JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2025MS004977

Offline Performance of a Nonlocal Deep Learning Parameterization for Climate Model Representation of Atmospheric Gravity Waves

Aman Gupta[1], **Aditi Sheshadri**[1], **Sujit Roy**[2,3], **and Valentine Anantharaj**[4]

[1]Department of Earth System Science, Stanford University, Stanford, CA, USA, [2]Earth System Science Center, The University of Alabama in Huntsville, Huntsville, AL, USA, [3]NASA Marshall Space Flight Center, Huntsville, AL, USA, [4]Oak Ridge National Laboratory, Oak Ridge, TN, USA

Key Points:

- Three machine learning schemes are trained on global, resolved wave fluxes, embedding three different levels of horizontal nonlocality
- The three neural networks accurately reproduce both time-averaged statistics and transient flux variability over prominent gravity wave hotspots
- Transfer learning on a 1.4 km climate model improves flux prediction and variability around the tropical quasi-biennial oscillation and Antarctic final warming

Supporting Information:

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

Correspondence to:

A. Gupta,
ag4680@stanford.edu

Citation:

Gupta, A., Sheshadri, A., Roy, S., & Anantharaj, V. (2025). Offline performance of a nonlocal deep learning parameterization for climate model representation of atmospheric gravity waves. *Journal of Advances in Modeling Earth Systems*, *17*, e2025MS004977. https://doi.org/10.1029/2025MS004977

Received 23 JAN 2025
Accepted 28 SEP 2025

Author Contributions:

Conceptualization: Aman Gupta, Aditi Sheshadri
Data curation: Aman Gupta, Valentine Anantharaj
Formal analysis: Aman Gupta, Valentine Anantharaj
Funding acquisition: Aditi Sheshadri, Valentine Anantharaj
Investigation: Aman Gupta
Methodology: Aman Gupta, Aditi Sheshadri, Sujit Roy
Project administration: Aditi Sheshadri

Abstract

Gravity waves (GWs) make crucial contributions to the middle atmospheric circulation. Yet, their climate model representation remains inaccurate, leading to key circulation biases. This study introduces a set of three neural networks (NNs) that learn to predict GW fluxes (GWFs) from multiple years of high-resolution ERA5 reanalysis. The three NNs: a 1 × 1 ANN, a 3 × 3 ANN-CNN, and an Attention UNet embed different levels of horizontal nonlocality in their architecture and are capable of representing nonlocal GW effects that are missing from current operational GW parameterizations. The NNs are evaluated offline on both time-averaged statistics and time-evolving flux variability. All NNs, especially the Attention UNet, accurately recreate the global GWF distribution in both the troposphere and the stratosphere. Moreover, the Attention UNet most skillfully predicts the transient evolution of GWFs over prominent orographic and nonorographic hotspots, with the 1 × 1 model being a close second. Since even ERA5 does not resolve a substantial portion of GWFs, this deficiency is compensated by subsequently applying transfer learning on the ERA5-trained ML models for GWFs from a 1.4 km global climate model. It is found that the re-trained models both (a) preserve their learning from ERA5, and (b) learn to appropriately scale the predicted fluxes to account for ERA5's limited resolution. Our results highlight the importance of embedding nonlocal information for a more accurate GWF prediction and establish strategies to complement abundant reanalysis data with limited high-resolution data to develop machine learning-driven parameterizations for missing mesoscale processes in climate models.

Plain Language Summary

Gravity waves (GWs) are ubiquitous atmospheric oscillations generated by flow disturbances in the atmosphere. Since they operate on smaller scales than a climate model can resolve, their effects are mostly unresolved in coarse-resolution climate models. So, climate models typically parameterize/approximate their effects, but these parameterizations can often be oversimplified, leading to physical inaccuracies in models. We propose a set of three fully machine learning (ML)-based parameterizations whose architectures are chosen to capture both horizontal and vertical GW propagation: single column, multiple columns, and globally nonlocal, to learn GW effects from data. Following training on multiple (four) years of modern reanalysis, these purely data-driven schemes generate accurate flux statistics, time evolution, and variability. The globally nonlocal ML model offers the best prediction, indicating the importance of nonlocality for data-driven GW schemes. We subsequently re-train the models on 4 months of a 1.4 km climate model and find that iteratively training on high-volume, low-resolution reanalysis and low-volume, high-resolution climate model output allows the model to learn GW effects from both data sets effectively. Our results establish the capability of ML-based schemes to learn essential GW physics from a mix of data, to represent these missing effects in climate models, and improve their prediction.

1. Introduction

Atmospheric gravity waves (GWs) are ubiquitous multiscale oscillations generated by a myriad of atmospheric disturbances, including strong convective storms, flow over mountains, storm tracks, fronts, etc. They manifest over spatial scales ranging from $\mathcal{O}(1)$ km to $\mathcal{O}(1{,}000)$ km and evolve over timescales ranging from ~5 min (high-frequency oscillations) to over a day (near-inertial oscillations).

GWs dynamically couple the different layers of the atmosphere and are among the key drivers of the meridional overturning circulation in the middle atmosphere (Fritts & Alexander, 2003). They provide an important contribution to the driving of the tropical quasi-biennial oscillation (QBO) (Giorgetta et al., 2002). They influence

Resources: Aditi Sheshadri, Sujit Roy, Valentine Anantharaj
Software: Aman Gupta, Sujit Roy, Valentine Anantharaj
Supervision: Aditi Sheshadri
Validation: Aman Gupta, Aditi Sheshadri, Sujit Roy
Visualization: Aman Gupta, Sujit Roy
Writing – original draft: Aman Gupta
Writing – review & editing: Aman Gupta, Aditi Sheshadri, Sujit Roy, Valentine Anantharaj

the springtime breakdown of the Antarctic polar vortex (Gupta et al., 2021), and, in turn, Antarctic surface temperatures (Choi et al., 2024). They also potentially contribute to rapid breakdowns of the wintertime polar vortex, that is, sudden warmings (Albers & Birner, 2014; Song et al., 2020), eventually influencing tropospheric storm tracks (Domeisen & Butler, 2020; Kidston et al., 2015). In the mesosphere, they play a critical role in driving the pole-to-pole mesospheric circulation and in maintaining the observed equator-to-pole temperature gradient (Becker, 2012; Holton, 1982).

Climate models, such as those used in CMIP6 ScenarioMIP, typically have a grid resolution of 100–200 km, and thus fail to explicitly capture the effects of most of the GW spectrum. As a result, these effects are approximated using *numerical parameterizations* (Achatz et al., 2024; Kim et al., 2003).

Depending on the source, GW parameterizations can be broadly classified as orographic (e.g., Garner, 2005; Lott & Miller, 1997; van Niekerk & Vosper, 2021) or nonorographic (e.g., Alexander & Dunkerton, 1999; Garcia et al., 2017; Hines, 1997; Lott & Guez, 2013; Orr et al., 2010). Similar to parameterizations for other atmosphere-ocean processes, these GW parameterizations are often developed as single-column reduced-order analytical models and are, therefore, subject to various simplifying assumptions. For instance, all single-column GW parameterizations assume purely vertical propagation and steady-state GW dissipation. Additionally, orographic parameterizations often assume idealized ellipsoidal subgrid-scale topography. Likewise, nonorographic parameterizations—in the absence of sufficient observations and source/sink information—assume a highly idealized source spectrum of monochromatic GWs. As argued by McLandress et al. (2012), such total neglect of horizontal GW propagation is the leading hypothesis for the occurrence of the prominent "cold-pole" bias (delays in seasonal wind and temperature transitions) associated with unrealistically low springtime ozone concentration over the Antarctic. Practically all operational GW parameterizations neglect these observed GW properties (Plougonven et al., 2020).

Developing GW parameterizations that capture the effects of lateral propagation and transient evolution is challenging. To date, only a few of such parameterizations exist (Amemiya & Sato, 2016; Eichinger et al., 2023; Voelker et al., 2023). However, none of them are operational yet due to their own set of limitations. To this end, machine learning (ML)-based, a.k.a. data-driven parameterizations, present a fast, promising approach to improving the climate model representation of GW effects. An ML scheme can learn GWF evolution (generation, propagation, dissipation) directly from data (empirically) without relying on analytical models. This approach has been applied to develop data-driven schemes for atmosphere-ocean processes ranging from precipitation, turbulence, and radiation, to ocean heat transport, ice sheet modeling, and vegetation (see Eyring et al., 2024; Mansfield et al., 2023 for a review).

The data-driven approach is also being increasingly used to develop fast GWF emulators for numerical weather prediction models and climate models of varying complexity (Chantry et al., 2021; Connelly & Gerber, 2024; Espinosa et al., 2022; Hardiman et al., 2023; Lu et al., 2024; Sun et al., 2024; Ukkonen & Chantry, 2024). While quite effective, a major limitation of these works lies in the fact that all these ML models have been trained on parameterized GWFs, implying that these emulators learn the biases and assumptions of the underlying parameterizations used to train them. As a result, the emulators offer little to improve the physics aspects of GW representation in climate models. In this work, we move beyond parameterized training data and use modern advances in deep learning to develop ML models (NNs or NNs) that learn GW evolution from *resolved* GWFs. Training on resolved GWFs allows the NNs to learn key physical properties of GWs directly from multiple GW-resolving data sets.

We present a set of three NNs that learn to predict GWF for a given background atmospheric state using three different degrees of horizontal nonlocality: a single-column artificial NN motivated by single-column parameterization design, a multiple-column artificial NN inspired by Wang et al. (2022), and a globally nonlocal Attention UNet NN. The models are first trained on multiple years of high-resolution global reanalysis, which partly resolves the mesoscale spectrum of GWs, and then re-trained on GWF from merely months of a kilometer-scale global model, which resolves a greater part of the spectrum of GWs. This allows the model to learn from a mix of high-volume, low-fidelity data and low-volume, high-fidelity data and use the blend to provide accurate predictions of GWFs. A glimpse of the offline performance of our scheme is shown in Figure 1, which illustrates our Attention UNet NN's considerable skill in predicting GW excitation from multiple sources scattered across the Southern Ocean. While not apparent in this figure, the NNs also represent lateral propagation of the GWFs

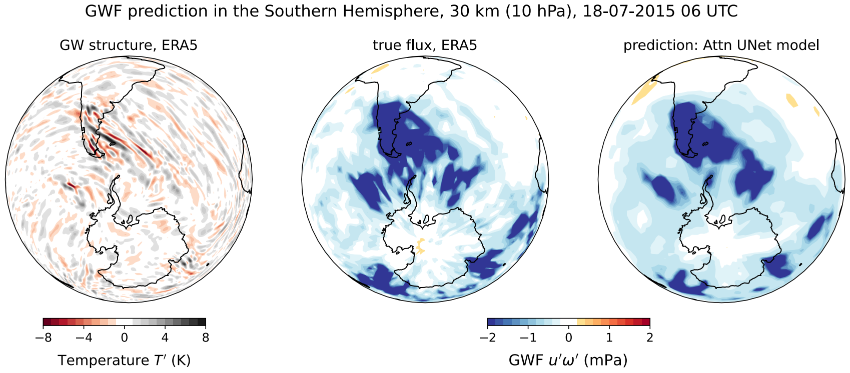
GWF prediction in the Southern Hemisphere, 30 km (10 hPa), 18-07-2015 06 UTC
GW structure, ERA5
true flux, ERA5
prediction: Attn UNet model
−8 −4 0 4 8
Temperature T' (K)
−2 −1 0 1 2
GWF $u'\omega'$ (mPa)

Figure 1. (left) Temperature perturbations (in K) associated with gravity waves (GWs) over the Drake Passage and the Southern Ocean on 18 July 2015 06 UTC, as resolved in ERA5, (middle) the momentum flux $u'\omega'$ (units mPa) associated with the excited GWs, and (right) the momentum flux predicted using an Attention UNet convolutional neural network trained on 3 years of ERA5 data.

well. Moreover, the NN predicted fluxes exhibit temporal coherence despite no explicit embedding of the temporal structure (recurrence) in the ML architecture.

The paper is structured as follows: Section 2 outlines the procedure to compute GWF from GW-resolving reanalysis and climate model output using Helmholtz Decomposition. The experimental setup, including the three different kinds of NNs and the TL approach to learn from multiple data sets, is described in Section 3. Following that, a complete description of the ML training data and the recipe to scale it prior to ML training is provided in Section 4. Section 5 presents the key results. Finally, a summary of the key findings along with a discussion of the next steps, related to coupling the neural nets to a climate model, is provided in Section 6.

2. Computing Resolved Gravity Wave Momentum Fluxes

Per the EP-Flux theory, the momentum fluxes associated with GWs can be estimated using wind and temperature covariances. The covariances $u'\omega'$ and $v'\omega'$ respectively represent the vertical flux of zonal and meridional momentum. Here, u', v', and ω' are the small-scale wind perturbations associated with GWs. The vertical derivative of the covariances equates to the total acceleration/deceleration of the zonal and meridional flow provided by these waves upon dissipation. Since climate models cannot resolve the small-scale perturbations, model parameterizations focus on emulating these covariances and their derivatives to represent the missing GW effects (excitation, propagation, and dissipation) in the atmosphere.

Multiple approaches exist to retrieve the small-scale GW perturbations from high-resolution data. Inspired by recent studies that compute GW fluxes from observations and high-resolution climate models (Köhler et al., 2023; Lindborg, 2015), we use Helmholtz decomposition (hereafter HD) to extract the covariances from global, high-resolution wind output.

HD is often used in fluid dynamics research to decompose the horizontal flow into purely rotational and purely divergent flow components. The rotational part is associated with the large-scale balanced flow, whereas the divergent part, which typically has a finer-scale structure, is associated with small-scale GWs. Mathematically, the decomposition can be expressed as:

$$\vec{u} = (u, v) = -\nabla\phi + \nabla\times\psi \tag{1}$$

where (u, v) is the full horizontal flow, ϕ is the potential function such that $\nabla\phi$ is irrotational, that is, the curl of $\nabla\phi$ is 0. Similarly, ψ is the rotational streamfunction such that $\nabla \times \psi$ is non-divergent, that is, the divergence of $\nabla \times \psi$ is zero. Thus, HD provides ϕ and ψ, which, following an application of inverse spectral transforms, yield the divergent and rotational parts of the horizontal flow as:

$$\vec{u} = (u, v) \xrightarrow{HD} (u_{div}, v_{div}) + (u_{rot}, v_{rot}) \tag{2}$$

Next, to ensure that the large-scale background is completely removed from the divergent flow, an additional fixed-wavenumber high-pass filter is applied by removing the T21 truncated divergent velocity, $(u_{div,T21}, v_{div,T21})$, from the divergent flow. This operation is expressed as:

$$\left(u'_{div}, v'_{div}\right) = \left(u_{div} - u_{div,T21}, v_{div} - v_{div,T21}\right) \tag{3}$$

Finally, the high-pass filtered divergent velocities were used to compute the fluxes by multiplying with the eddy vertical velocity (ω'). ω' was obtained by removing the zonal mean component $\overline{\omega}$ of the full velocity (ω), that is $\omega' = \omega - \overline{\omega}$. The zonal and meridional components of the vertical momentum flux were then computed as:

$$\vec{F} = (F_x, F_y) = g^{-1}\left(u'_{div}\omega', v'_{div}\omega'\right) \tag{4}$$

Here, $g = -9.81$ m/s^2 is the acceleration due to gravity. We refer the readers to Köhler et al. (2023) and Gupta, Sheshadri, and Anantharaj (2024) for more technical details on the momentum flux computations.

Comparing differences among GW flux distributions: To compare histograms of GW fluxes predicted by different neural nets, we use the Hellinger distance, which measures the distances between two probability distributions.

First proposed by Hellinger (1909), the Hellinger distance $(\mathcal{H})$ between two probability densities p and q is a measure of their statistical distance and is defined as:

$$\mathcal{H}(p,q) = \frac{1}{2}\int_{x\in X}\left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx = 1 - \int_{x\in X}\sqrt{p(x)q(x)}dx. \tag{5}$$

By definition, $\mathcal{H} \in [0,1]$. A Hellinger distance of 0 means the distributions are identical almost everywhere, while a Hellinger distance of 1 implies the distributions are disjoint, that is, p is non-zero wherever q is zero, and vice versa.

3. Deep Learning Model Description

To learn the nonlocal horizontal propagation of atmospheric GWs and to contrast it with the traditional single-column parameterization approach, we create three different ML architectures or NNs (or NNs in short) that employ three varying degrees of horizontal nonlocality. A schematic outlining the three nonlocal architectures is shown in Figure 2.

3.1. The Three Neural Network (NN) Architectures

M1: A single-column Artificial Neural Network (ANN)-based ML model that takes a single column of input conditions to predict the GWF within that single column—hereafter referred to as M1 or simply 1 × 1. The model comprises one input layer, followed by 6 hidden layers, and then one output layer, which outputs the fluxes $u'\omega'$ and $v'\omega'$.

M2: A multiple-column ANN with a convolutional layer with a 3 × 3 filter preceding the ANN, which takes the input background state of the atmosphere over nine neighboring vertical columns to predict the GWF within one central column. The 3 × 3 filter reduces the input from nine columns into a single column. This design choice is based on the nonlocal parameterization setup proposed in Wang et al. (2022), with the main difference being the additional 3 × 3 convolutional layer. This NN is hereafter interchangeably referred to as M2 or simply 3 × 3.

M3: A globally nonlocal Attention U-Net that takes the full three-dimensional background state of the atmosphere as input and predicts the fluxes over the whole domain. The architecture is adopted from Oktay et al. (2018) and is hereafter referred to as M3 or UNet. The NN comprises four downsampling blocks and a bottleneck, followed by four upsampling blocks. A skip connection and a learnable attention gate are added between the downsampling and upsampling blocks at each level. The attention gate allows the UNet to

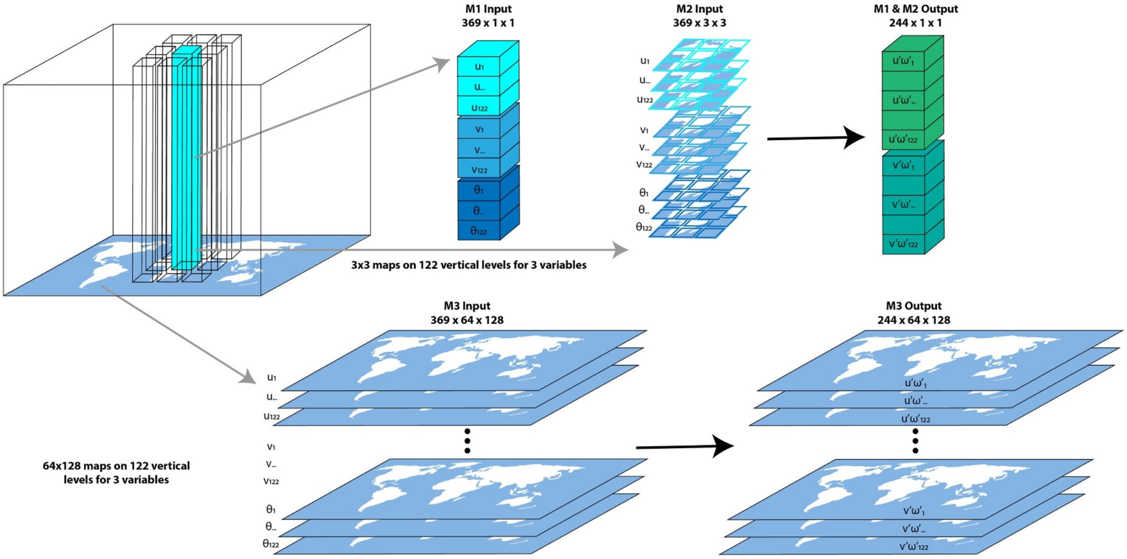

Figure 2. A schematic of the three machine learning architectures used in the study. Inspired by the three-dimensional propagation of atmospheric gravity waves, we define three neural net architectures with different extents of horizontal nonlocality: a (local) single-column artificial neural network M1, a (locally nonlocal) 3 × 3 columns artificial neural network (NN) M2 with one preceding convolutional layer, and a (globally nonlocal) Attention UNet convolutional NN M3. As shown, M1 and M2 respectively take 1 and 9 columns of input data and predict GWFs $u'\omega'$ and $v'\omega'$ for a single vertical column, while M3 takes input over the full spatial domain to predict the fluxes over the full spatial domain. The NN architectures are shown in more detail in Figure 3.

localize the parts of the domain that contain information critical to making accurate predictions. Finally, a convolutional layer at the end reshapes and projects the output to the appropriate number of output channels.

The model architectures are illustrated in detail in Figure 3.

3.2. Runs

For each of the three NN architectures described above, we train two different networks. The two networks differ in terms of their prediction domain in the vertical: global (i.e., troposphere + stratosphere) and stratosphere. A summary of the runs is also provided in Table 1.

- *Global*: networks which use the scaled dynamical variables over all levels in the troposphere and the stratosphere as input to predict the fluxes at all levels.
- *Stratosphere*: networks which use the scaled dynamical variables for all levels in the troposphere and the stratosphere as input, but predict the fluxes only for the 60 levels in the stratosphere (1 to 200 hPa).

Since the tropospheric small-scale fluxes can have some unwanted divergent contributions from strong convective fluxes and not GWs, predicting fluxes only in the stratosphere allows us to eliminate this source of uncertainty.

3.3. Model Hyperparameters

For brevity, the model hyperparameters are outlined in the Appendix in Table S1 in Supporting Information S1.

The respective ML model hyperparameters for all the runs, such as the number of hidden layers, optimizers, learning rates, etc., remain unchanged. For M3, only the input and output layers' dimensions change, and for M1 and M2, both the input and output dimensions and the hidden layers' widths change.

For the most discussed NN configuration in this study, that is, the stratosphere only NN with input features $\{u, v, \theta, \omega\}$, M1 has approximately 17 million, M2 has 19 million, and M3 has 38 million learnable parameters, respectively.

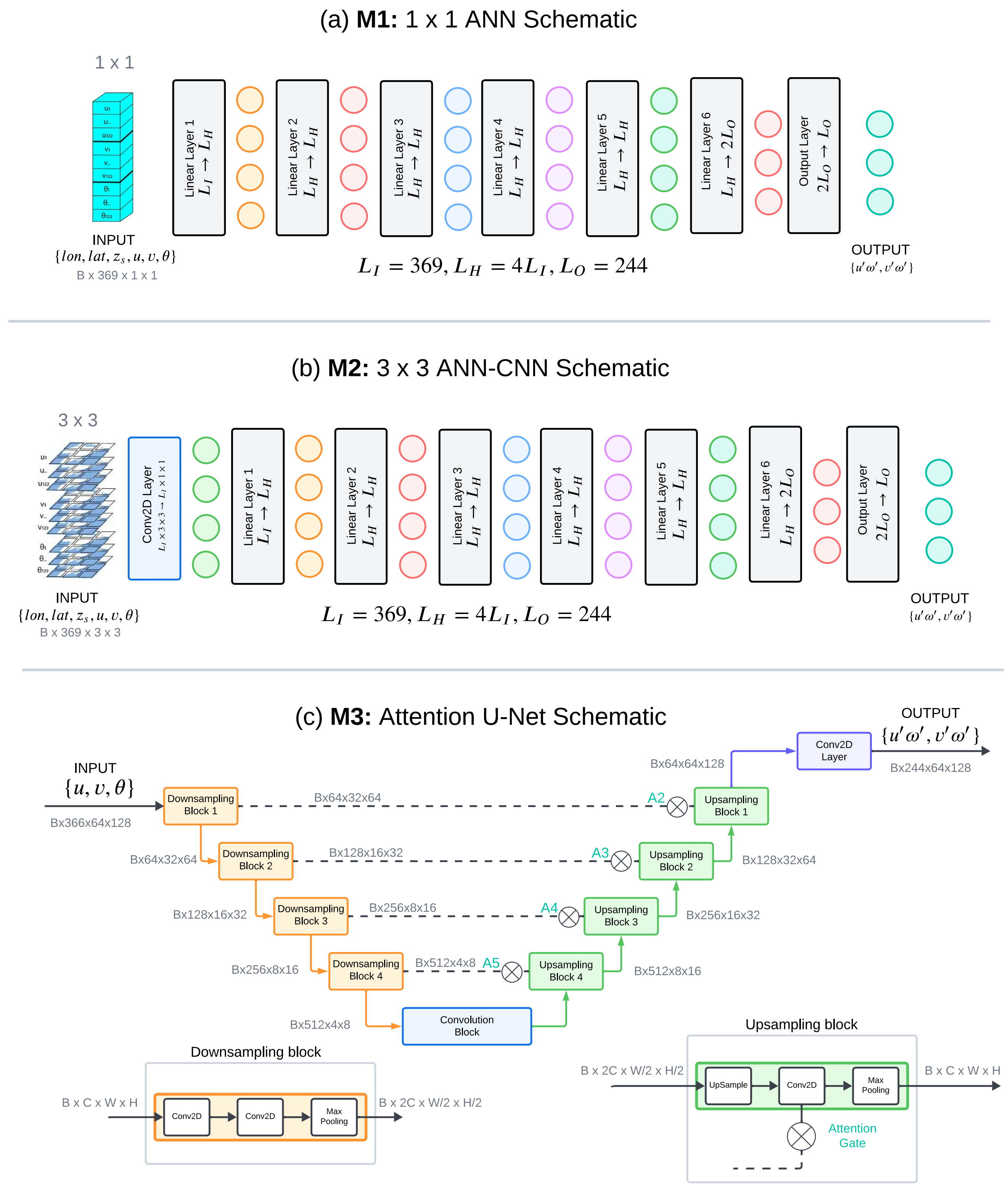
(a) M1: 1 x 1 ANN Schematic
1 x 1
INPUT
$\{lon, lat, z_s, u, v, \theta\}$
B x 369 x 1 x 1
Linear Layer 1 $L_I \rightarrow L_H$
Linear Layer 2 $L_H \rightarrow L_H$
Linear Layer 3 $L_H \rightarrow L_H$
Linear Layer 4 $L_H \rightarrow L_H$
Linear Layer 5 $L_H \rightarrow L_H$
Linear Layer 6 $L_H \rightarrow 2L_O$
Output Layer $2L_O \rightarrow L_O$
$L_I = 369, L_H = 4L_I, L_O = 244$
OUTPUT
$\{u'\omega', v'\omega'\}$
(b) M2: 3 x 3 ANN-CNN Schematic
3 x 3
Conv2D Layer $L_I \times 3 \times 3 \rightarrow L_I \times 1 \times 1$
INPUT
$\{lon, lat, z_s, u, v, \theta\}$
B x 369 x 3 x 3
Linear Layer 1 $L_I \rightarrow L_H$
Linear Layer 2 $L_H \rightarrow L_H$
Linear Layer 3 $L_H \rightarrow L_H$
Linear Layer 4 $L_H \rightarrow L_H$
Linear Layer 5 $L_H \rightarrow L_H$
Linear Layer 6 $L_H \rightarrow 2L_O$
Output Layer $2L_O \rightarrow L_O$
$L_I = 369, L_H = 4L_I, L_O = 244$
OUTPUT
$\{u'\omega', v'\omega'\}$
(c) M3: Attention U-Net Schematic
INPUT
$\{u, v, \theta\}$
Bx366x64x128
Downsampling Block 1
Bx64x32x64
Downsampling Block 2
Bx128x16x32
Downsampling Block 3
Bx256x8x16
Downsampling Block 4
Bx512x4x8
Convolution Block
Bx64x32x64
Bx128x16x32
Bx256x8x16
Bx512x4x8
A2
A3
A4
A5
Upsampling Block 1
Upsampling Block 2
Upsampling Block 3
Upsampling Block 4
Bx512x8x16
Bx256x16x32
Bx128x32x64
Bx64x64x128
Conv2D Layer
OUTPUT
$\{u'\omega', v'\omega'\}$
Bx244x64x128
Downsampling block
B x C x W x H
Conv2D
Conv2D
Max Pooling
B x 2C x W/2 x H/2
Upsampling block
B x 2C x W/2 x H/2
UpSample
Conv2D
Max Pooling
B x C x W x H
Attention Gate

Figure 3.

3.4. Transfer Learning

At 25 km resolution, ERA5 resolves only a portion of the mesoscale GW spectrum. Recent studies have shown that even a 9 km global model might only resolve roughly half of the actual GW forcing in the midlatitudes as compared to a 1 km model (Polichtchouk et al., 2023). This means that ERA5 either underestimates the amplitude of some resolved wave packets or does not resolve finer-scale components of some wave packets at all. As measured by Gupta, Reichert, et al. (2024), the fluxes in ERA5 could even be a factor of 2–2.5 weaker than in a full mesoscale-resolving NWP model. Ideally, a straightforward way to address this shortcoming would be to train the ML models on resolved GWFs from higher-resolution NWP and climate models. However, due to a lack of high-volume, high-resolution global climate model output/data, this approach is currently infeasible. Even prominent high-resolution climate modeling projects, including the DYAMOND (Stevens et al., 2019), GEOS (Gelaro, 2015), and IFS (Wedi et al., 2020), publicly provide not more than several months of high-frequency high-resolution output at best.

To alleviate this issue, we apply TL principles to train the NNs on both multiple years of low-resolution ERA5 data and limited high-fidelity data from a global kilometer-scale climate model, which resolves most of the GW spectrum. This is carried out in two steps, as illustrated in Figure 4.

First, all the NNs are trained on 3 years of (relatively) high-volume, low-fidelity ERA5 data. Then, *part of* the NNs are re-trained on low-volume high-fidelity IFS-1 km data. The data sets are described in detail in the following section. For M1 and M2, only the final two layers, that is, Linear Layer 6 and the Output Layer, are re-trained for TL. For M3, only the Upsampling Block 1 and the Conv2D Layer are re-trained for TL. All the other model weights and biases are frozen during TL. We also ran an ablation/sensitivity study by freezing the last two, last three, and last four layers of M1 and M3 (Linear Layers 4, 5, and 6, and the Output Layer for M1, and Upsampling Blocks 3, 2, and 1, and Conv2D layer for M3). To assess the improvements due to TL, all NNs were evaluated on both ERA5 and IFS both after regular training (Step 1), and TL (Step 2). Identical hyperparameters were used for both Step 1 and Step 2 trainings, and in Step 2, all the models were trained for 200 epochs. For more details and background on TL, see F. Zhuang et al. (2020) and references therein.

To summarize, each NN in this study is characterized by its:

1. ML Architecture: what architecture the NN uses—[1 × 1, 3 × 3, UNet]
2. Feature set: what feature set the NN uses as input—[$\{u, v, w\}, \{u, v, \theta\}, \{u, v, w, \theta\}$]
3. Vertical domain: the vertical extent domain for the input-output pairs—[global, stratosphere]
4. Training procedure: whether the NN was trained on just ERA5 data, or it was re-trained on IFS Fluxes as well.

4. Preparing ML Training Data

The ML training data used in this work were sourced from a combination of modern reanalysis (ERA5) and ultra-high-resolution climate model outputs (IFS-1 km), both from the European Center for Medium-Range Weather Forecasts (ECMWF).

ERA5 Reanalysis: Most of the training data was computed using the publicly available hourly reanalysis, ECMWF's ERA5 (Hersbach et al., 2020). ERA5 is originally produced by assimilating observations with a forecast model that uses 639 spherical harmonics ($\sim 0.3° \times 0.3°$) in the horizontal at 137 hybrid σ-p levels in the vertical, ranging from the surface to 0.01 hPa. Among all current publicly available reanalysis products, ERA5 offers the finest resolution. Accounting for grid-scale hyperdiffusion and other numerical effects, ERA5 still resolves GWs with wavelengths 200 km and longer.

To numerically "absorb" vertically propagating GWs, stratospheric and mesospheric sponges are applied at pressures less than 10 hPa and 1 hPa, respectively. To prevent the NNs from learning the strong artificial damping

Figure 3. A schematic for the three neural networks (M1–M3) introduced in Figure 2 and described in Section 3. (a, b) M1 and M2 both have 1 input layer, 6 hidden layers, and 1 output layer. Since the input to M2 is a 3 × 3 image, its input layer is a 3 × 3 learnable convolutional layer that transforms the 3 × 3 column data into a single column. The input and output dimensions vary based on the vertical domain. For the feature set shown here ($lat, lon, z_s, u, v, \theta, \omega$), the input dimension is $491 = 3 \times 1 + 4 \times 122$, and the output dimension is $244 = 2 \times 122$. B is the minibatch size. M3 is inspired from Wang et al. (2022). (c) For M3, the input dimension is 491 vertical channels × 64 latitudes × 128 longitudes, while the output dimensions are 244 vertical channels × 64 latitudes × 128 longitudes. The Attention gate, not shown here, is identical to the one defined in Oktay et al. (2018). The PyTorch implementation of the models is provided at: https://doi.org/10.5281/zenodo.16415113.

Table 1
The Six Models Were Trained for Different Feature Sets: $\{u, v, \theta\}$, $\{u, v, \omega\}$, and $\{u, v, \theta, \omega\}$

Neural net type	Prediction domain in the vertical
M1: 1 × 1 ANN	Troposphere + Stratosphere (surface to 1.5 hPa)
M2: 3 × 3 ANN-CNN	Troposphere + Stratosphere (surface to 1.5 hPa)
M3: Attention-UNet	Troposphere + Stratosphere (surface to 1.5 hPa)
M1: 1 × 1 ANN	Stratosphere only (200–1.5 hPa)
M2: 3 × 3 ANN-CNN	Stratosphere only (200–1.5 hPa)
M3: Attention-UNet	Stratosphere only (200–1.5 hPa)

Note. Each of the models were also re-trained on IFS-1 km data using TL. For the 1 × 1 and 3 × 3 neural networks, positional encodings $\{\text{lat}, \text{lon}, z_s\}$ were also used as input features to embed spatial information.

in the mesosphere and above, we discard the topmost 15 levels and only use model levels 16 (1.5 hPa ∼ 45 km) to 137 (surface), that is, 122 levels, for training purposes.

IFS-1 km: The high-fidelity training data for TL experiments (as described in the next section) was created using the 1.4-km experimental nature runs performed using ECMWF's IFS model (Wedi et al., 2020). The hydrostatic model simulates global atmospheric evolution at an unprecedented horizontal resolution of ∼1 km for November-February Boreal Winter 2018–2019. It does so by employing a total of 8,000 spherical harmonics to solve the primitive equations of fluid flow. The grid resolution is at least a factor of 2 higher than any existing high-resolution simulations conducted to study GWs, and provides a glimpse into global GW activity and fluxes in unprecedented detail.

Numerically, the model has a design similar to ERA5 but it is a free-running model tuned to provide forecasts at a 1.4 km resolution without using any explicit GW parameterizations. Grid-scale hyperdiffusion and other numerical method choices in the model reduce its effective grid resolution from $\Delta x = 1.4$ km to about $6\Delta x$–$8\Delta x$ (Klaver et al., 2020; Skamarock, 2004). Thus, the model resolves the complete mesoscale GW spectrum (wavelengths ≥ 10 km). We use 3-hourly instantaneous fields on model levels to calculate the small-scale momentum flux due to resolved GWs. The model configuration is described in detail in Polichtchouk et al. (2022, 2023) and the procedure to compute the fluxes used in this study is described in Gupta, Sheshadri, and Anantharaj (2024).

The same set of input and output features are computed from both ERA5 and IFS-1 km:

Input features: We extract horizontal and vertical winds, temperature, pressure, and potential temperature from the data sets. The features extracted from the high-resolution data sets are conservatively coarse-grained to a coarse 2.8° ≈ 280 km Gaussian grid, which is the typical resolution at which climate models resolve these quantities. The coarse-graining was accomplished using a first-order conservative regridding function using Python's xESMF library (J. Zhuang et al., 2024).

We explore three different feature combinations, viz., $\{u, v, \theta\}$, $\{u, v, \omega\}$, and $\{u, v, \omega, \theta\}$ to train the NNs. This allows for assessing the relative importance of different input features toward flux prediction. For each feature set for ANNs, positional variables, including latitude, longitude, and surface geopotential, were also appended to the

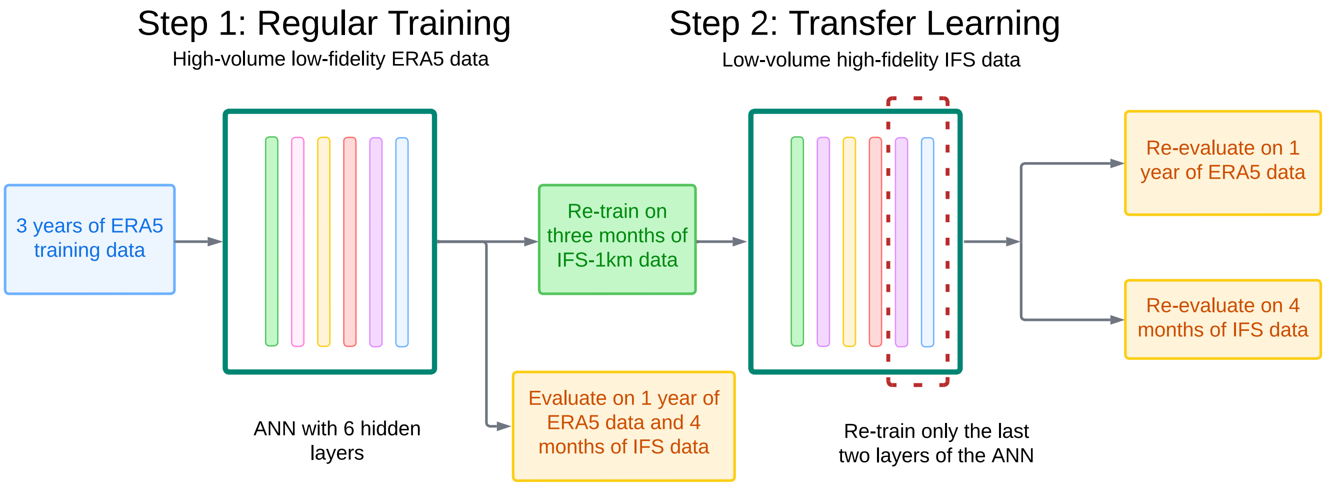
Step 1: Regular Training
High-volume low-fidelity ERA5 data
Step 2: Transfer Learning
Low-volume high-fidelity IFS data
3 years of ERA5 training data
ANN with 6 hidden layers
Re-train on three months of IFS-1km data
Evaluate on 1 year of ERA5 data and 4 months of IFS data
Re-train only the last two layers of the ANN
Re-evaluate on 1 year of ERA5 data
Re-evaluate on 4 months of IFS data

Figure 4. Schematic for the Transfer Learning (TL) experiment. The three types of neural nets M1–M3, which were pre-trained on 3 years of ERA5 data, were re-trained on 4 months of IFS-1 km data. Only the final couple of layers of each model were trained. For M1 and M2, the Linear Layer 6's and Output Layer's weights and biases were re-trained; all other weights were frozen. For M3, the Upsampling Block 1 and the final Conv2D Layer were re-trained; all other weights were frozen. Following both regular training (Step 1) and TL (Step 2), the three models were tested on ERA5 2015 data and the IFS-1 km data.

input variables. We note that positional variables were not included as inputs to the globally nonlocal Attention UNet.

Output features: The momentum fluxes $u'\omega'$ and $v'\omega'$ form the output of our ML models. For a given set of background dynamic variables, the ML models predict the momentum fluxes associated with the state. Similar to the input variables, the fluxes computed from the high-resolution data set are conservatively coarse-grained to a ≈280 km grid. This ensures that the fluxes are averaged over the largest resolved wavenumbers associated with GWs.

Data scaling: Proper scaling of the input-output data is crucial to ensuring ML model stability and efficient training. Since different physical variables have different magnitudes, a lack of proper scaling can lead to imprecise or spurious weight updates during backpropagation. The use of nonlinear activation functions would worsen the issue as large values, for instance, near-surface pressure in Pascals, would tend to get more importance during model weight updates than, say, near-surface winds. We use different scaling parameters for different variables, based on their underlying distributions, to ensure that a bulk of the input and output training data have values $\in [-2, 2]$ and $[-6, 6]$ respectively. We use the following scaling:

(a) *Positional variables*: longitude (λ) and latitude (ϕ) were normalized by dividing by 360° and 90° respectively. The surface geopotential (z_s) was scaled by 5×10^4 $\mathrm{m^2/s^2}$.
(b) *Horizontal winds u and v*: were normalized using their 4-year global mean (μ) and global standard deviation (σ) computed from ERA5 as $X \rightarrow \left(X - \mu_X\right)/3\sigma_X$, X being the respective variable. The mean and standard deviation were computed only for data between the surface and 50 km height.
(c) *Potential temperature θ*: ranges from ∼300 K near the surface to ∼2,000 K near the stratopause. Thus, θ was normalized by dividing by 1,000 K.
(d) *Vertical wind ω and momentum fluxes $u'\omega'$, and $v'\omega'$*: have more fine-scale variations and exhibit a Laplace distribution centered around 0. To scale them, the quantities were first scaled using the mean and standard deviation following which a cube root was applied, that is, $X \rightarrow \left(\left(X - \mu_X\right)/\sigma_X\right)^{1/3}$, X being the respective variable. Applying the cube root moves both close-to-zero and large values toward 1.

For the 4 years of ERA5 considered, $\mu_u = 6.3954$ m/s, $\sigma_u = 22.1755$ m/s, $\mu_v = 0.0203$ m/s, $\sigma_v = 9.8414$ m/s, $\mu_{u'\omega'} = -0.51$ mPa, $\sigma_{u'\omega'} = 5.07$ mPa, $\mu_{v'\omega'} = -0.298$ mPa, and $\sigma_{v'\omega'} = 3.79$ mPa. To prevent scaling inconsistencies across data sets, for all the variables, the same variable normalizations were used for both ERA5 and IFS-1 km.

Training-test split: for all experiments in this study, the years 2010, 2012, and 2014 were used for training, and the year 2015 was used for testing. Having a whole year for validation allows for testing how well the models learn seasonal variations. As a result, M1 and M2 have approximately 215 million and 72 million (single-column) training and validation samples, respectively, while M3 has approximately 27,000 and 9,000 (three-dimensional) training samples, respectively. For TL experiments, all 4 months of 3-hourly data from IFS-1 km were used for retraining.

Training time: The NNs were trained on a single NVIDIA A100 GPU. M1 and M2 require roughly 80 and 120 hrs to train over 100 epochs on global ERA5 data with four features $\{u, v, \theta, \omega\}$. Likewise, M3 requires 14 hrs to train over 100 epochs.

The TL training times are relatively faster. For instance, for M1, TL over 200 epochs on an A100 GPU (though for four months of IFS-1 km data) takes 1.5 hrs when only the final two layers are retrained, 1.75 hrs when the final three layers are retrained, and 2 hrs when the final four layers are retrained. All corresponding variations of the UNet took roughly 40–50 min of training time on the same GPU.

5. Results

We first show the predictions from the vertically global simulations. Following that, we exclusively focus on stratosphere-only runs. For simplicity, we only show the results from the best-performing feature set, $\{u, v, \theta, \omega\}$, unless specified. Results for other feature sets are shared as Supporting Information S1 wherever appropriate.

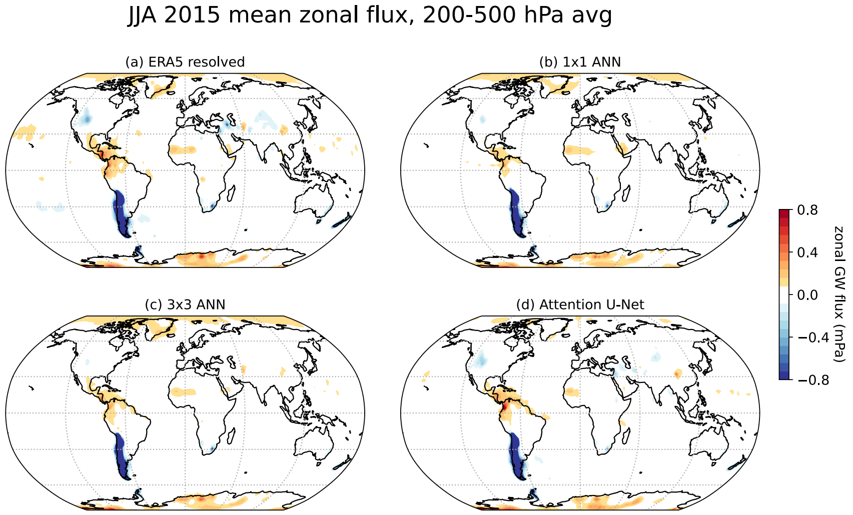

Figure 5. JJA 2015 mean resolved momentum fluxes for (a) ERA5, (b) M1, (c) M2, and (d) M3, in the 200–500 hPa troposphere. Fluxes are shown for the vertically global configuration with feature set $\{u, v, \theta, \omega\}$.

5.1. Global Simulation

The time-averaged global prediction of GWF in the troposphere from ERA5 and the three types of NNs is shown in Figure 5 and the difference in model predictions w.r.t. ERA5 is shown in Figure 6. All of M1–M3 demonstrate substantial skill in predicting the mean GWFs. The difference between ERA5 resolved fluxes and the ML predicted fluxes is the lowest for M3 (Figure 6c). The fluxes predicted from M1 and M2 have similar biases, with the largest biases around the Andes, North America, and Western Asia. Moreover, the true fluxes from ERA5 and the deviations in these regions have an order-of-magnitude similar to the fluxes.

In contrast, in the stratosphere, the prediction skill for the vertically global M3 is poorer than M1 and M2 at all latitudes, as shown in Figures 7 and 8, which show the fluxes and the difference w.r.t. ERA5, respectively. In particular, while all models contain biases over the Southern Ocean and the tropics, the biases for Attention UNet over these regions are much stronger. A similar comparison of the DJF troposphere and stratosphere for vertically global NNs is shown in Figures S1 and S2 in Supporting Information S1. Similar conclusions are drawn for the DJF period as well, but now the biases are concentrated in the Northern Hemisphere.

In a seasonally-averaged sense, the ML models generate an extremely accurate prediction of the global GWF distribution, including over well-known hotspots. These predictions are also qualitatively consistent with the GWF climatology presented in past studies (Gupta, Sheshadri, Alexander, & Birner, 2024; Hindley et al., 2020; Wei et al., 2022). This is further corroborated by the global flux distribution for all four seasonal averages in Figure 9 (left column). All NNs capture the tails of the distribution with considerable accuracy. The similarity between the predicted and the ERA5 flux distribution is quantified using the Hellinger distance defined in Section 2. The three NNs consistently generate low Hellinger distances and show impressive skill in capturing the tails of the GWF distribution. The Hellinger distances are lowest for the 1 × 1 model, but the distances are low enough for these differences to be considered small.

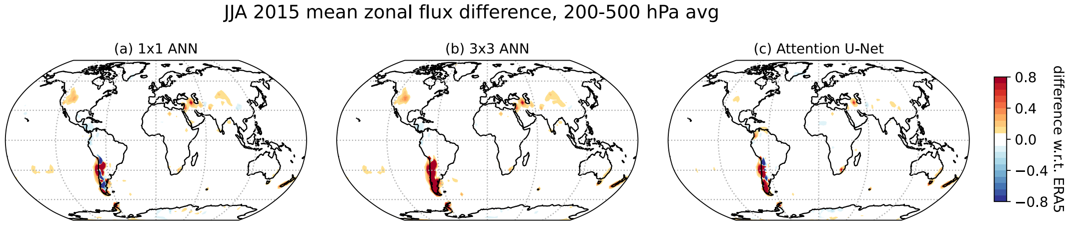

Figure 6. Similar to Figure 5 but shows the difference (w.r.t. ERA5) of the JJA fluxes in the troposphere for the three models (a) M1, (b) M2, and (c) M3.

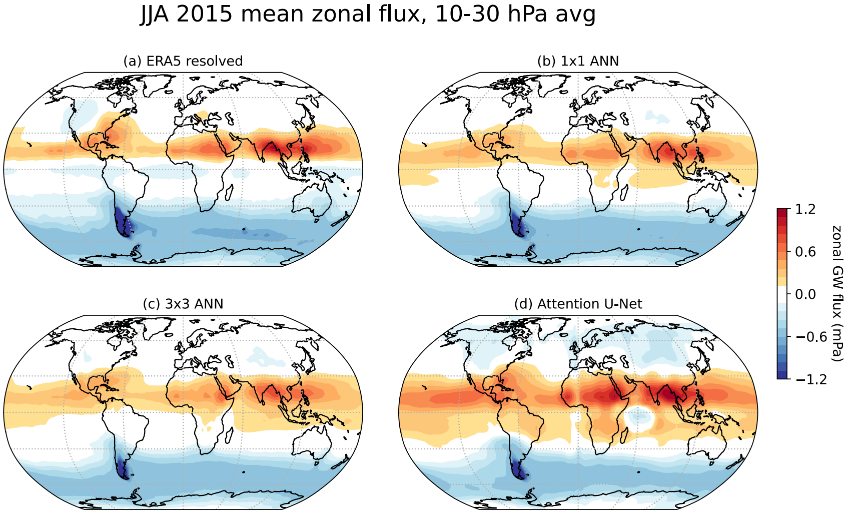

Figure 7. Same as Figure 5 but for the stratosphere (10–30 hPa or 30–45 km average). Mean resolved fluxes for (a) ERA5, (b) M1, (c) M2, and (d) M3. Fluxes are shown for the global configuration with feature set $\{u, v, \theta, \omega\}$.

5.2. Stratosphere-Only Simulation

Despite outperforming M1 and M2 in the troposphere, the M3 Attention UNet's sub-par performance in the stratosphere can be understood in terms of data imbalance and receptive fields, as discussed by Sun et al. (2024) and Pahlavan et al. (2024), respectively. ERA5 has double the number of vertical levels (channels) in the troposphere than in the stratosphere. Moreover, each vertical level/channel in the UNet model is treated independently due to which it has a much more restricted receptive field in the vertical than the ANNs. This negatively impacts the predictions in the stratosphere, away from the dominant source of GW excitation, that is, the troposphere.

To overcome these limitations, we now restrict the prediction domain to just the stratosphere, using stratosphere-specific ML models described in Section 3, that take the input dynamical conditions over both the troposphere and stratosphere (as before) but only predict fluxes in the stratosphere (1 to 200 hPa). This does not increase the vertical receptive field, but it reduces the data imbalance between the stratospheric and tropospheric fluxes, which could have quite different magnitudes. Since the GW fluxes retrieved using Helmholtz decomposition can also contain some contributions from strong convective fluxes not associated with GWs, this also allows us to focus exclusively on predicting GWFs. Such a strategy is also somewhat consistent with the present treatment of nonorographic drag in coarse-climate models where a fixed launch level (~200–300 hPa) is assumed for non-orographic GW packets.

Predicting fluxes only in the stratosphere results in much-improved prediction skills for all models in the Drake Passage (Figure 10). While earlier the UNet failed to model the JJA belt of GW activity over the Southern Ocean (top row), predicting fluxes only in the stratosphere alleviates this bias (bottom row). Reevaluating the JJA averages for the stratosphere-specific models, we find that while M1 and M2 models offer similar performance as in the global case (Figures 11b and 11c), the UNet's performance now surpasses both the ANNs (Figure 11d). Most importantly, we now obtain lowest biases over the Southern Ocean for M3. A similar comparison for the DJF period is shown in Figure S3 in Supporting Information S1.

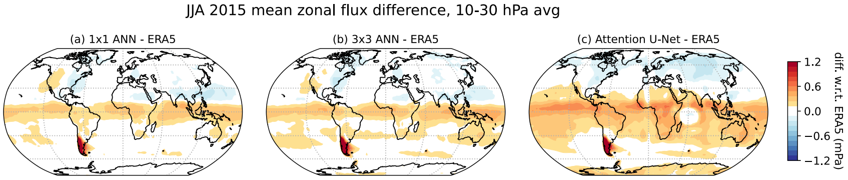
JJA 2015 mean zonal flux difference, 10-30 hPa avg
(a) 1x1 ANN - ERA5
(b) 3x3 ANN - ERA5
(c) Attention U-Net - ERA5
diff. w.r.t. ERA5 (mPa)
1.2
0.6
0.0
−0.6
−1.2

Figure 8. Similar to Figure 7 but shows the difference (w.r.t. ERA5) of the JJA fluxes in the stratosphere for the three models (a) M1, (b) M2, and (c) M3.

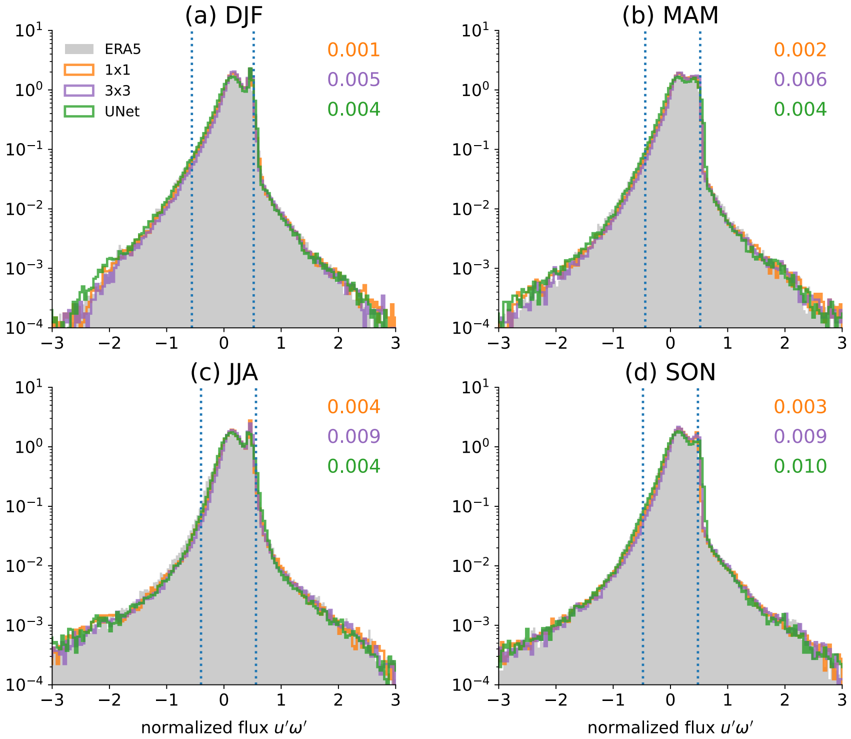

Figure 9. Histogram of the seasonally averaged predicted zonal flux $u'\omega'$ from ERA5 (gray shading), M1:1 × 1 M1 (orange), M2:3 × 3 (violet), and M3:Attention UNet (Green), for (a) December-January-February, (b) March-April-May, (c) June-July-August, and (d) September-October-November 2015. The numbers on the top right corner show the Hellinger distance for a corresponding predicted distribution w.r.t. ERA5 distribution. The vertical axis shows the flux density. Distribution is shown for the vertically global configuration with feature set $\{u, v, \theta, \omega\}$. The dotted vertical bars mark the 2.5th and 97.5th percentile.

We break down the time-averaged flux distribution by latitude and height (Figure 12). For visual clarity, we only show ERA5, M1, and M3. In many regions, most notably the tropics and the midlatitudes, the UNet provides a better prediction than the 1 × 1 ANN, as determined by the Hellinger distance. In some regions, like the upper stratospheric northern hemisphere polar regions, both models fail to capture the underlying flux distribution

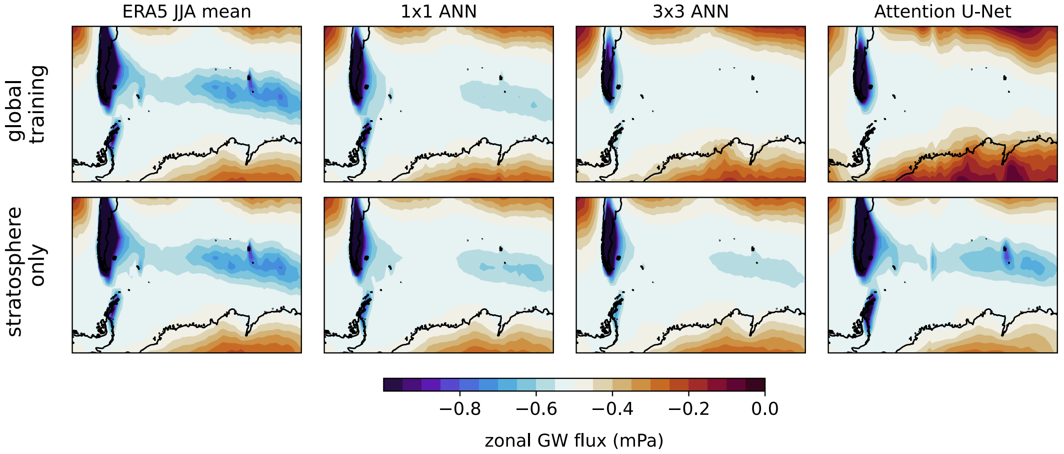

Figure 10. The first column shows JJA 2015 mean zonal flux $u'\omega'$ in ERA5 at 10 hPa over the Drake Passage and the Southern Ocean. The second to fourth columns show the JJA 2015 fluxes as predicted by M1–M3, respectively. The first row shows the true and predicted fluxes at 10 hPa from the vertically global configuration, while the second row shows the prediction from the stratosphere-only models.

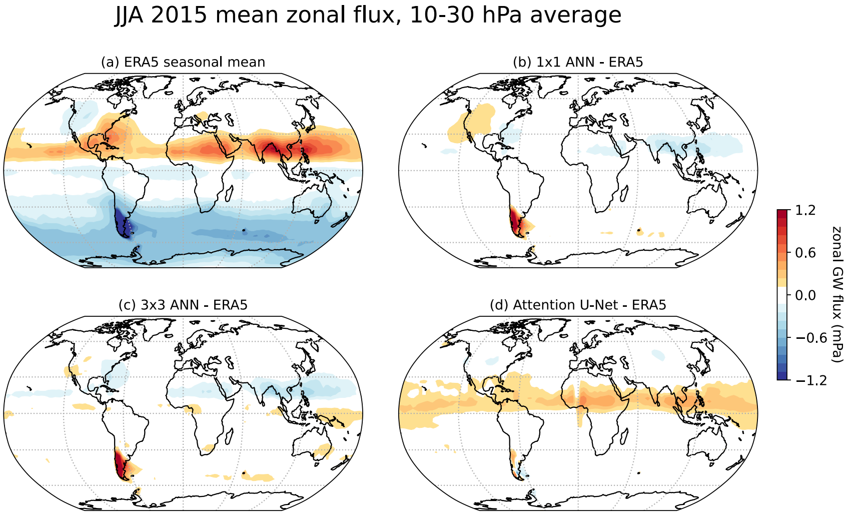

Figure 11. Same as Figure 7 but for the stratosphere-only model configuration and $\{u, v, \theta, \omega\}$ feature set.

accurately. Even in regions where the ANN has a lower Hellinger distance, we note its tendency to underestimate the range of flux values, leading to a higher concentration of fluxes around the mode, that is, its density around the mode is often higher than ERA5 in the summer hemisphere. This is not the case for UNet. As with the global ML models, the Hellinger distances for daily sampled fluxes are consistently lower for UNet, indicating its prowess in predicting the tails more accurately than M1 (Figure 17).

From a deep learning perspective, alternative strategies to obtain improved performance could be to either (a) increase the vertical receptive field by introducing vertical convolutions as well (Pahlavan et al., 2024), or (b) perhaps use an iterative recurrence-based approach introduced in Ukkonen and Chantry (2024) to level-wise predict the GWF in the vertical.

We now focus on the time evolution of the predicted fluxes. To analyze their transient wintertime evolution, we select six distinct orographic + nonorographic GW hotspots (Figure 13, center). The choice is based on the ERA5-derived lateral flux climatology presented in Gupta, Sheshadri, Alexander, and Birner (2024).

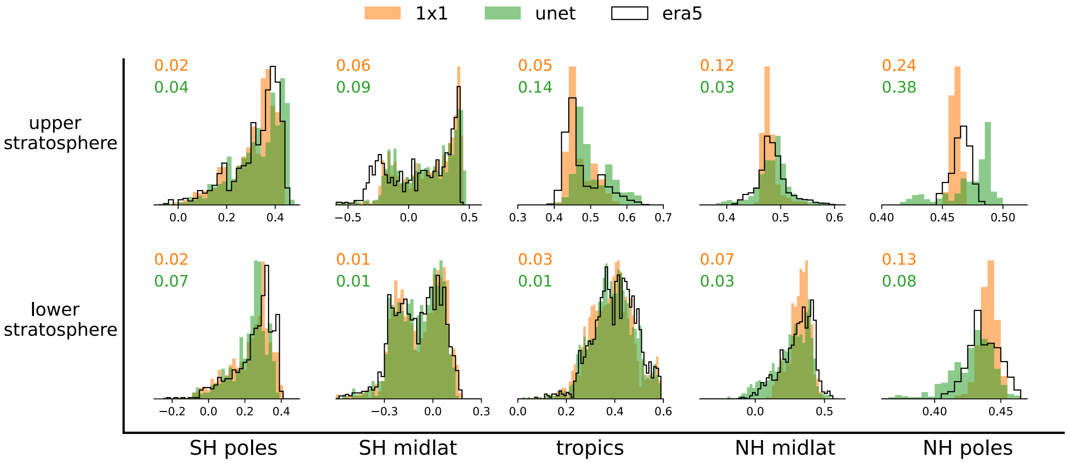

Figure 12. Histogram of the true versus predicted normalized fluxes $u'\omega'$ in the stratosphere for JJA 2015 partitioned into five latitude bands and two altitude levels. M1 is shown in orange, M3 is shown in green, and ERA5 is shown in black. Latitudes 75° to 90° are treated as polar regions, latitudes 30° to 60° as midlatitudes, and latitudes −20° to 20° as the tropics. The stratosphere between 1 and 30 hPa (30 km) is considered the upper stratosphere, and between 30 hPa (30 km) and 100 hPa (15 km) is considered the lower stratosphere. The numbers in the top-left show the Hellinger distance w.r.t. ERA5 distribution. Fluxes are shown for the stratosphere-only vertical configuration and features $\{u, v, \theta, \omega\}$.

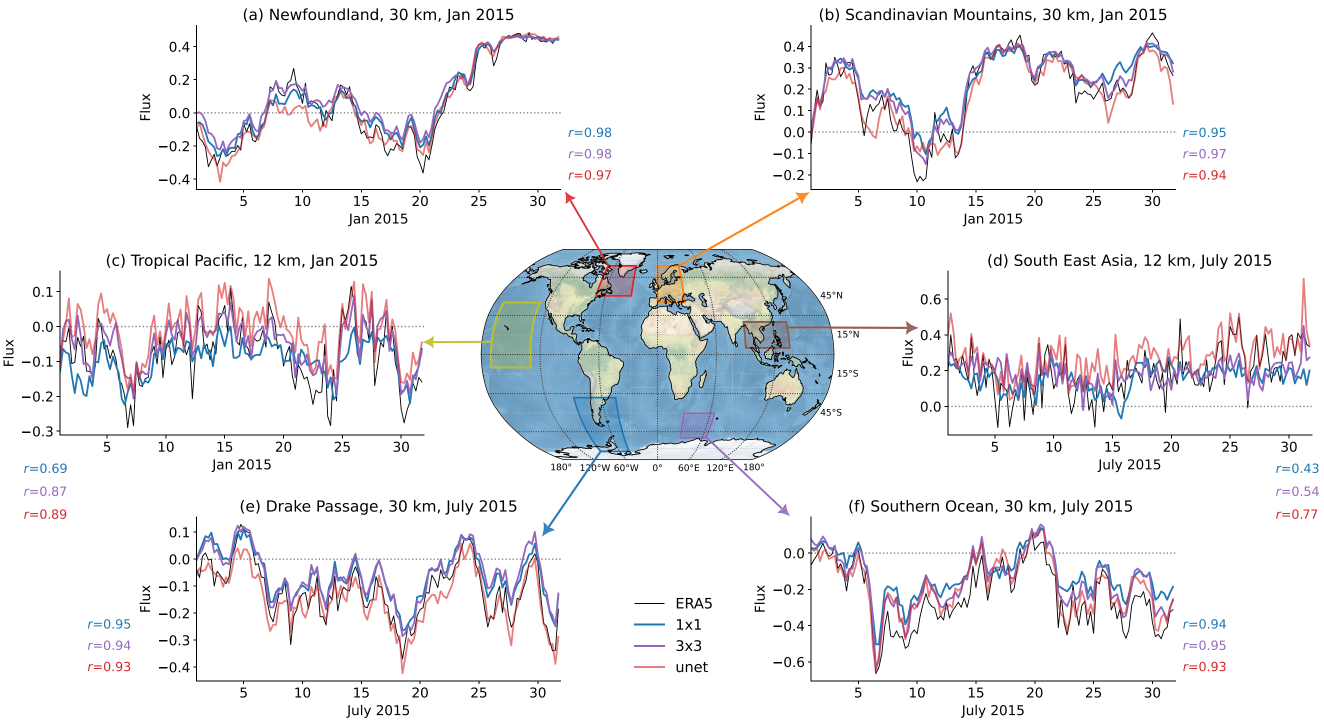

Figure 13. Timeseries of the normalized zonal momentum flux $u'\omega'$ over six hotspots highlighted over the map projection: (a) Newfoundland, (b) Scandinavian Mountains, (c) Tropical Pacific Ocean, (d) Southeast Asia, (e) Drake Passage, and (f) Southern Ocean. The fluxes are shown for January 2015 for (a–c), and for July 2015 for (d–f). For the midlatitudes, the fluxes are shown in the upper stratosphere (30 km), while for the tropics, the fluxes are shown in the upper troposphere (12 km). ERA5 is shown in black, M1 in blue, M2 in violet, and M3 in red. The Pearson correlation coefficients for each model w.r.t. ERA5 are provided next to each plot.

First focusing on the orographic hotspots, we note that all models predict the transient evolution of the GWF over Newfoundland, the Scandinavian Mountains, and the Drake Passage with high accuracy (Figures 13a, 13b, and 13e). In most cases, the Pearson correlation coefficient measures above 0.95, explaining more than 90% of the variance. In contrast, a mixed performance is obtained over nonorographic hotspots. Over the Southern Ocean away from the Drake Passage, the models exhibit exceptional prediction skill, with correlation coefficients around 0.95, correctly predicting the intermittent rise and decay of the GWF (Figure 13f). The correlation coefficients are much weaker in the tropics where the (presumably convective) GWFs are more prevalent and intermittent than in the extratropics due to the inherent stochasticity of convection. In the lower stratosphere, over the Tropical Pacific, Attention UNet has the highest correlation (0.89), 3 × 3 being a close second (0.87), and the 1 × 1 model the lowest (0.69) (Figure 13c). The correlation coefficients are even lower over the Southeast Asian region, with the correlation being 0.77 for the UNet and merely 0.43 for the 1 × 1 ANN (Figure 13d). A similar pattern is noticed for the meridional momentum flux $v'\omega'$ (Figure S4 in Supporting Information S1).

No clear pattern in prediction skill over the nonorographic hotspots is obtained, but the skill appears to be the poorest in the tropics. While this has not been explored here, because a bulk of the GW activity in the tropics is convectively generated, it is possible that adding specific humidity as a feature could alleviate some of these biases. Here, vertical velocity serves as a proxy to represent tropical convection.

The zonal mean flux profiles of the predicted fluxes, and the vertical profile of flux variability are shown in Figures S5 and S6 in Supporting Information S1, respectively. The NNs generate strong predictions of the zonal mean profile and flux variability of the predicted fluxes. Moreover, the variability generated by M3 matches the variability in ERA5 much more closely than the variability generated by M1. The deviations are the strongest in the upper stratosphere.

Based on the analysis, a stratosphere-only Attention UNet-based architecture with a global receptive field in the horizontal and a limited receptive field in the vertical consistently predicts the GWF skillfully over all latitudes. Many of these differences are not apparent when analyzing the time-averaged flux maps and are only revealed while analyzing the GWF's transient evolution.

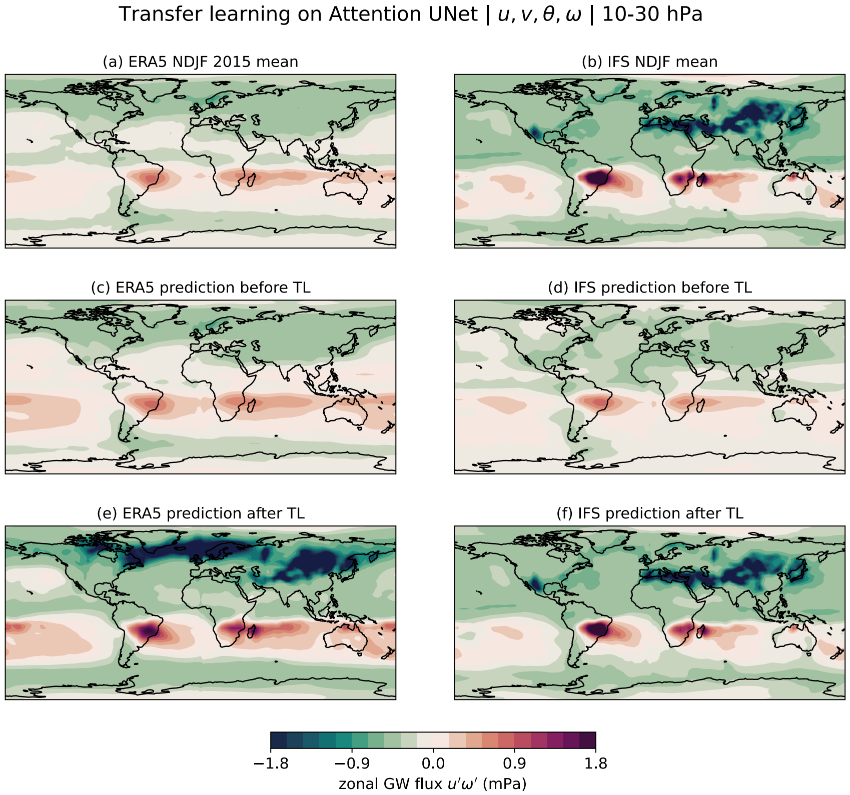
Transfer learning on Attention UNet | u, v, θ, ω | 10-30 hPa
(a) ERA5 NDJF 2015 mean
(b) IFS NDJF mean
(c) ERA5 prediction before TL
(d) IFS prediction before TL
(e) ERA5 prediction after TL
(f) IFS prediction after TL
−1.8
−0.9
0.0
0.9
1.8
zonal GW flux $u'\omega'$ (mPa)

Figure 14. (a, b) True resolved gravity wave flux $u'\omega'$ (units mPa) for the NDJF 2015 period and the NDJF 2018–2019 period for ERA5 and IFS-1 km, respectively. Predicted fluxes for (c) ERA5 NDJF 2015 and (d) IFS-1 km NDJF 2018–2019 input from the stratosphere-only Attention UNet model before transfer learning (TL). (e, f) Respective neural network predicted fluxes after TL on IFS-1 km. Fluxes are shown for the model with input features $\{u, v, \theta, \omega\}$.

5.3. Correcting Flux Magnitudes Using Transfer Learning (TL)

To systematically correct the magnitudes of the predicted GWF, we apply TL, as described in Section 3 and illustrated in Figure 4. All the NNs were re-trained on IFS-1 km fluxes using TL, but for clarity here we only discuss the results for the stratosphere-only Attention UNet model. The models were re-trained by "unfreezing" the last 2 layers, the last 3 layers, and the last 4 layers. The losses were the lowest when only the last 4 layers were unfrozen for re-training (while keeping the preceding layers unchanged) (Figure S9 in Supporting Information S1). Here, we only show the results for these 4-layer runs and share the results from the other runs as Supporting Information S1.

To assess the efficacy of TL, we first evaluate/validate the ERA5-trained NNs on input-output pairs from IFS-1 km. Then, following TL, we evaluate the re-trained neural nets on input-output pairs from both ERA5 and IFS-1 km. This serves two key purposes. First, it allows assessing the model's predictive capability on out-of-set data from IFS-1 km. Before retraining, one can expect the ERA5-trained models to underestimate the IFS-1 km GWF. Second, it allows a direct assessment of the improvements in model prediction due to TL. After retraining, the models should ideally predict stronger GWF for ERA5 and provide magnitudes that are comparable to IFS-1 km.

The time-averaged predictions from the TL experiments are shown in Figure 14. Before TL, the predictions from UNet strongly match the GWF from ERA5 (Figures 14a–14c). In addition, the UNet predicts significantly weaker GWF for IFS-1 km: the positive fluxes in the tropics, the Southern Ocean belt, the fluxes over Central Asia are all weaker (Figures 14b–14d). Following TL, however, a striking increase in the flux magnitudes is noted for both ERA5 and IFS inputs. The predicted flux map qualitatively agrees with ERA5 prediction prior to TL but has much stronger magnitudes (Figures 14c–14e). Also, the predicted fluxes for IFS input now largely agree with the GWF mean from IFS-1 km (Figures 14d–14f). Thus, the models preserve their learning from the low-resolution ERA5 data and update the model parameters to appropriately scale it, while also matching the fluxes from the high-

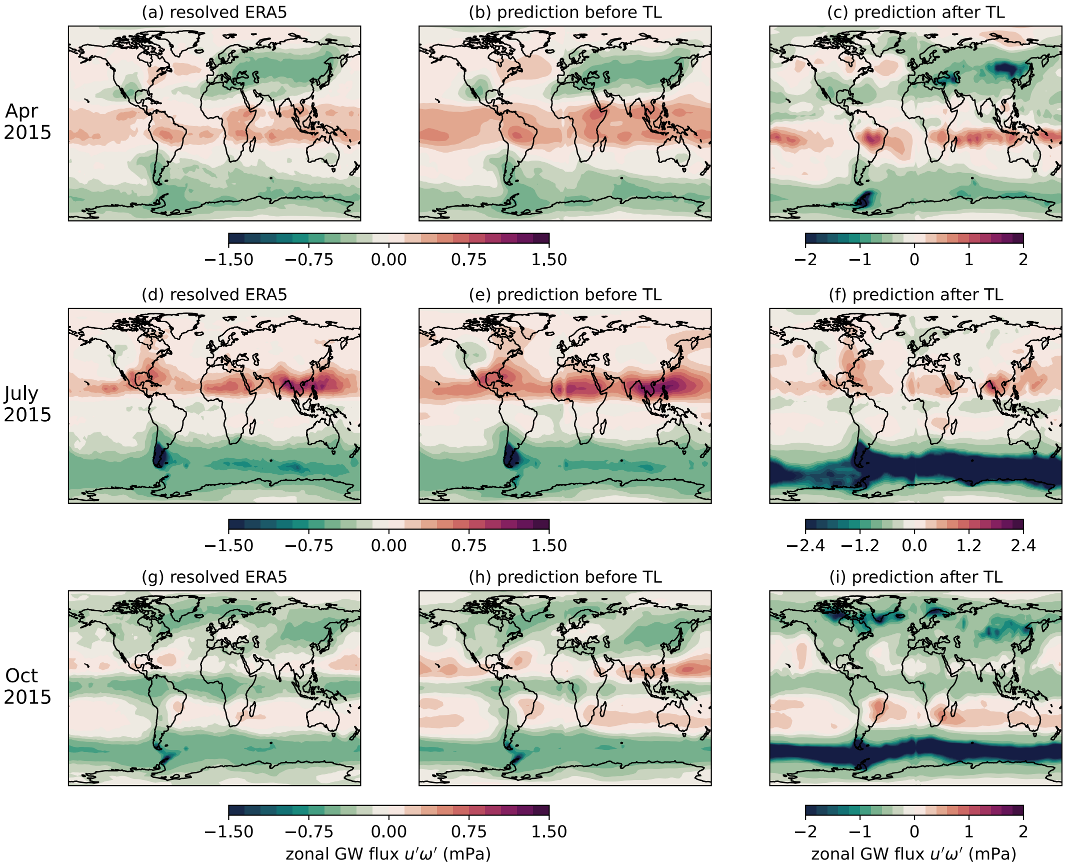

Figure 15. (left column) Resolved ERA5 flux $u'\omega'$ (units mPa), (middle column) its prediction on ERA5 test set before transfer learning (TL), and (right column) its prediction on ERA5 test set after TL. The fluxes are shown for three different months: (top row) April 2015, (middle row) July 2015, and (bottom row) October 2015. The color bars for the third column are differently scaled than those for the first two columns. Predictions are shown for stratosphere-only M3 with input features $\{u, v, \theta, \omega\}$.

resolution IFS-1 km. Sequentially using the two data sets thus allows learning GWFs from one data set and then using another, better-quality data set, to enhance learning and provide improved GWF magnitudes. The predictions generated by models when only the last two and three layers were re-trained are shown in Figures S10 and S11 in Supporting Information S1, respectively.

Similar conclusions are drawn for the out-of-set months of April, July, and October 2015 (Figure 15). Note that since we had access to only 4 months (NDJF) of IFS-1 km data, it is only possible to test the re-trained models for these months on ERA5. For all 3 months, we find that before TL, the UNet predicts a global GWF distribution similar to ERA5 (Figures 15b,e and 15h), but following TL, the UNet preserves the global features/hotspots for GW activity while yielding much stronger GWFs, correcting for ERA5's low resolution (Figures 15c, 15f, and 15i). Most notably, the positive fluxes in the tropics are enhanced by a factor of 2. Likewise, the wintertime belts of midlatitude GW activity identified in Hendricks et al. (2014) and Gupta, Sheshadri, Alexander, and Birner (2024) are enhanced by at least a factor of 2. For April, the fluxes are scaled similarly in both hemispheres. For July and October, the fluxes are scaled more strongly in the winter (Southern) hemisphere. The predictions generated by models when only the last two and three layers were re-trained are shown in Figures S12 and S13 in Supporting Information S1, respectively.

Finally, we illustrate the time evolution of the fluxes from the TL models on two prominent features of variability in the stratosphere: the tropical QBO and the Southern final warmings (Figure 16). Both the features are strongly influenced by GW forcings. For the QBO, M3 UNet correctly learns the GWF transition around easterly-to-westerly phase transitions (Figures 16a,b,c). Following TL, the magnitude of the positive GWF change significantly between Jan 2015 to July 2015, and is accompanied by an increase in the negative fluxes from August 2015

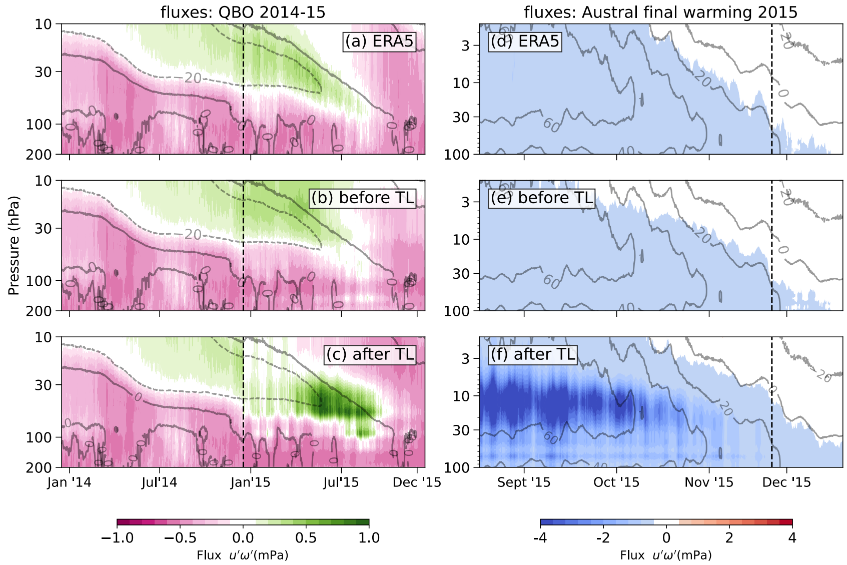

Figure 16. Evolution of the true resolved flux $u'\omega'$ (units mPa) from (a, d) ERA5, (b, e) predicted flux from stratosphere-only Attention UNet before transfer learning (TL), and (c, f) predicted flux after TL for two prominent patterns of variability in the stratosphere: (left column) tropical quasi-biennial oscillation and (right column) Antarctic final warmings. Tropical fluxes are averaged over 10°S–10°N and the winter midlatitude fluxes are averaged over 55°S–65°S. In (a–c), the black dashed line separates the training period from the testing period, and thus, fluxes to the left of the bar are identical for all three plots, while the fluxes to the right of the bar in (b) and (c) show the predicted fluxes. In contrast, in (d)–(f), the black dashed line shows the final warming date for the SON 2015 period. For all plots, the fluxes are shown in color, and the solid black curves show the zonal mean zonal wind with a contour interval of 20 m/s. Fluxes are shown for TL-updated stratosphere only M3 Neural Network with input features $\{u, v, \theta, \omega\}$.

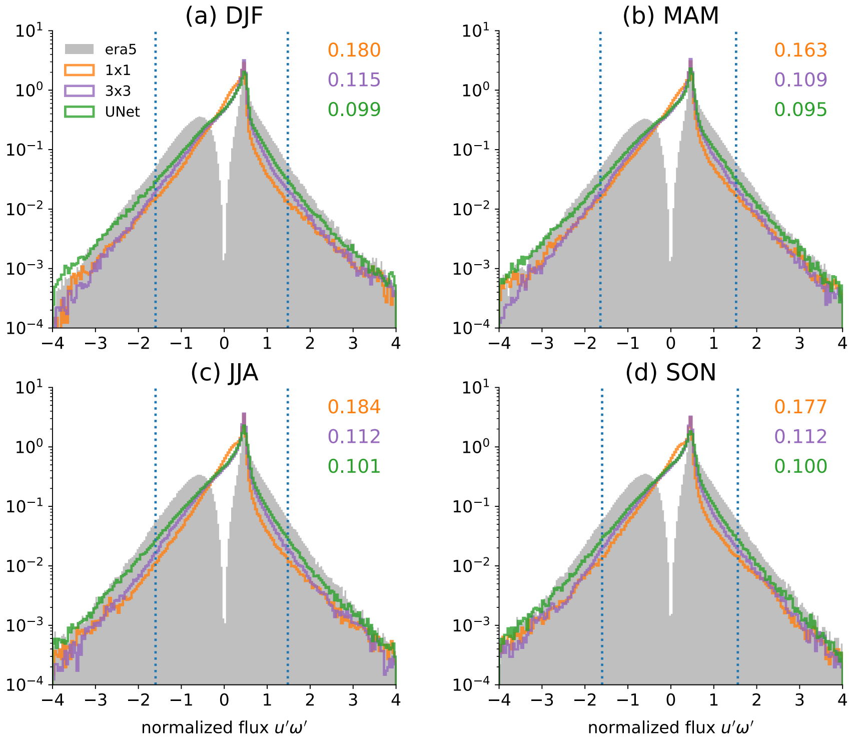

Figure 17. Same as Figure 9 but for daily averages instead of seasonal averages. (a) December-January-February, (b) March-April-May, (c) June-July-August, and (d) September-October-November 2015. For daily averages, the Attention UNet model consistently has the lowest Hellinger distances, and offers better predictions over both the distribution bulk and tails.

to December 2015. Likewise, a striking increase in the negative GWF flux magnitudes is obtained during the final warming period following TL (Figures 16d,e,f). Re-training on IFS-1 km fluxes enhances the GWF around 60°S by a factor of 2–3 during the final warming period. These magnitudes are consistent with the calculations in Gupta et al. (2021), which revealed a 60% underestimation of the resolved fluxes by ERA5 during the final warming period.

6. Conclusion and Discussion

Climate prediction models represent GW effects using simplified parameterizations that often neglect their horizontal propagation, leading to key biases. We developed three deep-learning NNs to represent atmospheric GWs in coarse-climate models. The three NNs represent three different degrees of horizontal nonlocality: single-column, 3×3 neighboring columns, and globally nonlocal. All NNs take background winds and temperature as input and produce the vertical momentum fluxes $u'\omega'$ and $v'\omega'$ as output. The NNs were trained and tested on 4 years of resolved GW fluxes from modern reanalysis, ERA5, using three different feature sets, and evaluated in both the troposphere and the stratosphere. Since the 25 km ERA5 only resolves a fraction of the mesoscale GW spectrum, the NNs were subsequently partially re-trained using transfer learning on mesoscale-resolving GWF data extracted from a 1.4 km global climate model, to compensate for and correct the fluxes underestimated in ERA5.

The offline performance of the NNs was assessed on both the seasonal mean and transient evolution. The NNs provide a reasonable prediction of the DJF and JJA global fluxes in the troposphere and the stratosphere. Using the tropospheric and stratospheric background conditions to predict the GWFs only in the stratosphere led to better model performance and accurately modeled the belt of GW activity in the midlatitude stratosphere. The NNs, especially the Attention UNet, also demonstrated proficiency in predicting the fluxes over both orographic and nonorographic hotspots, with some remaining biases in the tropics. This is despite the Attention UNet being trained on a substantially lower number of training samples (215 million single columns vs. 27k global time slices). Our experiments, thus, demonstrate the importance of embedding horizontal nonlocality in the deep learning architectures for a more accurate ML-based flux prediction.

Following TL trials, the NNs preserved their learning from training on low-fidelity ERA5, while also learning to correct the resolved flux magnitudes from re-training on high-fidelity IFS-1 km. As a result, the final NNs showed skill both in (a) identifying GWFs in space and time, and (b) consistently amplifying (or correcting) the GWF magnitudes for all seasons. The NNs also predicted consistent momentum fluxes in the tropics and the mid-latitudes associated with the QBO and the Austral final warmings, respectively.

Stable offline performance of NNs does not always equate to stable online performance (Brenowitz et al., 2020). Given the promising offline performance of the data-driven "schemes," efforts are underway to couple them to a coarse-climate model (National Center for Atmospheric Research's Community Atmosphere Model (CAM7)) and evaluate its online performance. Due to computational feasibility, particular focus will be on coupling the single-column and the globally nonlocal UNet to CAM and assessing their contributions to overall stratospheric variability, stratospheric extremes, and their predictions for various global warming scenarios. This will present a rigorous test of the scheme's generalizability to unseen dynamical conditions.

The results obtained indicate the following regarding the development of data-driven sub-grid scale parameterizations:

i. Embedding nonlocal dynamical information can be crucial to improving deep learning-based process representation and flux predictions. The increase in performance from M1 to M2 to M3 is not monotonic, as M2 has a higher bias than M1 and M3, despite having comparable or lower training errors. Ultimately, the model with the best performance (Attention UNet) was the one that embedded the most nonlocal information to make the prediction. The training errors too decreased more rapidly for the nonlocal models (M2 and M3) than for the single-column model (M1).

ii. Since GWFs have a fine-scale structure, having vertical velocity as an input feature improves model prediction. NNs with just $\{u, v, \omega\}$ as input fail to predict the finer-scale flux features. This contrasts the findings from the parameterized drag emulator (WaveNet) of Espinosa et al. (2022), where training on just u leads to R^2 values ≥ 0.9 and adding ω as a feature results in marginal improvements.

iii. High-resolution climate model output, while scarce, can serve as invaluable training data for developing data-driven parameterizations of subgrid-scale processes. As argued in Parthipan and Wischik (2022), high-resolution climate data from multiple projects and initiatives can be combined to prepare a unified high-volume, high-fidelity training set covering a broad range of scenarios and numerics. Our TL experiments strengthen their argument. The ERA5 + IFS-1 km approach adopted here highlights a pathway to use a combination of heterogeneous high-resolution GW data sets to develop future data-driven parameterizations, also potentially improving their generalizability.

Multiple avenues exist to further improve the deep learning schemes developed in this work, and efforts are underway to address some of the following limitations:

- *Predicting small values*: Unlike the seasonally-averaged GWFs with Hellinger distances less than 0.01 (Figure 9 and Figure S7 in Supporting Information S1), the histogram of daily averaged fluxes (Figure 17 and Figure S8 in Supporting Information S1) reveals the NNs' tendency to underestimate GWFs. With Hellinger distances of 0.1 and higher, the daily predictions underestimate a portion of the bulk of the bimodal spectrum and instead generate a unimodal spectrum centered around 0. This reveals the models' inability to distinguish between weak atmospheric variability and noise.
- *Validating TL*: GWF corrections following TL appear to be consistent with our understanding of GW modeling. Yet, there is no direct way to rigorously validate the GWFs except either through (a) testing on other high-resolution climate model outputs, or (b) by coupling the NNs to a coarse-climate model and assessing the resolved wind fields and variability like the QBO period, sudden stratospheric warming frequency, and final warming dates.
- *Training data*: Extracting GWF from high-resolution data sets is computationally intensive. For this reason, we only used 4 years of ERA5 data and complemented it with 4 months of IFS-1 km data. Given the remarkable performance on 4 years of data, expanding the training set can undoubtedly lead to key performance gains, especially in the case of extreme events and low-frequency climate variability. However, due to significant latency in model-level data retrieval from the Copernicus Climate Data Store with added issues related to storage, we have restricted the analysis to only 4 years. These years, however, do capture almost two complete QBO cycles and major portions of ENSO-related variability and serve as a robust, if not exhaustive, data set for training. Part of the ongoing work addresses this issue of data (in-)sufficiency head-on by developing data-driven schemes with sparse training data but superior skill by leveraging large AI foundation models already pre-trained on high volumes of global weather and climate data.
- *Feature set*: The neural architectures and feature sets used in this study were inspired by traditional parameterizations. Because the ML model size scales proportionally to the square of the number of features, we have restricted ourselves to using up to four features. While not amply explored in this study due to data limitations, we surmise that adding other relevant fine-scale variables, like specific humidity and convective fluxes, can lead to notable performance gains.

In conclusion, our analysis demonstrates the strong capability of machine learning methods to learn physically consistent GW fluxes from a blend of high-volume low-resolution and low-volume high-resolution climate data to skillfully represent their missing effects in coarse-climate models; especially nonlocal horizontal propagation and transience. Ours being the first-ever deep learning model to be trained on globally resolved GW fluxes, we identify both the strengths of this approach and avenues for future improvements, while also providing a benchmark for future studies. The models provide a remarkable prediction of both the seasonally averaged fluxes and their time evolution over prominent hotspots. Upon coupling with coarse-climate models, these NNs can potentially replace existing GW parameterizations and serve as fast emulators to represent missing GW effects. These missing small-scale effects present as one of the leading sources of structural uncertainty in future climate projections. Adopting this approach for a broader cohort of sub-grid scale processes can potentially improve the representation of global circulation in climate models and open avenues to generate more precise future climate projections.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

ECMWF's ERA5 data (Hersbach et al., 2023) can be freely accessed from https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels. The code to compute the GW momentum fluxes and to conservatively coarse-grain the fluxes, along with the modified PySpharm functions can be accessed at: https://doi.org/10.17605/OSF.IO/GX32S in the *python_scripts* folder. Gravity wave fluxes extracted from ECMWF's IFS-1 km run are available at native grid resolution at https://doi.ccs.ornl.gov/ui/doi/475. The default WindSpharm Python package is publicly available at https://ajdawson.github.io/windspharm/, and the PySpharm Python package is publicly available at https://pypi.org/project/pyspharm/. The xESMF package used for conservative coarse-graining is publicly available at https://xesmf.readthedocs.io/en/stable/.

The code for all the machine learning models, along with the jobscripts and inference scripts, is publicly available at Gupta et al. (2025): https://doi.org/10.5281/zenodo.16415113, and the ML model checkpoints can be accessed at https://huggingface.co/amangupta2/nonlocal_gwfluxes.

Acknowledgments

Aditi Sheshadri and Aman Gupta are supported by Schmidt Sciences, LLC, a philanthropic initiative founded by Eric and Wendy Schmidt, as part of the Virtual Earth System Research Institute (VESRI). Aditi Sheshadri also acknowledges support from the National Science Foundation through Grant OAC-2004492. The work was also supported by NASA's Office of Chief Science Data Officer and the Earth Science Division's Earth Science Scientific Computing, Earth Science Data Systems Program, and the Earth Science Modeling and Analysis Program. The 1.4 km IFS runs were performed by Nils Wedi and Inna Polichtchouk (ECMWF) on the Summit Supercomputer at the Oak Ridge National Laboratory, using resources of the Oak Ridge Leadership Computing Facility (OLCF), which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. We thank M. Joan Alexander, Edwin P. Gerber, Pedram Hassanzadeh, Laura Mansfield, Hamid Pahlavan, Y. Qiang Sun, Brian Green, Robert King, Manil Maskey, and Rahul Ramachandran for insightful discussions. We also thank two anonymous reviewers for helpful comments on the transfer learning experiments.

References

Achatz, U., Alexander, M. J., Becker, E., Chun, H.-Y., Dörnbrack, A., Holt, L., et al. (2024). *Atmospheric gravity waves: Processes and parameterization*. JAS. https://doi.org/10.1175/JAS-D-23-0210.1

Albers, J. R., & Birner, T. (2014). Vortex preconditioning due to planetary and gravity waves prior to sudden stratospheric warmings. *Journal of the Atmospheric Sciences*, *71*(11), 4028–4054. https://doi.org/10.1175/JAS-D-14-0026.1

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

Amemiya, A., & Sato, K. (2016). A new gravity wave parameterization including three-dimensional propagation. *Journal of the Meteorological Society of Japan. Ser. II*, *94*(3), 237–256. https://doi.org/10.2151/jmsj.2016-013

Becker, E. (2012). Dynamical control of the middle atmosphere. *Space Science Reviews*, *168*(1), 283–314. https://doi.org/10.1007/s11214-011-9841-5

Brenowitz, N. D., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., et al. (2020). Machine learning climate model dynamics: Offline versus online performance. https://doi.org/10.48550/arXiv.2011.03081

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

Choi, H., Kwon, H., Kim, S.-J., & Kim, B.-M. (2024). Warmer Antarctic summers in recent decades linked to earlier stratospheric final warming occurrences. *Communications Earth & Environment*, *5*(1), 1–9. https://doi.org/10.1038/s43247-024-01221-0

Connelly, D. S., & Gerber, E. P. (2024). Regression forest approaches to gravity wave parameterization for climate projection. *Journal of Advances in Modeling Earth Systems*, *16*(7), e2023MS004184. https://doi.org/10.1029/2023MS004184

Domeisen, D. I. V., & Butler, A. H. (2020). Stratospheric drivers of extreme events at the Earth's surface. *Communications Earth & Environment*, *1*(1), 1–8. https://doi.org/10.1038/s43247-020-00060-z

Eichinger, R., Rhode, S., Garny, H., Preusse, P., Pisoft, P., Kuchař, A., et al. (2023). Emulating lateral gravity wave propagation in a global chemistry–climate model (EMAC v2.55.2) through horizontal flux redistribution. *Geoscientific Model Development*, *16*(19), 5561–5583. https://doi.org/10.5194/gmd-16-5561-2023

Espinosa, Z. I., Sheshadri, A., Cain, G. R., Gerber, E. P., & DallaSanta, K. J. (2022). Machine learning gravity wave parameterization generalizes to capture the QBO and response to increased CO_2. *Geophysical Research Letters*, *49*(8), e2022GL098174. https://doi.org/10.1029/2022GL098174

Eyring, V., Collins, W. D., Gentine, P., Barnes, E. A., Barreiro, M., Beucler, T., et al. (2024). Pushing the frontiers in climate modelling and analysis with machine learning. *Nature Climate Change*, *14*(9), 1–13. https://doi.org/10.1038/s41558-024-02095-y

Fritts, D. C., & Alexander, M. J. (2003). Gravity wave dynamics and effects in the middle atmosphere. *Reviews of Geophysics*, *41*(1), 1003. https://doi.org/10.1029/2001RG000106

Garcia, R. R., Smith, A. K., Kinnison, D. E., de la Cámara, Á., & Murphy, D. J. (2017). Modification of the gravity wave parameterization in the whole atmosphere community climate model: Motivation and results. *Journal of the Atmospheric Sciences*, *74*(1), 275–291. https://doi.org/10.1175/JAS-D-16-0104.1

Garner, S. T. (2005). A topographic drag closure built on an analytical base flux. *Journal of the Atmospheric Sciences*, *62*(7), 2302–2315. https://doi.org/10.1175/JAS3496.1

Gelaro, R. (2015). *Evaluation of the 7-km GEOS-5 nature run*. National Aeronautics and Space Administration, Goddard Space Flight Center.

Giorgetta, M. A., Manzini, E., & Roeckner, E. (2002). Forcing of the quasi-biennial oscillation from a broad spectrum of atmospheric waves. *Geophysical Research Letters*, *29*(8), 86-1–86-4. https://doi.org/10.1029/2002GL014756

Gupta, A., Birner, T., Dörnbrack, A., & Polichtchouk, I. (2021). Importance of gravity wave forcing for springtime Southern polar vortex breakdown as revealed by ERA5. *Geophysical Research Letters*, *48*(10), e2021GL092762. https://doi.org/10.1029/2021GL092762

Gupta, A., Reichert, R., Dörnbrack, A., Garny, H., Eichinger, R., Polichtchouk, I., et al. (2024). Estimates of Southern hemispheric gravity wave momentum fluxes across observations, reanalyses, and kilometer-scale Numerical weather prediction model. *Journal of the Atmospheric Sciences*, *81*(aop), 583–604. https://doi.org/10.1175/JAS-D-23-0095.1

Gupta, A., Sheshadri, A., Alexander, M. J., & Birner, T. (2024). Insights on lateral gravity wave propagation in the extratropical stratosphere from 44 years of ERA5 data. *Geophysical Research Letters*, *51*(14), e2024GL108541. https://doi.org/10.1029/2024GL108541

Gupta, A., Sheshadri, A., & Anantharaj, V. (2024). Gravity Wave Momentum Fluxes from 1 km Global ECMWF Integrated forecast system. *Scientific Data*, *11*(1), 903. https://doi.org/10.1038/s41597-024-03699-x

Gupta, A., Tommelt, & Emberton, J. (2025). DataWaveProject/nonlocal_gwfluxes: Publication share. *Zenodo*. https://doi.org/10.5281/zenodo.16415113

Hardiman, S. C., Scaife, A. A., van Niekerk, A., Prudden, R., Owen, A., Adams, S. V., et al. (2023). Machine learning for nonorographic gravity waves in a climate model. *Artificial Intelligence for the Earth Systems*, *2*(4), e220081. https://doi.org/10.1175/AIES-D-22-0081.1

Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die Reine und Angewandte Mathematik*, *1909*(136), 210–271. https://doi.org/10.1515/crll.1909.136.210

Hendricks, E. A., Doyle, J. D., Eckermann, S. D., Jiang, Q., & Reinecke, P. A. (2014). What is the source of the stratospheric gravity wave belt in Austral winter? *Journal of the Atmospheric Sciences*, *71*(5), 1583–1592. https://doi.org/10.1175/JAS-D-13-0332.1

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Munoz-Sabater, J., et al. (2023). ERA5 hourly data on pressure levels from 1940 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. https://doi.org/10.24381/CDS.BD0915C6

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hindley, N. P., Wright, C. J., Hoffmann, L., Moffat-Griffin, T., & Mitchell, N. J. (2020). An 18-Year climatology of directional stratospheric gravity wave momentum flux from 3-D satellite observations. *Geophysical Research Letters*, *47*(22), e2020GL089557. https://doi.org/10.1029/2020GL089557

Hines, C. O. (1997). Doppler-spread parameterization of gravity-wave momentum deposition in the middle atmosphere. Part 2: Broad and quasi monochromatic spectra, and implementation. *Journal of Atmospheric and Solar-Terrestrial Physics*, *59*(4), 387–400. https://doi.org/10.1016/S1364-6826(96)00080-6

Holton, J. R. (1982). The role of gravity wave induced drag and diffusion in the momentum budget of the mesosphere. *Journal of the Atmospheric Sciences*, *39*(4), 791–799. https://doi.org/10.1175/1520-0469(1982)039⟨0791:TROGWI⟩2.0.CO;2

Kidston, J., Scaife, A. A., Hardiman, S. C., Mitchell, D. M., Butchart, N., Baldwin, M. P., & Gray, L. J. (2015). Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. *Nature Geoscience*, *8*(6), 433–440. https://doi.org/10.1038/ngeo2424

Kim, Y.-J., Eckermann, S. D., & Chun, H.-Y. (2003). An overview of the past, present and future of gravity-wave drag parametrization for numerical climate and weather prediction models. *Atmosphere-Ocean*, *41*(1), 65–98. https://doi.org/10.3137/ao.410105

Klaver, R., Haarsma, R., Vidale, P. L., & Hazeleger, W. (2020). Effective resolution in high resolution global atmospheric models for climate studies. *Atmospheric Science Letters*, *21*(4), e952. https://doi.org/10.1002/asl.952

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. https://doi.org/10.1029/2023JD038549

Lindborg, E. (2015). A Helmholtz decomposition of structure functions and spectra calculated from aircraft data. *Journal of Fluid Mechanics*, *762*, R4. https://doi.org/10.1017/jfm.2014.685

Lott, F., & Guez, L. (2013). A stochastic parameterization of the gravity waves due to convection and its impact on the equatorial stratosphere. *Journal of Geophysical Research: Atmospheres*, *118*(16), 8897–8909. https://doi.org/10.1002/jgrd.50705

Lott, F., & Miller, M. J. (1997). A new subgrid-scale orographic drag parametrization: Its formulation and testing. *Quarterly Journal of the Royal Meteorological Society*, *123*(537), 101–127. https://doi.org/10.1002/qj.49712353704

Lu, Y., Xu, X., Wang, L., Liu, Y., Wu, T., Jie, W., & Sun, J. (2024). Machine learning emulation of subgrid-scale orographic gravity wave drag in a general circulation model with middle atmosphere extension. *Journal of Advances in Modeling Earth Systems*, *16*(3), e2023MS003611. https://doi.org/10.1029/2023MS003611

Mansfield, L. A., Gupta, A., Burnett, A. C., Green, B., Wilka, C., & Sheshadri, A. (2023). Updates on model hierarchies for understanding and simulating the climate System: A focus on data-informed methods and climate change impacts. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003715. https://doi.org/10.1029/2023MS003715

McLandress, C., Shepherd, T. G., Polavarapu, S., & Beagley, S. R. (2012). Is missing orographic gravity wave drag near 60°S the cause of the stratospheric zonal wind biases in chemistry–climate models? *Journal of the Atmospheric Sciences*, *69*(3), 802–818. https://doi.org/10.1175/JAS-D-11-0159.1

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention U-Net: Learning where to look for the pancreas (No. arXiv:1804.03999). *arXiv*. https://doi.org/10.48550/arXiv.1804.03999

Orr, A., Bechtold, P., Scinocca, J., Ern, M., & Janiskova, M. (2010). Improved middle atmosphere climate and forecasts in the ECMWF model through a nonorographic gravity wave drag parameterization. *Journal of Climate*, *23*(22), 5905–5926. https://doi.org/10.1175/2010JCLI3490.1

Pahlavan, H. A., Hassanzadeh, P., & Alexander, M. J. (2024). On the importance of learning non-local dynamics for stable data-driven climate modeling: A 1D gravity wave-QBO testbed (No. arXiv:2407.05224). *arXiv*. https://doi.org/10.48550/arXiv.2407.05224

Parthipan, R., & Wischik, D. J. (2022). Don't waste data: Transfer learning to leverage all data for machine-learnt climate model emulation. Retrieved from https://arxiv.org/abs/2210.04001v2

Plougonven, R., de la Cámara, A., Hertzog, A., & Lott, F. (2020). How does knowledge of atmospheric gravity waves guide their parameterizations? *Quarterly Journal of the Royal Meteorological Society*, *146*(728), 1529–1543. https://doi.org/10.1002/qj.3732

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. https://doi.org/10.1175/JAS-D-22-0138.1

Polichtchouk, I., Wedi, N., & Kim, Y.-H. (2022). Resolved gravity waves in the tropical stratosphere: Impact of horizontal resolution and deep convection parametrization. *Quarterly Journal of the Royal Meteorological Society*, *148*(742), 233–251. https://doi.org/10.1002/qj.4202

Skamarock, W. C. (2004). Evaluating mesoscale NWP models using kinetic energy spectra. *Monthly Weather Review*, *132*(12), 3019–3032. https://doi.org/10.1175/MWR2830.1

Song, B.-G., Chun, H.-Y., & Song, I.-S. (2020). Role of gravity waves in a vortex-split sudden stratospheric warming in January 2009. *Journal of the Atmospheric Sciences*, *77*(10), 3321–3342. https://doi.org/10.1175/JAS-D-20-0039.1

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., et al. (2019). DYAMOND: The DYnamics of the atmospheric general circulation modeled on Non-hydrostatic domains. *Progress in Earth and Planetary Science*, *6*(1), 61. https://doi.org/10.1186/s40645-019-0304-z

Sun, Y. Q., Pahlavan, H. A., Chattopadhyay, A., Hassanzadeh, P., Lubis, S. W., Alexander, M. J., et al. (2024). Data imbalance, uncertainty quantification, and transfer learning in data-driven parameterizations: Lessons from the emulation of gravity wave momentum transport in WACCM. *Journal of Advances in Modeling Earth Systems*, *16*(7), e2023MS004145. https://doi.org/10.1029/2023MS004145

Ukkonen, P., & Chantry, M. (2024). Representing sub-grid processes in weather and climate models via sequence learning. https://doi.org/10.22541/essoar.172098075.51621106/v1

van Niekerk, A., & Vosper, S. (2021). Towards a more "scale-aware" orographic gravity wave drag parametrization: Description and initial testing. *Quarterly Journal of the Royal Meteorological Society*, *147*(739), 3243–3262. https://doi.org/10.1002/qj.4126

Voelker, G. S., Bölöni, G., Kim, Y.-H., Zängl, G., & Achatz, U. (2023). MS-GWaM: A 3-dimensional transient gravity wave parametrization for atmospheric models (no. arXiv:2309.11257). *arXiv*.

Wang, P., Yuval, J., & O'Gorman, P. A. (2022). Non-local parameterization of atmospheric subgrid processes with neural networks. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS002984. https://doi.org/10.1029/2022MS002984

Wedi, N. P., Polichtchouk, I., Dueben, P., Anantharaj, V. G., Bauer, P., Boussetta, S., et al. (2020). A baseline for global weather and climate simulations at 1 km resolution. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002192. https://doi.org/10.1029/2020MS002192

Wei, J., Zhang, F., Richter, J. H., Alexander, M. J., & Sun, Y. Q. (2022). Global distributions of tropospheric and stratospheric gravity wave momentum fluxes resolved by the 9-km ECMWF experiments. *Journal of the Atmospheric Sciences*, *79*(10), 2621–2644. https://doi.org/10.1175/JAS-D-21-0173.1

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., et al. (2020). A comprehensive survey on transfer learning (no. arXiv:1911.02685). *arXiv*. https://doi.org/10.48550/arXiv.1911.02685

Zhuang, J., Dussin, R., Huard, D., Bourgault, P., Banihirwe, A., Raynaud, S., et al. (2024). Pangeo-data/xESMF: V0.8.8. *Zenodo*. https://doi.org/10.5281/zenodo.14025505

References From the Supporting Information

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization (no. arXiv:1412.6980). *arXiv*. https://doi.org/10.48550/arXiv.1412.6980

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.