of 0.1, which is incorporated in all hidden layers, but we also investigate the sensitivity of the DNN to different dropout rates, as later shown in Appendix B. Note that the random dropping out is also used during inference, leading to a distribution for each prediction.

Variational auto-encoder (VAE): A typical VAE (Kingma & Welling, 2014) consists of two NNs (Figure 2b): an encoder that transforms the input into a lower-dimensional latent space, parameterized by a normal probability distribution, and a decoder that inverts this transformation and produces the original input. The difference between the decoder's output and the original input drives the learning process of the encoder and decoder, while the parameterized lower-dimensional latent space provides the uncertainty of this transformation. The VAE was developed for generative reconstructions of data by simply drawing samples from the latent space. The VAE is basically a dimension-reduction method. Many variants, however, have been proposed for more specific purposes. In this study, following Foster et al. (2021), we add a third NN, as illustrated in Figure 2b, that randomly draws samples from the parameterized latent space as inputs, and predicts the zonal and meridional GWDs as outputs. The difference between the predicted GWDs and the true GWDs drives the learning of the third NN. Consequently, the loss for the entire network consists of three components: the loss between the reconstructed input and the original input, the Kullback–Leibler (KL) divergence between the distribution of the latent space and a normal distribution, and the loss between the predicted GWDs by the third NN and the true GWDs.

For a specific input, each of these three UQ methods discussed above can be run multiple times, generating an ensemble of predictions with different realizations of the weights by drawing from the trained distribution. This is in contrast to the deterministic NN that provides just a single-valued prediction for a given input. These ensembles can then be used to quantify the uncertainty associated with that prediction. We expect that the RMSE of the ensemble mean should exhibit approximately a 1-1 relationship with the ensemble spread (i.e., the standard deviation of the ensemble members). To investigate this relationship, we use the spread-skill plot (Delle Monache et al., 2013). Detailed calculations behind the spread-skill plot can be found in Appendix C, where we also introduce two metrics: spread-

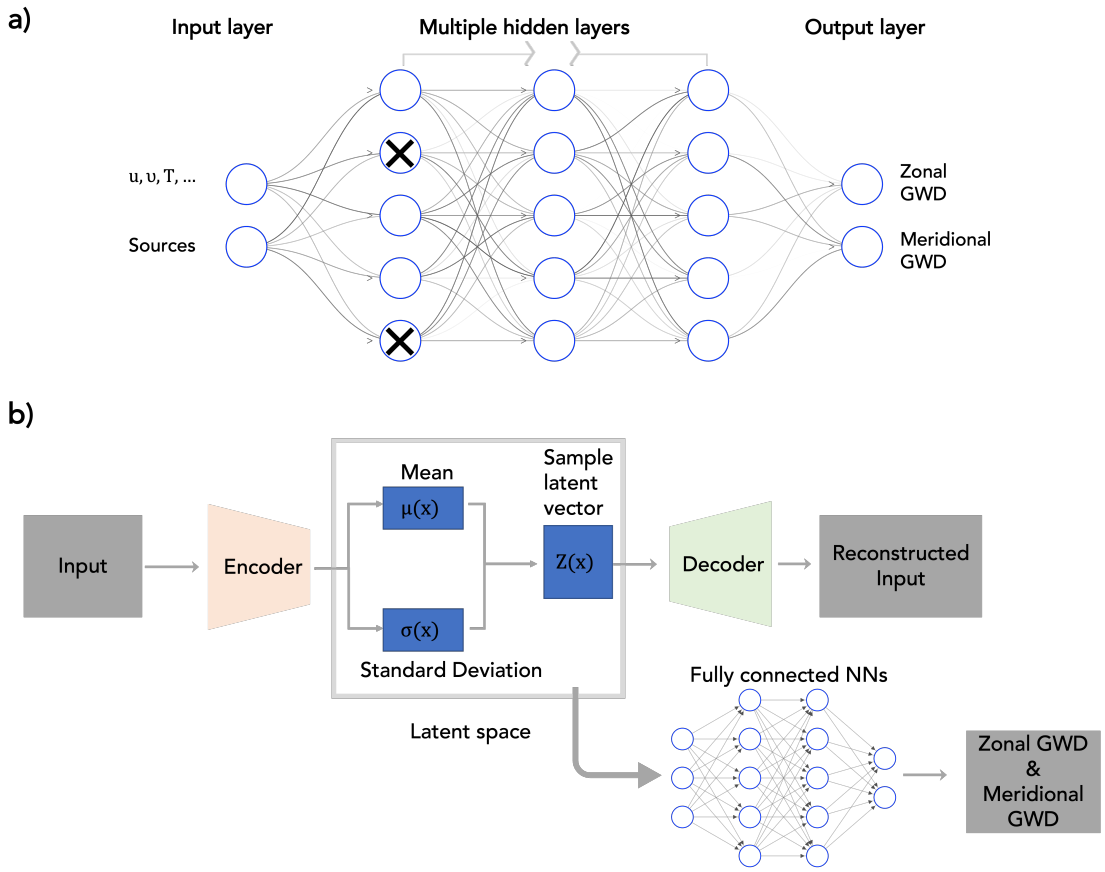

Figure 2. Schematics of the NN-based emulators and different training/re-training strategies used in this study. (a) Schematic for the MLP and DNN. The inputs of the NN are connected through successive layers of neurons (blue circles) to the output (GWDs). A fully connected MLP NN is trained from randomly initialized weights and biases in all layers. A DNN is the same but some connections are randomly eliminated during training and inference (black crosses). In TL, only some of the layers of a previously trained MLP are re-trained using new data. (b) Schematic for the VAE. A low-dimensional latent space is constructed and then used as the input for the additional fully connected NNs, which is similar to the one in (a).

skill reliability (SSREL) and overall spread-skill ratio (SSRAT), both of which summarize the information presented in the spread-skill plot.

2.3 Transfer Learning

Transfer learning refers to leveraging/reusing information (weights) from an already well-trained base NN to effectively build a new NN for a different system from which only a small amount of training data is available (Yosinski et al., 2014; Tan et al., 2018; Chattopadhyay, Subel, & Hassanzadeh, 2020). For our purpose, which is improving OOD generalization to the warmer climate, the TL procedure is as follows. For any of the NNs described earlier (e.g., the one in Figure 2a), we train them from randomly initialized weights and biases with data from the control simulations. The NN will work well during inference for test

samples from the control but not from future (warmer climate) simulations (as shown in the Results section). To address this, TL is applied wherein most of the NN's weights are kept constant, and only one or two hidden layers are re-trained using a limited dataset from the future simulation. Although this small dataset is insufficient for training an entire NN from random initialization, careful and correct selection of hidden layers for re-training, as discussed in Subel et al. (2023), allows the development of an NN that accurately adapts to the new system, i.e., the future climate conditions.

Here, we re-train the NN-based emulator that was initially trained on the control data with new data from only 1 month (30 consecutive days) of integration (1.4% of 6 years simulation for the initial training) of WACCM model under future forcing ($4\times CO_2$). We have explored different choices of layers to re-train with the same amount of new data and found that re-training the first hidden layer yields the best results, consistent with Subel et al. (2023). Therefore, the results with only re-training the first hidden layer are shown in Section 3 unless stated otherwise.

3 Results

3.1 Data Imbalance

As discussed earlier, the physics-based GWP schemes in WACCM are directly linked to their sources. This means they only produce non-zero values when their respective sources are active. For example, in a specific grid box, CGWs only register non-zero values when there is active convection within that box. The heterogeneous and sometimes intermittent nature of these sources leads to a dataset that is significantly imbalanced. Figure 3 shows global maps of the occurrence frequency of non-zero GWD for CGWs and FGWs. On average, only 7.6% of all GCM columns yield non-zero CGWs, primarily located in the tropics. Similarly, for FGWs, only 8.5% of all columns have non-zero outputs, but unlike CGWs, the majority of these are located in mid-to-high latitudes, particularly along the storm track region. For the OGWs in WACCM, data imbalance presents a greater challenge, to be discussed in a later section. While it is possible to simply separate zero and non-zero columns for emulation work where we know the truth, this approach falls short with real-world data, which is the main purpose of this study.

In addition to their sources, several other factors specific to GWD data exacerbate the data imbalance problem. In the case of each GCM column with non-zero GW activity, momentum fluxes are generally concentrated at a few critical height levels rather than being smoothly distributed throughout the entire column. This further restricts the effective occurrence frequency of non-zero values. Moreover, GWs exhibit significant intermittency, where a small portion of large-amplitude GWs often dominates the morphology of the observed global GW momentum flux distribution (Hertzog et al., 2012; Geller et al., 2013). Therefore, it is crucial for NNs to not only accurately identify the columns that produce GWDs but also to effectively learn and recognize rare and extreme GWDs.

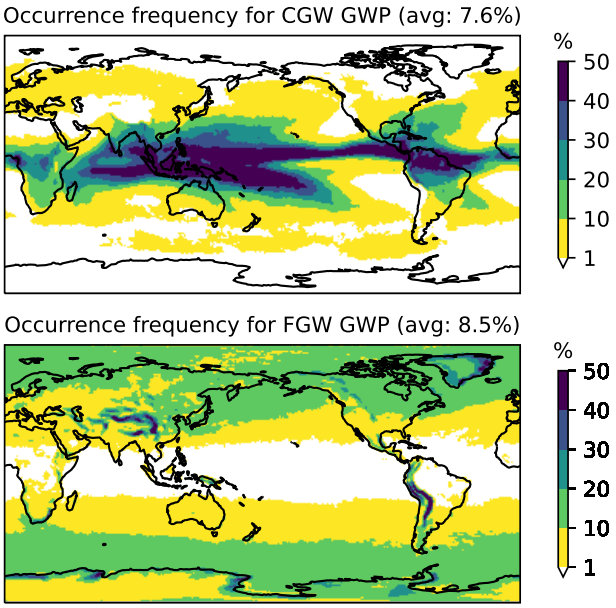

Figure 3. Distribution of occurrence frequency for CGWs (top) and FGWs (bottom) in the WACCM pre-industrial control simulations, based on the average of the 6-year training dataset.

Given the complexity of the GWD dataset, different normalization methods are considered in this study. The first method, dubbed "NORM1", is the typical normalization used in ML practices, which calculates elemental means and standard deviations for each feature (i.e., input variable at a given model level) and normalizes both inputs and outputs by these values (e.g., Espinosa et al. (2022)). With this approach, the same relative changes in wind at each level are treated equally in the input. The loss function in Equation (2) also penalizes the same relative error in GWD at each level equally. The second method, referred to as "NORM2", is designed with the physics of GWD in mind. For the velocity inputs (u, v) and the tendency outputs (GWD), each column is normalized by one single value, which is the largest standard deviation from all model levels. Additionally, the mean values for these variables, are retained (e.g., $u_{norm2}(x, y, z, t) = u(x, y, z, t)/max(std(u))$). Unlike NORM1, the original wind profile's structure is preserved in NORM2, and large GWD values at certain heights maintain a relatively larger value after this normalization. For all other input variables, NORM2 is identical to NORM1. Compared to NORM1, NORM2 places more emphasis on large GWD values and penalizes the NN more for missing these significant tendencies. These two normalization methods are also employed in Chantry et al. (2021), who found similar performance from these methods with the non-orographic GWPs.

Figure 4 shows the performance of the emulations for CGWs with the two normalization methods. When employing NORM1, the conventional approach seen in prior ML practices, and also our initial attempts, the emulator's performance is poor. Although the NN demonstrates some skill, its predictions tend to cluster around zero. However, when the second normalization method (NORM2) is employed, the emulation results show significant

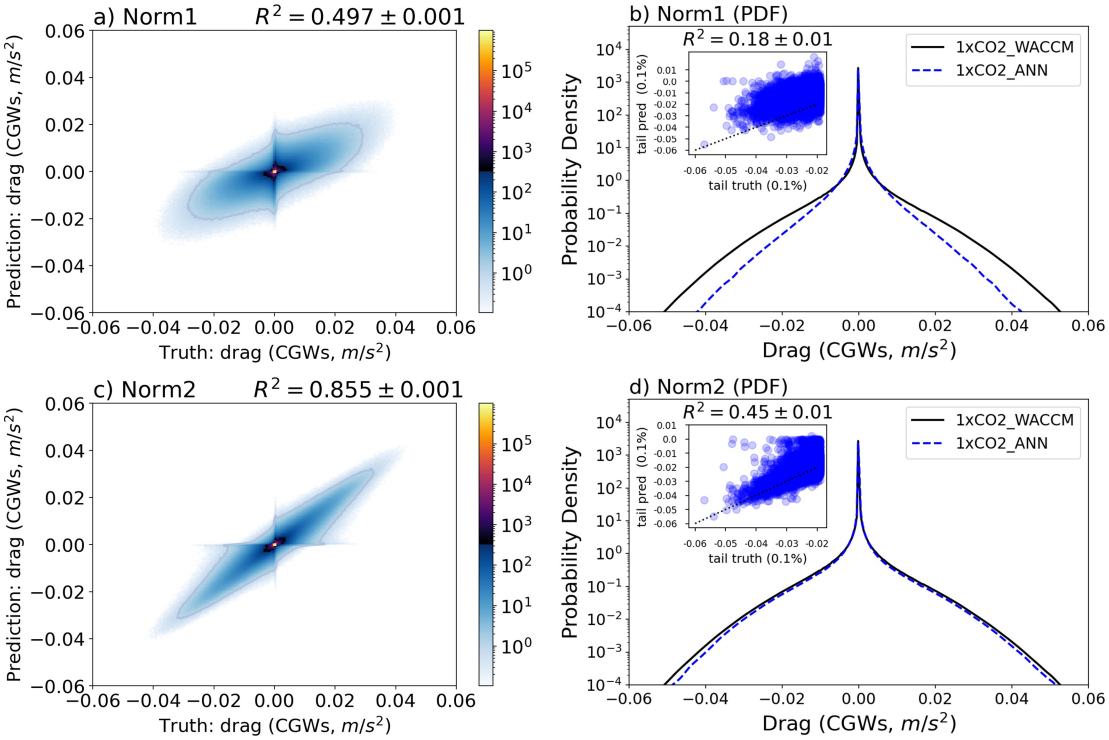

Figure 4. Data imbalance for GWD due to CGWs and the emulation results with two different normalization methods. a) A 2D histogram displaying the emulated GWD due to CGWs and the truth, with the training dataset normalized using NORM1; b) Distribution of the original convective GWD (black line) and the predicted values (blue line) with NORM2. The scatter plot in the corner represents the tail part only, including points with the top 0.1% amplitudes; c) Similar to a), but for NORM2; d) Similar to b), but for NORM2. The R^2 uncertainty range is estimated by dividing the test data into 10 segments, calculating the metric for each segment, and then computing the standard deviation (STD).

improvement, in contrast to the findings of Chantry et al. (2021). We attribute this improvement to the more pronounced data imbalance in our dataset, and it is likely a consequence of NORM2's emphasis on modeling the large GWD values. Nonetheless, emulating the tail of the probability density function (PDF) (rare events) remains poor, as evidenced by the tails in Figure 4c, primarily due to the predominance of zero GWD columns in the training dataset. To more effectively address the data imbalance issue in these regression tasks, we further propose two approaches here:

1. Resampling the data (ReSAM): In this approach, we limit the number of training sample pairs with zero GWD to be equal to the number of samples with non-zero GWD. This significantly reduces the number of columns with zero GWD, thus mitigating the data imbalance issue. Additionally, this sub-sampling reduces the total size of the training dataset, which, in turn, enhances the training speed (approximately sevenfold). While resampling methods have been well-established in the ML

literature, they have mainly been used for classification problems. Their application to regression problems in climate research has not been extensively explored.

2. Weighted loss function (WeLoss): Instead of assigning the same weight to all sample pairs in the loss function, we modify the weight for each column based on the PDF of its maximum GWD amplitude. This adjustment allows us to re-formulate the loss function defined in Equation (2) as

$$\mathcal{L}(\Theta) = \frac{1}{n}\sum_{i=1}^{n} \left\| \mathbf{W_i}\{\mathbf{NN}\left(x_i, \Theta\right) - y_i\} \right\|_2^2 \tag{4}$$

where

$$\mathbf{W_i} = \frac{1}{PDF(max(|y_i(z)|))} \tag{5}$$

Note that, in practice, we lack knowledge of the precise PDF for the maximum GWD within each column. Therefore, we employ a histogram with 20 bins as an alternative. Given the fact that large-amplitude GW events are rare, the WeLoss approach incentivizes the NN to prioritize these significant events.

When we apply the ReSAM approach to balance the training dataset (after normalization with NORM1 or NORM2), the emulation results significantly improve, as shown in Figure 5. In fact, when considering the R-squared value between the NN prediction and the ground truth, the ReSAM approach with NORM2 yields the best results. However, as the training dataset is still predominantly composed of zeros and small GWD values due to the intermittence of the GWs, examining the emulation results for only large amplitude GW events (e.g., the top 0.1% in Figure 5d) reveals less satisfactory performance ($R^2 = 0.72$). Regarding the WeLoss approach, it has a more limited impact on improving the R-squared value of the emulation (as shown in Figure 5e). However, it proves valuable in capturing the tails of the PDF and, thus, rare events (as depicted in Figure 5f). Moreover, as ReSAM and WeLoss represent distinct operations, they can be effectively combined when constructing a NN. The result of this combined approach for emulating the CGWs can be found in Figures 5g and 5h. While the R-squared value for the entire distribution only marginally changes (0.925 vs. 0.931 with ReSAM only), the performance of the emulation for the tail part has been improved (R^2 increased to 0.77).

Similarly, Figure 6 presents the offline emulation results for the FGWs. The conclusions drawn for CGWs generally hold true. However, data imbalance in FGWs is less pronounced compared to CGWs, which simplifies the task of emulating FGWs. Even without any resampling or changes to the normalization or (see Figure 6a), we achieve reasonable emulation results ($R^2 = 0.9$). One contributing factor is the wider spatial distribution of FGWs compared to CGWs (refer to Figure 3). Additionally, the source of FGWs (frontogenesis function) in WACCM exhibits a much more continuous nature compared to precipitation and diabatic heating. As the data imbalance issue is less severe for FGWs, the performance with different normalization methods becomes more similar, echoing findings from Chantry et al. (2021) who emulated non-orographic GWs (including convective and frontal GWs) together.

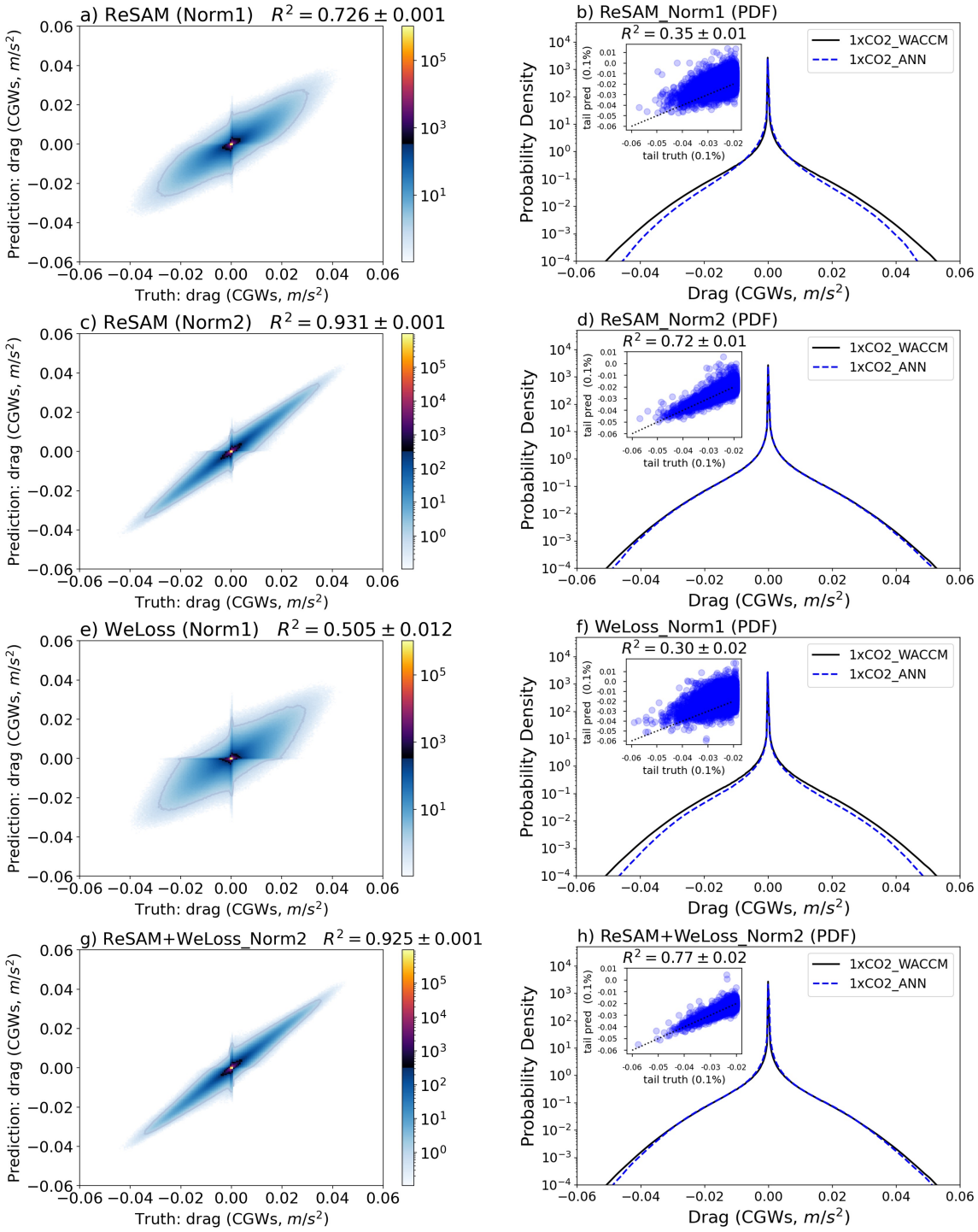

Figure 5. Similar to Figure 4, but for CGWs with the proposed ReSAM and WeLoss methods. a) A 2D histogram for the emulation with resampled data (ReSAM) after using Norm1; b) Distribution of the emulated GWD due to CGWs similar to Figure 4b, but with ReSAM applied; c) Similar to a), with training data normalized using Norm2; d) Similar to b), with training data normalized using Norm2; e) Similar to a), but with the WeLoss approach; f) Similar to b), but with the WeLoss approach; g) Similar to c), after applying both ReSAM and WeLoss methods together; h) Similar to d), after applying both ReSAM and WeLoss methods.

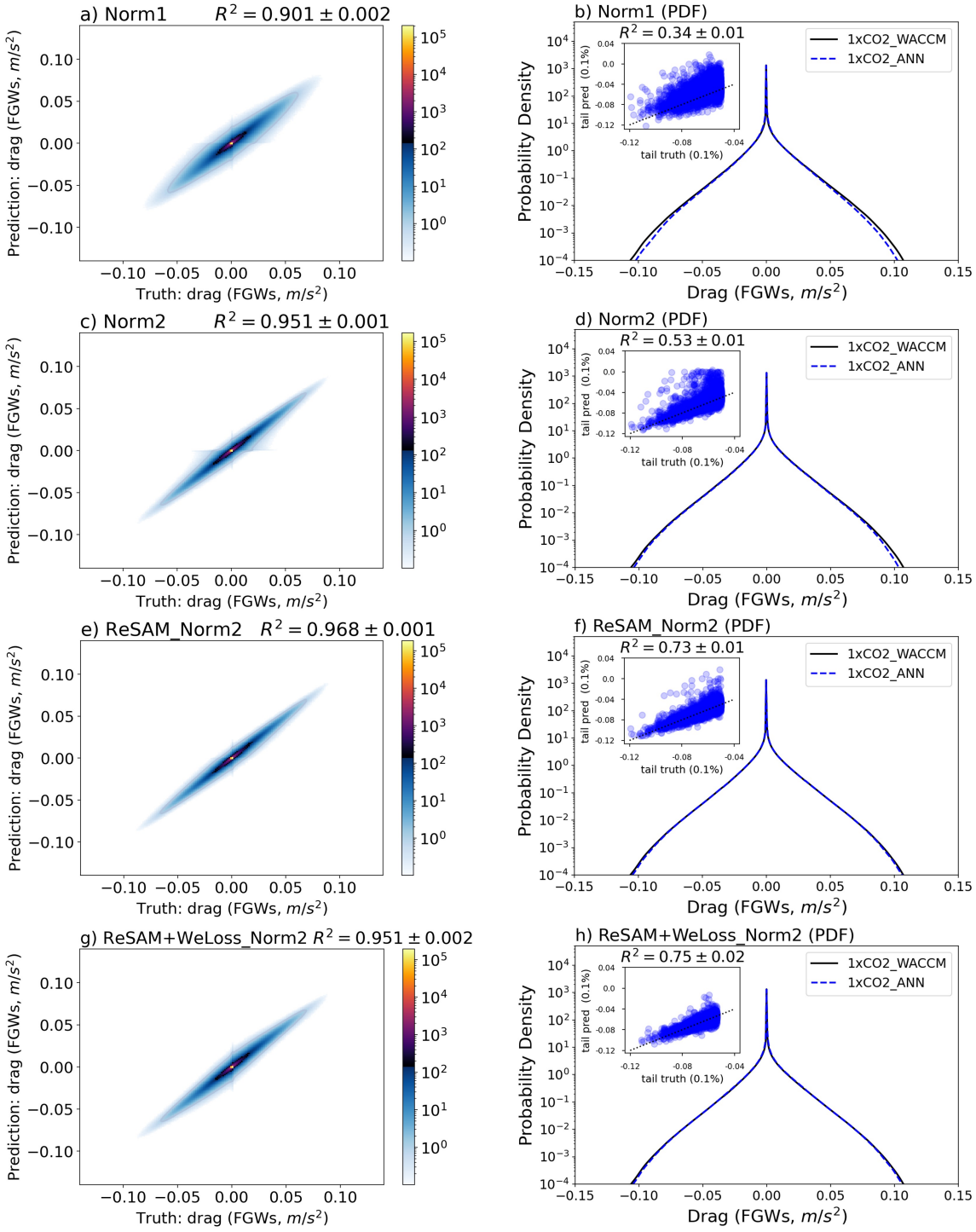

Figure 6. Similar to Figure 5, except for FGWs. a) A 2D histogram for the emulation using Norm1, without ReSAM or WeLoss; b) Distribution of the GWD due to FGWs with NORM1, similar to Figure 4b; c) Similar to a), with training data normalized using Norm2; d) Similar to b), with training data normalized using Norm2; e) Similar to c), but with the ReSAM approach; f) Similar to d), but with the ReSAM approach; g) Similar to Figure 5g, applying both ReSAM and WeLoss methods to the FGWs; h) Similar to Figure 5h, applying both ReSAM and WeLoss methods to the FGWs.

In summary, data imbalance can pose challenges when learning from data that closely resembles real-world data (further discussed in the subsequent section on emulating OGWs). Proper resampling techniques can significantly enhance the NNs' performance by improving dataset balance. Furthermore, modifying the loss function to penalize the NNs more for missing extreme values can further improve performance at the tails of the PDF. For the remainder of the paper, unless otherwise specified, we continue to employ the ReSAM approach and the standard loss function with NORM2 unless stated otherwise.

3.2 Uncertainty Quantification

As outlined in subsection 2.2.2, we employ three different methods (i.e., BNN, DNN, and VAE) to quantify the uncertainty of predictions during inference (testing). For this purpose, an ensemble of 1000 members is generated by running each UQ-equipped NN 1000 times for each input from the testing set. Figure 7 presents sample profiles of zonal GWD derived from the deterministic NN (ANN) and the three UQ-equipped NNs, alongside the true GWD profiles from WACCM. Note that these examples have not been used in the training or validation process. It is evident from the figure that all three UQ-equipped NNs show reasonable skill in predicting the complex profiles of GWD due to CGWs and FGWs (also reflected in R-squared in Table 1), albeit with a slight decrease in accuracy compared to ANN. As discussed earlier, a valuable uncertainty estimate should correspond closely with the NN's test accuracy, providing insights into when to trust the NN's prediction during inference. Such a relationship can be seen in a few randomly chosen GWD profiles that's shown in Figure 7. In each pair of CGW and FGW profiles, the left column shows the estimated uncertainty is also low when the prediction error is low, indicating the NN's confidence in its accurate predictions. In contrast, the right column, which generally represents more complex profiles, exhibits the NN's less accurate predictions, and increased uncertainty, highlighted by the wider confidence intervals.

While Figure 7 demonstrates the performance of the UQ methods for just a few GWD profiles, the spread-skill plots shown in Figure 8 offer a broader perspective based on 60,000 profiles, following the calculations detailed in Appendix C. It is evident from the plots that all three UQ methods produce reasonably informative uncertainty estimates, as their curves closely align with the 1-to-1 line. In the case of CGWs, all data points are above the 1:1 line, indicating a slight overconfidence (underdispersiveness) across all three UQ methods, with the DNN being slightly closer to the 1-to-1 line. For the FGWs, the DNN demonstrates slightly better performance, although it marginally drops below the 1-to-1 line in the first few bins, indicating a slight underconfidence. Notably, it can be seen from the spread frequency inset that the vast majority of the data points are within the first few bins, for which both spread and skill values are small, and they are generally closer to the 1-to-1 line.

It should also be noted that for the large values of model spread ($\overline{\mathrm{SD}}$), there is only a very limited number of data points, as is evident from the inset histograms. Consequently,

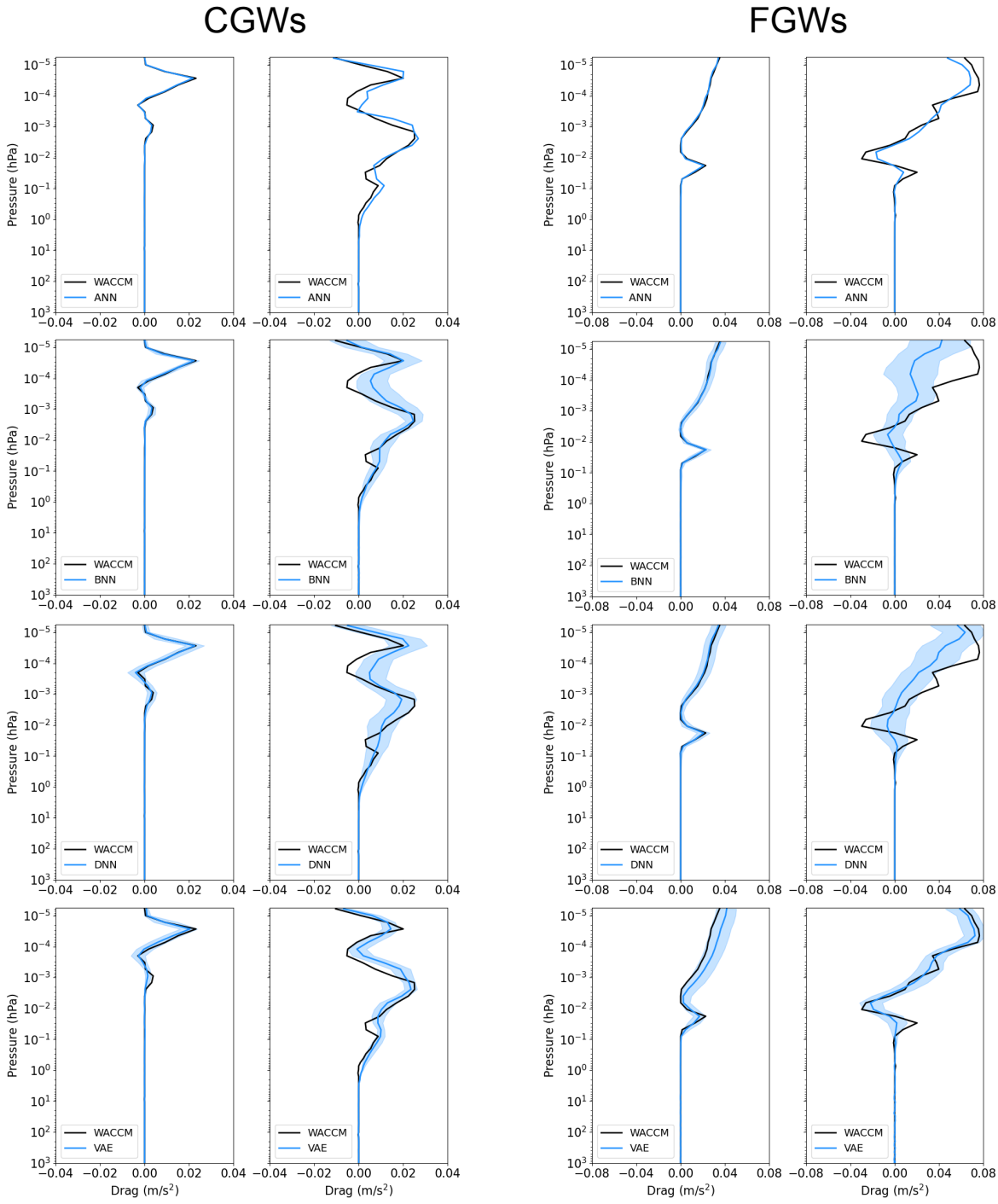

Figure 7. Sample profiles of zonal GWD as predicted by various NNs, as indicated. The true profile is shown by the black line, while the blue solid line represents the mean of 1000 ensemble members. The shaded region indicates the 95% confidence interval. In each pair of CGWs and FGWs profiles, the left column provides examples with low estimated uncertainty, corresponding to instances of low error. Conversely, the right column illustrates cases with high uncertainty when the error is high.

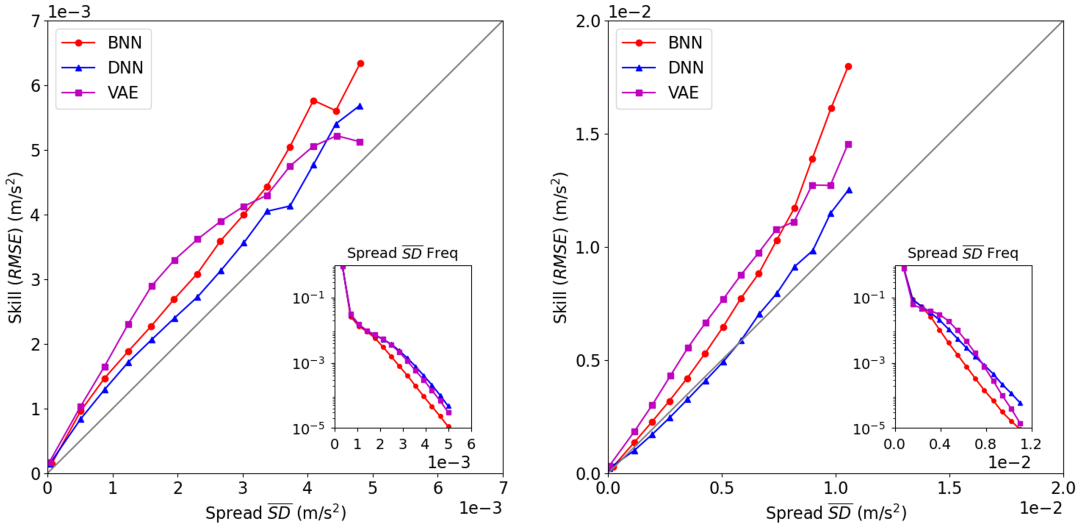

Figure 8. Spread-skill plot for GWD due to (left) CGWs, and (right) FGWs. The diagonal 1:1 line represents the perfect spread-skill line. Points above (below) this line correspond to spread values where the model is overconfident (underconfident). The inset histogram shows how often each spread value occurs. See Appendix C for a detailed discussion on the calculations of the spread-skill plot.

the standard deviation (STD) can become a misleading measure of spread because of the non-normal distributions.

To summarize the quality of the spread-skill plots for the three UQ methods, we explore the metrics introduced in subsection 2.2.2 and Appendix C (see Table 1). The R-squared value for the ensemble mean prediction is also given to show the accuracy of each UQ method. Based on SSREL, whose ideal value is zero, BNN shows the best performance for both CGWs and FGWs. However, if we check SSRAT, where 1 is the optimal number, DNN is the best among these three methods. This discrepancy can be explained by a closer look at the Equations (C2) and (C3). SSREL, which is a bin-weighted average difference, is most sensitive to the performance of the NN in the first bin, where the vast majority of the data points are located (see the inset histograms in Figure 8), while SSRAT is more influenced by larger values of spread and skill. Accordingly, the VAE shows the highest values of SSREL, which is indicative of its sub-optimal performance in the first bin, where there are small values of spread and skill.

In the results presented in Figure 8 and Table 1, each height level of a GWD profile is considered as an individual sample. A zonal GWD profile, with its 70 vertical levels, thus constitutes 70 distinct samples. While analyzing these samples offers insights into the NN's overall performance by averaging statistics across numerous profiles, our primary interest is often in the uncertainty associated with an individual GWD profile. This uncertainty can then aid in determining whether to trust/use the NN's prediction for that particular GWD

profile. Therefore, we use Equation (C4) to assess the relationship between uncertainty and test accuracy for each GWD profile. Furthermore, to estimate uncertainty, here we use the interquartile range (IQR) to reduce the influence of outliers.

Figure 9 shows the Gaussian kernel density of spread against RMSE for all 60,000 profiles, as indicated by the color shading. The x-axis represents the IQR of each GWD profile, while the y-axis denotes its corresponding RMSE. A strong correlation between the two is observed across all three UQ methods. Consequently, GWD profiles with larger uncertainties often coincide with larger errors. Figure 9 also shows a close similarity between BNN and DNN. In contrast, VAE tends to provide marginally larger uncertainties, especially for FGWs. This is consistent with VAE's slightly reduced accuracy as indicated in Table 1. Overall, given the monotonic relationship between the uncertainty and test error, these results show that all three UQ methods provide useful and informative uncertainty for with-distribution test samples. A user can set a threshold on uncertainty based on their tolerance for error (RMSE) and decide whether they trust the NN for a given input sample.

The results presented so far show the performance of the UQ methods based on the testing data, i.e., data from the current climate. However, the effective performance of UQ methods can also be tested (perhaps more meaningfully) on OOD data, e.g., data from a warmer climate. This is particularly relevant for climate change studies. Accordingly, we evaluate the performance of these trained NNs with input data from the future climate, as depicted by the black lines in Figure 9. For FGWs, the spread-skill relationship remains largely similar, especially for BNN and DNN. This suggests that, based on their uncertainties, we can still reliably estimate the error in the NN predictions for FGWs for the warming climate. A similar pattern is observed for the VAE, though it exhibits increased uncertainties and higher errors with OOD data. As shown in a later section, for FGWs, the NNs generalize to the warmer climate without any further effort.

In contrast, for CGWs, given the same level of uncertainty, the error in NN predictions increases significantly for the OOD data compared to that from the current climate, which means the spread-skill relationship, especially for the BNN and DNN, fails to generalize to the OOD data. From this perspective, VAE performs better, showing that for the same level of uncertainty, the increase in error is not as substantial as in BNN and DNN. The VAE also yields considerably higher uncertainty estimates for future climate, which may aid in the detection of OOD data. The observed discrepancies in the performance of the NNs for CGWs and FGWs hint at different levels of their generalizability, a topic we will delve into more deeply in the following subsection.

In summary, while the three UQ methods provide credible and valuable uncertainty estimates for the current climate, the BNN and DNN are confidently wrong in estimating CGWs in a warmer climate although VAE shows some promising results. This problem is common among various UQ techniques as pointed out in the ML literature: they frequently show overconfidence when assessed with OOD data (e.g., Ovadia et al., 2019). The optimal UQ method selection depends on the specific metric of interest and the intended application. While BNN is more broadly used in the literature and gives the best accuracy, DNN could

Table 1. Evaluation scores for the three UQ methods. See Section 2 for more details.

	CGWs			FGWs		
	BNN	DNN	VAE	BNN	DNN	VAE
SSREL (1e-4)	1.29	1.48	2.14	1.20	1.69	5.21
SSRAT	0.73	0.82	0.72	0.69	0.93	0.69
R-squared	0.90	0.86	0.87	0.94	0.92	0.89

achieve similar performance and is often more practical given its simplicity. On the other hand, VAE seems to perform better when applied to OOD data, at least in the one test case here. These observations warrant further research in the future using multiple test cases and climate-relevant applications. We also note here that each method has multiple tuning hyperparameters to optimize its uncertainty quantification. Consequently, the discrepancies among the three methods could potentially be mitigated with proper hyperparameter tuning (as discussed in Appendix B).

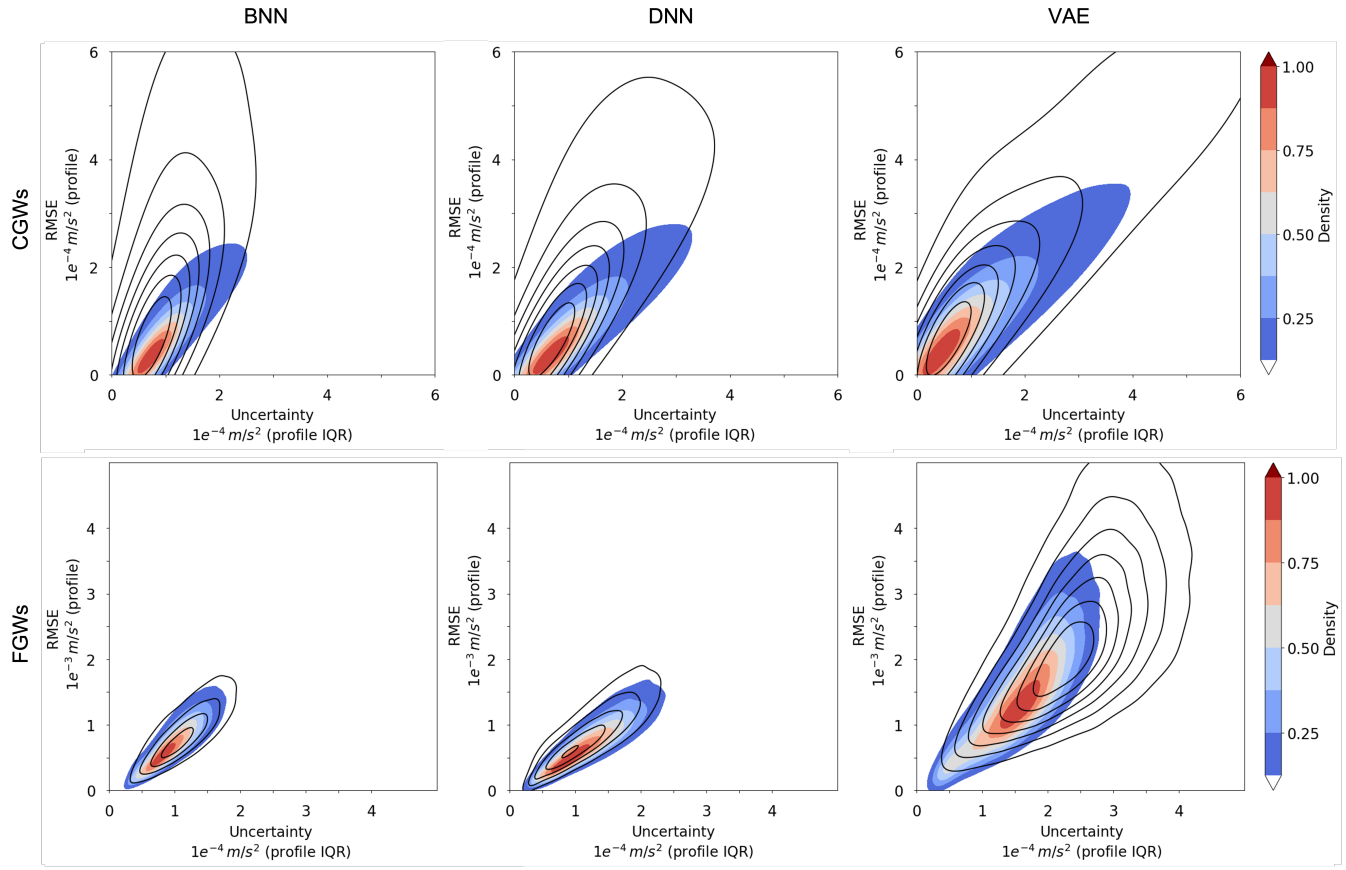

Figure 9. Gaussian kernel density for spread, defined as interquartile range (IQR) versus RMSE for (top) CGWs, and (bottom) FGWs, when validated against out-of-sample data from the current climate (color shading) and out-of-distribution data from the future, warmer climate (black lines). Results are shown for BNN (left), DNN (middle), and VAE (right). All NNs are exclusively trained on data from the current (control) climate.

3.3 Out-of-distribution (OOD) Generalization via Transfer Learning

As previously discussed, the GWP schemes in WACCM are coupled to their sources, which might change in a warmer climate. Specifically, under $4\times CO_2$ forcing, we expect changes in both the amplitude and the phase speed distribution of GWs, in particular for the CGWs, due to their built-in sensitivities to changes in the convection. Consequently, the physics scheme in WACCM produces slightly stronger GWD for CGWs, especially in the tail of the distribution. This intensified GWD results in a shorter quasi-biennial oscillation (QBO) period in WACCM. However, it is important to recognize that the response of the QBO to climate change differs across various general circulation models (Richter et al., 2022).

The intensification of the CGWs in future climate simulations presents an opportunity to study how NNs handle the OOD data. Our findings in the UQ section already suggest increased prediction errors when testing NNs with OOD data, which raises concerns about their applicability in climate change studies. To more thoroughly investigate this issue, we conduct additional evaluations on our ANNs, by applying them to data samples from future climate simulations, as illustrated in Figure 10. It is clear that the ANN for the CGWs does not generalize well, evidenced by a decrease in R^2 from 0.93 to 0.79. The ANN particularly struggles to capture the increase in GWD in the tail, with R^2 for the tails decreasing from 0.72 to 0.36. As a result, it seems unlikely that this emulator will accurately reproduce changes in the circulation under different climate conditions, such as the shorter QBO period resulting from future warming in WACCM.

In contrast to CGWs, the amplitude of FGWs shows a less marked increase in the future climate, and their PDF distribution closely resembles that of the control simulations. As a result, the ANN demonstrates better generalizability for FGWs when it is tested against future climate data, as seen in Figure 10d. There is only a slight decrease in the ANN's performance, with R^2 dropping from 0.97 to 0.95.

Two factors can contribute to the considerable OOD generalization errors in an NN when applied across two distinct systems. First, the input-output relationship might vary between the two systems. Second, the input variables in the new system could originate from a distribution different from that of the original system (regardless of whether the input-output relationship remains the same or changes). The former is hard to quantify in a high-dimensional dataset. The latter can be quantified using similarity distances. To help us better understand these differences between the OOD generalizability of CGWs and FGWs, we assess the similarity between their input and output data distributions from control and future climate simulations using the Mahalanobis distance (D). The Mahalanobis distance is a measure of the distance between a data point and a distribution (Ling & Templeton, 2015). Specifically, it is a multi-dimensional generalization of the idea of measuring how many standard deviations away a point is from the mean of the distribution. The application of Mahalanobis distance in understanding the source of OOD generalization errors in data-driven parameterization was previously demonstrated in Guan et al. (2022) for a simple turbulent system.

Table 2. Change of Mahalanobis distance based on the ratio of the average distance of the points that are more than 3 standard deviations away from the mean. The choice of the variables here is based on Appendix A, showing u, v, T, and source function contain most of the information needed for the NN.

Variables	u	v	T	Source (diabatic heating for CGWs, frontogenesis for FGWs)	Zonal drag	Meridional drag
Distance (Convection)	1.03	1.00	1.19	3.62	1.42	1.44
Distance (Front)	1.03	0.96	1.50	1.10	1.00	1.00

To use the Mahalanobis distance, we first calculate the mean and covariance matrix of the training data from the control run. We then analyze the distribution of Mahalanobis distances in this training data, setting a baseline value, referred to as D_{ctrl}. This baseline is the average distance for data points that deviate by more than 3 standard deviations from the mean. This choice aims to focus on outliers for which extrapolation is more challenging. Following this, we apply the same process to the data points in the future climate dataset, denoted as D_{warm}. Table 2 presents the ratio of D_{warm} for the warming scenario to D_{ctrl} for the control scenario for selected variables. When this ratio is close to 1.0, it suggests minimal changes in this variable's distribution under a warming scenario. Note that the NNs trained based only on these variables demonstrate performance comparable to NNs trained on all variables (not shown), which is why we only focus on these few key variables.

The results reveal that among the various variables significantly contributing to the emulation of CGWs, diabatic heating (source of CGWs) is the sole variable that exhibits substantial changes from the control to the warming scenario. Conversely, changes in variables used to emulate FGWs are considerably smaller. This outcome suggests that the likely reason for the better generalizability of FGWs is that the input distribution remains almost unchanged (and the input-output relationship, which is the physics scheme, remains the same too).

To improve the generalizability of the emulator for CGWs, we explore TL, a technique introduced earlier and proven to be a powerful tool for improving the OOD generalizability of data-driven parameterization in canonical turbulent flows (e.g., Guan et al., 2022; Subel et al., 2023). Rather than re-training the entire NN for future climate scenarios, we only re-train, follwoing Subel et al. (2023), just a portion of the NN, thereby requiring only a small fraction of the data. Figure 10e showcases the emulation results after only re-training the first hidden layer of ANN using data from the first month of the WACCM simulation in the $4\times CO_2$ scenario, which amounts to approximately 1% of the original training dataset. After applying TL, the performance of the emulator in the warming scenario significantly improves, with R^2 rising from 0.79 to 0.91, nearly matching its performance in the control simulations ($R^2 = 0.93$). However, the improvement in the PDF tails is less pronounced, showing only a modest increase in R^2 from 0.36 to 0.51. This is likely due to the limited number of large-amplitude GW events within the one-month period. Instead of using more data

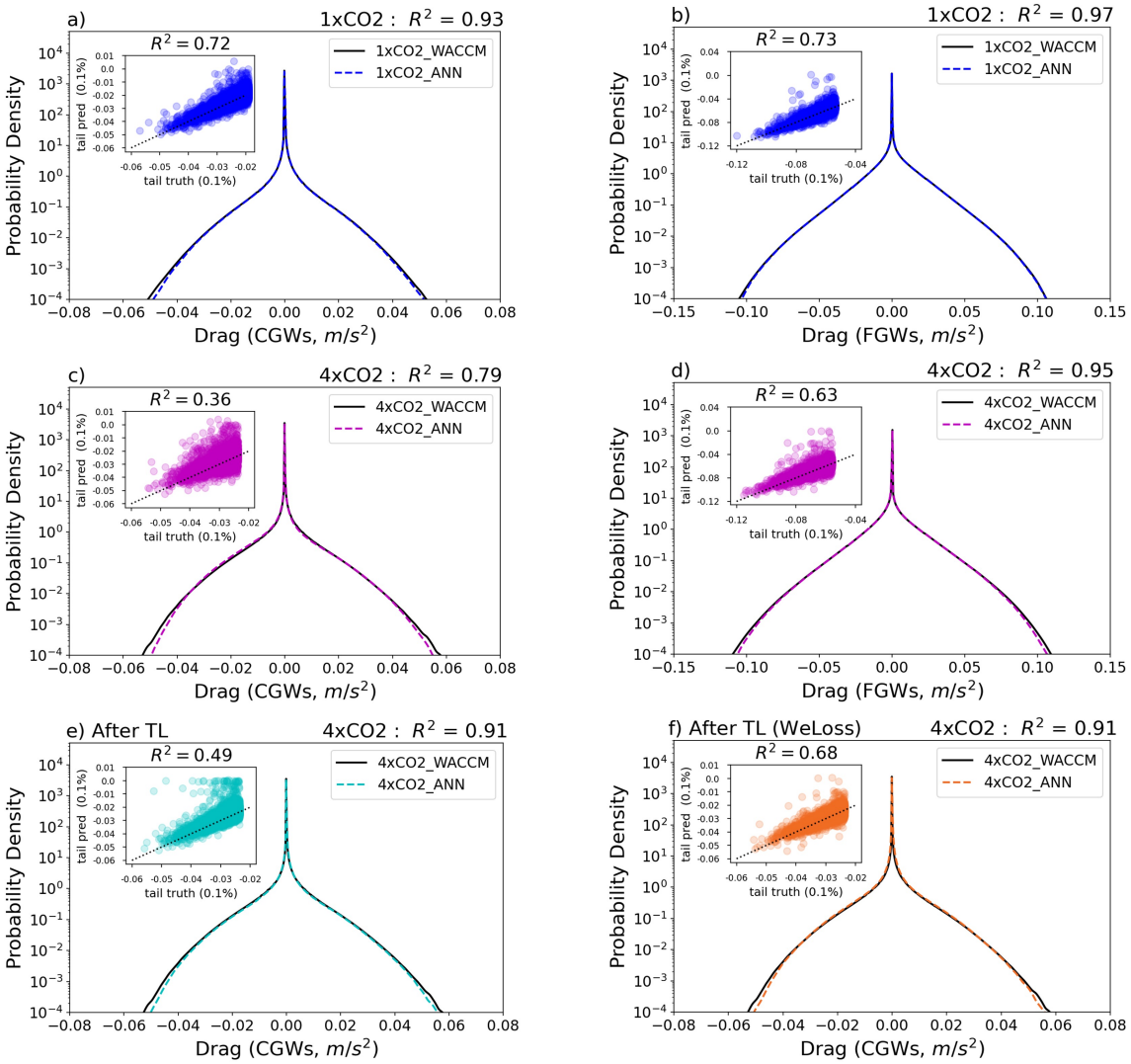

Figure 10. NN performance for pre-industrial and warming scenarios for different sources (a,c,e,f: CGWs ; b,d: FGWs). **a)** PDF of GWD due to CGWs in WACCM simulation and the predicted CGWs using NN emulator, scatter plot shows points for the tail part only. **b)** same as (a), but for FGWs. **c)** same as (a), but for the warming scenario, **d)** same as (b) but for the warming scenario. **e)** same as (c) but after applying transfer learning to the first hidden layer of the NN with 1-month WACCM simulation data under warming scenario ($\sim$ 1% of the size of the training data) **f)** same as (e) but with the weighted loss function used when we conduct transfer learning (WeLoss).

from the future climate (which is challenging to obtain in a realistic situation), we leverage the WeLoss approach, described earlier, during re-training. This modification results in a significant improvement in the tail, with R^2 increasing from 0.51 to 0.68. Note that this improvement in the tail is critical, as inadequate learning of these rare but large-amplitude GWDs can result in significant errors and instabilities.

We would like to point out that during the TL experiments, we have examined the effects of re-training each individual hidden layer of the NN. Our findings indicate that re-training the first layer yields the best results, which aligns with the findings in Subel et al. (2023). Re-training the last layer only brings marginal improvements to the NN (not shown). Notably, our experiments involving re-training the first two layers did not result in further performance enhancements, suggesting that the number of neurons is not the primary factor contributing to the varied performance observed when re-training different layers.

Similar results regarding TL are also observed with other NNs used in this study. For instance, Figure 11 presents the same plot as Figure 10, but for the BNN. It is evident that BNN also struggles with generalization to OOD data, as could also be interpreted based on the results presented in section 3.2. It is also the case for DNN and VAE (not shown). Overall, when these NNs are tested against the $4\times CO_2$ future climate data, their accuracy is not better than the deterministic ANN. However, methods with UQ, especially the VAE (see Figure 9), could potentially indicate the increased uncertainty when testing with input data from the $4\times CO_2$ integration. These results underscore the necessity of re-training the NNs using TL.

4 Emulation of Orographic GWs (OGWs)

Similar to Chantry et al. (2021), our initial attempts to emulate OGWs did not succeed, primarily due to the presence of a pronounced data imbalance. Notably, the physics-based scheme responsible for OGW generation operates exclusively over terrestrial regions. However, it is surprising that the issue of data imbalance continues to persist, even when we limit our NN training and testing exclusively to columns located over land (Figure 12a). Still, the emulated OGW drag often remains close to zero and completely fails to predict the rare events (Figure 12b), which poses a considerable hurdle for the emulator's performance. Further investigations reveal that the key to this problem lies in the highly localized nature of orographic GWD, where significant drag is observed only at a handful of specific locations. Furthermore, even within these limited regions, GWD exhibits a significant intermittent behavior. To help our understanding, we also conducted a K-means clustering analysis, categorizing GWD data for all land-based columns (Table 3). Among the 6 clusters, cluster 4 accounts for a staggering 97.51% of the dataset. Remarkably, all samples within this cluster exhibit exceptionally weak orographic GWD, as evidenced by the cluster center's maximum GWD amplitude, which is two orders of magnitude smaller than that of other clusters.

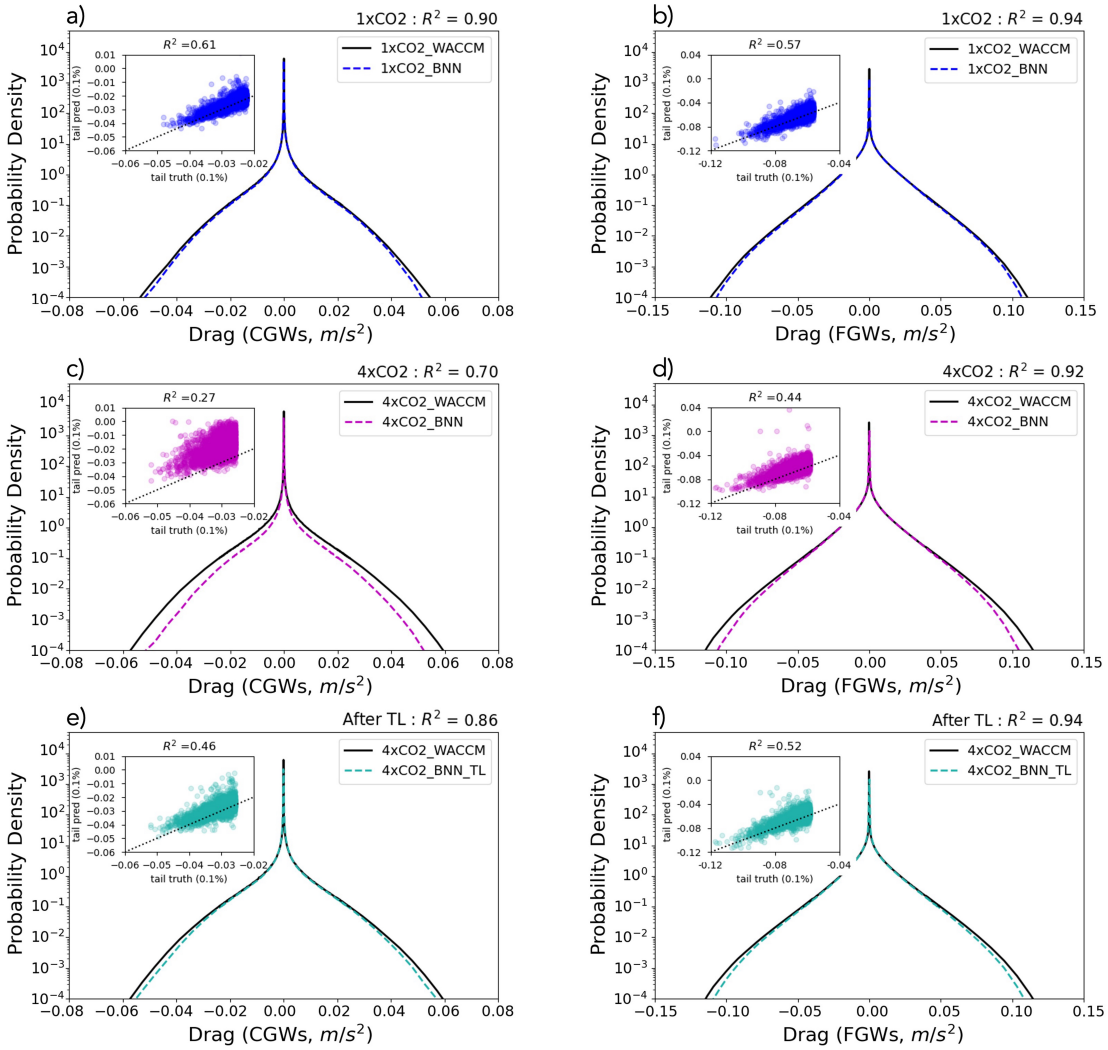

Figure 11. Panels (a) to (e) are the same as those in Figure 10 but for BNN. Panel (f) shows emulation for FGWs under warming scenario after applying transfer learning to the first hidden layer of the NN.

Table 3. Clustering analysis for OGWs. Analysis is done for all columns over land in the training data.

Cluster	Frequency (%) in the training data	Maximum GWD amplitude of cluster center
c1	0.18	8.7 e-3 m/s^2
c2	0.13	4.4 e-3 m/s^2
c3	0.93	3.6 e-3 m/s^2
c4	**97.51**	2.8 **e-5** m/s^2
c5	0.15	2.1 e-3 m/s^2
c6	1.10	4.3 e-3 m/s^2

To overcome this persistent data imbalance in the OGWs, we first separate all columns over land into large-drag columns (with column maximum greater than one STD of all GWD from OGWs) and small-drag columns. We then perform subsampling on the latter group only to create a more balanced dataset. To improve NN training, we also include all columns from the 6-year simulation to augment the sample size of the large-drag columns. Figures 12c and 12d illustrate the performance after re-balancing the dataset. Notably, the result represents a substantial improvement, evidenced by an R^2 increase from 0.29 to 0.80, and also a significant improvement in the accuracy for rare events. While we acknowledge that this skill remains lower than what is achieved for CGWs and FGWs, it already signifies a reasonable NN. Furthermore, we posit that by incorporating additional training data (either by extending the WACCM model integration or simply augmenting the data with OGWs scheme only), we can further improve our emulation results. The possibility of achieving superior emulation outcomes through the adoption of an alternative NN architecture is also possible, although such exploration is beyond the scope of this paper.

5 Summary and Discussion

Through the emulation of complex GWPs in a state-of-the-art atmospheric model (WACCM), we have elucidated and explored solutions for three critical challenges in the development of ML-based data-driven SGS schemes for climate applications: data imbalance, UQ, and OOD generalizability under different climates. A brief summary is provided below:

1. In the presence of non-stationary, and highly imbalanced datasets, such as those encountered in WACCM, specialized approaches (e.g., resampling and weighted loss function) are essential to enhance the performance of data-driven models. Through resampling, we have successfully trained a robust NN emulator for OGWs, a challenging task as demonstrated in Chantry et al. (2021). The effectiveness of the trained emulator is also significantly influenced by the choice of the loss function used during training. In our case, while a weighted loss function (WeLoss) does not improve the overall R^2 score, it yields significant improvements in the emulation results for

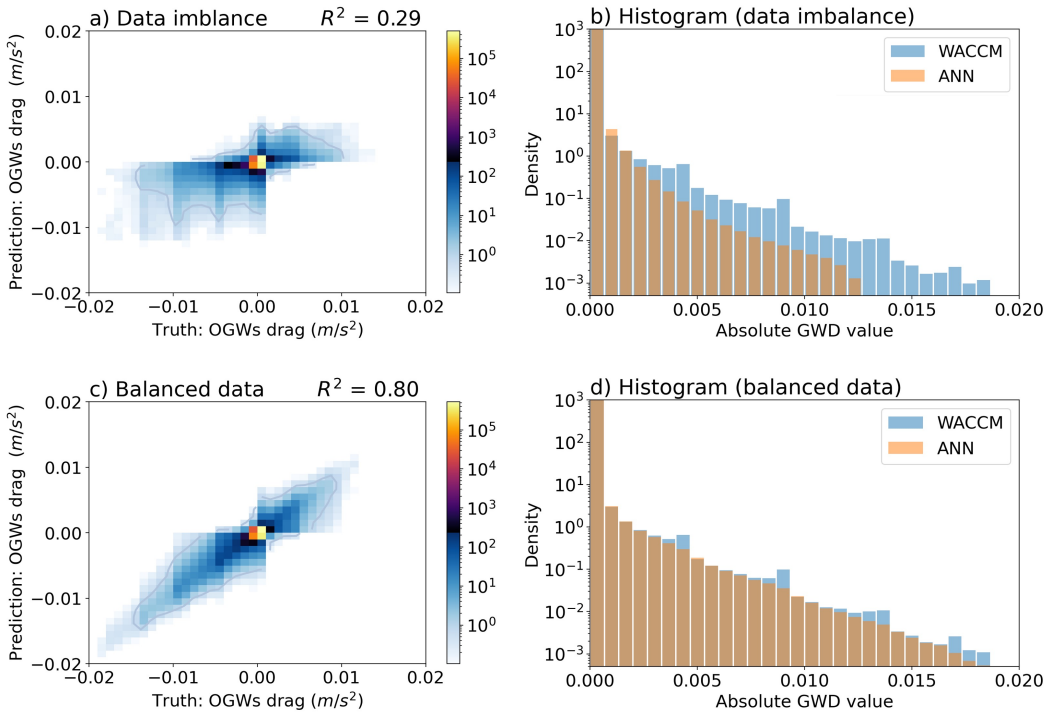

Figure 12. Performance of the emulator for OGWs when trained with all columns over land (panel a) & panel b)) and balanced training data with a balanced number of large-drag columns (column maximum > 1 STD of all GWD from OGWs) and small-drag columns (panel c) & panel d).

the PDF tails of the GWD. This finding aligns with those in Lopez-Gomez et al. (2022), where their custom loss function, tailored to emphasize extreme events, led to substantial improvements in predicting heatwaves.

2. All three UQ methods employed in this study provide reasonable uncertainty estimates for GWD prediction for the current climate. The spread-skill plots (refer to Figures 8 and 9) indicate that greater uncertainty corresponds to a larger prediction error. Yet, the reliability of UQ methods diminishes when they are challenged with OOD data. Both BNN and DNN used in this study tend to be overconfident in estimating CGWs in a warmer climate, thereby struggling to identify OOD samples. The VAE, on the other hand, yields rather promising results in providing useful UQ for OOD data. Given the variations in different methods, the metrics selected to assess the SGS model will play a significant role in determining the choice for the UQ methods. We also note that further optimization of tunable parameters within each UQ method could affect their performance (refer to Appendix C).
3. Our findings illustrate the challenges SGS schemes face in generalizing to OOD data and extrapolating to new climates. Nonetheless, the TL approach has proven highly effective in aiding an NN to extrapolate to different climates. For CGWs in WACCM, the physics-based scheme exhibits larger GWD under $4\times CO_2$ forcing, primarily due to an increase in diabatic heating from convection. With only one month of sim-

ulation data from this future warming scenario (representing approximately 1% of the original training dataset), we successfully reduce its OOD generalization error through re-training the first layer of the NN, following the findings of Subel et al. (2023). Additionally, we have illustrated the value of metrics like the Mahalanobis distance in assessing the potential OOD generalizability of NNs.

We would like to emphasize that these challenges are often intertwined. For instance, addressing data imbalance in CGWs is a prerequisite for obtaining an accurate NN model, which, in turn, impacts UQ and OOD generalizability assessments. Moreover, there exists a strong link between UQ and OOD generalizability evaluations: if the NN struggles with OOD generalization, performing poorly with OOD data, the reliability of UQ for such data (e.g., data from a warmer climate) also becomes questionable. This presents a substantial challenge for UQ methods, especially for climate change research where reliable UQ methods are crucial.

This study has primarily focused on offline skill assessment. We acknowledge that good offline performance (at least in terms of common metrics such as R^2) is not necessarily an indicator of stable and accurate online (coupled to climate model) performance (Ross et al., 2022; Guan et al., 2022), though more strict metrics such as R^2 of the PDF tails might better connect the offline and online performance (Pahlavan et al., 2023). However, for the purpose of this study, which is to provide a testbed to test ideas for data imbalance, UQ, and OOD generalization with transfer learning, the offline tests, particularly using the several metrics we have used, suffice. That said, the main reason that we have not provided online results is that coupling various complex NNs, with the same framework, to complex climate models (e.g., WACCM) without slowing down the model is a challenging and time-consuming task (Espinosa et al., 2022), and this is work in progress.

Emulating complex GWPs within the WACCM provided a unique opportunity to address three critical challenges in developing ML-based, data-driven SGS schemes for climate science applications. However, it is crucial to acknowledge that such emulated schemes essentially adopt the limitations inherent in the physics-based schemes. Addressing these limitations, the next step is to harness high-resolution data from GW-resolving simulations, which are carefully validated against observational data. A library of such high-resolution simulations, notably of convectively generated GWs using the Weather Research and Forecasting (WRF) model, is now established (Sun et al., 2023), alongside additional global high-resolution simulations (Wedi et al., 2020; Polichtchouk et al., 2023; Köhler et al., 2023). The next phase involves integrating the approaches outlined in this study with the data from these GW-resolving simulations to develop a stable, trustworthy, and generalizable data-driven GWP scheme. This scheme is then expected to overcome the limitations of physics-based GWPs and potentially incorporate features like the transient effect (Bölöni et al., 2021; Kim et al., 2021) and lateral propagation of GWs (e.g., Sato et al., 2009)—marking a significant advancement towards next-generation GWP schemes.

Appendix A Input/output variables for the physics-based GWP schemes and their emulators

We use the exact same inputs as those of each GWP scheme in the WACCM for the training of the NN-based emulator of that scheme. These inputs are listed in Table A1. As for the outputs, we only consider the zonal and meridional drag forcings. The GWPs in WACCM also estimate additional effects of the GWs that result in changes of temperature profile and vertical diffusion. These outputs are not considered in our emulations.

Table A1. List of the input and output variables for the NNs trained as emulators of the GWP schemes in WACCM. The numbers in parentheses in front of each variable are the number of vertical levels for that variable. Note that each input and output is a 1D column at a given latitude/longitude grid point. Diabatic heating in WACCM is provided by the cumulus scheme. The topography variables listed in the table are mxdis (height estimates for ridges), hwdth (width of ridges), clngt (length of ridges), angll (orientation of ridges), and anixy (anisotropy of ridges).

<table>
<tr><th rowspan="2">GWP</th><th colspan="3">Input</th><th rowspan="2">Output</th></tr>
<tr><th>pressure levels</th><th>surface level</th><th>forcing</th></tr>
<tr><td>CGWs</td><td rowspan="4">u(70),
v(70),
T(70),
z(70),
ρ(71),
Brunt–Väisälä frequency N (70),
dry static energy DSE (70)</td><td rowspan="4">lat (1),
lon (1),
$P_{surface}$ (1),</td><td>diabatic heating (70)</td><td rowspan="4">zonal drag
GWD$_x$ (70),
meridional drag
GWD$_y$ (70),</td></tr>
<tr><td>FGWs</td><td>frontogenesis
function (70)</td></tr>
<tr><td rowspan="2">OGWs</td><td rowspan="2">mxdis (16),
hwdth (16),
clngt (16),
angll (16),
anixy (16),</td></tr>
</table>

From Table A1, one can guess that some input variables are correlated with each other. Consequently, it is plausible that the trained NNs may have spurious connections. Preliminary tests further support this notion, indicating that employing only u, v, T, and the forcing function as inputs yields comparable offline skill (results not presented here).

Appendix B Tuning UQ-equipped NNs

In addition to the hyperparameters of the deterministic NNs, designing an architecture for UQ often demands additional hyperparameter optimization. For instance, for the DNN, decisions need to be made regarding the number of neurons to drop out (dropout rate). While less common, one can also choose whether to apply dropout to all hidden layers or only selected ones. Variations in the dropout rate and the layers to which dropout is applied can influence the final configuration and performance of the DNN. Figure B1 illustrates these effects. As we increase the number of dropped neurons (whether through a higher dropout rate or by subjecting more layers to dropout), the uncertainty in the DNN predictions tends to rise. Yet, there is a persistent pattern in the relationship between spread (IQR) and RMSE across the various plots in Figure B1. Specifically, as spread increases, RMSE concurrently grows, consistent with the insights highlighted in Figure 9.

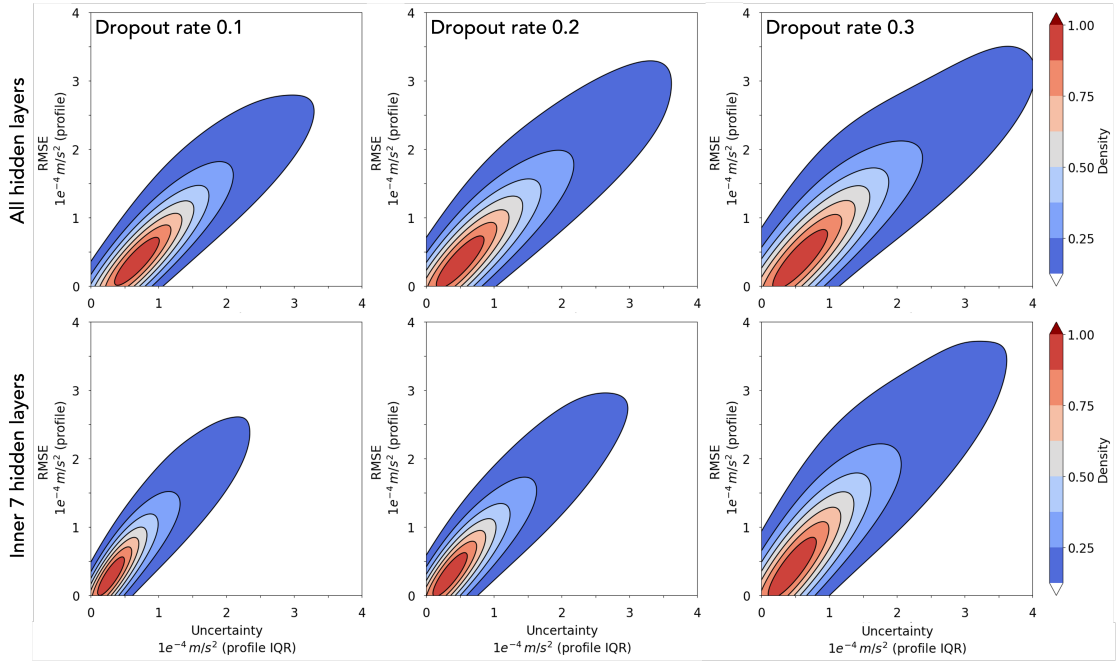

Figure B1. Similar to Figure 9 but for DNN only with different dropout rates, which are applied to different numbers of hidden layers.

In the case of BNN or VAE, even though there is no dropout rate, there are distinct tuning opportunities available. For instance, with the VAE, one might consider applying dropout to the NN emulator. Moreover, given that the loss function in VAE comprises three components, decisions can be made regarding which component to penalize more heavily, allowing for nuanced adjustments to its performance.

Appendix C The UQ metrics

Each point in the spread-skill plot corresponds to one specific bin of ensemble spread ($\overline{\mathrm{SD}_k}$), which is defined as the average standard deviation of the ensemble members. We first separate the spread using a pre-selected number of bins (a subjective choice of 15 is used here). Then for the k^{th} bin:

$$\begin{cases} \mathrm{RMSE}_k = \left[\frac{1}{N_k}\sum_{i=1}^{N_k}(\hat{y_i} - \overline{y_i})^2\right]^{\frac{1}{2}} \\ \overline{\mathrm{SD}_k} = \frac{1}{N_k}\sum_{i=1}^{N_k}\left[\frac{1}{M-1}\sum_{j=1}^{M}(\overline{y_i} - y_{ij})^2\right]^{\frac{1}{2}} \\ \overline{y_i} = \frac{1}{M}\sum_{j=1}^{M} y_{ij} \end{cases} \tag{C1}$$

$\hat{y_i}$ is the observed value for the i^{th} example, $\overline{y_i}$ is the mean prediction for the i^{th} example, y_{ij} is the j^{th} prediction for the i^{th} example, N_k is the total number of examples in the k^{th} bin, and M is the ensemble size. Following Haynes et al. (2023), we summarize the quality of the

spread-skill plot by two measures: spread-skill reliability (SSREL) and overall spread-skill ratio (SSRAT). SSREL is the bin-weighted mean distance from the 1-to-1 line:

$$\mathrm{SSREL} = \sum_{k=1}^{K} \frac{N_k}{N} |\mathrm{RMSE}_k - \overline{\mathrm{SD}_k}| \tag{C2}$$

where N is the total number of examples, K is the total number of bins, and other variables are as in Equation C1. SSREL varies from $[0, \infty)$, and the ideal value is 0. On the other hand, SSRAT is averaged over the whole dataset:

$$\mathrm{SSRAT} = \frac{\overline{\mathrm{SD}}}{\mathrm{RMSE}} \tag{C3}$$

SSRAT also varies from $[0, \infty)$, and the ideal value is 1. SSRAT > 1 indicates the model is under-confident on average, while SSRAT < 1 indicates that the model is overconfident on average.

In Equation (C1), each level of a GWD profile is considered as an individual sample. As discussed earlier, while these samples help assess the model's overall performance, our main interest is often the uncertainty of individual GWD profiles. Such uncertainty informs the trustworthiness of the model's prediction for that specific profile. Accordingly, for each profile, we can compute:

$$\begin{cases} \mathrm{RMSE}_{\mathrm{profile}} = \left[\frac{1}{N_z} \sum_{z=1}^{N_z} (\hat{y_z} - \overline{y_z})^2 \right]_{\mathrm{profile}}^{\frac{1}{2}} \\ \mathrm{IQR}_{\mathrm{profile}} = \left[\frac{1}{N_z} \sum_{z=1}^{N_z} (y_{z,75th} - y_{z,25th})^2 \right]_{\mathrm{profile}}^{\frac{1}{2}} \\ \overline{y_z} = \left[\frac{1}{M} \sum_{j=1}^{M} y_{zj} \right]_{\mathrm{profile}} \end{cases} \tag{C4}$$

where N_z is the number of vertical levels for each profile, and $\mathrm{IQR}_{\mathrm{profile}}$ is its interquartile range: $y_{z,25th}$ corresponds with the 25th percentile, and $y_{z,75th}$ corresponds with the 75th percentile.

Open Research

The data for all the analyses in the main text are available at `https://doi.org/10.5281/zenodo.10019987`. The emulator code is available at `https://github.com/yqsun91/WACCM-Emulation`. All the raw WACCM output data are available on request from authors.

Acknowledgments

We thank Andre Souza for insightful discussions. This work was supported by grants from the NSF OAC CSSI program (#2005123 , #2004512, #2004492, #2004572), and by the

generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program to PH, MJA, EG, and AS. PH is also supported by the Office of Naval Research (ONR) Young Investigator Award N00014-20-1-2722. SL is supported by the Office of Science, U.S. Department of Energy Biological and Environmental Research as part of the Regional and Global Climate Model Analysis program area. Computational resources were provided by NSF XSEDE (allocation ATM170020) and NCAR's CISL (allocation URIC0009).

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., ... Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243-297. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1566253521001081` doi: https://doi.org/10.1016/j.inffus.2021.05.008

Achatz, U. (2022). Gravity waves and their impact on the atmospheric flow. In *Atmospheric dynamics* (pp. 407–505). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from `https://doi.org/10.1007/978-3-662-63941-2_10` doi: 10.1007/978-3-662-63941-2_10

Alexander, M. J., Geller, M., McLandress, C., Polavarapu, S., Preusse, P., Sassi, F., ... Watanabe, S. (2010). Recent developments in gravity-wave effects in climatemodels and the global distribution of gravity-wavemomentum flux from observations and models. *Quarterly Journal of the Royal Meteorological Society*, *136*. doi: 10.1002/qj.637

Amiramjadi, M., Plougonven, R., Mohebalhojeh, A. R., & Mirzaei, M. (2022). Using machine learning to estimate non-orographic gravity wave characteristics at source levels. *Journal of the Atmospheric Sciences*. doi: 10.1175/JAS-D-22-0021.1

Ando, S., & Huang, C. Y. (2017). Deep over-sampling framework for classifying imbalanced data. In *Machine learning and knowledge discovery in databases: European conference, ecml pkdd 2017, skopje, macedonia, september 18–22, 2017, proceedings, part i 10* (pp. 770–785).

Bacmeister, J. T., Newman, P. A., Gary, B. L., & Chan, K. R. (1994). An algorithm for forecasting mountain wave-related turbulence in the stratosphere. *Weather Forecasting*, *9*. doi: 10.1175/1520-0434(1994)009⟨0241:AAFFMW⟩2.0.CO;2

Balaji, V. (2021). Climbing down charney's ladder: machine learning and the post-dennard era of computational climate science. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200085.

Baldi, P., Sadowski, P., & Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, *5*(1), 4308.

Ballnus, B., Hug, S., Hatz, K., Görlitz, L., Hasenauer, J., & Theis, F. J. (2017). Comprehensive benchmarking of markov chain monte carlo methods for dynamical systems. *BMC Systems Biology*, *11*(1), 1–18.

Barnes, E. A., Barnes, R. J., & DeMaria, M. (2023). Sinh-arcsinh-normal distributions to add uncertainty to neural network regression tasks: Applications to tropical cyclone

intensity forecasts. *Environmental Data Science*, *2*, e15.

Beck, A., Flad, D., & Munz, C.-D. (2019). Deep neural networks for data-driven LES closure models. *Journal of Computational Physics*, *398*, 108910.

Beljaars, A. C., Brown, A. R., & Wood, N. (2004). A new parametrization of turbulent orographic form drag. *Quarterly Journal of the Royal Meteorological Society*, *130*. doi: 10.1256/qj.03.73

Belochitski, A., & Krasnopolsky, V. (2021). Robustness of neural network emulations of radiative transfer parameterizations in a state-of-the-art general circulation model. *Geoscientific Model Development*, *14*(12), 7425–7437. Retrieved from `https://gmd.copernicus.org/articles/14/7425/2021/` doi: 10.5194/gmd-14-7425-2021

Beucler, T., Ebert-Uphoff, I., Rasp, S., Pritchard, M., & Gentine, P. (2021, 05). *Machine learning for clouds and climate (invited chapter for the agu geophysical monograph series "clouds and climate").* doi: 10.1002/essoar.10506925.1

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). *Weight uncertainty in neural networks.*

Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*(1), 376–399.

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. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/77/12/jas-d-20-0082.1.xml` doi: 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. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001711` doi: https://doi.org/10.1029/2019MS001711

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249-259. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0893608018302107` doi: https://doi.org/10.1016/j.neunet.2018.07.011

Bölöni, G., Kim, Y. H., Borchert, S., & Achatz, U. (2021). Toward transient subgrid-scale gravity wave representation in atmospheric models. part i: Propagation model including nondissipative wave mean-flow interactions. *Journal of the Atmospheric Sciences*, *78*. doi: 10.1175/JAS-D-20-0065.1

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I., & Palmer, T. (2021). Machine learning emulation of gravity wave drag in numerical weather forecasting. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002477. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002477` (e2021MS002477 2021MS002477) doi: https://doi.org/10.1029/2021MS002477

Chattopadhyay, A., Nabizadeh, E., & Hassanzadeh, P. (2020). Analog forecasting of extreme-causing weather patterns using deep learning. *Journal of Advances in Model-*

ing Earth Systems, *12*(2), e2019MS001958.

Chattopadhyay, A., Subel, A., & Hassanzadeh, P. (2020). Data-driven super-parameterization using deep learning: Experimentation with multiscale Lorenz 96 systems and transfer learning. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002084.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004, jun). Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, *6*(1), 1–6. Retrieved from `https://doi.org/10.1145/1007730.1007733` doi: 10.1145/1007730.1007733

Chen, N., & Majda, A. J. (2019). A new efficient parameter estimation algorithm for high-dimensional complex nonlinear turbulent dynamical systems with partial observations. *Journal of Computational Physics*, *397*, 108836.

Clare, M. C., Sonnewald, M., Lguensat, R., Deshayes, J., & Balaji, V. (2022). Explainable artificial intelligence for bayesian neural networks: Towards trustworthy predictions of ocean dynamics. *arXiv preprint arXiv:2205.00202*.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., & Searight, K. (2013). Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, *141*(10), 3498–3516.

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. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023GL104668` (e2023GL104668 2023GL104668) doi: https://doi.org/10.1029/2023GL104668

Espinosa, Z. I., Sheshadri, A., Cain, G. R., Gerber, E. P., & DallaSanta, K. J. (2022). Machine Learning Gravity Wave Parameterization Generalizes to Capture the QBO and Response to Increased CO_2. , *49*(8), e98174. doi: 10.1029/2022GL098174

Finkel, J., Gerber, E. P., Abbot, D. S., & Weare, J. (2023). Revealing the statistics of extreme events hidden in short weather forecast data. *AGU Advances*, *4*(2), e2023AV000881. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023AV000881` (e2023AV000881 2023AV000881) doi: https://doi.org/10.1029/2023AV000881

Foster, D., Gagne, D. J., & Whitt, D. B. (2021, 12). Probabilistic machine learning estimation of ocean mixed layer depth from dense satellite and sparse in situ observations. *Journal of Advances in Modeling Earth Systems*, *13*. doi: 10.1029/2021MS002474

Frezat, H., Sommer, J. L., Fablet, R., Balarac, G., & Lguensat, R. (2022). A posteriori learning for quasi-geostrophic turbulence parametrization. *arXiv preprint arXiv:2204.03911*. doi: 10.1029/2022MS003124

Gagne, 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.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing

model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

Garcia, R. R., Smith, A. K., Kinnison, D. E., Álvaro de la Cámara, & 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. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/74/1/jas-d-16-0104.1.xml` doi: https://doi.org/10.1175/JAS-D-16-0104.1

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.

Gettelman, A., Gagne, D. J., Chen, C. C., Christensen, M. W., Lebo, Z. J., Morrison, H., & Gantos, G. (2021). Machine learning the warm rain process. *Journal of Advances in Modeling Earth Systems*, *13*. doi: 10.1029/2020MS002268

Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., ... Randel, W. J. (2019, 12). The whole atmosphere community climate model version 6 (waccm6). *Journal of Geophysical Research: Atmospheres*, *124*, 12380-12403. doi: 10.1029/2019JD030943

Gordon, E. M., & Barnes, E. A. (2022, 8). Incorporating uncertainty into a regression neural network enables identification of decadal state-dependent predictability in cesm2. *Geophysical Research Letters*, *49*. doi: 10.1029/2022GL098635

Guan, Y., Chattopadhyay, A., Subel, A., & Hassanzadeh, P. (2022). Stable a posteriori LES of 2D turbulence using convolutional neural networks: Backscattering analysis and generalization to higher Re via transfer learning. *Journal of Computational Physics*, *458*, 111090.

Guan, Y., Subel, A., Chattopadhyay, A., & Hassanzadeh, P. (2023). Learning physics-constrained subgrid-scale closures in the small-data regime for stable and accurate les. *Physica D: Nonlinear Phenomena*, *443*, 133568. doi: https://doi.org/10.1016/j.physd.2022.133568

Guillaumin, A. P., & Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002534.

Hardiman, S. C., Scaife, A. A., Niekerk, A. v., Prudden, R., Owen, A., Adams, S. V., ... Madge, S. (2023). Machine learning for non-orographic gravity waves in a climate model. *Artificial Intelligence for the Earth Systems*.

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*, 1–58.

Hertzog, A., Alexander, M. J., & Plougonven, R. (2012). On the intermittency of gravity wave momentum flux in the stratosphere. *Journal of the Atmospheric Sciences*, *69*(11), 3433–3448.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., ... others (2017). The art and science of climate model tuning. *Bulletin of the American Meteoro-*

logical Society, *98*(3), 589–602.

Huang, C., Li, Y., Loy, C. C., & Tang, X. (2016). Learning deep representation for imbalanced classification. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5375–5384).

Iglesias-Suarez, F., Gentine, P., Solino-Fernandez, B., Beucler, T., Pritchard, M., Runge, J., & Eyring, V. (2023). *Causally-informed deep learning to improve climate models and projections.*

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*, 429-449. Retrieved from `https://doi.org/10.3233/IDA-2002-6504` (5) doi: 10.3233/IDA-2002-6504

Johnson, J. M., & Khoshgoftaar, T. M. (2019, Mar 19). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 27. Retrieved from `https://doi.org/10.1186/s40537-019-0192-5` doi: 10.1186/s40537-019-0192-5

Kim, Y., Bölöni, G., Borchert, S., Chun, H. Y., & Achatz, U. (2021). Toward transient subgrid-scale gravity wave representation in atmospheric models. part ii: Wave intermittency simulated with convective sources. *Journal of the Atmospheric Sciences*, *78*. doi: 10.1175/JAS-D-20-0066.1

Kim, Y., Eckermann, S. D., & Chun, H. (2003). An overview of the past, present and future of gravity-wave drag parametrization for numerical climate and weather prediction models. *Atmosphere-Ocean*, *41*, 65-98. Retrieved from `https://doi.org/10.3137/ao.410105` doi: 10.3137/ao.410105

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*.

Köhler, L., Green, B., & Stephan, C. C. (2023). Comparing loon superpressure balloon observations of gravity waves in the tropics with global storm-resolving models. *Journal of Geophysical Research: Atmospheres*, *128*(15), e2023JD038549.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Chalikov, D. V. (2005). New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Monthly Weather Review*, *133*(5), 1370 - 1383. Retrieved from `https://journals.ametsoc.org/view/journals/mwre/133/5/mwr2923.1.xml` doi: https://doi.org/10.1175/MWR2923.1

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., ... Courville, A. (2021). Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning* (pp. 5815–5826).

Kruse, C. G., Alexander, M. J., Hoffmann, L., Niekerk, A. V., Polichtchouk, I., Bacmeister, J. T., ... Stein, O. (2022). Observed and modeled mountain waves from the surface to the mesosphere near the drake passage. *Journal of the Atmospheric Sciences*, *79*. doi: 10.1175/JAS-D-21-0252.1

Li, X., Dai, Y., Ge, Y., Liu, J., Shan, Y., & Duan, L.-Y. (2022). Uncertainty modeling for out-of-distribution generalization. *arXiv preprint arXiv:2202.03958*.

Ling, J., & Templeton, J. (2015, 08). Evaluation of machine learning algorithms for prediction of regions of high Reynolds averaged Navier Stokes uncertainty. *Physics of*

Fluids, *27*(8). Retrieved from https://doi.org/10.1063/1.4927765 (085103) doi: 10.1063/1.4927765

Liu, Y., Racah, E., Prabhat, M., Correa, J., Khosrowshahi, A., Lavers, D., ... Collins, W. (2016, 05). Application of deep convolutional neural networks for detecting extreme weather in climate datasets.

Lopez-Gomez, I., McGovern, A., Agrawal, S., & Hickey, J. (2022). Global extreme heat forecasting using neural weather models. *Artificial Intelligence for the Earth Systems*, 1 - 41. Retrieved from https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-22-0035.1/AIES-D-22-0035.1.xml doi: 10.1175/AIES-D-22-0035.1

Lu, L., Jin, P., Pang, G., Zhang, Z., & Karniadakis, G. E. (2021). Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, *3*(3), 218–229.

Maalouf, M., & Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems*, *59*, 142–148.

Maalouf, M., & Trafalis, T. B. (2011). Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics & Data Analysis*, *55*(1), 168–183.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., & Wilson, A. G. (2019). A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, *32*.

Mamalakis, A., Barnes, E. A., & Ebert-Uphoff, I. (2022). Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artificial Intelligence for the Earth Systems*, *1*(4), e220012.

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. , *47*(19), e89436. doi: 10.1029/2020GL089436

Maulik, R., San, O., Rasheed, A., & Vedula, P. (2019). Subgrid modelling for two-dimensional turbulence using neural networks. *Journal of Fluid Mechanics*, *858*, 122–144.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*. doi: 10.1175/BAMS-D-18-0195.1

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., ... Schmidt, L. (2021). Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning* (pp. 7721–7735).

Miloshevich, G., Cozian, B., Abry, P., Borgnat, P., & Bouchet, F. (2023, Apr). Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data. *Phys. Rev. Fluids*, *8*, 040501. Retrieved from https://link.aps.org/doi/10.1103/PhysRevFluids.8.040501 doi: 10.1103/PhysRevFluids.8.040501

Nagarajan, V., Andreassen, A., & Neyshabur, B. (2020). Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*.

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. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001351 doi: https://doi.org/10.1029/2018MS001351

Oh, S. M., Rehg, J. M., Balch, T., & Dellaert, F. (2005). Data-driven mcmc for learning and inference in switching linear dynamic systems. In *Proceedings of the national conference on artificial intelligence* (Vol. 20, p. 944).

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, *32*.

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 preprint arXiv:2309.09024*.

Palmer, T. (2019). Stochastic weather and climate models. *Nature Reviews Physics*, *1*(7), 463–471.

Polichtchouk, I., Van Niekerk, A., & 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.

Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., ... others (2015). A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Reviews of geophysics*, *53*(2), 323–361.

Psaros, A. F., Meng, X., Zou, Z., Guo, L., & Karniadakis, G. E. (2023). Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics*, *477*, 111902.

Qi, D., & Majda, A. J. (2020). Using machine learning to predict extreme events in complex systems. *Proceedings of the National Academy of Sciences*, *117*(1), 52-59. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1917285117 doi: 10.1073/pnas.1917285117

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Rasp, S. (2020). Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: general algorithms and lorenz 96 case study (v1.0). *Geoscientific Model Development*, *13*(5), 2185–2196. Retrieved from https://gmd.copernicus.org/articles/13/2185/2020/ doi: 10.5194/gmd-13-2185-2020

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.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019, Feb 01). Deep learning and process understanding for data-driven earth system science. *Nature*, *566*(7743), 195-204. Retrieved from https://doi.org/10.1038/s41586-019-0912-1 doi: 10.1038/s41586-019-0912-1

Richter, J. H., Butchart, N., Kawatani, Y., Bushell, A. C., Holt, L., Serva, F., ... Yukimoto, S. (2022, 4). Response of the quasi-biennial oscillation to a warming climate in global climate models. *Quarterly Journal of the Royal Meteorological Society*, *148*, 1490-1518. doi: 10.1002/qj.3749

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*. doi: 10.1175/2009JAS3112.1

Ross, A. S., Li, Z., Perezhogin, P., Fernandez-Granda, C., & Zanna, L. (2022). Benchmarking of machine learning ocean subgrid parameterizations in an idealized model.

Sato, K., Watanabe, S., Kawatani, Y., Tomikawa, Y., Miyazaki, K., & Takahashi, M. (2009). On the origins of mesospheric gravity waves. *Geophysical Research Letters*, *36*. doi: 10.1029/2009GL039908

Schneider, T., Jeevanjee, N., & Socolow, R. (2021). Accelerating progress in climate science. *Physics Today*, *74*(6), 44–51.

Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, *44*. doi: 10.1002/2017GL076101

Scinocca, J. F., & McFarlane, N. A. (2000). The parametrization of drag induced by stratified flow over anisotropic orography. *Quarterly Journal of the Royal Meteorological Society*, *126*. doi: 10.1002/qj.49712656802

Seifert, A., & Rasp, S. (2020). Potential and limitations of machine learning for modeling warm-rain cloud microphysical processes. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002301. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002301` (e2020MS002301 10.1029/2020MS002301) doi: https://doi.org/10.1029/2020MS002301

Shamekh, S., Lamb, K. D., Huang, Y., & Gentine, P. (2023). Implicit learning of convective organization explains precipitation stochasticity. *Proceedings of the National Academy of Sciences*, *120*(20), e2216158120. Retrieved from `https://www.pnas.org/doi/abs/10.1073/pnas.2216158120` doi: 10.1073/pnas.2216158120

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., & Cui, P. (2021). Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Song, H.-J., & Roh, S. (2021). Improved weather forecasting using neural network emulation for radiation parameterization. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002609. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002609` (e2021MS002609 2021MS002609) doi: https://doi.org/10.1029/2021MS002609

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Stensrud, D. J. (2007). *Parameterization schemes: Keys to understanding numerical weather prediction models*. Cambridge University Press. Retrieved from `https://www.cambridge.org/core/product/identifier/9780511812590/type/book` doi: 10

.1017/CBO9780511812590

Subel, A., Chattopadhyay, A., Guan, Y., & Hassanzadeh, P. (2021). Data-driven subgrid-scale modeling of forced Burgers turbulence using deep learning with generalization to higher Reynolds numbers via transfer learning. *Physics of Fluids*, *33*(3), 031702.

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

Sun, Y., Hassanzadeh, P., Alexander, M., & Kruse, C. (2023, 12). Quantifying 3d gravity wave drag in a library of tropical convection-permitting simulations for data-driven parameterizations.
doi: 10.1029/2022MS003585

Sun, Y., Wong, A., & Kamel, M. S. (2009, 11). Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*, 687-719. doi: 10.1142/S0218001409007326

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. , 270–279.

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

Wu, D., Gao, L., Xiong, X., Chinazzi, M., Vespignani, A., Ma, Y.-A., & Yu, R. (2021). *Quantifying uncertainty in deep spatiotemporal forecasting.*

Wu, G., & Chang, E. Y. (2003). Adaptive feature-space conformal transformation for imbalanced-data learning. In *Proceedings of the twentieth international conference on international conference on machine learning* (p. 816–823). AAAI Press.

Ye, H., Xie, C., Cai, T., Li, R., Li, Z., & Wang, L. (2021). Towards a theoretical framework of out-of-distribution generalization. *Advances in Neural Information Processing Systems*, *34*, 23519–23531.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, *27*.

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), 1–10.

Zhang, C., Perezhogin, P., Gultekin, C., Adcroft, A., Fernandez-Granda, C., & Zanna, L. (2023). *Implementation and evaluation of a machine learned mesoscale eddy parameterization into a numerical ocean circulation model.*

Zhang, D., Ahuja, K., Xu, Y., Wang, Y., & Courville, A. (2021). Can subnetwork structure be the key to out-of-distribution generalization? In *International conference on machine learning* (pp. 12356–12367).

Zhu, Y., Zabaras, N., Koutsourelakis, P.-S., & Perdikaris, P. (2019). Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *Journal of Computational Physics*, *394*, 56–81.

Data Imbalance, Uncertainty Quantification, and Generalization via Transfer Learning in Data-driven Parameterizations: Lessons from the Emulation of Gravity Wave Momentum Transport in WACCM

Y. Qiang Sun[1], Hamid A. Pahlavan[1], Ashesh Chattopadhyay[1], Pedram Hassanzadeh[1], Sandro W. Lubis[1,3], M. Joan Alexander[2], Edwin Gerber[4], Aditi Sheshadri[5], Yifei Guan[1]

[1]Rice University, Houston, TX, 77005
[2]NorthWest Research Associates, Boulder, CO, 80301, USA
[3]Pacific Northwest National Laboratory, Richland, WA, 99354, USA
[4]New York University, New York, NY 10012, USA
[5]Stanford University, Stanford, CA, 94305, USA

Key Points:

- WACCM's orographic, convective, and frontal gravity wave parameterizations are emulated using neural nets to inform future modeling efforts
- Out-of-distribution generalization (extrapolation) of the neural nets under $4\times CO_2$ forcing is enabled via transfer learning with 1% new data
- Data imbalance is addressed via resampling and weighted loss; uncertainty quantification via Bayesian, dropout, and variational methods

Corresponding author: Y. Qiang Sun and Pedram Hassanzadeh, ys91@rice.edu and pedramh@uchicago.edu

Abstract

Neural networks (NNs) are increasingly used for data-driven subgrid-scale parameterization in weather and climate models. While NNs are powerful tools for learning complex nonlinear relationships from data, there are several challenges in using them for parameterizations. Three of these challenges are 1) data imbalance related to learning rare (often large-amplitude) samples; 2) uncertainty quantification (UQ) of the predictions to provide an accuracy indicator; and 3) generalization to other climates, e.g., those with higher radiative forcing. Here, we examine performance of methods for addressing these challenges using NN-based emulators of the Whole Atmosphere Community Climate Model (WACCM) physics-based gravity wave (GW) parameterizations as the test case. WACCM has complex, state-of-the-art parameterizations for orography-, convection- and frontal-driven GWs. Convection- and orography-driven GWs have significant data imbalance due to the absence of convection or orography in many grid points. We address data imbalance using resampling and/or weighted loss functions, enabling the successful emulation of parameterizations for all three sources. We demonstrate that three UQ methods (Bayesian NNs, variational auto-encoders, and dropouts) provide ensemble spreads that correspond to accuracy during testing, offering criteria on when a NN gives inaccurate predictions. Finally, we show that accuracy of these NNs decreases for a warmer climate ($4\times CO_2$). However, the generalization accuracy is significantly improved by applying transfer learning, e.g., re-training only one layer using $\sim 1\%$ new data from the warmer climate. The findings of this study offer insights for developing reliable and generalizable data-driven parameterizations for various processes, including (but not limited) to GWs.

Plain Language Summary

Scientists are increasingly using machine learning methods, especially neural networks (NNs), to improve weather and climate models. However, it can be challenging for a NN to learn rare, large-amplitude events, because they are infrequent in training data. Also, NNs need to express their confidence (certainty) about a prediction and work effectively across different climates, e.g., warmer climates due to increased CO_2. Traditional NNs often struggle with these challenges. Here, we share insights gained from emulating the complex physics-based parameterization schemes for gravity waves in a state-of-the-art climate model. We propose specific strategies for addressing imbalanced data, uncertainty quantification (UQ), and making accurate predictions across various climates. For instance, to manage data balance, one such strategy involves amplifying the impact of infrequent events in the training data. We also demonstrate that several UQ methods could be useful in determining the accuracy of predictions. Furthermore, we show that NNs trained on simulations of the historical period do not perform as well in warmer climates. However, we improve the NNs' performance by employing transfer learning using limited data from warmer climates. This study provides lessons for developing robust and generalizable approaches for using NNs to improve models in the future.

1 Introduction

Small-scale processes such as moist convection, gravity waves, and turbulence are key players in the variability of the climate system and its response to increased greenhouse gases. However, as these processes cannot be resolved, entirely or partially, by the coarse-resolution general circulation models (GCMs), they need to be represented as functions of the resolved dynamics via subgrid-scale (SGS) parameterization schemes (e.g., Kim et al., 2003; Stensrud, 2007; Prein et al., 2015). Many of these parameterization schemes are based on heuristic approximations and simplifications, introducing large parametric and epistemic uncertainties in GCMs (Schneider et al., 2017; Hourdin et al., 2017; Palmer, 2019).

Recently, there has been a growing interest in developing data-driven SGS parameterizations for different complex processes in the Earth system using machine learning (ML) techniques, particularly deep neural networks (NNs). Promising results have been demonstrated in a wide range of idealized applications, including prototype systems (Maulik et al., 2019; Gagne et al., 2020; Rasp, 2020; Chattopadhyay, Subel, & Hassanzadeh, 2020; Frezat et al., 2022; Guan et al., 2022; Pahlavan et al., 2023), ocean turbulent processes (Bolton & Zanna, 2019; C. Zhang et al., 2023), moist convection in the atmosphere (O'Gorman & Dwyer, 2018; Brenowitz & Bretherton, 2019; Yuval & O'Gorman, 2020; Beucler et al., 2021; Iglesias-Suarez et al., 2023), radiation (Krasnopolsky et al., 2005; Belochitski & Krasnopolsky, 2021; Song & Roh, 2021), and microphysics (Seifert & Rasp, 2020; Gettelman et al., 2021). The ultimate promise of data-driven parameterizations, learned from observation-derived data and/or high-fidelity high-resolution simulations, is that they might have smaller parametric/structural errors, thus reducing the biases of GCMs and producing more reliable climate change projections (e.g., Schneider et al., 2017; Reichstein et al., 2019; Schneider et al., 2021).

However, there are major challenges in developing trustworthy, interpretable, stable, and generalizable data-driven parameterizations that can be used for such climate change projection efforts. Discussing and even listing all of these challenges is well beyond the scope of this paper. Well-known challenges such as interpretability and stability have been extensively discussed in a number of recent studies (e.g., McGovern et al., 2019; Beck et al., 2019; Brenowitz et al., 2020; Balaji, 2021; Clare et al., 2022; Mamalakis et al., 2022; Guan et al., 2022; Subel et al., 2023; Pahlavan et al., 2023). Here, we focus on three other key issues:

1. Data imbalance, and related to that, learning rare/extreme events,
2. Uncertainty quantification (UQ) of the NN-based SGS parameterization outputs,
3. Out-of-distribution (OOD) generalization (e.g., extrapolation to climates with higher radiative forcings).

Below we briefly discuss the importance of 1-3 and the current state-of-the-art methods in addressing them in the climate and ML literature. Data imbalance is a well-known problem in the ML literature, especially in the context of classification tasks (e.g., Japkowicz &

Stephen, 2002; G. Wu & Chang, 2003; Chawla et al., 2004; Sun et al., 2009; Huang et al., 2016; Ando & Huang, 2017; Buda et al., 2018; Johnson & Khoshgoftaar, 2019). The problem becomes particularly significant when one aims to learn rare/extreme events (Maalouf & Trafalis, 2011; Maalouf & Siddiqi, 2014; Baldi et al., 2014; Liu et al., 2016; O'Gorman & Dwyer, 2018; Qi & Majda, 2020; Chattopadhyay, Nabizadeh, & Hassanzadeh, 2020; Miloshevich et al., 2023; Finkel et al., 2023; Shamekh et al., 2023). For example, suppose we aim to learn the binary classification of the 99 percentile of temperature anomalies using a NN. In this case, label 0 (no extreme) will constitute 99% of the training (or testing) set while label 1 (extreme) will be just 1%. With many common loss functions such as mean squared error (MSE) or root-mean squared-error (RMSE), training a NN will result in one that predicts 0 for any sample (extreme or no extreme) while having a seemingly high accuracy of 99% (of course, other metrics such as precision/recall will show the shortcoming, see Chattopadhyay, Nabizadeh, & Hassanzadeh (2020)). The most common remedy to this problem for classification tasks is resampling. An example is down-sampling non-extreme cases by a factor of 100, which effectively balances the dataset.

In addition to *classification* tasks, Data imbalance also presents a significant challenge in *regression* tasks required for parameterization schemes in climate models. As highlighted by Chantry et al. (2021), such imbalances contributed to the unsuccessful emulation of their orographic gravity wave parameterization (GWP) scheme, largely because orography affects the gravity wave (GW) drag in only a fraction of the grid columns. This challenge also persists in emulating GWP for non-orographic GWs, especially when GWs are intricately linked to their sources. For instance, the presence of zero convective GW drag at numerous grid points due to the absence of convection creates a notably imbalanced dataset. This issue will be explored further in the results section. In regression tasks, data imbalance may also manifest in the form of difficulty in learning large-amplitude (extreme) outputs, which are rare and constitute only a small fraction of the training set. In the case of GWs, Observations have shown that gravity wave amplitudes are highly intermittent such that the largest 10% events alone can contribute more than 50% of the total momentum flux (Hertzog et al., 2012), so the extreme events will contribute an outsized fraction of the total drag. Nonetheless, poorly learning these large-amplitude outputs, like drag forces, can result in instabilities (e.g., Guan et al., 2022). Addressing data imbalance in climate applications has received relatively limited attention. In this study, we propose several remedies based on resampling techniques and weighted loss functions, demonstrating their advantages in enabling successful emulations of all GWP schemes and improving the learning of rare/extreme events.

Quantifying the uncertainties in outputs from NN-based parameterization schemes is essential when employing these schemes, particularly for high-stakes decision-making tasks such as climate change projections. Crucially, during testing when we are unable to directly determine a prediction's accuracy, we need a UQ method that can provide a credible *confidence level* for each prediction, serving as a reliable indicator of its accuracy. During inference, the output of an NN can be inaccurate for various reasons, including poor approximation (e.g., due to poor NN architecture), poor within-distribution generalization (e.g.,

for inputs that are rare events), or poor optimization (collectively referred to as *epistemic uncertainty*), as well as because of OOD generalization errors due to input samples from a distribution different from that of the training set (Abdar et al., 2021; Lu et al., 2021; Krueger et al., 2021; Miller et al., 2021; Shen et al., 2021; D. Wu et al., 2021; Ye et al., 2021; D. Zhang et al., 2021; Subel et al., 2023). Quantifying the level of uncertainty would then allow us to avoid using a data-driven parameterization scheme when it is inaccurate due to one of the aforementioned reasons (Maddox et al., 2019; Zhu et al., 2019; Li et al., 2022; Psaros et al., 2023). In the context of data-driven parameterization in climate modeling, the two most challenging sources of uncertainty are rare/extreme events and OOD generalization errors. The latter is a concern, particularly when the GCM is used for climate change studies (see below for more discussions).

Developing UQ methods for NNs is an active area of research in the ML community, and there is not a generally applicable rigorous method yet. For instance, techniques like Markov-Chain Monte Carlo can be prohibitively expensive, especially when dealing with high-dimensional systems (Oh et al., 2005; Ballnus et al., 2017; Chen & Majda, 2019). For a comprehensive review in the context of scientific ML, refer to Psaros et al. (2023). The topic has also started to increasingly gain attention in the climate literature (Guillaumin & Zanna, 2021; Gordon & Barnes, 2022; Haynes et al., 2023; Barnes et al., 2023). In this study, we will assess the performance of three common UQ methods (Bayesian, dropout, and variational NNs) by analyzing the relationships between uncertainty and accuracy during inference testing. We will also consider scenarios involving OOD generalization errors resulting from global warming.

As already mentioned above, OOD generalization (extrapolation to a test data distribution different from that of the training set) is a major challenge for applications involving non-stationarity, like a changing climate. Studies have already shown that the lack of OOD generalization in data-driven parameterizations leads to inaccurate and unstable simulation (Rasp et al., 2018; O'Gorman & Dwyer, 2018; Chattopadhyay, Subel, & Hassanzadeh, 2020; Guan et al., 2022; Nagarajan et al., 2020). A general and powerful method for improving the OOD generalization capability of NNs is transfer learning (TL), which involves re-training a few or all of the layers of a NN using a small amount of data from the new system (Yosinski et al., 2014). This approach has already shown remarkable success in enabling data-driven parameterization schemes to extrapolate across the parameter space (e.g., to 100$\times$ higher Reynolds number) in canonical test cases (Chattopadhyay, Subel, & Hassanzadeh, 2020; Subel et al., 2021; Guan et al., 2023; Subel et al., 2023; C. Zhang et al., 2023). In particular, Subel et al. (2023) introduced SpArK (Spectral Analysis of Regression Kernels and Activations) showing that re-training even one layer can lead to successful OOD generalization, although this optimal layer, unlike the rule of thumb in the ML literature, may not be the deepest but the shallowest hidden layer. Here, we further leverage these studies and show how TL can enable OOD generalization of data-driven parameterization schemes in state-of-the-art GCMs.

The methods used in this study and the learned lessons apply to a broad range of processes and applications in climate modeling. However, the results are presented for a single test case, that is based on the emulation of complex physics-based GWP schemes in version 6 of the Whole Atmosphere Community Climate Model (WACCM), a state-of-the-art GCM (Gettelman et al., 2019). Here, we use the emulations of current physics-based parameterization schemes as a stepping stone towards learning data-driven parameterizations from observations and high-fidelity simulations by testing ideas for addressing items 1-3 listed earlier. Furthermore, developing better representations of un- and under-resolved GWs in GCMs is an important problem on its own (Kim et al., 2003; Alexander et al., 2010; Achatz, 2022). A number of recent studies have taken the first steps in learning data-driven GWP from observations and high-resolution simulations (Matsuoka et al., 2020; Amiramjadi et al., 2022; Sun et al., 2023; Dong et al., 2023), though careful and time-consuming steps are needed in producing, analyzing, and using such data. Furthermore, two recent studies focused on emulators of simpler GWP schemes in a forecast model and idealized GCM have readily shown the usefulness of lessons learned from emulators (Chantry et al., 2021; Espinosa et al., 2022; Hardiman et al., 2023). This further motivates the focus on using emulators for testing ideas for addressing data imbalance, UQ, and OOD generalization.

This paper is structured as follows. Section 2 introduces the WACCM simulations and the NN architectures used in this study. The findings, detailed in Section 3, emphasize the insights gained in addressing data imbalance and UQ, alongside OOD generalization of the emulators under warmer climate conditions. Consistent with Chantry et al. (2021), we find that using an NN to emulate the parameterization of orographic GWs is significantly more challenging than non-orographic GWs. This necessitated additional steps to achieve reasonable offline performance, as detailed in Section 4. To the best of our knowledge, this stands as the first NN-based emulation of orographic GWs to address the challenges in Chantry et al. (2021). Finally, we provide a concluding summary in Section 5.

2 Data and Methods

2.1 The Whole Atmosphere Community Climate Model (WACCM)

The NCAR's WACCM version 6 introduced in Gettelman et al. (2019) is used in this study. WACCM has state-of-the-art GWP schemes for GWs from three different sources: orography (OGWs), convection (CGWs), and fronts (FGWs). These complex sources make the emulation of the GWP schemes in WACCM a challenging task. This is, therefore, a suitable test case to investigate ideas for learning rare events, UQ, and OOD generalization to benefit the future efforts for the much more complex task, that is learning data-driven GWP schemes from observations and/or high-resolution GW-resolving simulations (Amiramjadi et al., 2022; Sun et al., 2023).

The configuration of the WACCM used in this study is identical to the public version in Gettelman et al. (2019), with a horizontal resolution of $0.95^\circ \times 1.25^\circ$ and 70 vertical levels. The two non-orographic GWP schemes in WACCM both follow Richter et al. (2010), yet

allow separate specifications of FGW and CGW sources. For OGWs, WACCM uses an updated planetary boundary layer form drag scheme from Beljaars et al. (2004), near-surface nonlinear drag processes following Scinocca & McFarlane (2000), and a ridge-finding algorithm to define orographic sources based on Bacmeister et al. (1994). A full documentation of WACCM OGWs can also be found in Kruse et al. (2022).

We conduct two sets of simulations: A 10-year pre-industrial "control" run, and a 10-year pseudo-global-warming "future" run with $4\times CO_2$ and uniform +4 K sea-surface temperature increases. In each run, we save, on the native grid, all the inputs and outputs for each of the three GWPs every 3 hours to capture the diurnal cycle. A complete list of these inputs/outputs, which are used in the training of the NN-based emulators, is presented in Appendix A.

We train separate NNs for emulating the three GWP schemes that have different sources. We use the first 6 years of the control run for training and the last 4 years for validation (years 7 and 8) and testing (years 9 and 10). With a grid resolution of $\sim 1^\circ$, there are 55,296 columns for each time snapshot, resulting in approximately 960 million input/output columns during the 6-year training period. Given the strong temporal correlation between the 3-hourly samples, we perform sub-sampling on both the training and validation data to reduce the dataset size. To accomplish this, we begin by shuffling all the input/output column pairs in time at each latitude/longitude grid point. Then, we randomly select 2,000 input/output pairs at each location for training and 500 pairs for validation.

To give the readers a general idea of the parameterized GWs and large-scale circulation in WACCM, Figure 1 shows the zonal-mean climatology for zonal GW drag/forcing, hereinafter referred to as GWD, arises from the divergence of gravity wave momentum transport (fluxes), from all three sources, computed from the 6-year training period in the control run. The zonal-mean zonal wind climatology is also shown. Seasonal dependency for both the GWD and the circulation is observed in the simulations. At levels below 100 hPa, the tendencies of non-orographic GW are relatively small compared to those from OGWs; however, their amplitudes increase significantly at higher altitudes. While the parameterized effect of GWs is generally to decelerate the zonal flow, there are exceptions, notably in regions like the equatorward flanks of the stratospheric polar night jets, where FGWs can accelerate the flow. For more information on the GWP schemes and circulations in WACCM, see Garcia et al. (2017) and Gettelman et al. (2019).

2.2 The NNs and UQ

2.2.1 The Deterministic Fully Connected NN

Here we briefly describe the general structure of the NN-based regression models trained as emulators for GWP schemes. For the deterministic artificial NN, denoted as ANN in this study, we use multilayer perceptrons (MLP). MLPs, which are feedforward fully connected NNs, take inputs through successive layers of linear transformation and non-linear activation

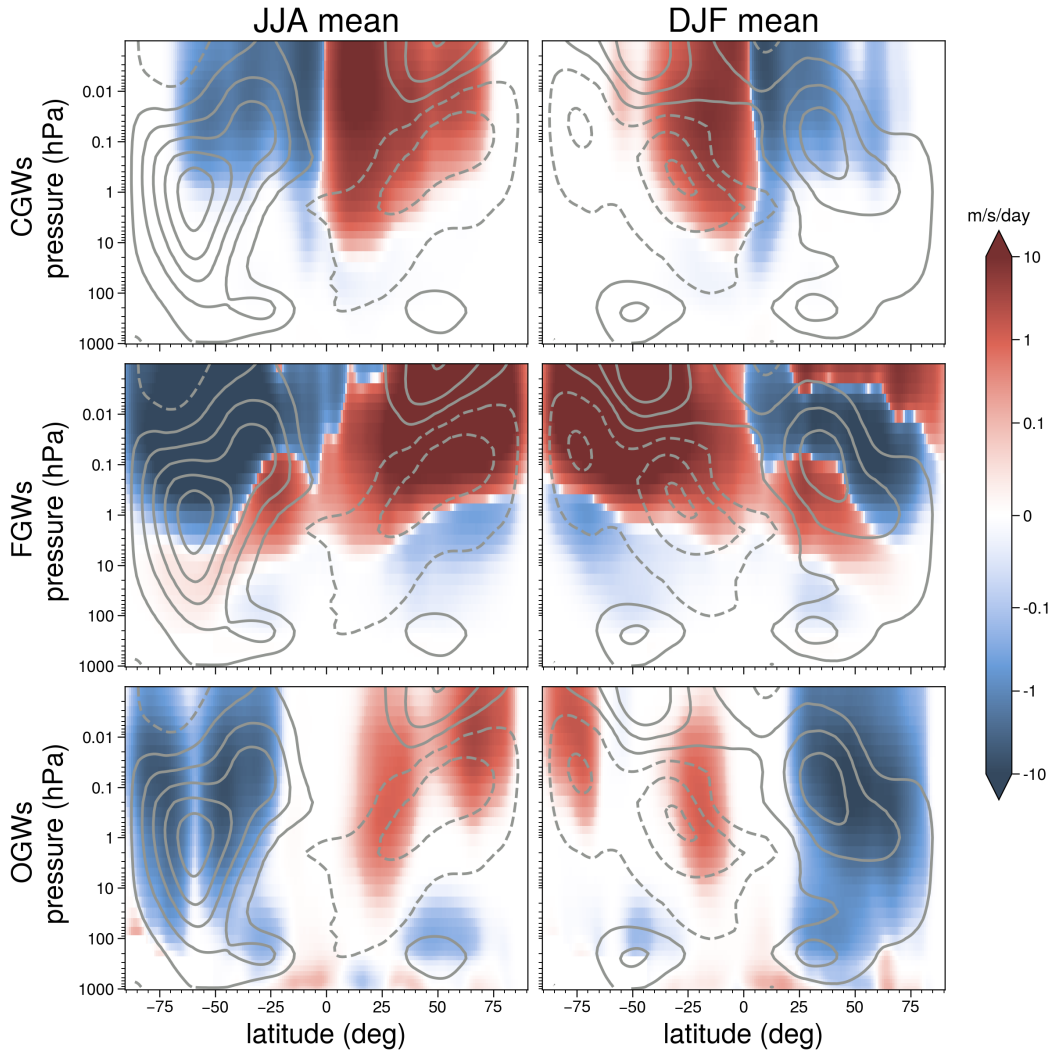

Figure 1. Climatology of zonal-mean GWD during summers (JJA) and winters (DJF) from all 3 sources in the control (pre-industrial) WACCM simulations. Top: CGWs; middle: FGWs; bottom: OGWs. The climatology of the zonal-mean zonal wind is also shown (grey lines), with an interval of 20 m/s. Dashed lines indicate negative values. Zero lines are omitted.

functions to produce an output, so as to learn a functional relationship between the input and output (Figure 2a). Deep MLPs have multiple layers of weights, which are optimized over many samples of input-output data pairs. Such MLPs are thus very powerful in terms of learning complicated functional relationships. Generally, we can write the governing equations of an MLP as

$$z^{\ell} = \sigma\left(W^{\ell} z^{\ell-1} + b^{\ell}\right), \tag{1}$$

where z^{ℓ} is the activation (output) of layer ℓ, W^{ℓ} is the weight matrix connecting layers ℓ and $\ell - 1$, and b^{ℓ} is the bias at layer ℓ, which allows the network to fit the data even when all input features are equal to 0. σ is the non-linear activation function.

In this study, we employ the same NN structure while training three distinct NNs, each for GWP originating from one of the three unique GW sources. The input layer contains the same input variables (see Appendix A) used by the WACCM GWPs across all vertical levels. There are 10 hidden layers in total (Figure 2a), and there are 500 neurons in each hidden layer. In the output layer, both zonal and meridional GWD are predicted. The activation function in each layer, σ, is chosen to be swish (Ramachandran et al., 2017), except for the output layer, where it is linear. During training, W^ℓ and b^ℓ are randomly initialized and learned by minimizing a loss function using an ADAM optimizer, with a fixed learning rate of $\alpha = 0.0001$. One of the loss functions used here is the common MSE, i.e.,

$$\mathcal{L}(\Theta) = \frac{1}{n}\sum_{i=1}^{n} \left\| \mathbf{NN}\left(x_i, \Theta\right) - y_i \right\|_2^2 \tag{2}$$

Here, n is the number of training samples and $\|.\|_2$ is the L_2 norm. For training sample i, vector x_i contains all the inputs to the NN (Appendix A), vector y_i contains the true zonal and meridional GWD at each vertical level, and $\Theta = \{\theta_j\}_{j=1\cdots p}$ denotes the trainable parameters, i.e., the weights ($p \approx 3 \times 10^6$).

2.2.2 The UQ Methods and Metrics

Although deterministic NNs are powerfully expressive and can exhibit high out-of-sample predictive skills, they do not provide estimates of the uncertainty associated with their predictions. As mentioned earlier, currently there is no rigorous method to estimate the uncertainty of an NN prediction. That said, a variety of techniques have been developed for UQ in NNs, though the validity and usefulness of the estimated uncertainty for scientific applications remain subjects of ongoing investigations (e.g., Psaros et al., 2023; Haynes et al., 2023). In this paper, we use three different and widely used approaches to perform UQ from the ML literature: Bayesian neural network (BNN), dropout neural network (DNN), and variational auto-encoder (VAE). A brief overview of these approaches is provided below.

Bayesian neural network (BNN): A BNN combines the deterministic NN described earlier and in Figure 2a with Bayesian inference (Blundell et al., 2015). Simply speaking, a BNN estimates distributions of the weights, rather than point values (as in a deterministic NN). The posterior distributions in the BNN (i.e., the distributions of the weights and biases) are calculated using the Bayes rule. In this study, we follow the standard practice and assume that all variational forms of the posterior are normal distributions. Furthermore, to accelerate the training process, we use the normal distribution $\mathcal{N}(\mu, 1)$ for all the priors in the BNN (where μ is obtained from parameters of the trained deterministic NN). Note that while we are assuming normal distributions for the trainable parameters, the predictions generated by BNN can fit different distributions due to the use of nonlinear activation functions. The resulting distribution of the predictions during inference gives an estimate of their uncertainty.

Dropout neural network (DNN): A DNN is developed by randomly eliminating all outgoing connections from some of the nodes (Figure 2a) in each hidden layer of a deterministic

NN during the training and the inference (Srivastava et al., 2014). The fraction of nodes "dropped" on average in each layer is called the dropout ratio. Mathematically, Equation (1) can be reformulated for a DNN as:

$$z^{\ell} = \sigma\left(D^{\ell} W^{\ell} z^{\ell-1} + b^{\ell}\right), \tag{3}$$

where the dropout matrix D^{ℓ} is a square diagonal binary matrix of integers 0 or 1. The diagonal elements of D^{ℓ} follow a Bernoulli distribution where the probability of zero is the dropout ratio.

Dropout was initially developed as a regularization technique to prevent over-fitting in NNs. However, Gal & Ghahramani (2016) showed that training a NN with the dropout technique approximates a Bayesian NN. In this study, we use a dropout rate of 0.1, which is incorporated in all hidden layers, but we also investigate the sensitivity of the DNN to different dropout rates, as later shown in Appendix B. Note that the random dropping out is also used during inference, leading to a distribution for each prediction.

Variational auto-encoder (VAE): A typical VAE (Kingma & Welling, 2014) consists of two NNs (Figure 2b): an encoder that transforms the input into a lower-dimensional latent space, parameterized by a normal probability distribution, and a decoder that inverts this transformation and produces the original input. The difference between the decoder's output and the original input drives the learning process of the encoder and decoder, while the parameterized lower-dimensional latent space provides the uncertainty of this transformation. The VAE was developed for generative reconstructions of data by simply drawing samples from the latent space. The VAE is basically a dimension-reduction method. Many variants, however, have been proposed for more specific purposes. In this study, following Foster et al. (2021), we add a third NN, as illustrated in Figure 2b, that randomly draws samples from the parameterized latent space as inputs, and predicts the zonal and meridional GWDs as outputs. The difference between the predicted GWDs and the true GWDs drives the learning of the third NN. Consequently, the loss for the entire network consists of three components: the loss between the reconstructed input and the original input, the Kullback–Leibler (KL) divergence between the distribution of the latent space and a normal distribution, and the loss between the predicted GWDs by the third NN and the true GWDs.

For a specific input, each of these three UQ methods discussed above can be run multiple times, generating an ensemble of predictions with different realizations of the weights by drawing from the trained distribution. This is in contrast to the deterministic NN that provides just a single-valued prediction for a given input. These ensembles can then be used to quantify the uncertainty associated with that prediction. We expect that the RMSE of the ensemble mean should exhibit approximately a 1-1 relationship with the ensemble spread (i.e., the standard deviation of the ensemble members). To investigate this relationship, we use the spread-skill plot (Delle Monache et al., 2013). Detailed calculations behind the spread-skill plot can be found in Appendix C, where we also introduce two metrics: spread-

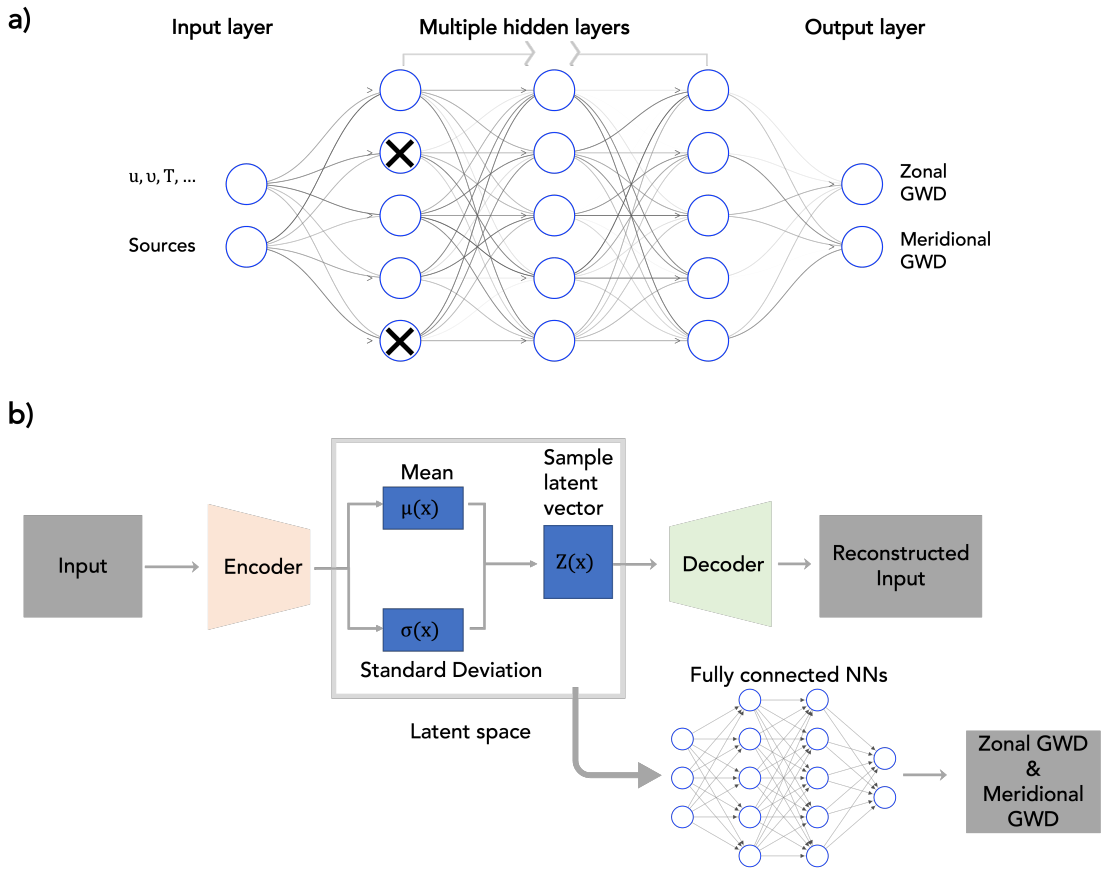

Figure 2. Schematics of the NN-based emulators and different training/re-training strategies used in this study. (a) Schematic for the MLP and DNN. The inputs of the NN are connected through successive layers of neurons (blue circles) to the output (GWDs). A fully connected MLP NN is trained from randomly initialized weights and biases in all layers. A DNN is the same but some connections are randomly eliminated during training and inference (black crosses). In TL, only some of the layers of a previously trained MLP are re-trained using new data. (b) Schematic for the VAE. A low-dimensional latent space is constructed and then used as the input for the additional fully connected NNs, which is similar to the one in (a).

skill reliability (SSREL) and overall spread-skill ratio (SSRAT), both of which summarize the information presented in the spread-skill plot.

2.3 Transfer Learning

Transfer learning refers to leveraging/reusing information (weights) from an already well-trained base NN to effectively build a new NN for a different system from which only a small amount of training data is available (Yosinski et al., 2014; Tan et al., 2018; Chattopadhyay, Subel, & Hassanzadeh, 2020). For our purpose, which is improving OOD generalization to the warmer climate, the TL procedure is as follows. For any of the NNs described earlier (e.g., the one in Figure 2a), we train them from randomly initialized weights and biases with data from the control simulations. The NN will work well during inference for test

samples from the control but not from future (warmer climate) simulations (as shown in the Results section). To address this, TL is applied wherein most of the NN's weights are kept constant, and only one or two hidden layers are re-trained using a limited dataset from the future simulation. Although this small dataset is insufficient for training an entire NN from random initialization, careful and correct selection of hidden layers for re-training, as discussed in Subel et al. (2023), allows the development of an NN that accurately adapts to the new system, i.e., the future climate conditions.

Here, we re-train the NN-based emulator that was initially trained on the control data with new data from only 1 month (30 consecutive days) of integration (1.4% of 6 years simulation for the initial training) of WACCM model under future forcing ($4\times CO_2$). We have explored different choices of layers to re-train with the same amount of new data and found that re-training the first hidden layer yields the best results, consistent with Subel et al. (2023). Therefore, the results with only re-training the first hidden layer are shown in Section 3 unless stated otherwise.

3 Results

3.1 Data Imbalance

As discussed earlier, the physics-based GWP schemes in WACCM are directly linked to their sources. This means they only produce non-zero values when their respective sources are active. For example, in a specific grid box, CGWs only register non-zero values when there is active convection within that box. The heterogeneous and sometimes intermittent nature of these sources leads to a dataset that is significantly imbalanced. Figure 3 shows global maps of the occurrence frequency of non-zero GWD for CGWs and FGWs. On average, only 7.6% of all GCM columns yield non-zero CGWs, primarily located in the tropics. Similarly, for FGWs, only 8.5% of all columns have non-zero outputs, but unlike CGWs, the majority of these are located in mid-to-high latitudes, particularly along the storm track region. For the OGWs in WACCM, data imbalance presents a greater challenge, to be discussed in a later section. While it is possible to simply separate zero and non-zero columns for emulation work where we know the truth, this approach falls short with real-world data, which is the main purpose of this study.

In addition to their sources, several other factors specific to GWD data exacerbate the data imbalance problem. In the case of each GCM column with non-zero GW activity, momentum fluxes are generally concentrated at a few critical height levels rather than being smoothly distributed throughout the entire column. This further restricts the effective occurrence frequency of non-zero values. Moreover, GWs exhibit significant intermittency, where a small portion of large-amplitude GWs often dominates the morphology of the observed global GW momentum flux distribution (Hertzog et al., 2012; Geller et al., 2013). Therefore, it is crucial for NNs to not only accurately identify the columns that produce GWDs but also to effectively learn and recognize rare and extreme GWDs.

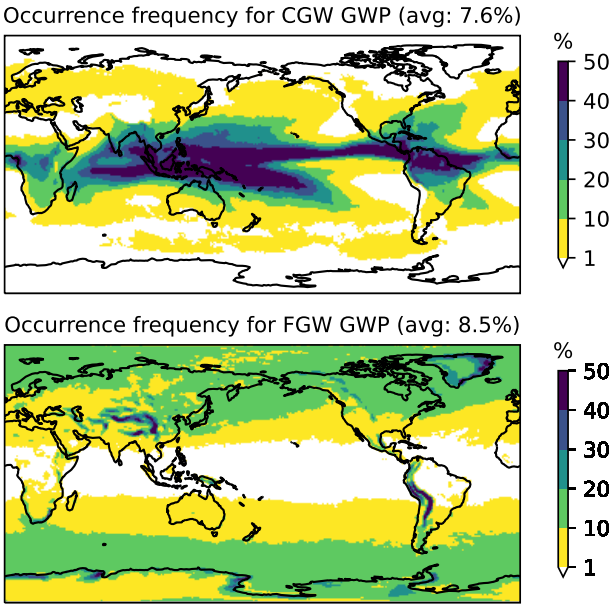

Figure 3. Distribution of occurrence frequency for CGWs (top) and FGWs (bottom) in the WACCM pre-industrial control simulations, based on the average of the 6-year training dataset.

Given the complexity of the GWD dataset, different normalization methods are considered in this study. The first method, dubbed "NORM1", is the typical normalization used in ML practices, which calculates elemental means and standard deviations for each feature (i.e., input variable at a given model level) and normalizes both inputs and outputs by these values (e.g., Espinosa et al. (2022)). With this approach, the same relative changes in wind at each level are treated equally in the input. The loss function in Equation (2) also penalizes the same relative error in GWD at each level equally. The second method, referred to as "NORM2", is designed with the physics of GWD in mind. For the velocity inputs (u, v) and the tendency outputs (GWD), each column is normalized by one single value, which is the largest standard deviation from all model levels. Additionally, the mean values for these variables, are retained (e.g., $u_{norm2}(x, y, z, t) = u(x, y, z, t)/max(std(u))$). Unlike NORM1, the original wind profile's structure is preserved in NORM2, and large GWD values at certain heights maintain a relatively larger value after this normalization. For all other input variables, NORM2 is identical to NORM1. Compared to NORM1, NORM2 places more emphasis on large GWD values and penalizes the NN more for missing these significant tendencies. These two normalization methods are also employed in Chantry et al. (2021), who found similar performance from these methods with the non-orographic GWPs.

Figure 4 shows the performance of the emulations for CGWs with the two normalization methods. When employing NORM1, the conventional approach seen in prior ML practices, and also our initial attempts, the emulator's performance is poor. Although the NN demonstrates some skill, its predictions tend to cluster around zero. However, when the second normalization method (NORM2) is employed, the emulation results show significant

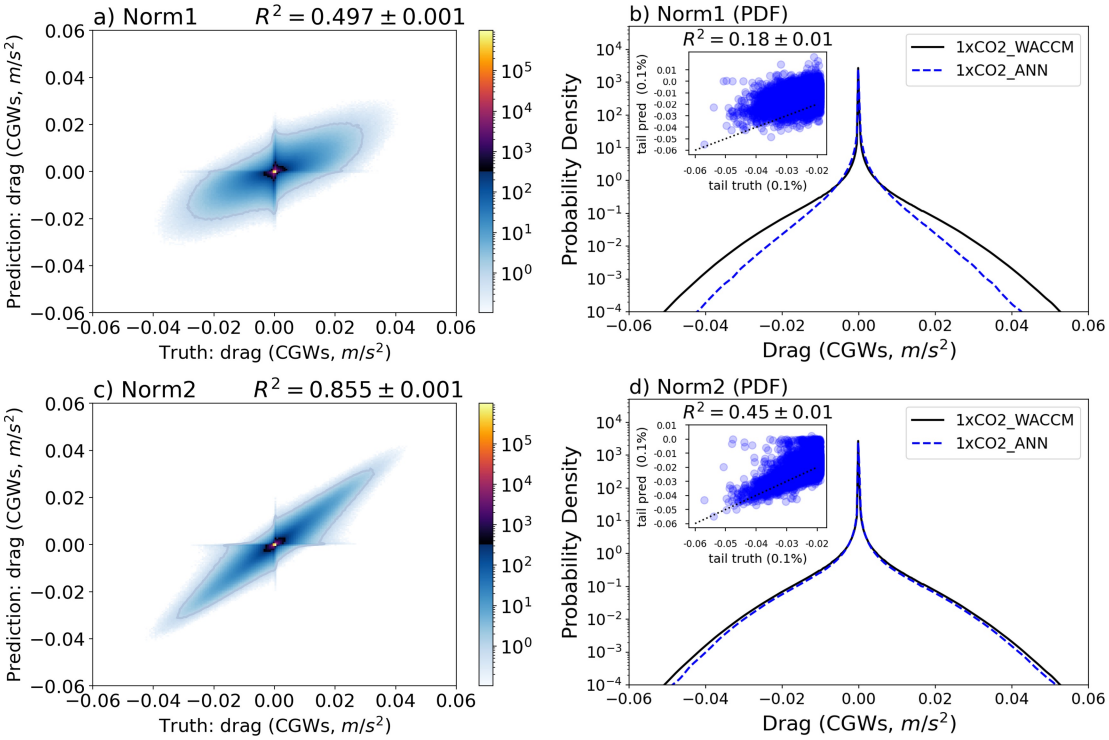

Figure 4. Data imbalance for GWD due to CGWs and the emulation results with two different normalization methods. a) A 2D histogram displaying the emulated GWD due to CGWs and the truth, with the training dataset normalized using NORM1; b) Distribution of the original convective GWD (black line) and the predicted values (blue line) with NORM2. The scatter plot in the corner represents the tail part only, including points with the top 0.1% amplitudes; c) Similar to a), but for NORM2; d) Similar to b), but for NORM2. The R^2 uncertainty range is estimated by dividing the test data into 10 segments, calculating the metric for each segment, and then computing the standard deviation (STD).

improvement, in contrast to the findings of Chantry et al. (2021). We attribute this improvement to the more pronounced data imbalance in our dataset, and it is likely a consequence of NORM2's emphasis on modeling the large GWD values. Nonetheless, emulating the tail of the probability density function (PDF) (rare events) remains poor, as evidenced by the tails in Figure 4c, primarily due to the predominance of zero GWD columns in the training dataset. To more effectively address the data imbalance issue in these regression tasks, we further propose two approaches here:

1. Resampling the data (ReSAM): In this approach, we limit the number of training sample pairs with zero GWD to be equal to the number of samples with non-zero GWD. This significantly reduces the number of columns with zero GWD, thus mitigating the data imbalance issue. Additionally, this sub-sampling reduces the total size of the training dataset, which, in turn, enhances the training speed (approximately sevenfold). While resampling methods have been well-established in the ML

literature, they have mainly been used for classification problems. Their application to regression problems in climate research has not been extensively explored.

2. Weighted loss function (WeLoss): Instead of assigning the same weight to all sample pairs in the loss function, we modify the weight for each column based on the PDF of its maximum GWD amplitude. This adjustment allows us to re-formulate the loss function defined in Equation (2) as

$$\mathcal{L}(\Theta) = \frac{1}{n}\sum_{i=1}^{n} \left\| \mathbf{W_i}\{\mathbf{NN}\left(x_i, \Theta\right) - y_i\} \right\|_2^2 \tag{4}$$

where

$$\mathbf{W_i} = \frac{1}{PDF(max(|y_i(z)|))} \tag{5}$$

Note that, in practice, we lack knowledge of the precise PDF for the maximum GWD within each column. Therefore, we employ a histogram with 20 bins as an alternative. Given the fact that large-amplitude GW events are rare, the WeLoss approach incentivizes the NN to prioritize these significant events.

When we apply the ReSAM approach to balance the training dataset (after normalization with NORM1 or NORM2), the emulation results significantly improve, as shown in Figure 5. In fact, when considering the R-squared value between the NN prediction and the ground truth, the ReSAM approach with NORM2 yields the best results. However, as the training dataset is still predominantly composed of zeros and small GWD values due to the intermittence of the GWs, examining the emulation results for only large amplitude GW events (e.g., the top 0.1% in Figure 5d) reveals less satisfactory performance ($R^2 = 0.72$). Regarding the WeLoss approach, it has a more limited impact on improving the R-squared value of the emulation (as shown in Figure 5e). However, it proves valuable in capturing the tails of the PDF and, thus, rare events (as depicted in Figure 5f). Moreover, as ReSAM and WeLoss represent distinct operations, they can be effectively combined when constructing a NN. The result of this combined approach for emulating the CGWs can be found in Figures 5g and 5h. While the R-squared value for the entire distribution only marginally changes (0.925 vs. 0.931 with ReSAM only), the performance of the emulation for the tail part has been improved (R^2 increased to 0.77).

Similarly, Figure 6 presents the offline emulation results for the FGWs. The conclusions drawn for CGWs generally hold true. However, data imbalance in FGWs is less pronounced compared to CGWs, which simplifies the task of emulating FGWs. Even without any resampling or changes to the normalization or (see Figure 6a), we achieve reasonable emulation results ($R^2 = 0.9$). One contributing factor is the wider spatial distribution of FGWs compared to CGWs (refer to Figure 3). Additionally, the source of FGWs (frontogenesis function) in WACCM exhibits a much more continuous nature compared to precipitation and diabatic heating. As the data imbalance issue is less severe for FGWs, the performance with different normalization methods becomes more similar, echoing findings from Chantry et al. (2021) who emulated non-orographic GWs (including convective and frontal GWs) together.

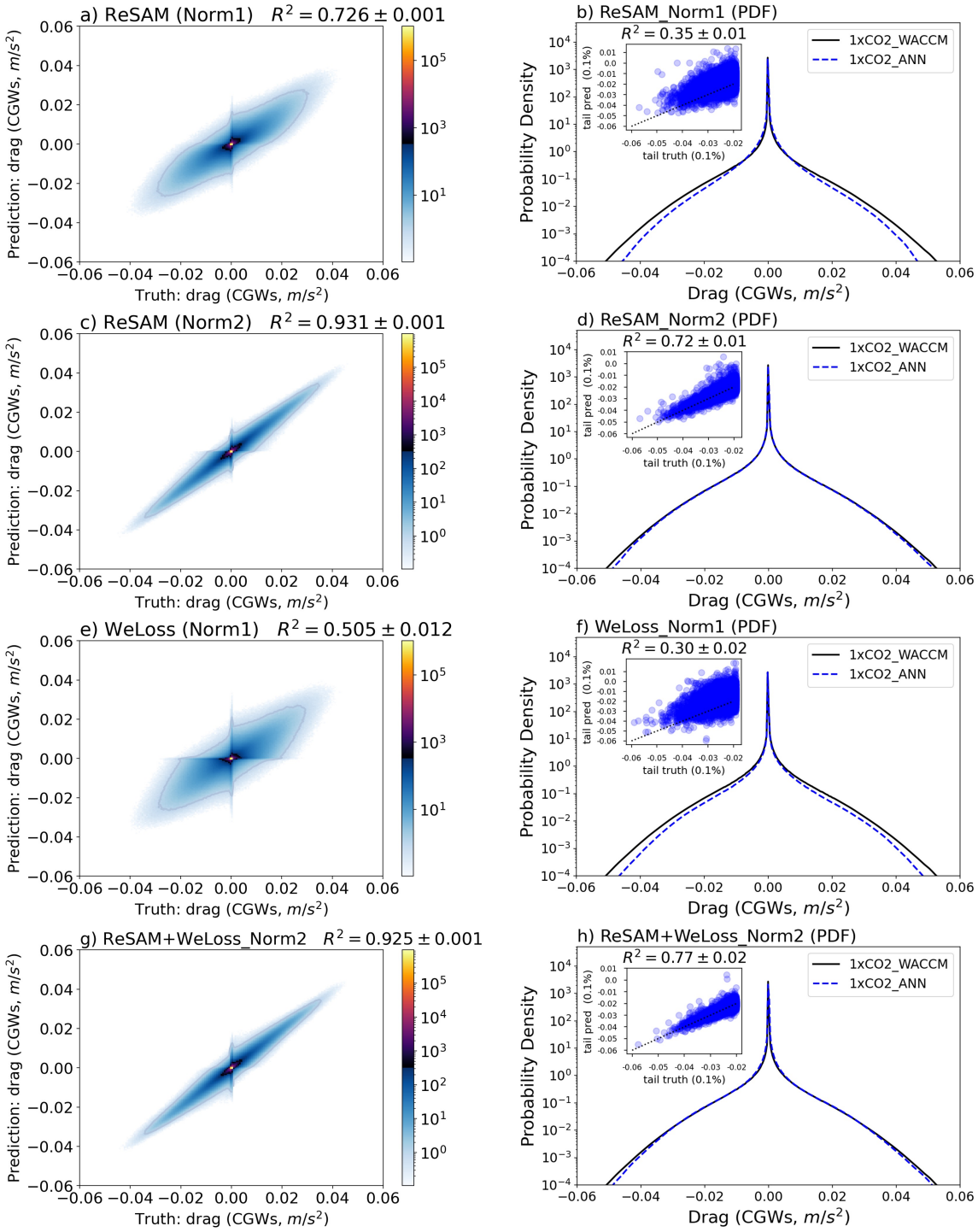

Figure 5. Similar to Figure 4, but for CGWs with the proposed ReSAM and WeLoss methods. a) A 2D histogram for the emulation with resampled data (ReSAM) after using Norm1; b) Distribution of the emulated GWD due to CGWs similar to Figure 4b, but with ReSAM applied; c) Similar to a), with training data normalized using Norm2; d) Similar to b), with training data normalized using Norm2; e) Similar to a), but with the WeLoss approach; f) Similar to b), but with the WeLoss approach; g) Similar to c), after applying both ReSAM and WeLoss methods together; h) Similar to d), after applying both ReSAM and WeLoss methods.

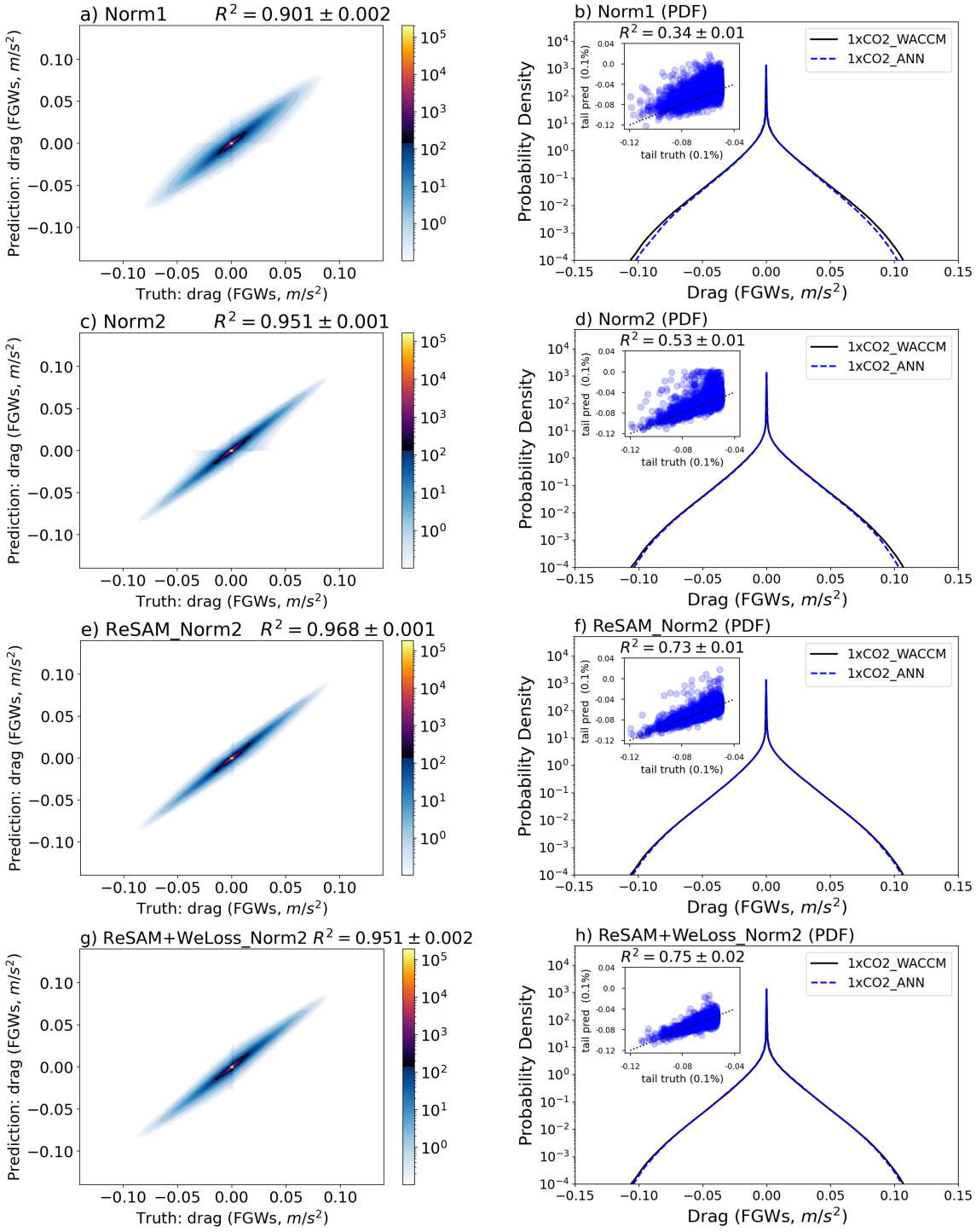

Figure 6. Similar to Figure 5, except for FGWs. a) A 2D histogram for the emulation using Norm1, without ReSAM or WeLoss; b) Distribution of the GWD due to FGWs with NORM1, similar to Figure 4b; c) Similar to a), with training data normalized using Norm2; d) Similar to b), with training data normalized using Norm2; e) Similar to c), but with the ReSAM approach; f) Similar to d), but with the ReSAM approach; g) Similar to Figure 5g, applying both ReSAM and WeLoss methods to the FGWs; h) Similar to Figure 5h, applying both ReSAM and WeLoss methods to the FGWs.

In summary, data imbalance can pose challenges when learning from data that closely resembles real-world data (further discussed in the subsequent section on emulating OGWs). Proper resampling techniques can significantly enhance the NNs' performance by improving dataset balance. Furthermore, modifying the loss function to penalize the NNs more for missing extreme values can further improve performance at the tails of the PDF. For the remainder of the paper, unless otherwise specified, we continue to employ the ReSAM approach and the standard loss function with NORM2 unless stated otherwise.

3.2 Uncertainty Quantification

As outlined in subsection 2.2.2, we employ three different methods (i.e., BNN, DNN, and VAE) to quantify the uncertainty of predictions during inference (testing). For this purpose, an ensemble of 1000 members is generated by running each UQ-equipped NN 1000 times for each input from the testing set. Figure 7 presents sample profiles of zonal GWD derived from the deterministic NN (ANN) and the three UQ-equipped NNs, alongside the true GWD profiles from WACCM. Note that these examples have not been used in the training or validation process. It is evident from the figure that all three UQ-equipped NNs show reasonable skill in predicting the complex profiles of GWD due to CGWs and FGWs (also reflected in R-squared in Table 1), albeit with a slight decrease in accuracy compared to ANN. As discussed earlier, a valuable uncertainty estimate should correspond closely with the NN's test accuracy, providing insights into when to trust the NN's prediction during inference. Such a relationship can be seen in a few randomly chosen GWD profiles that's shown in Figure 7. In each pair of CGW and FGW profiles, the left column shows the estimated uncertainty is also low when the prediction error is low, indicating the NN's confidence in its accurate predictions. In contrast, the right column, which generally represents more complex profiles, exhibits the NN's less accurate predictions, and increased uncertainty, highlighted by the wider confidence intervals.

While Figure 7 demonstrates the performance of the UQ methods for just a few GWD profiles, the spread-skill plots shown in Figure 8 offer a broader perspective based on 60,000 profiles, following the calculations detailed in Appendix C. It is evident from the plots that all three UQ methods produce reasonably informative uncertainty estimates, as their curves closely align with the 1-to-1 line. In the case of CGWs, all data points are above the 1:1 line, indicating a slight overconfidence (underdispersiveness) across all three UQ methods, with the DNN being slightly closer to the 1-to-1 line. For the FGWs, the DNN demonstrates slightly better performance, although it marginally drops below the 1-to-1 line in the first few bins, indicating a slight underconfidence. Notably, it can be seen from the spread frequency inset that the vast majority of the data points are within the first few bins, for which both spread and skill values are small, and they are generally closer to the 1-to-1 line.

It should also be noted that for the large values of model spread ($\overline{\mathrm{SD}}$), there is only a very limited number of data points, as is evident from the inset histograms. Consequently,

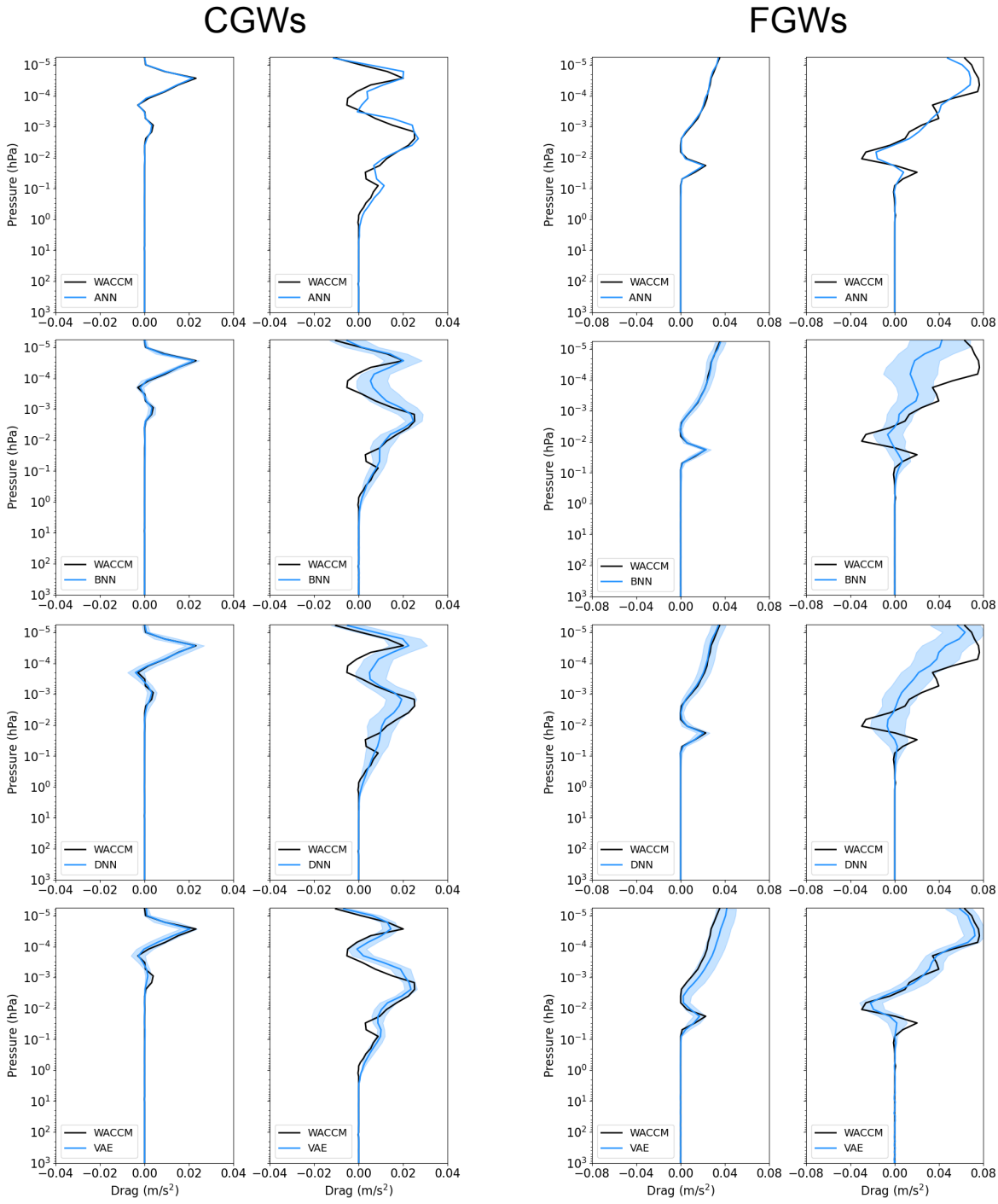

Figure 7. Sample profiles of zonal GWD as predicted by various NNs, as indicated. The true profile is shown by the black line, while the blue solid line represents the mean of 1000 ensemble members. The shaded region indicates the 95% confidence interval. In each pair of CGWs and FGWs profiles, the left column provides examples with low estimated uncertainty, corresponding to instances of low error. Conversely, the right column illustrates cases with high uncertainty when the error is high.

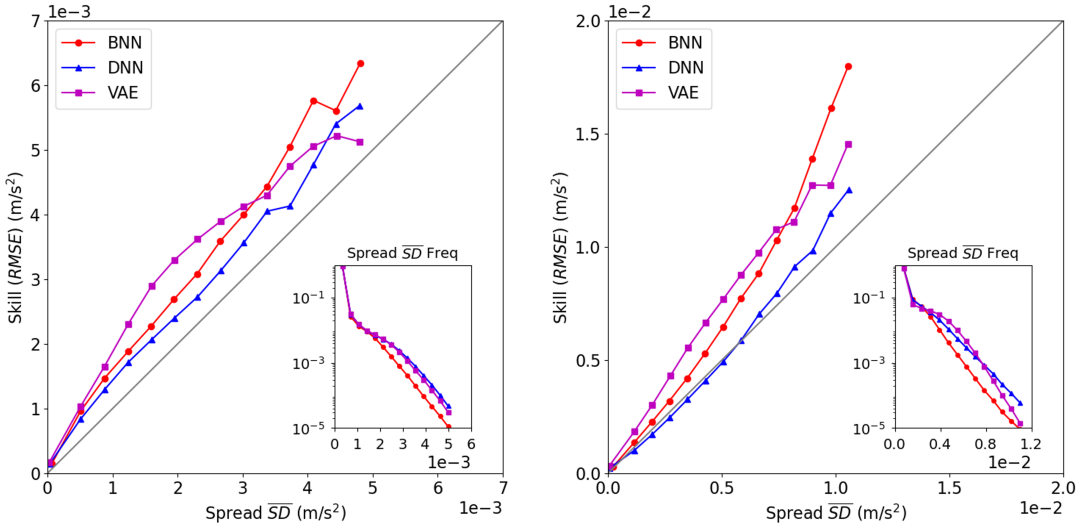

Figure 8. Spread-skill plot for GWD due to (left) CGWs, and (right) FGWs. The diagonal 1:1 line represents the perfect spread-skill line. Points above (below) this line correspond to spread values where the model is overconfident (underconfident). The inset histogram shows how often each spread value occurs. See Appendix C for a detailed discussion on the calculations of the spread-skill plot.

the standard deviation (STD) can become a misleading measure of spread because of the non-normal distributions.

To summarize the quality of the spread-skill plots for the three UQ methods, we explore the metrics introduced in subsection 2.2.2 and Appendix C (see Table 1). The R-squared value for the ensemble mean prediction is also given to show the accuracy of each UQ method. Based on SSREL, whose ideal value is zero, BNN shows the best performance for both CGWs and FGWs. However, if we check SSRAT, where 1 is the optimal number, DNN is the best among these three methods. This discrepancy can be explained by a closer look at the Equations (C2) and (C3). SSREL, which is a bin-weighted average difference, is most sensitive to the performance of the NN in the first bin, where the vast majority of the data points are located (see the inset histograms in Figure 8), while SSRAT is more influenced by larger values of spread and skill. Accordingly, the VAE shows the highest values of SSREL, which is indicative of its sub-optimal performance in the first bin, where there are small values of spread and skill.

In the results presented in Figure 8 and Table 1, each height level of a GWD profile is considered as an individual sample. A zonal GWD profile, with its 70 vertical levels, thus constitutes 70 distinct samples. While analyzing these samples offers insights into the NN's overall performance by averaging statistics across numerous profiles, our primary interest is often in the uncertainty associated with an individual GWD profile. This uncertainty can then aid in determining whether to trust/use the NN's prediction for that particular GWD

profile. Therefore, we use Equation (C4) to assess the relationship between uncertainty and test accuracy for each GWD profile. Furthermore, to estimate uncertainty, here we use the interquartile range (IQR) to reduce the influence of outliers.

Figure 9 shows the Gaussian kernel density of spread against RMSE for all 60,000 profiles, as indicated by the color shading. The x-axis represents the IQR of each GWD profile, while the y-axis denotes its corresponding RMSE. A strong correlation between the two is observed across all three UQ methods. Consequently, GWD profiles with larger uncertainties often coincide with larger errors. Figure 9 also shows a close similarity between BNN and DNN. In contrast, VAE tends to provide marginally larger uncertainties, especially for FGWs. This is consistent with VAE's slightly reduced accuracy as indicated in Table 1. Overall, given the monotonic relationship between the uncertainty and test error, these results show that all three UQ methods provide useful and informative uncertainty for with-distribution test samples. A user can set a threshold on uncertainty based on their tolerance for error (RMSE) and decide whether they trust the NN for a given input sample.

The results presented so far show the performance of the UQ methods based on the testing data, i.e., data from the current climate. However, the effective performance of UQ methods can also be tested (perhaps more meaningfully) on OOD data, e.g., data from a warmer climate. This is particularly relevant for climate change studies. Accordingly, we evaluate the performance of these trained NNs with input data from the future climate, as depicted by the black lines in Figure 9. For FGWs, the spread-skill relationship remains largely similar, especially for BNN and DNN. This suggests that, based on their uncertainties, we can still reliably estimate the error in the NN predictions for FGWs for the warming climate. A similar pattern is observed for the VAE, though it exhibits increased uncertainties and higher errors with OOD data. As shown in a later section, for FGWs, the NNs generalize to the warmer climate without any further effort.

In contrast, for CGWs, given the same level of uncertainty, the error in NN predictions increases significantly for the OOD data compared to that from the current climate, which means the spread-skill relationship, especially for the BNN and DNN, fails to generalize to the OOD data. From this perspective, VAE performs better, showing that for the same level of uncertainty, the increase in error is not as substantial as in BNN and DNN. The VAE also yields considerably higher uncertainty estimates for future climate, which may aid in the detection of OOD data. The observed discrepancies in the performance of the NNs for CGWs and FGWs hint at different levels of their generalizability, a topic we will delve into more deeply in the following subsection.

In summary, while the three UQ methods provide credible and valuable uncertainty estimates for the current climate, the BNN and DNN are confidently wrong in estimating CGWs in a warmer climate although VAE shows some promising results. This problem is common among various UQ techniques as pointed out in the ML literature: they frequently show overconfidence when assessed with OOD data (e.g., Ovadia et al., 2019). The optimal UQ method selection depends on the specific metric of interest and the intended application. While BNN is more broadly used in the literature and gives the best accuracy, DNN could

Table 1. Evaluation scores for the three UQ methods. See Section 2 for more details.

	CGWs			FGWs		
	BNN	DNN	VAE	BNN	DNN	VAE
SSREL (1e-4)	1.29	1.48	2.14	1.20	1.69	5.21
SSRAT	0.73	0.82	0.72	0.69	0.93	0.69
R-squared	0.90	0.86	0.87	0.94	0.92	0.89

achieve similar performance and is often more practical given its simplicity. On the other hand, VAE seems to perform better when applied to OOD data, at least in the one test case here. These observations warrant further research in the future using multiple test cases and climate-relevant applications. We also note here that each method has multiple tuning hyperparameters to optimize its uncertainty quantification. Consequently, the discrepancies among the three methods could potentially be mitigated with proper hyperparameter tuning (as discussed in Appendix B).

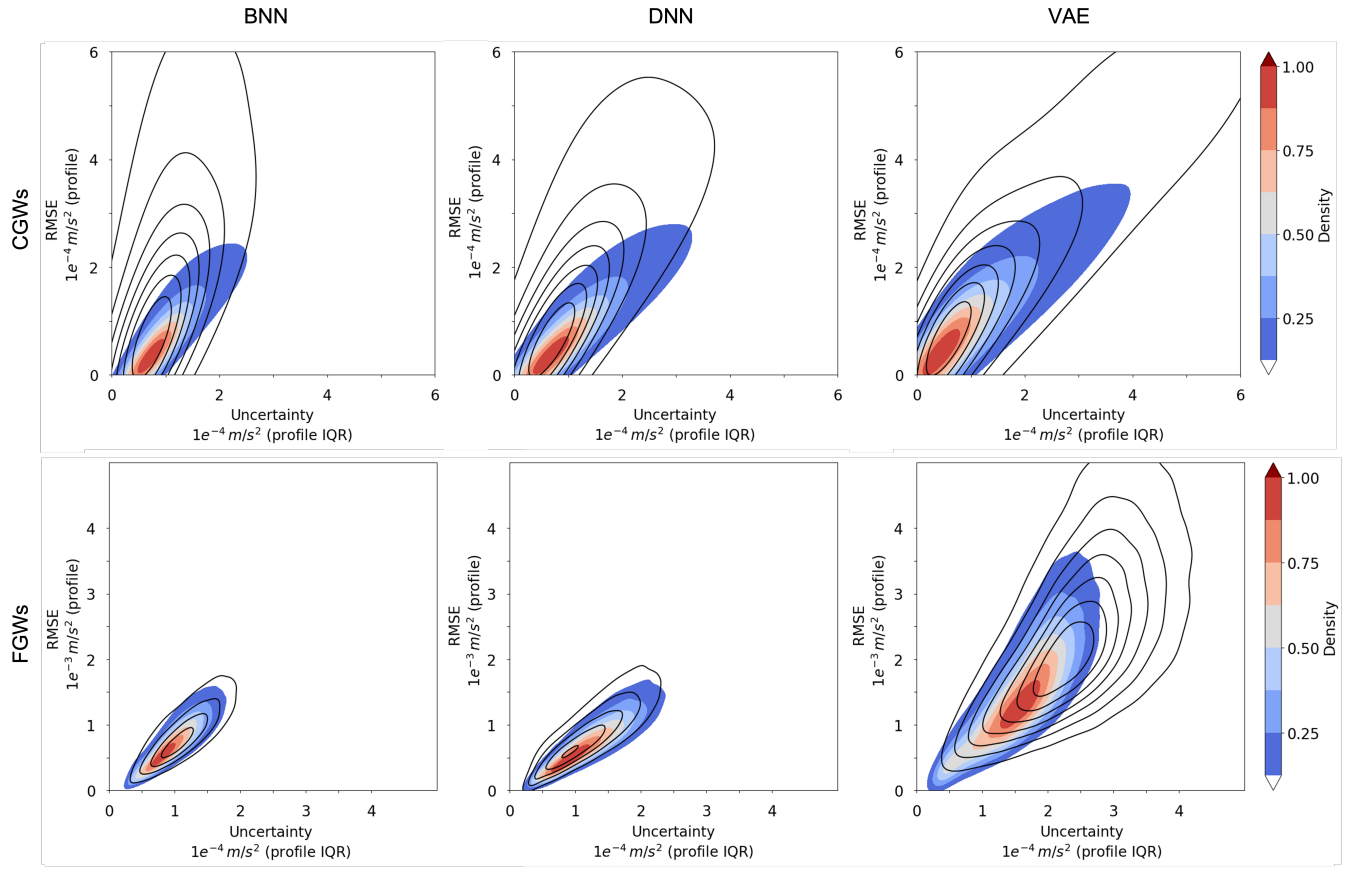

Figure 9. Gaussian kernel density for spread, defined as interquartile range (IQR) versus RMSE for (top) CGWs, and (bottom) FGWs, when validated against out-of-sample data from the current climate (color shading) and out-of-distribution data from the future, warmer climate (black lines). Results are shown for BNN (left), DNN (middle), and VAE (right). All NNs are exclusively trained on data from the current (control) climate.

3.3 Out-of-distribution (OOD) Generalization via Transfer Learning

As previously discussed, the GWP schemes in WACCM are coupled to their sources, which might change in a warmer climate. Specifically, under $4\times CO_2$ forcing, we expect changes in both the amplitude and the phase speed distribution of GWs, in particular for the CGWs, due to their built-in sensitivities to changes in the convection. Consequently, the physics scheme in WACCM produces slightly stronger GWD for CGWs, especially in the tail of the distribution. This intensified GWD results in a shorter quasi-biennial oscillation (QBO) period in WACCM. However, it is important to recognize that the response of the QBO to climate change differs across various general circulation models (Richter et al., 2022).

The intensification of the CGWs in future climate simulations presents an opportunity to study how NNs handle the OOD data. Our findings in the UQ section already suggest increased prediction errors when testing NNs with OOD data, which raises concerns about their applicability in climate change studies. To more thoroughly investigate this issue, we conduct additional evaluations on our ANNs, by applying them to data samples from future climate simulations, as illustrated in Figure 10. It is clear that the ANN for the CGWs does not generalize well, evidenced by a decrease in R^2 from 0.93 to 0.79. The ANN particularly struggles to capture the increase in GWD in the tail, with R^2 for the tails decreasing from 0.72 to 0.36. As a result, it seems unlikely that this emulator will accurately reproduce changes in the circulation under different climate conditions, such as the shorter QBO period resulting from future warming in WACCM.

In contrast to CGWs, the amplitude of FGWs shows a less marked increase in the future climate, and their PDF distribution closely resembles that of the control simulations. As a result, the ANN demonstrates better generalizability for FGWs when it is tested against future climate data, as seen in Figure 10d. There is only a slight decrease in the ANN's performance, with R^2 dropping from 0.97 to 0.95.

Two factors can contribute to the considerable OOD generalization errors in an NN when applied across two distinct systems. First, the input-output relationship might vary between the two systems. Second, the input variables in the new system could originate from a distribution different from that of the original system (regardless of whether the input-output relationship remains the same or changes). The former is hard to quantify in a high-dimensional dataset. The latter can be quantified using similarity distances. To help us better understand these differences between the OOD generalizability of CGWs and FGWs, we assess the similarity between their input and output data distributions from control and future climate simulations using the Mahalanobis distance (D). The Mahalanobis distance is a measure of the distance between a data point and a distribution (Ling & Templeton, 2015). Specifically, it is a multi-dimensional generalization of the idea of measuring how many standard deviations away a point is from the mean of the distribution. The application of Mahalanobis distance in understanding the source of OOD generalization errors in data-driven parameterization was previously demonstrated in Guan et al. (2022) for a simple turbulent system.

Table 2. Change of Mahalanobis distance based on the ratio of the average distance of the points that are more than 3 standard deviations away from the mean. The choice of the variables here is based on Appendix A, showing u, v, T, and source function contain most of the information needed for the NN.

Variables	u	v	T	Source (diabatic heating for CGWs, frontogenesis for FGWs)	Zonal drag	Meridional drag
Distance (Convection)	1.03	1.00	1.19	3.62	1.42	1.44
Distance (Front)	1.03	0.96	1.50	1.10	1.00	1.00

To use the Mahalanobis distance, we first calculate the mean and covariance matrix of the training data from the control run. We then analyze the distribution of Mahalanobis distances in this training data, setting a baseline value, referred to as D_{ctrl}. This baseline is the average distance for data points that deviate by more than 3 standard deviations from the mean. This choice aims to focus on outliers for which extrapolation is more challenging. Following this, we apply the same process to the data points in the future climate dataset, denoted as D_{warm}. Table 2 presents the ratio of D_{warm} for the warming scenario to D_{ctrl} for the control scenario for selected variables. When this ratio is close to 1.0, it suggests minimal changes in this variable's distribution under a warming scenario. Note that the NNs trained based only on these variables demonstrate performance comparable to NNs trained on all variables (not shown), which is why we only focus on these few key variables.

The results reveal that among the various variables significantly contributing to the emulation of CGWs, diabatic heating (source of CGWs) is the sole variable that exhibits substantial changes from the control to the warming scenario. Conversely, changes in variables used to emulate FGWs are considerably smaller. This outcome suggests that the likely reason for the better generalizability of FGWs is that the input distribution remains almost unchanged (and the input-output relationship, which is the physics scheme, remains the same too).

To improve the generalizability of the emulator for CGWs, we explore TL, a technique introduced earlier and proven to be a powerful tool for improving the OOD generalizability of data-driven parameterization in canonical turbulent flows (e.g., Guan et al., 2022; Subel et al., 2023). Rather than re-training the entire NN for future climate scenarios, we only re-train, follwoing Subel et al. (2023), just a portion of the NN, thereby requiring only a small fraction of the data. Figure 10e showcases the emulation results after only re-training the first hidden layer of ANN using data from the first month of the WACCM simulation in the $4\times CO_2$ scenario, which amounts to approximately 1% of the original training dataset. After applying TL, the performance of the emulator in the warming scenario significantly improves, with R^2 rising from 0.79 to 0.91, nearly matching its performance in the control simulations ($R^2 = 0.93$). However, the improvement in the PDF tails is less pronounced, showing only a modest increase in R^2 from 0.36 to 0.51. This is likely due to the limited number of large-amplitude GW events within the one-month period. Instead of using more data

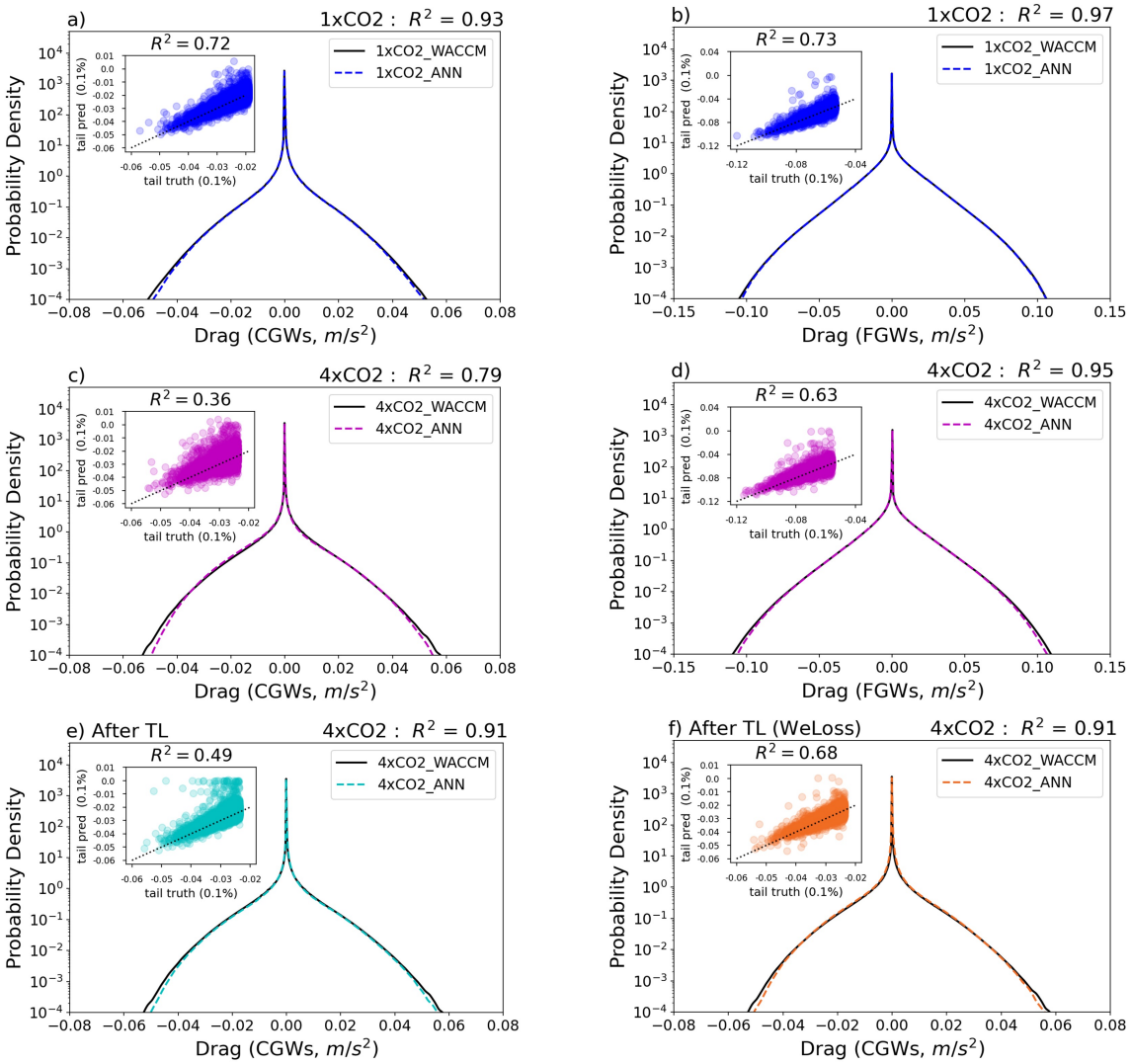

Figure 10. NN performance for pre-industrial and warming scenarios for different sources (a,c,e,f: CGWs ; b,d: FGWs). **a)** PDF of GWD due to CGWs in WACCM simulation and the predicted CGWs using NN emulator, scatter plot shows points for the tail part only. **b)** same as (a), but for FGWs. **c)** same as (a), but for the warming scenario, **d)** same as (b) but for the warming scenario. **e)** same as (c) but after applying transfer learning to the first hidden layer of the NN with 1-month WACCM simulation data under warming scenario ($\sim$ 1% of the size of the training data) **f)** same as (e) but with the weighted loss function used when we conduct transfer learning (WeLoss).

from the future climate (which is challenging to obtain in a realistic situation), we leverage the WeLoss approach, described earlier, during re-training. This modification results in a significant improvement in the tail, with R^2 increasing from 0.51 to 0.68. Note that this improvement in the tail is critical, as inadequate learning of these rare but large-amplitude GWDs can result in significant errors and instabilities.

We would like to point out that during the TL experiments, we have examined the effects of re-training each individual hidden layer of the NN. Our findings indicate that re-training the first layer yields the best results, which aligns with the findings in Subel et al. (2023). Re-training the last layer only brings marginal improvements to the NN (not shown). Notably, our experiments involving re-training the first two layers did not result in further performance enhancements, suggesting that the number of neurons is not the primary factor contributing to the varied performance observed when re-training different layers.

Similar results regarding TL are also observed with other NNs used in this study. For instance, Figure 11 presents the same plot as Figure 10, but for the BNN. It is evident that BNN also struggles with generalization to OOD data, as could also be interpreted based on the results presented in section 3.2. It is also the case for DNN and VAE (not shown). Overall, when these NNs are tested against the $4\times CO_2$ future climate data, their accuracy is not better than the deterministic ANN. However, methods with UQ, especially the VAE (see Figure 9), could potentially indicate the increased uncertainty when testing with input data from the $4\times CO_2$ integration. These results underscore the necessity of re-training the NNs using TL.

4 Emulation of Orographic GWs (OGWs)

Similar to Chantry et al. (2021), our initial attempts to emulate OGWs did not succeed, primarily due to the presence of a pronounced data imbalance. Notably, the physics-based scheme responsible for OGW generation operates exclusively over terrestrial regions. However, it is surprising that the issue of data imbalance continues to persist, even when we limit our NN training and testing exclusively to columns located over land (Figure 12a). Still, the emulated OGW drag often remains close to zero and completely fails to predict the rare events (Figure 12b), which poses a considerable hurdle for the emulator's performance. Further investigations reveal that the key to this problem lies in the highly localized nature of orographic GWD, where significant drag is observed only at a handful of specific locations. Furthermore, even within these limited regions, GWD exhibits a significant intermittent behavior. To help our understanding, we also conducted a K-means clustering analysis, categorizing GWD data for all land-based columns (Table 3). Among the 6 clusters, cluster 4 accounts for a staggering 97.51% of the dataset. Remarkably, all samples within this cluster exhibit exceptionally weak orographic GWD, as evidenced by the cluster center's maximum GWD amplitude, which is two orders of magnitude smaller than that of other clusters.

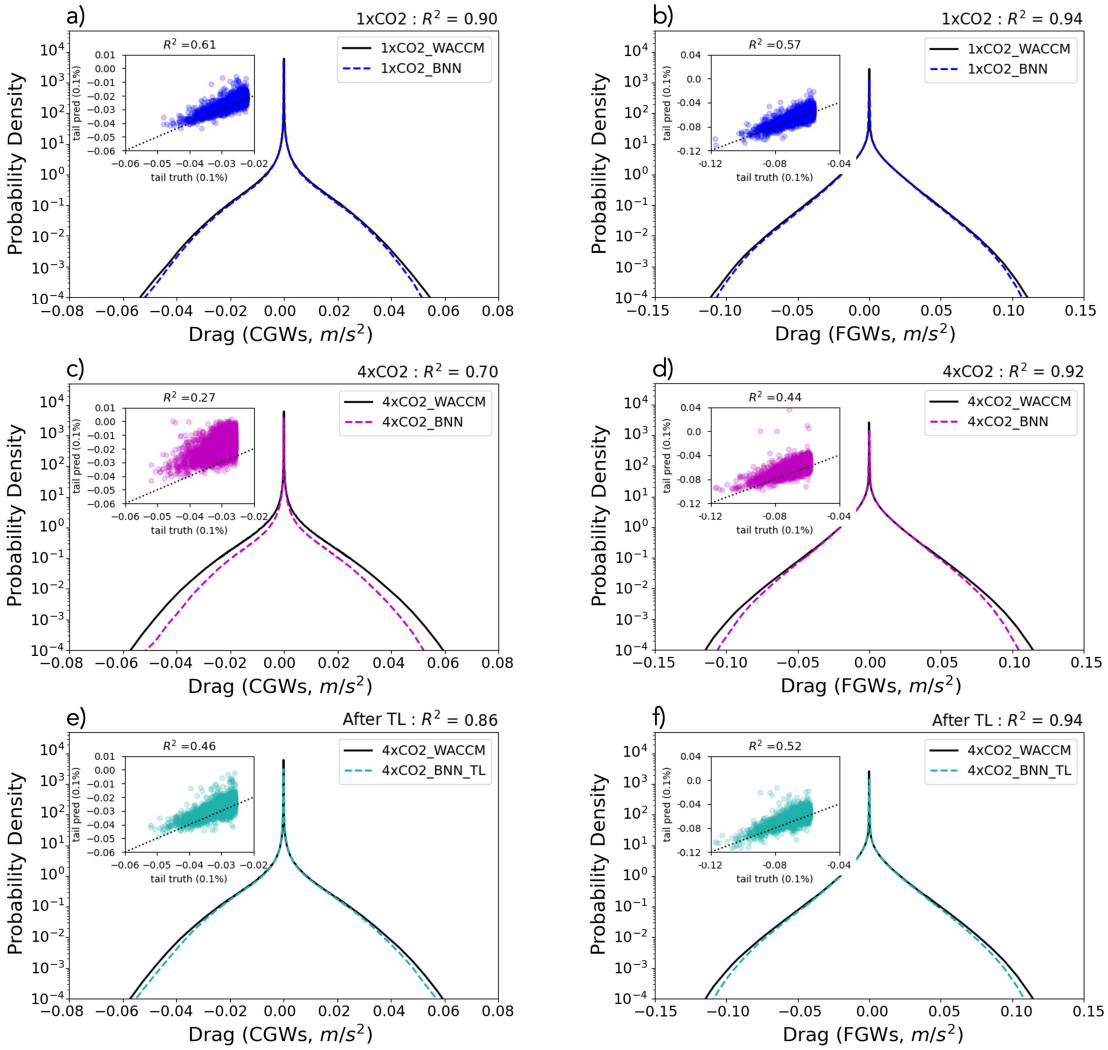

Figure 11. Panels (a) to (e) are the same as those in Figure 10 but for BNN. Panel (f) shows emulation for FGWs under warming scenario after applying transfer learning to the first hidden layer of the NN.

Table 3. Clustering analysis for OGWs. Analysis is done for all columns over land in the training data.

Cluster	Frequency (%) in the training data	Maximum GWD amplitude of cluster center
c1	0.18	8.7 e-3 m/s^2
c2	0.13	4.4 e-3 m/s^2
c3	0.93	3.6 e-3 m/s^2
c4	**97.51**	2.8 **e-5** m/s^2
c5	0.15	2.1 e-3 m/s^2
c6	1.10	4.3 e-3 m/s^2

To overcome this persistent data imbalance in the OGWs, we first separate all columns over land into large-drag columns (with column maximum greater than one STD of all GWD from OGWs) and small-drag columns. We then perform subsampling on the latter group only to create a more balanced dataset. To improve NN training, we also include all columns from the 6-year simulation to augment the sample size of the large-drag columns. Figures 12c and 12d illustrate the performance after re-balancing the dataset. Notably, the result represents a substantial improvement, evidenced by an R^2 increase from 0.29 to 0.80, and also a significant improvement in the accuracy for rare events. While we acknowledge that this skill remains lower than what is achieved for CGWs and FGWs, it already signifies a reasonable NN. Furthermore, we posit that by incorporating additional training data (either by extending the WACCM model integration or simply augmenting the data with OGWs scheme only), we can further improve our emulation results. The possibility of achieving superior emulation outcomes through the adoption of an alternative NN architecture is also possible, although such exploration is beyond the scope of this paper.

5 Summary and Discussion

Through the emulation of complex GWPs in a state-of-the-art atmospheric model (WACCM), we have elucidated and explored solutions for three critical challenges in the development of ML-based data-driven SGS schemes for climate applications: data imbalance, UQ, and OOD generalizability under different climates. A brief summary is provided below:

1. In the presence of non-stationary, and highly imbalanced datasets, such as those encountered in WACCM, specialized approaches (e.g., resampling and weighted loss function) are essential to enhance the performance of data-driven models. Through resampling, we have successfully trained a robust NN emulator for OGWs, a challenging task as demonstrated in Chantry et al. (2021). The effectiveness of the trained emulator is also significantly influenced by the choice of the loss function used during training. In our case, while a weighted loss function (WeLoss) does not improve the overall R^2 score, it yields significant improvements in the emulation results for

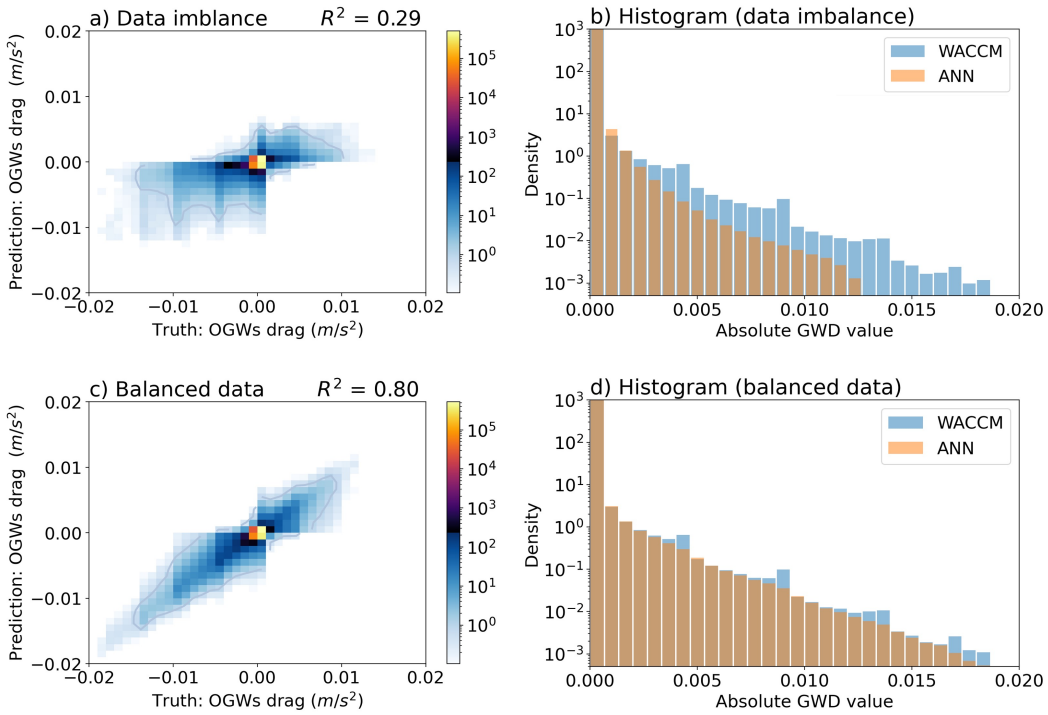

Figure 12. Performance of the emulator for OGWs when trained with all columns over land (panel a) & panel b)) and balanced training data with a balanced number of large-drag columns (column maximum > 1 STD of all GWD from OGWs) and small-drag columns (panel c) & panel d).

the PDF tails of the GWD. This finding aligns with those in Lopez-Gomez et al. (2022), where their custom loss function, tailored to emphasize extreme events, led to substantial improvements in predicting heatwaves.

2. All three UQ methods employed in this study provide reasonable uncertainty estimates for GWD prediction for the current climate. The spread-skill plots (refer to Figures 8 and 9) indicate that greater uncertainty corresponds to a larger prediction error. Yet, the reliability of UQ methods diminishes when they are challenged with OOD data. Both BNN and DNN used in this study tend to be overconfident in estimating CGWs in a warmer climate, thereby struggling to identify OOD samples. The VAE, on the other hand, yields rather promising results in providing useful UQ for OOD data. Given the variations in different methods, the metrics selected to assess the SGS model will play a significant role in determining the choice for the UQ methods. We also note that further optimization of tunable parameters within each UQ method could affect their performance (refer to Appendix C).
3. Our findings illustrate the challenges SGS schemes face in generalizing to OOD data and extrapolating to new climates. Nonetheless, the TL approach has proven highly effective in aiding an NN to extrapolate to different climates. For CGWs in WACCM, the physics-based scheme exhibits larger GWD under $4\times CO_2$ forcing, primarily due to an increase in diabatic heating from convection. With only one month of sim-

ulation data from this future warming scenario (representing approximately 1% of the original training dataset), we successfully reduce its OOD generalization error through re-training the first layer of the NN, following the findings of Subel et al. (2023). Additionally, we have illustrated the value of metrics like the Mahalanobis distance in assessing the potential OOD generalizability of NNs.

We would like to emphasize that these challenges are often intertwined. For instance, addressing data imbalance in CGWs is a prerequisite for obtaining an accurate NN model, which, in turn, impacts UQ and OOD generalizability assessments. Moreover, there exists a strong link between UQ and OOD generalizability evaluations: if the NN struggles with OOD generalization, performing poorly with OOD data, the reliability of UQ for such data (e.g., data from a warmer climate) also becomes questionable. This presents a substantial challenge for UQ methods, especially for climate change research where reliable UQ methods are crucial.

This study has primarily focused on offline skill assessment. We acknowledge that good offline performance (at least in terms of common metrics such as R^2) is not necessarily an indicator of stable and accurate online (coupled to climate model) performance (Ross et al., 2022; Guan et al., 2022), though more strict metrics such as R^2 of the PDF tails might better connect the offline and online performance (Pahlavan et al., 2023). However, for the purpose of this study, which is to provide a testbed to test ideas for data imbalance, UQ, and OOD generalization with transfer learning, the offline tests, particularly using the several metrics we have used, suffice. That said, the main reason that we have not provided online results is that coupling various complex NNs, with the same framework, to complex climate models (e.g., WACCM) without slowing down the model is a challenging and time-consuming task (Espinosa et al., 2022), and this is work in progress.

Emulating complex GWPs within the WACCM provided a unique opportunity to address three critical challenges in developing ML-based, data-driven SGS schemes for climate science applications. However, it is crucial to acknowledge that such emulated schemes essentially adopt the limitations inherent in the physics-based schemes. Addressing these limitations, the next step is to harness high-resolution data from GW-resolving simulations, which are carefully validated against observational data. A library of such high-resolution simulations, notably of convectively generated GWs using the Weather Research and Forecasting (WRF) model, is now established (Sun et al., 2023), alongside additional global high-resolution simulations (Wedi et al., 2020; Polichtchouk et al., 2023; Köhler et al., 2023). The next phase involves integrating the approaches outlined in this study with the data from these GW-resolving simulations to develop a stable, trustworthy, and generalizable data-driven GWP scheme. This scheme is then expected to overcome the limitations of physics-based GWPs and potentially incorporate features like the transient effect (Bölöni et al., 2021; Kim et al., 2021) and lateral propagation of GWs (e.g., Sato et al., 2009)—marking a significant advancement towards next-generation GWP schemes.

Appendix A Input/output variables for the physics-based GWP schemes and their emulators

We use the exact same inputs as those of each GWP scheme in the WACCM for the training of the NN-based emulator of that scheme. These inputs are listed in Table A1. As for the outputs, we only consider the zonal and meridional drag forcings. The GWPs in WACCM also estimate additional effects of the GWs that result in changes of temperature profile and vertical diffusion. These outputs are not considered in our emulations.

Table A1. List of the input and output variables for the NNs trained as emulators of the GWP schemes in WACCM. The numbers in parentheses in front of each variable are the number of vertical levels for that variable. Note that each input and output is a 1D column at a given latitude/longitude grid point. Diabatic heating in WACCM is provided by the cumulus scheme. The topography variables listed in the table are mxdis (height estimates for ridges), hwdth (width of ridges), clngt (length of ridges), angll (orientation of ridges), and anixy (anisotropy of ridges).

<table>
<tr><th rowspan="2">GWP</th><th colspan="3">Input</th><th rowspan="2">Output</th></tr>
<tr><th>pressure levels</th><th>surface level</th><th>forcing</th></tr>
<tr><td>CGWs</td><td rowspan="4">u(70),
v(70),
T(70),
z(70),
ρ(71),
Brunt–Väisälä frequency N (70),
dry static energy DSE (70)</td><td rowspan="4">lat (1),
lon (1),
$P_{surface}$ (1),</td><td>diabatic heating (70)</td><td rowspan="4">zonal drag
GWD$_x$ (70),
meridional drag
GWD$_y$ (70),</td></tr>
<tr><td>FGWs</td><td>frontogenesis
function (70)</td></tr>
<tr><td rowspan="2">OGWs</td><td rowspan="2">mxdis (16),
hwdth (16),
clngt (16),
angll (16),
anixy (16),</td></tr>
</table>

From Table A1, one can guess that some input variables are correlated with each other. Consequently, it is plausible that the trained NNs may have spurious connections. Preliminary tests further support this notion, indicating that employing only u, v, T, and the forcing function as inputs yields comparable offline skill (results not presented here).

Appendix B Tuning UQ-equipped NNs

In addition to the hyperparameters of the deterministic NNs, designing an architecture for UQ often demands additional hyperparameter optimization. For instance, for the DNN, decisions need to be made regarding the number of neurons to drop out (dropout rate). While less common, one can also choose whether to apply dropout to all hidden layers or only selected ones. Variations in the dropout rate and the layers to which dropout is applied can influence the final configuration and performance of the DNN. Figure B1 illustrates these effects. As we increase the number of dropped neurons (whether through a higher dropout rate or by subjecting more layers to dropout), the uncertainty in the DNN predictions tends to rise. Yet, there is a persistent pattern in the relationship between spread (IQR) and RMSE across the various plots in Figure B1. Specifically, as spread increases, RMSE concurrently grows, consistent with the insights highlighted in Figure 9.

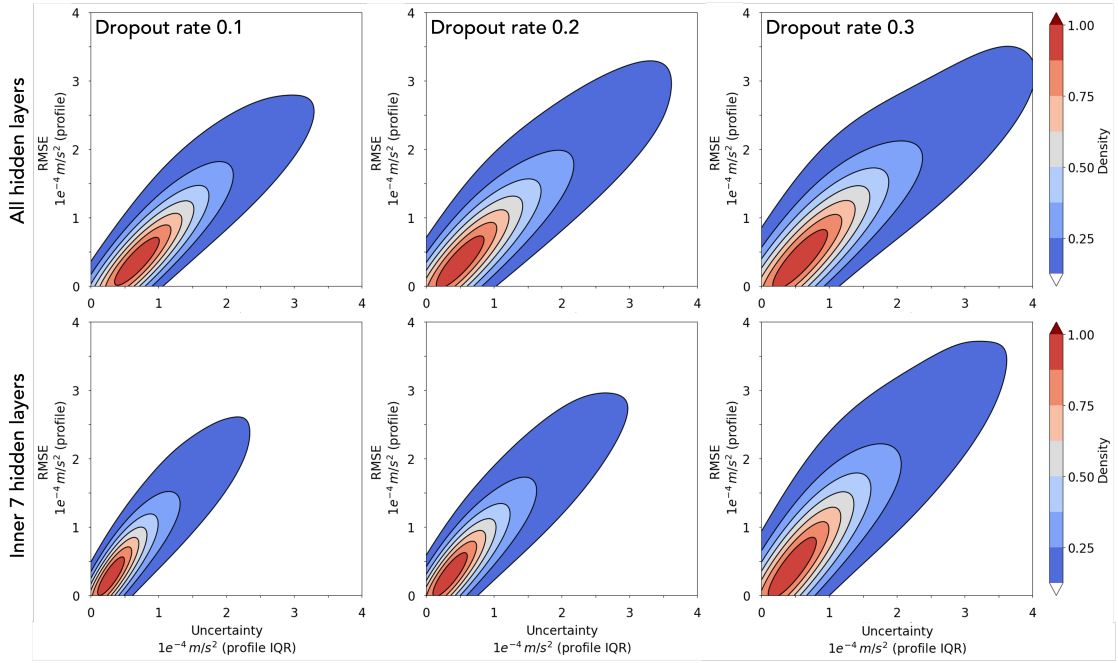

Figure B1. Similar to Figure 9 but for DNN only with different dropout rates, which are applied to different numbers of hidden layers.

In the case of BNN or VAE, even though there is no dropout rate, there are distinct tuning opportunities available. For instance, with the VAE, one might consider applying dropout to the NN emulator. Moreover, given that the loss function in VAE comprises three components, decisions can be made regarding which component to penalize more heavily, allowing for nuanced adjustments to its performance.

Appendix C The UQ metrics

Each point in the spread-skill plot corresponds to one specific bin of ensemble spread ($\overline{\mathrm{SD}_k}$), which is defined as the average standard deviation of the ensemble members. We first separate the spread using a pre-selected number of bins (a subjective choice of 15 is used here). Then for the k^{th} bin:

$$\begin{cases} \mathrm{RMSE}_k = \left[\frac{1}{N_k}\sum_{i=1}^{N_k}(\hat{y_i} - \overline{y_i})^2\right]^{\frac{1}{2}} \\ \overline{\mathrm{SD}_k} = \frac{1}{N_k}\sum_{i=1}^{N_k}\left[\frac{1}{M-1}\sum_{j=1}^{M}(\overline{y_i} - y_{ij})^2\right]^{\frac{1}{2}} \\ \overline{y_i} = \frac{1}{M}\sum_{j=1}^{M} y_{ij} \end{cases} \tag{C1}$$

$\hat{y_i}$ is the observed value for the i^{th} example, $\overline{y_i}$ is the mean prediction for the i^{th} example, y_{ij} is the j^{th} prediction for the i^{th} example, N_k is the total number of examples in the k^{th} bin, and M is the ensemble size. Following Haynes et al. (2023), we summarize the quality of the

spread-skill plot by two measures: spread-skill reliability (SSREL) and overall spread-skill ratio (SSRAT). SSREL is the bin-weighted mean distance from the 1-to-1 line:

$$\mathrm{SSREL} = \sum_{k=1}^{K} \frac{N_k}{N} |\mathrm{RMSE}_k - \overline{\mathrm{SD}_k}| \tag{C2}$$

where N is the total number of examples, K is the total number of bins, and other variables are as in Equation C1. SSREL varies from $[0, \infty)$, and the ideal value is 0. On the other hand, SSRAT is averaged over the whole dataset:

$$\mathrm{SSRAT} = \frac{\overline{\mathrm{SD}}}{\mathrm{RMSE}} \tag{C3}$$

SSRAT also varies from $[0, \infty)$, and the ideal value is 1. SSRAT > 1 indicates the model is under-confident on average, while SSRAT < 1 indicates that the model is overconfident on average.

In Equation (C1), each level of a GWD profile is considered as an individual sample. As discussed earlier, while these samples help assess the model's overall performance, our main interest is often the uncertainty of individual GWD profiles. Such uncertainty informs the trustworthiness of the model's prediction for that specific profile. Accordingly, for each profile, we can compute:

$$\begin{cases} \mathrm{RMSE}_{\mathrm{profile}} = \left[\frac{1}{N_z} \sum_{z=1}^{N_z} (\hat{y_z} - \overline{y_z})^2 \right]_{\mathrm{profile}}^{\frac{1}{2}} \\ \mathrm{IQR}_{\mathrm{profile}} = \left[\frac{1}{N_z} \sum_{z=1}^{N_z} (y_{z,75th} - y_{z,25th})^2 \right]_{\mathrm{profile}}^{\frac{1}{2}} \\ \overline{y_z} = \left[\frac{1}{M} \sum_{j=1}^{M} y_{zj} \right]_{\mathrm{profile}} \end{cases} \tag{C4}$$

where N_z is the number of vertical levels for each profile, and $\mathrm{IQR}_{\mathrm{profile}}$ is its interquartile range: $y_{z,25th}$ corresponds with the 25th percentile, and $y_{z,75th}$ corresponds with the 75th percentile.

Open Research

The data for all the analyses in the main text are available at `https://doi.org/10.5281/zenodo.10019987`. The emulator code is available at `https://github.com/yqsun91/WACCM-Emulation`. All the raw WACCM output data are available on request from authors.

Acknowledgments

We thank Andre Souza for insightful discussions. This work was supported by grants from the NSF OAC CSSI program (#2005123 , #2004512, #2004492, #2004572), and by the

generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program to PH, MJA, EG, and AS. PH is also supported by the Office of Naval Research (ONR) Young Investigator Award N00014-20-1-2722. SL is supported by the Office of Science, U.S. Department of Energy Biological and Environmental Research as part of the Regional and Global Climate Model Analysis program area. Computational resources were provided by NSF XSEDE (allocation ATM170020) and NCAR's CISL (allocation URIC0009).

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., ... Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243-297. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1566253521001081` doi: https://doi.org/10.1016/j.inffus.2021.05.008

Achatz, U. (2022). Gravity waves and their impact on the atmospheric flow. In *Atmospheric dynamics* (pp. 407–505). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from `https://doi.org/10.1007/978-3-662-63941-2_10` doi: 10.1007/978-3-662-63941-2_10

Alexander, M. J., Geller, M., McLandress, C., Polavarapu, S., Preusse, P., Sassi, F., ... Watanabe, S. (2010). Recent developments in gravity-wave effects in climatemodels and the global distribution of gravity-wavemomentum flux from observations and models. *Quarterly Journal of the Royal Meteorological Society*, *136*. doi: 10.1002/qj.637

Amiramjadi, M., Plougonven, R., Mohebalhojeh, A. R., & Mirzaei, M. (2022). Using machine learning to estimate non-orographic gravity wave characteristics at source levels. *Journal of the Atmospheric Sciences*. doi: 10.1175/JAS-D-22-0021.1

Ando, S., & Huang, C. Y. (2017). Deep over-sampling framework for classifying imbalanced data. In *Machine learning and knowledge discovery in databases: European conference, ecml pkdd 2017, skopje, macedonia, september 18–22, 2017, proceedings, part i 10* (pp. 770–785).

Bacmeister, J. T., Newman, P. A., Gary, B. L., & Chan, K. R. (1994). An algorithm for forecasting mountain wave-related turbulence in the stratosphere. *Weather Forecasting*, *9*. doi: 10.1175/1520-0434(1994)009⟨0241:AAFFMW⟩2.0.CO;2

Balaji, V. (2021). Climbing down charney's ladder: machine learning and the post-dennard era of computational climate science. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200085.

Baldi, P., Sadowski, P., & Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, *5*(1), 4308.

Ballnus, B., Hug, S., Hatz, K., Görlitz, L., Hasenauer, J., & Theis, F. J. (2017). Comprehensive benchmarking of markov chain monte carlo methods for dynamical systems. *BMC Systems Biology*, *11*(1), 1–18.

Barnes, E. A., Barnes, R. J., & DeMaria, M. (2023). Sinh-arcsinh-normal distributions to add uncertainty to neural network regression tasks: Applications to tropical cyclone

intensity forecasts. *Environmental Data Science*, *2*, e15.

Beck, A., Flad, D., & Munz, C.-D. (2019). Deep neural networks for data-driven LES closure models. *Journal of Computational Physics*, *398*, 108910.

Beljaars, A. C., Brown, A. R., & Wood, N. (2004). A new parametrization of turbulent orographic form drag. *Quarterly Journal of the Royal Meteorological Society*, *130*. doi: 10.1256/qj.03.73

Belochitski, A., & Krasnopolsky, V. (2021). Robustness of neural network emulations of radiative transfer parameterizations in a state-of-the-art general circulation model. *Geoscientific Model Development*, *14*(12), 7425–7437. Retrieved from `https://gmd.copernicus.org/articles/14/7425/2021/` doi: 10.5194/gmd-14-7425-2021

Beucler, T., Ebert-Uphoff, I., Rasp, S., Pritchard, M., & Gentine, P. (2021, 05). *Machine learning for clouds and climate (invited chapter for the agu geophysical monograph series "clouds and climate").* doi: 10.1002/essoar.10506925.1

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). *Weight uncertainty in neural networks.*

Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*(1), 376–399.

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. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/77/12/jas-d-20-0082.1.xml` doi: 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. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001711` doi: https://doi.org/10.1029/2019MS001711

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249-259. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0893608018302107` doi: https://doi.org/10.1016/j.neunet.2018.07.011

Bölöni, G., Kim, Y. H., Borchert, S., & Achatz, U. (2021). Toward transient subgrid-scale gravity wave representation in atmospheric models. part i: Propagation model including nondissipative wave mean-flow interactions. *Journal of the Atmospheric Sciences*, *78*. doi: 10.1175/JAS-D-20-0065.1

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I., & Palmer, T. (2021). Machine learning emulation of gravity wave drag in numerical weather forecasting. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002477. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002477` (e2021MS002477 2021MS002477) doi: https://doi.org/10.1029/2021MS002477

Chattopadhyay, A., Nabizadeh, E., & Hassanzadeh, P. (2020). Analog forecasting of extreme-causing weather patterns using deep learning. *Journal of Advances in Model-*

ing Earth Systems, *12*(2), e2019MS001958.

Chattopadhyay, A., Subel, A., & Hassanzadeh, P. (2020). Data-driven super-parameterization using deep learning: Experimentation with multiscale Lorenz 96 systems and transfer learning. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002084.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004, jun). Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, *6*(1), 1–6. Retrieved from `https://doi.org/10.1145/1007730.1007733` doi: 10.1145/1007730.1007733

Chen, N., & Majda, A. J. (2019). A new efficient parameter estimation algorithm for high-dimensional complex nonlinear turbulent dynamical systems with partial observations. *Journal of Computational Physics*, *397*, 108836.

Clare, M. C., Sonnewald, M., Lguensat, R., Deshayes, J., & Balaji, V. (2022). Explainable artificial intelligence for bayesian neural networks: Towards trustworthy predictions of ocean dynamics. *arXiv preprint arXiv:2205.00202*.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., & Searight, K. (2013). Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, *141*(10), 3498–3516.

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. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023GL104668` (e2023GL104668 2023GL104668) doi: https://doi.org/10.1029/2023GL104668

Espinosa, Z. I., Sheshadri, A., Cain, G. R., Gerber, E. P., & DallaSanta, K. J. (2022). Machine Learning Gravity Wave Parameterization Generalizes to Capture the QBO and Response to Increased CO_2. , *49*(8), e98174. doi: 10.1029/2022GL098174

Finkel, J., Gerber, E. P., Abbot, D. S., & Weare, J. (2023). Revealing the statistics of extreme events hidden in short weather forecast data. *AGU Advances*, *4*(2), e2023AV000881. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023AV000881` (e2023AV000881 2023AV000881) doi: https://doi.org/10.1029/2023AV000881

Foster, D., Gagne, D. J., & Whitt, D. B. (2021, 12). Probabilistic machine learning estimation of ocean mixed layer depth from dense satellite and sparse in situ observations. *Journal of Advances in Modeling Earth Systems*, *13*. doi: 10.1029/2021MS002474

Frezat, H., Sommer, J. L., Fablet, R., Balarac, G., & Lguensat, R. (2022). A posteriori learning for quasi-geostrophic turbulence parametrization. *arXiv preprint arXiv:2204.03911*. doi: 10.1029/2022MS003124

Gagne, 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.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing

model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

Garcia, R. R., Smith, A. K., Kinnison, D. E., Álvaro de la Cámara, & 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. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/74/1/jas-d-16-0104.1.xml` doi: https://doi.org/10.1175/JAS-D-16-0104.1

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.

Gettelman, A., Gagne, D. J., Chen, C. C., Christensen, M. W., Lebo, Z. J., Morrison, H., & Gantos, G. (2021). Machine learning the warm rain process. *Journal of Advances in Modeling Earth Systems*, *13*. doi: 10.1029/2020MS002268

Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., ... Randel, W. J. (2019, 12). The whole atmosphere community climate model version 6 (waccm6). *Journal of Geophysical Research: Atmospheres*, *124*, 12380-12403. doi: 10.1029/2019JD030943

Gordon, E. M., & Barnes, E. A. (2022, 8). Incorporating uncertainty into a regression neural network enables identification of decadal state-dependent predictability in cesm2. *Geophysical Research Letters*, *49*. doi: 10.1029/2022GL098635

Guan, Y., Chattopadhyay, A., Subel, A., & Hassanzadeh, P. (2022). Stable a posteriori LES of 2D turbulence using convolutional neural networks: Backscattering analysis and generalization to higher Re via transfer learning. *Journal of Computational Physics*, *458*, 111090.

Guan, Y., Subel, A., Chattopadhyay, A., & Hassanzadeh, P. (2023). Learning physics-constrained subgrid-scale closures in the small-data regime for stable and accurate les. *Physica D: Nonlinear Phenomena*, *443*, 133568. doi: https://doi.org/10.1016/j.physd.2022.133568

Guillaumin, A. P., & Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002534.

Hardiman, S. C., Scaife, A. A., Niekerk, A. v., Prudden, R., Owen, A., Adams, S. V., ... Madge, S. (2023). Machine learning for non-orographic gravity waves in a climate model. *Artificial Intelligence for the Earth Systems*.

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*, 1–58.

Hertzog, A., Alexander, M. J., & Plougonven, R. (2012). On the intermittency of gravity wave momentum flux in the stratosphere. *Journal of the Atmospheric Sciences*, *69*(11), 3433–3448.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., ... others (2017). The art and science of climate model tuning. *Bulletin of the American Meteoro-*

logical Society, *98*(3), 589–602.

Huang, C., Li, Y., Loy, C. C., & Tang, X. (2016). Learning deep representation for imbalanced classification. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5375–5384).

Iglesias-Suarez, F., Gentine, P., Solino-Fernandez, B., Beucler, T., Pritchard, M., Runge, J., & Eyring, V. (2023). *Causally-informed deep learning to improve climate models and projections.*

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*, 429-449. Retrieved from `https://doi.org/10.3233/IDA-2002-6504` (5) doi: 10.3233/IDA-2002-6504

Johnson, J. M., & Khoshgoftaar, T. M. (2019, Mar 19). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 27. Retrieved from `https://doi.org/10.1186/s40537-019-0192-5` doi: 10.1186/s40537-019-0192-5

Kim, Y., Bölöni, G., Borchert, S., Chun, H. Y., & Achatz, U. (2021). Toward transient subgrid-scale gravity wave representation in atmospheric models. part ii: Wave intermittency simulated with convective sources. *Journal of the Atmospheric Sciences*, *78*. doi: 10.1175/JAS-D-20-0066.1

Kim, Y., Eckermann, S. D., & Chun, H. (2003). An overview of the past, present and future of gravity-wave drag parametrization for numerical climate and weather prediction models. *Atmosphere-Ocean*, *41*, 65-98. Retrieved from `https://doi.org/10.3137/ao.410105` doi: 10.3137/ao.410105

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*.

Köhler, L., Green, B., & Stephan, C. C. (2023). Comparing loon superpressure balloon observations of gravity waves in the tropics with global storm-resolving models. *Journal of Geophysical Research: Atmospheres*, *128*(15), e2023JD038549.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Chalikov, D. V. (2005). New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Monthly Weather Review*, *133*(5), 1370 - 1383. Retrieved from `https://journals.ametsoc.org/view/journals/mwre/133/5/mwr2923.1.xml` doi: https://doi.org/10.1175/MWR2923.1

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., ... Courville, A. (2021). Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning* (pp. 5815–5826).

Kruse, C. G., Alexander, M. J., Hoffmann, L., Niekerk, A. V., Polichtchouk, I., Bacmeister, J. T., ... Stein, O. (2022). Observed and modeled mountain waves from the surface to the mesosphere near the drake passage. *Journal of the Atmospheric Sciences*, *79*. doi: 10.1175/JAS-D-21-0252.1

Li, X., Dai, Y., Ge, Y., Liu, J., Shan, Y., & Duan, L.-Y. (2022). Uncertainty modeling for out-of-distribution generalization. *arXiv preprint arXiv:2202.03958*.

Ling, J., & Templeton, J. (2015, 08). Evaluation of machine learning algorithms for prediction of regions of high Reynolds averaged Navier Stokes uncertainty. *Physics of*

Fluids, *27*(8). Retrieved from https://doi.org/10.1063/1.4927765 (085103) doi: 10.1063/1.4927765

Liu, Y., Racah, E., Prabhat, M., Correa, J., Khosrowshahi, A., Lavers, D., ... Collins, W. (2016, 05). Application of deep convolutional neural networks for detecting extreme weather in climate datasets.

Lopez-Gomez, I., McGovern, A., Agrawal, S., & Hickey, J. (2022). Global extreme heat forecasting using neural weather models. *Artificial Intelligence for the Earth Systems*, 1 - 41. Retrieved from https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-22-0035.1/AIES-D-22-0035.1.xml doi: 10.1175/AIES-D-22-0035.1

Lu, L., Jin, P., Pang, G., Zhang, Z., & Karniadakis, G. E. (2021). Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, *3*(3), 218–229.

Maalouf, M., & Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems*, *59*, 142–148.

Maalouf, M., & Trafalis, T. B. (2011). Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics & Data Analysis*, *55*(1), 168–183.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., & Wilson, A. G. (2019). A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, *32*.

Mamalakis, A., Barnes, E. A., & Ebert-Uphoff, I. (2022). Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artificial Intelligence for the Earth Systems*, *1*(4), e220012.

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. , *47*(19), e89436. doi: 10.1029/2020GL089436

Maulik, R., San, O., Rasheed, A., & Vedula, P. (2019). Subgrid modelling for two-dimensional turbulence using neural networks. *Journal of Fluid Mechanics*, *858*, 122–144.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*. doi: 10.1175/BAMS-D-18-0195.1

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., ... Schmidt, L. (2021). Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning* (pp. 7721–7735).

Miloshevich, G., Cozian, B., Abry, P., Borgnat, P., & Bouchet, F. (2023, Apr). Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data. *Phys. Rev. Fluids*, *8*, 040501. Retrieved from https://link.aps.org/doi/10.1103/PhysRevFluids.8.040501 doi: 10.1103/PhysRevFluids.8.040501

Nagarajan, V., Andreassen, A., & Neyshabur, B. (2020). Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*.

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. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001351 doi: https://doi.org/10.1029/2018MS001351

Oh, S. M., Rehg, J. M., Balch, T., & Dellaert, F. (2005). Data-driven mcmc for learning and inference in switching linear dynamic systems. In *Proceedings of the national conference on artificial intelligence* (Vol. 20, p. 944).

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, *32*.

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 preprint arXiv:2309.09024*.

Palmer, T. (2019). Stochastic weather and climate models. *Nature Reviews Physics*, *1*(7), 463–471.

Polichtchouk, I., Van Niekerk, A., & 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.

Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., ... others (2015). A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Reviews of geophysics*, *53*(2), 323–361.

Psaros, A. F., Meng, X., Zou, Z., Guo, L., & Karniadakis, G. E. (2023). Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics*, *477*, 111902.

Qi, D., & Majda, A. J. (2020). Using machine learning to predict extreme events in complex systems. *Proceedings of the National Academy of Sciences*, *117*(1), 52-59. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1917285117 doi: 10.1073/pnas.1917285117

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Rasp, S. (2020). Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: general algorithms and lorenz 96 case study (v1.0). *Geoscientific Model Development*, *13*(5), 2185–2196. Retrieved from https://gmd.copernicus.org/articles/13/2185/2020/ doi: 10.5194/gmd-13-2185-2020

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.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019, Feb 01). Deep learning and process understanding for data-driven earth system science. *Nature*, *566*(7743), 195-204. Retrieved from https://doi.org/10.1038/s41586-019-0912-1 doi: 10.1038/s41586-019-0912-1

Richter, J. H., Butchart, N., Kawatani, Y., Bushell, A. C., Holt, L., Serva, F., ... Yukimoto, S. (2022, 4). Response of the quasi-biennial oscillation to a warming climate in global climate models. *Quarterly Journal of the Royal Meteorological Society*, *148*, 1490-1518. doi: 10.1002/qj.3749

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*. doi: 10.1175/2009JAS3112.1

Ross, A. S., Li, Z., Perezhogin, P., Fernandez-Granda, C., & Zanna, L. (2022). Benchmarking of machine learning ocean subgrid parameterizations in an idealized model.

Sato, K., Watanabe, S., Kawatani, Y., Tomikawa, Y., Miyazaki, K., & Takahashi, M. (2009). On the origins of mesospheric gravity waves. *Geophysical Research Letters*, *36*. doi: 10.1029/2009GL039908

Schneider, T., Jeevanjee, N., & Socolow, R. (2021). Accelerating progress in climate science. *Physics Today*, *74*(6), 44–51.

Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, *44*. doi: 10.1002/2017GL076101

Scinocca, J. F., & McFarlane, N. A. (2000). The parametrization of drag induced by stratified flow over anisotropic orography. *Quarterly Journal of the Royal Meteorological Society*, *126*. doi: 10.1002/qj.49712656802

Seifert, A., & Rasp, S. (2020). Potential and limitations of machine learning for modeling warm-rain cloud microphysical processes. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002301. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002301` (e2020MS002301 10.1029/2020MS002301) doi: https://doi.org/10.1029/2020MS002301

Shamekh, S., Lamb, K. D., Huang, Y., & Gentine, P. (2023). Implicit learning of convective organization explains precipitation stochasticity. *Proceedings of the National Academy of Sciences*, *120*(20), e2216158120. Retrieved from `https://www.pnas.org/doi/abs/10.1073/pnas.2216158120` doi: 10.1073/pnas.2216158120

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., & Cui, P. (2021). Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Song, H.-J., & Roh, S. (2021). Improved weather forecasting using neural network emulation for radiation parameterization. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002609. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002609` (e2021MS002609 2021MS002609) doi: https://doi.org/10.1029/2021MS002609

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Stensrud, D. J. (2007). *Parameterization schemes: Keys to understanding numerical weather prediction models*. Cambridge University Press. Retrieved from `https://www.cambridge.org/core/product/identifier/9780511812590/type/book` doi: 10

.1017/CBO9780511812590

Subel, A., Chattopadhyay, A., Guan, Y., & Hassanzadeh, P. (2021). Data-driven subgrid-scale modeling of forced Burgers turbulence using deep learning with generalization to higher Reynolds numbers via transfer learning. *Physics of Fluids*, *33*(3), 031702.

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

Sun, Y., Hassanzadeh, P., Alexander, M., & Kruse, C. (2023, 12). Quantifying 3d gravity wave drag in a library of tropical convection-permitting simulations for data-driven parameterizations.
doi: 10.1029/2022MS003585

Sun, Y., Wong, A., & Kamel, M. S. (2009, 11). Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*, 687-719. doi: 10.1142/S0218001409007326

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. , 270–279.

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

Wu, D., Gao, L., Xiong, X., Chinazzi, M., Vespignani, A., Ma, Y.-A., & Yu, R. (2021). *Quantifying uncertainty in deep spatiotemporal forecasting.*

Wu, G., & Chang, E. Y. (2003). Adaptive feature-space conformal transformation for imbalanced-data learning. In *Proceedings of the twentieth international conference on international conference on machine learning* (p. 816–823). AAAI Press.

Ye, H., Xie, C., Cai, T., Li, R., Li, Z., & Wang, L. (2021). Towards a theoretical framework of out-of-distribution generalization. *Advances in Neural Information Processing Systems*, *34*, 23519–23531.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, *27*.

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), 1–10.

Zhang, C., Perezhogin, P., Gultekin, C., Adcroft, A., Fernandez-Granda, C., & Zanna, L. (2023). *Implementation and evaluation of a machine learned mesoscale eddy parameterization into a numerical ocean circulation model.*

Zhang, D., Ahuja, K., Xu, Y., Wang, Y., & Courville, A. (2021). Can subnetwork structure be the key to out-of-distribution generalization? In *International conference on machine learning* (pp. 12356–12367).

Zhu, Y., Zabaras, N., Koutsourelakis, P.-S., & Perdikaris, P. (2019). Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *Journal of Computational Physics*, *394*, 56–81.