JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2025MS005075

# Finetuning AI Foundation Models to Develop Subgrid-Scale Parameterizations: A Case Study on Atmospheric Gravity Waves

**Aman Gupta[1], Aditi Sheshadri[1], Sujit Roy[2,3], Johannes Schmude[4], Vishal Gaur[3], Wei Ji Leong[5], Manil Maskey[2], and Rahul Ramachandran[2]**

[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]IBM Research, Yorktown, NY, USA, [5]Development Seed, Washington, DC, USA

**Abstract** Global climate models parameterize a range of atmospheric-oceanic processes, including gravity waves (GWs), clouds, moist convection, and turbulence, that cannot be sufficiently resolved. These subgrid-scale closures for unresolved processes are a substantial source of model uncertainty. Here, we present a new approach to developing machine learning (ML) parameterizations of small-scale climate processes by fine-tuning a pre-trained AI foundation model (FM). FMs are largely unexplored in climate research. A pre-trained encoder-decoder from a 2.3 billion parameter FM (NASA and IBM Research's Prithvi WxC)—which contains a latent probabilistic representation of atmospheric evolution—is fine-tuned (or reused) to create a deep learning parameterization for atmospheric gravity waves (GWs); a process unseen during pre-training. The parameterization captures GW effects for a coarse-resolution climate model by learning the fluxes from an atmospheric reanalysis with 10 times finer resolution. A comparison of monthly averages and instantaneous evolution with a machine learning model baseline (an Attention U-Net) reveals superior predictive performance of the FM parameterization throughout the atmosphere, even in regions excluded during pre-training. This performance boost is quantified using the Hellinger distance, which is 0.11 for the baseline and 0.06 for the fine-tuned model. Our findings emphasize the versatility and reusability of FMs, which could be used to accomplish a range of atmosphere- and climate-related applications, leading the way for the creation of observations-driven and physically accurate parameterizations for more earth system processes.

**Plain Language Summary** Climate models struggle to accurately capture the physical effects of small-scale atmospheric processes like gravity waves, turbulence, and clouds, which are critical to accurately predicting future climate states. These processes evolve on scales finer than typical model grid resolutions. As a result, they continue to rely on approximations, known as physical parameterizations, to represent their missing effects. The use of parameterizations introduces uncertainty and makes climate predictions less reliable. Here, we propose a new approach to improving these parameterizations using modern advances in deep learning. Specifically, we use Prithvi WxC, a large AI model trained on multiple decades of one reanalysis, and fine-tune it using limited years of gravity wave (GW) data from another reanalysis to develop an emulator capable of predicting a physically consistent atmospheric GW flux evolution. The novel approach of leveraging a large AI model pre-trained on vast volumes of atmospheric data and augmenting it with limited process-specific data allows the creation of compact and easily trainable data-driven physical parameterizations. While we focus on gravity waves, our approach is flexible and can be generalized to developing data-driven parameterizations of other earth system processes.

Aman Gupta and Sujit Roy contributed equally to this work.

**Key Points:**

- AI weather foundation models can be fine-tuned to create climate model parameterizations for subgrid-scale processes like gravity waves
- The fine-tuned machine learning (ML) parameterization beats existing ML benchmarks, predicting accurate wave fluxes and variability
- Despite predicting accurate monthly averages and strong wave events, ML models continue to struggle with the prediction of small flux values

**Correspondence to:**
A. Gupta and S. Roy,
ag4680@stanford.edu;
sujit.roy@uah.edu

**Citation:**
Gupta, A., Sheshadri, A., Roy, S., Schmude, J., Gaur, V., Leong, W. J., et al. (2025). Finetuning AI foundation models to develop subgrid-scale parameterizations: A case study on atmospheric gravity waves. *Journal of Advances in Modeling Earth Systems*, *17*, e2025MS005075. https://doi.org/10.1029/2025MS005075

Received 8 MAR 2025
Accepted 22 OCT 2025

**Author Contributions:**
**Conceptualization:** Aman Gupta, Aditi Sheshadri
**Data curation:** Aman Gupta, Sujit Roy, Vishal Gaur, Wei Ji Leong
**Formal analysis:** Aman Gupta, Sujit Roy
**Funding acquisition:** Aditi Sheshadri, Manil Maskey, Rahul Ramachandran
**Investigation:** Aman Gupta
**Methodology:** Aman Gupta, Sujit Roy
**Project administration:** Aditi Sheshadri, Rahul Ramachandran



## 1. Introduction

Accurate prediction of future climate is a trillion-dollar challenge with critical consequences for the world economy, food security, global health, and urban planning. Currently, state-of-the-art climate projections are highly uncertain, and much of the inherent model uncertainty stems from approximations made in subgrid-scale parameterizations (Lee et al., 2023; Morrison & Lawrence, 2020). For instance, it has been suggested that model uncertainty accounts for 98% of the total uncertainty in precipitation projections (Wu et al., 2022). This study aims to demonstrate the untapped potential of AI foundation models (FMs) to improve traditional numerical climate models by facilitating the creation of subgrid-scale parameterizations.

**Resources:** Aditi Sheshadri, Sujit Roy, Johannes Schmude, Manil Maskey, Rahul Ramachandran
**Software:** Sujit Roy, Johannes Schmude, Vishal Gaur, Wei Ji Leong
**Supervision:** Aditi Sheshadri
**Validation:** Aman Gupta
**Visualization:** Aman Gupta, Sujit Roy
**Writing – original draft:** Aman Gupta, Sujit Roy
**Writing – review & editing:** Aman Gupta, Aditi Sheshadri, Sujit Roy, Rahul Ramachandran

FMs can be broadly defined as task-agnostic large AI models which are pre-trained using a self-supervised learning objective (Bommasani et al., 2022), such as learning weather evolution from time $t$ to $t + \Delta t$. Realizing the single-task limitation of existing AI weather forecasting models (Bi et al., 2023; Lam et al., 2023; Price et al., 2025), despite their massive compute requirements, FMs are developed to be versatile and present the next frontier in AI research. Pre-trained FMs are subsequently fine-tuned to perform a broad range of sub-tasks, a.k.a., downstream tasks. FMs are largely unexplored in climate science, and only a couple of weather and geospatial FMs exist to date: AtmoRep (Lessig et al., 2023), Aurora (Bodnar et al., 2025), and Prithvi HLS (Jakubik et al., 2023). To our knowledge, only weather-related downstream applications of FMs have been explored thus far, including hurricane track and intensity prediction, air quality predictions, downscaling, vegetation burn-scar detection, etc.

Here, we use a recently developed, state-of-the-art FM, Prithvi WxC (Schmude et al., 2024) (hereafter Prithvi), to demonstrate a climate-related application of FMs, that of developing deep learning parameterizations for unresolved earth system processes for climate models. The parameterization for atmospheric gravity waves (GWs) presented here is capable of representing the missing effects of atmospheric GWs in global climate models. We blend the pre-trained encoder-decoder pair from Prithvi with high-resolution GW momentum flux data (see Section 2) to create a fine-tuned AI model that skillfully predicts subgrid-scale GW activity and outperforms existing benchmarks (Gupta, Sheshadri, Roy, et al., 2024) for deep-learning-based GW flux prediction. The study motivates and calls for the strategic use of FMs for climate-related tasks by demonstrating how to leverage observations and FMs to efficiently achieve predictive tasks that might otherwise require much larger volumes of training data.

Atmospheric GWs are ubiquitous multiscale (spatial scale $\mathcal{O}(1)$-$\mathcal{O}(1{,}000)$ km) oscillations generated by atmospheric convection, jet stream disturbances, geostrophic imbalance, and flow over mountains (Fritts & Alexander, 2003). GWs dynamically couple different layers of the atmosphere by carrying near-surface momentum and energy to stratospheric and mesospheric heights. In the troposphere, GWs play a critical role in setting the location and strength of the jet streams (Palmer et al., 1986). In the stratosphere, they influence the quasi-biennial oscillation of tropical winds (Giorgetta et al., 2002), and the springtime breakdown of the Antarctic polar vortex (Gupta et al., 2021). In the mesosphere, GWs are the primary driver of the pole-to-pole overturning circulation (Becker, 2012). GW-induced cold anomalies in the polar winter stratosphere provide suitable conditions for the formation of polar stratospheric clouds, enabling reactions that promote the destruction of ozone (Dörnbrack et al., 1999; Hoffmann et al., 2017; Höpfner et al., 2006). Aside from their influence on climate variability, GW-induced clear air turbulence can influence commercial air travel and is believed to have caused the sudden plunging of Singapore Airlines flight SQ321 on 21 May 2024 (Hirschfeld, 2024).

The current climate model grid resolution (50–100 km) is insufficient to fully resolve dynamically important processes like GWs, clouds, and turbulence. The traditional approach to represent these missing processes has been to couple the numerical fluid solver with a suite of *sub-grid scale parameterizations* to approximately capture the unresolved effects of these processes (Alexander & Dunkerton, 1999; Lott & Miller, 1997; Bogenschutz et al., 2012; Iacono et al., 2000, to name a few).

Parameterizations are often not well constrained by observations and, for computational reasons, have simplified assumptions that compromise their physical accuracy. For GWs, these assumptions include an idealized source spectrum and, generally, complete neglect of their transient evolution and horizontal propagation (Achatz et al., 2024). Further, their parametric tuning is often sub-optimal because the parameters are optimized to replicate only certain atmospheric features of interest. These inductive biases (due to simplifying assumptions) often add up and result in inaccurate model dynamics, such as the prominent "cold-pole bias" (McLandress et al., 2012), leading to large uncertainties in future climate projections (Golaz et al., 2013; Mauritsen et al., 2012; Zhao et al., 2018).

Data-driven approaches are increasingly being used to develop fast GW flux (GWF) emulators for 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). These emulators complement existing efforts to develop nonlocal GW parameterization using physics-based approaches (Eichinger et al., 2023; Voelker et al., 2023). Despite being effective, these emulators are trained on parameterization data itself and do not offer an improved process physics representation. Here, we fine-tune the FM on *resolved* GWFs. Training on resolved GWFs allows the neural networks to learn key physical effects of GWs directly from fine-tuning data sets.

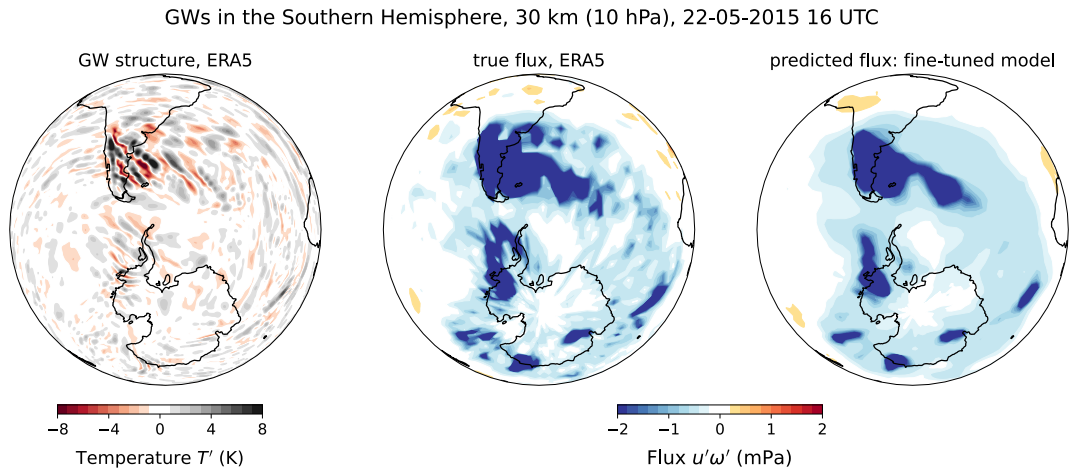


**Figure 1.** Predictions from the fine-tuned gravity wave (GW) parameterization. The left plot shows the temperature structure of GWs over the Drake Passage, as seen in ERA5 reanalysis (Hersbach et al., 2020). Temperature perturbations $T'$ were computed by removing the large scales, here defined as the first 21 total wavenumbers. The middle and right plots show the true and predicted momentum flux carried by the waves. 30 km is an approximate representative height since the fluxes are evaluated on a pure pressure level. Here (and throughout the study), "true" flux refers to the flux derived from the ERA5 reanalysis, and the predicted flux is the prediction from the machine learning models trained on ERA5. Almost all GWs in ERA5 are model-generated. Therefore, the GW structure and the inferred flux might not be a precise representation of the actual atmospheric conditions.

Our fine-tuned parameterization, created by blending Prithvi and ERA5 reanalysis skillfully predicts the resolved GW momentum fluxes for a provided background atmospheric state (as shown in Figure 1). The GW structure on 22 May 2015 and the momentum flux carried by the waves are shown in Figure 1. The fine-tuned model accurately predicts the fluxes over the Drake Passage and the Southern Ocean. The fluxes over the Andes extend sufficiently leeward (up to 80° longitude) over the Southern Ocean, indicating that the fine-tuned model can learn and represent the lateral propagation and transient evolution of the generated waves; a physical feature absent in most current GW parameterizations (Plougonven et al., 2020).

This fine-tuned parameterization for GWs can be coupled to a coarse-resolution climate model to represent "missing" GW effects. Since Prithvi was pre-trained on key atmosphere-ocean-land variables, the scope of this approach transcends GWs, and fosters and expedites the creation of physically accurate AI parameterizations of other small-scale earth system processes, ultimately contributing to the development of accurate and interpretable hybrid climate prediction systems.

## 2. Methods

### 2.1. The Prithvi WxC Foundation Model for Weather and Climate

Prithvi WxC, jointly developed by NASA and IBM Research, is a transformer-based deep learning architecture that combines features from several recent transformer architectures to effectively process regional and global dependencies of the input data and to efficiently process longer sequence lengths of tokens. Any image input to a transformer is broken down into smaller square patches that are then projected to a higher-dimensional space to represent the image in numerical space. These projections, which represent discrete amounts of information, are referred to as a token. This allows the model to, for instance, run in different spatial contexts or infuse additional tokens (i.e., adding more information as tokens into later stages of the model instead of the input to preserve or enhance context) from off-grid measurements into the model during fine-tuning. Prithvi has 2.3 billion trainable parameters and is trained on 160 data channels (10 variables over 14 pressure levels and 20 surface variables) using 40 years of 3-hourly MERRA-2 reanalysis (Gelaro et al., 2017) data at a $0.5° \times 0.625°$ spatial resolution. The channels include 20 surface variables (winds, pressure, latent heat flux, surface roughness, etc.) and 10 atmospheric variables (winds, clouds, humidity, etc.) on 14 vertical pressure levels each. These variables are tabulated in Tables 2 and 3 in Appendix A of Schmude et al. (2024).

The validation of Prithvi extends from zero-shot evaluations for reconstruction and forecasting to other downstream tasks, such as the downscaling of weather and climate models, the prediction of hurricane tracks, and climate model parameterization. The architecture of the pre-training backbone is shown in Figure 2a. As shown in the figure, Prithvi was trained on a masked reconstruction objective, which means that in addition to minimizing the root mean square for the predictions, the model also minimized reconstruction error from masked input data. A fixed fraction (50%) of the input cells were masked, and the model was tested on how well it could fully reconstruct the global field from the masked data. More details are provided in Schmude et al. (2024), where Equation 1 and Section 2.5.1 focus on masked reconstruction.

The fine-tuning task presented in this manuscript is identical to that presented in Section 3.2 of Schmude et al. (2024), that is, the "Climate Model Parameterization for GW Flux" task. Schmude et al. (2024) only briefly showcase it as one among many applications of an AI foundation model, but here, we delve deeper and provide a full detailed analysis of the task.

### 2.2. Preparing Training Data for GW Flux Prediction

The fine-tuning data for GW flux prediction was prepared using ERA5 global reanalysis data (Hersbach et al., 2020) retrieved at a $0.25° \times 0.25°$ horizontal resolution, 137 vertical levels, and at an hourly frequency. The effective resolution of GWs in ERA5 depends on the truncation of the underlying model in spectral space, which includes spherical harmonics up to total wavenumber 639, and its native N320 reduced Gaussian grid with ∼31 km resolution. Given the need to damp small-scale motions for numerical stability, waves in ERA5 are poorly represented on scales below ∼150 km. Since the model output is interpolated and publicly presented on a 0.25° latitude-longitude grid, which corresponds to roughly 25 km around the Equator, we refer to ERA5 as having 25 km resolution, but expect GWs to be accurate on scales of 150 km and larger.

We aim to represent the missing GW fluxes in a coarse-climate model by learning it from a higher resolution data set (ERA5) that resolves a substantial portion of the mesoscale GW spectrum. Therefore, we select a target model resolution of 280 km, which is an order of magnitude coarser than ERA5's underlying forecast model. A climate model at this resolution typically captures some large GWs and employs numerical parameterizations to represent the remaining unresolved GWs. So, we first filter out the large-scale non-divergent motions in ERA5, attempting to capture only the "unresolved" portion of the GW spectrum. Hence, we compute the resolved GW fluxes in ERA5 by applying a Helmholtz decomposition (HD) (Köhler et al., 2023; Lindborg, 2015) on the raw output as follows. First, the horizontal winds ($u$ and $v$) are decomposed into rotational and divergent components:

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

where $\phi$ is the potential function such that $\nabla\phi$ is irrotational. Similarly, $\psi$ is the rotational streamfunction such that $\nabla \times \psi$ is non-divergent. $\phi$ and $\psi$ are used to reconstruct the divergent (div) and rotational (rot) parts of the horizontal flow as:

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

The target climate model with ∼280 km resolution could resolve GWs with wavelengths greater than ∼1,400 km. To remove these "resolved" large GWs (including equatorial Kelvin waves) from the small-scale flux estimate, we apply an additional T21 high-pass filter on the divergent velocity field. This operation is expressed as:

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

These are multiplied with the zonal mean removed pressure velocity anomaly ($\omega'$) to compute the directional GW momentum fluxes:

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

which we aim to learn using the machine learning (ML) models. Here, $g = -9.81$ m/s$^2$ is the acceleration due to gravity. Hereafter, we use the shorthand notation $u'\omega'$ and $v'\omega'$ to denote the directional fluxes in Equation 4.

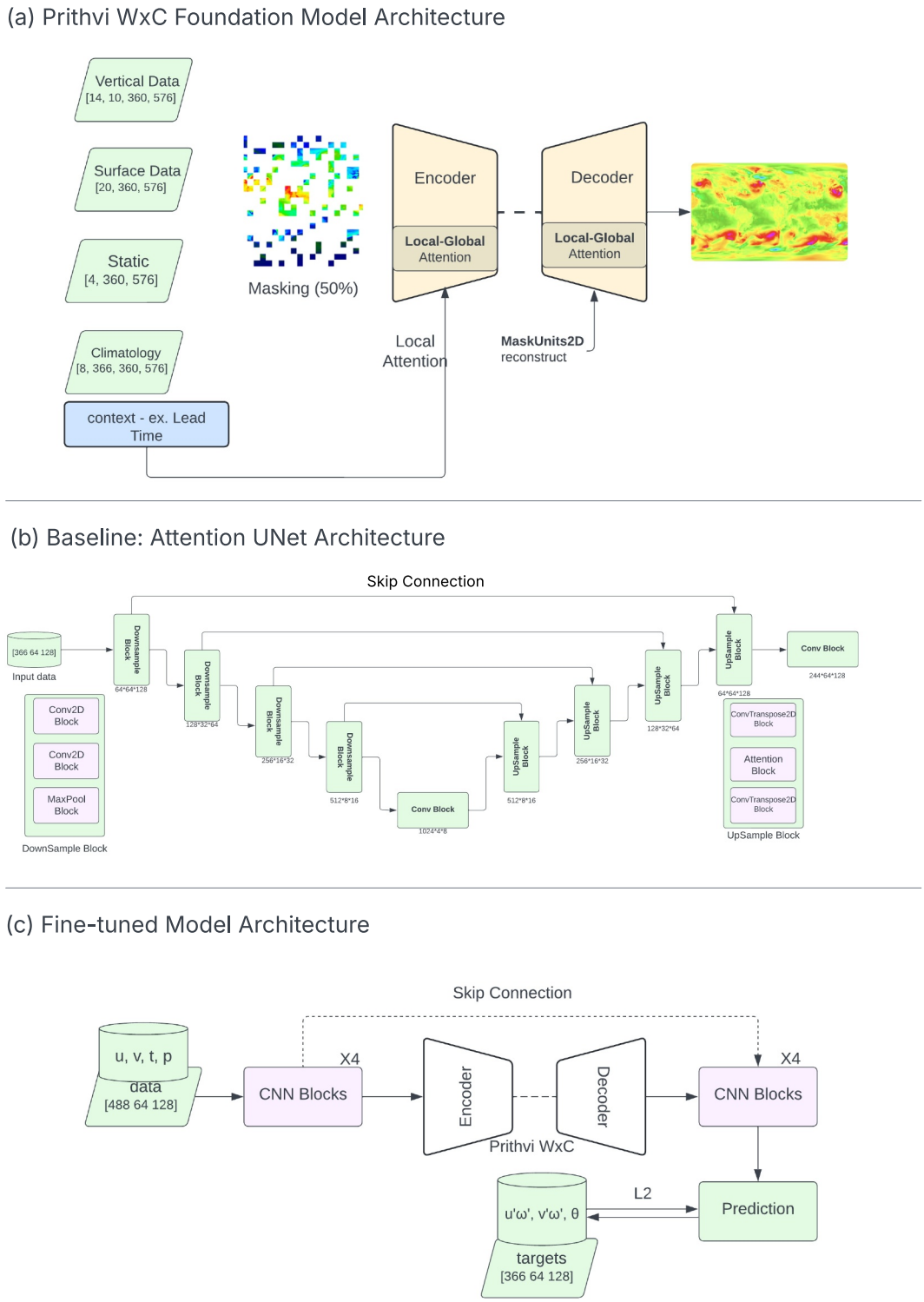


**Figure 2.** (a) Pre-training model architecture for Prithvi WxC. The encoder and decoder blocks from Prithvi are frozen and used for fine-tuning. 10 atmospheric variables on 14 vertical levels, 20 surface variables, 4 static variables, and 8 climatology variables for 366 days in a year, all on a 360 (lat) × 576 (lon) grid, for the input. (b) Model Architecture for attn unet using 3 input variables, each on 122 vertical levels and a 64 (lat) × 128 (lon) grid (schematically identical to Oktay et al. (2018)). (c) The foundation model (FM) fine-tuning architecture comprises (in order) 4 learnable convolutional layers, the frozen encoder, the frozen decoder, and 4 more learnable convolutional layers. A skip connection connects the former and latter convolutional layers. Takes four input variables, each on 122 vertical levels, and on a 64 (lat) × 128 (lon) grid. The blue block in the bottom left in (a) refers to the additional infused context or relevant information added at later stages in the hidden layers, for example, the lead time at which the FM makes the predictions.

The procedure is applied to create the fine-tuning training data. The top 15 of the 137 vertical levels are discarded due to artificial model damping. All input-output pairs are coarse-grained from a 0.25° km resolution to a 64 latitudes × 128 longitudes grid (roughly 2.8° ≈ 280 km resolution in the tropics) to obtain conservative wave averages (as the momentum flux carried by the waves is defined as an average over single or multiple wave cycles). The fluxes are computed for four years: 2010, 2012, 2014, and 2015. This corresponds to roughly 35k training + validation samples since one 64 × 128 × 122 hyperslab makes up 1 training sample. We are thus in a data-scarce regime, in which the number of observations is insufficient to cover the possible degrees of freedom.

*Variables for training the U-Net*: the input consists of winds $u$, $v$, and potential temperature $\theta$, which is a function of temperature $T$ and pressure $p$ (in hPa) as $\theta = T(p/1000)^{-0.286}$, each on 122 vertical levels, 64 latitudes and 128 longitudes. Similarly, the output comprises fluxes $u'\omega'$ and $v'\omega'$, each on 122 vertical levels, 64 latitudes, and 128 longitudes (Figure 2b).

*Variables for fine-tuning the FM*: this is slightly different from the baseline. The fine-tuning input consists of winds $u$, $v$, temperature $T$, and pressure $p$ (instead of $u$, $v$, and $\theta$), each on 122 vertical levels, 64 latitudes, and 128 longitudes. Similarly, the outputs are potential temperature $\theta$ (for validation) and fluxes $u'\omega'$ and $v'\omega'$, each on 122 vertical levels, 64 latitudes and 128 longitudes (Figure 2c). Using $T$ and $p$ as inputs and $\theta$ as an output also allowed us to test whether the fine-tuned model can learn a well-defined nonlinear analytical relation between the input and the output, arguably presenting a more stringent learning problem compared to the baseline.

*Variable Normalization*: Each variable is normalized differently. The zonal wind $u$ is normalized as: $u \rightarrow (u - u_{mean})/u_{std}$, where $u_{mean}$ and $u_{std}$ are the global mean and standard deviation. Similarly for $v$ and $T$. Pressure was scaled as $p \rightarrow \log_{10}(p)$, and potential temperature was scaled as $\theta \rightarrow \theta/1000$. Lastly, global mean and global standard deviations of $u'\omega'$ were used to scale the flux as $u'\omega' \rightarrow [(u'\omega' - u'\omega'_{mean})/u'\omega'_{std}]^{1/3}$. Here, applying a cuberoot helps constrain the range of flux magnitudes by shifting both inordinately large and inordinately small flux values toward 1. For instance, the cuberoot of 0.064 and 64 (normalized) GW flux is 0.4 and 4, respectively. The cuberoot makes it more convenient to constrain and learn extreme values by bringing them closer to 1.

All the data-driven models considered in this study are trained on resolved wave fluxes from ERA5. The objective is to reproduce the ERA5 fluxes as accurately as possible. For this reason, the fluxes in ERA5 are occasionally referred to as "true" fluxes, since they comprise the training and validation set.

ERA5 provides multi-decadal atmospheric coverage at a moderately high resolution; however, we caution against the limited GW representation in ERA5, due to which GW fluxes in ERA5 might not be a true representation of the actual GW fluxes in the atmosphere. Multiple recent studies have reported both substantial similarities and systematic differences between GWs in ERA5 and GWs in high-resolution models and observations (Gupta, Reichert, et al., 2024; Lear et al., 2024; Pahlavan et al., 2023; Yoshida et al., 2024). This could be due to multiple factors. First, with a resolution of 25 km, ERA5 does not resolve a portion of the atmospheric GWs with wavelengths shorter than 150 km. These waves likely make notable contributions to the large-scale atmospheric circulation (Polichtchouk et al., 2022, 2023). Second, while the large-scale winds and temperature are constrained by observations to some degree, small-scale GWs in ERA5 are model-generated in response to the constrained background state. Third, known biases in precipitation, clouds, land, and upper surface winds can result in biased GW generation in response to changes in these fields. This can be particularly important for small-scale convectively generated GWs which have a wide phase spectrum, and are likely to transport the momentum to mesospheric heights before dissipation (Achatz et al., 2024; Kim & Chun, 2015). Lastly, the use of a hydrostatic dynamical core to produce ERA5 means a compromised representation of non-hydrostatic GWs, potentially leading to an incorrect wave aspect ratio for a given angular frequency. Such differences could also exist among identically initialized high-resolution models with different underlying numerics, as noted by Stephan et al. (2019), Kruse et al. (2022), and Procházková et al. (2023).

### 2.3. Baseline Model

An advanced baseline was created by training an Attention U-Net model (hereafter attn unet) (Oktay et al., 2018) on the ERA5 data. The input is downsampled using four convolutional blocks and then upsampled using four convolutional blocks. The skip connection at each level comprises learnable attention layers. For every downsample (upsample), the number of channels increases (decreases) by a factor of 2, but all spatial dimensions

reduce (increase) by a factor of 2. As a result, the baseline model consists of over 35 million learnable parameters and provides a robust comparison benchmark for the fine-tuning model. The learning rate for the model was set to $10^{-4}$. On a single 80 GB A100 GPU, the model needed around 110 hr for 100 epochs of training.

### 2.4. Designing the Fine-Tuning Model

The architecture schematic for the fine-tuning is shown in Figure 2c. During fine-tuning, we freeze the encoder and decoder from Prithvi WxC. The frozen encoder is preceded by 4 learnable convolutional blocks, each with an increasing number of hidden channels, that is, $C$, $2C$, $4C$, and then $8C$, where $C = 160$. Likewise, the frozen decoder is succeeded by 4 new learnable convolutional blocks. For instantaneous prediction of GW fluxes, we fix Prithvi's lead time $\delta t$ to zero. The instantaneous model input for fine-tuning has the shape [1, 488, 64, 128] where the 488 channels comprise the four background variables $u$, $v$, $T$, and $p$ on 122 vertical levels each, and on a $64 \times 128$ horizontal grid, as discussed above. The model was fine-tuned to produce an output with shape [1, 366, 64, 128] comprising of the potential temperature $\theta$, and fluxes $u'\omega'$, and $v'\omega'$ on 122 vertical levels each. The model was trained for 26 hr on 2 nodes of 4 80 GB A100 GPUs for 100 epochs. However, the model error converged to lower than the final baseline model error after just 40 epochs of training.

### 2.5. Training Both Models

Both models use global information as input to predict global fluxes as output. This provides a strong contrast to traditional "single-column" parameterizations. Access to the global atmospheric state allows the models to learn spatio-temporal correlations and the effects of horizontal propagation of GWs.

Both models were trained and validated for 100 epochs on 4 years, that is, 48 months, of ERA5 background conditions and fluxes. The baseline model's global RMSE loss dropped from an epoch 1 loss of 0.38 to plateauing near 0.17 over 100 epochs (a 40% reduction). In contrast, the fine-tuned model showed much faster convergence, dropping from an epoch 1 loss of 0.275 to 0.16 (40% reduction) over just 5 epochs and finally converging to 0.106.

Since the main focus of the study is to highlight the application of FMs to make quick emulators, at present, only the month of May 2015 was used for validation; the remaining 47 months were used for training. Both models leveraged a *U-Net-like* architecture with skip connections to promote the extraction of high-frequency information from the source data. Both models were trained with an identical minibatch size of 4, that is, four randomly selected timeframes of each variable formed input during a single forward and backward pass of the model. We re-emphasize that Prithvi WxC was pre-trained on the MERRA-2 data set, but the fine-tuning was accomplished using ERA5 data instead. Both models yielded similar inference times on a single A100 GPU for an identical minibatch size of 4.

Both models were optimized using MSE Loss, which is defined as:

$$\mathcal{L}(\vec{x},\vec{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - y_i\right)^2 \tag{5}$$

where $x_i$ is the $i$th prediction compared against the $i$th measured sample $y_i$.

### 2.6. Hellinger Distance

Given two probability densities, $p$ and $q$, their Hellinger distance, $\mathcal{H}$ (Hellinger, 1909), is defined as:

$$\mathcal{H}(p,q) = 1 - \int_{x\in X}\sqrt{p(x)q(x)}dx. \tag{6}$$

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.

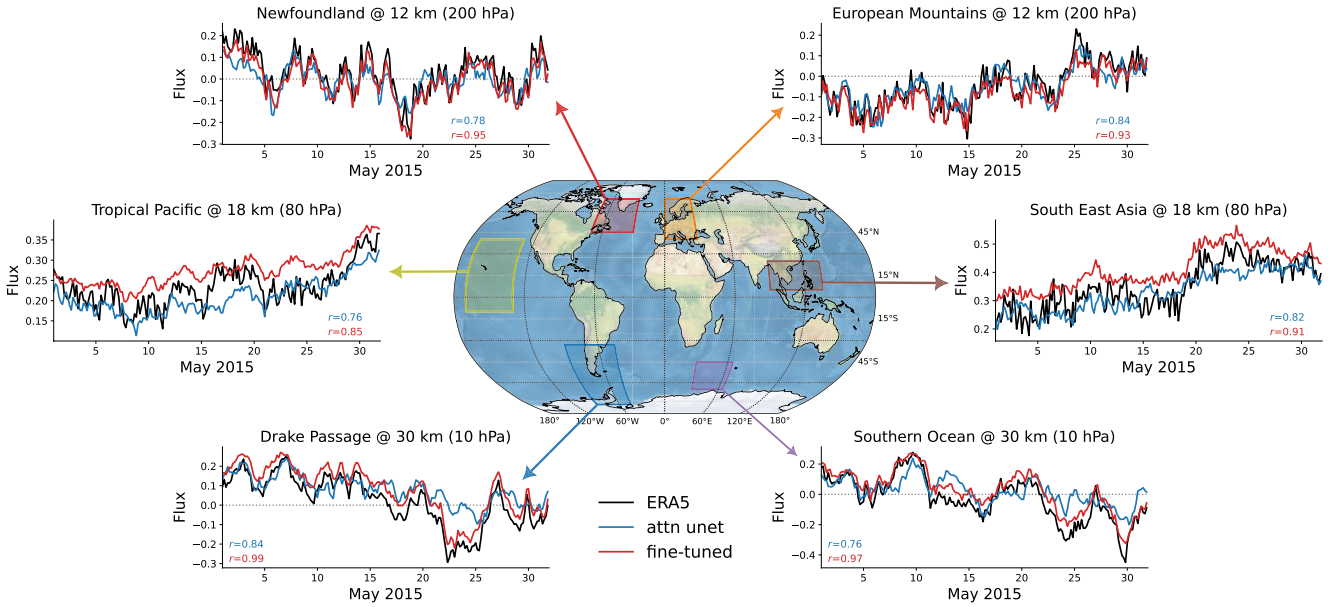


**Figure 3.** Instantaneous (non-dimensional/normalized) fluxes for May 2015 resolved in ERA5 reanalysis (black), predicted fluxes from attn unet (blue), and from the fine-tuned parameterization (red), over six well-known gravity wave (GW) hotspots. The numbers show the respective Pearson correlation coefficients with respect to ERA5. The fluxes in the winter hemisphere are shown at 30 km, whereas the fluxes in the summer hemisphere are shown at 12 km, as GW activity in the summer stratosphere is substantially weaker. Fluxes in the tropics are shown at 18 km. These altitudes are approximate representative heights since the fluxes are evaluated on pure-pressure and hybrid-pressure levels, respectively, which do not equate to similar geopotential heights throughout the domains. The Pearson correlation coefficient (between ERA5 and attn unet) is computed as the covariance between the ERA5 fluxes and attn unet fluxes divided by the product of ERA5 and attn unet standard deviations.

Hellinger distance measures the statistical distance between two distributions. In Section 3, Hellinger distance is used to quantify the difference (or statistical distance) between the flux distributions from ERA5 and the prediction flux distributions to estimate the quality of predictions by both the attn unet and the fine-tuned model.

## 3. Results

### 3.1. Instantaneous, Intermittent Evolution of Gravity Waves

We focus on predicting $u'\omega'$, which is the vertical flux of zonal momentum carried by GWs. Its vertical derivative equals the net forcing tendency (acceleration) exerted by GWs on the zonal wind. The findings are similar for the vertical flux of meridional momentum, $v'\omega'$, and equivalent plots for $v'\omega'$ are shared in the Appendix. In all instances, the predictions are compared to both the fluxes from ERA5 and to predictions from the existing benchmark, the attn unet model.

The time evolution of box-averaged fluxes for May 2015 over six well-known hotspots of GW activity is illustrated in Figure 3. The fine-tuned parameterization generates a remarkably accurate prediction of the intermittent generation and temporal coherence of GW packets, even though no explicit considerations were made to embed recurrence in the underlying fine-tuning architecture. The three predominantly orographic hotspots (Newfoundland, European Mountains, and Drake Passage) and three nonorographic hotspots (the tropical Pacific Ocean, Southeast Asia, and the Southern Ocean) were selected using the zonal GW flux and lateral GW flux climatology presented in Hindley et al. (2020), Wei et al. (2022), and Gupta, Sheshadri, Alexander and Birner (2024). Nonlocal propagation of GWs is more prominent in the winter stratosphere due to a stronger vertical shear (Gupta, Sheshadri, Alexander, & Birner, 2024; Sato et al., 2012), so wherever possible, the transient evolution is shown in the upper winter stratosphere (10 hPa $\sim$ 30 km), that is, the Southern Hemisphere for May. For regions in the summer/Northern hemisphere, the fluxes are instead analyzed in the upper troposphere (200 hPa $\sim$ 12 km). In the tropics, the GW fluxes are analyzed in the lower stratosphere (80 hPa $\sim$ 18 km) to ensure minimal contribution from convective fluxes.

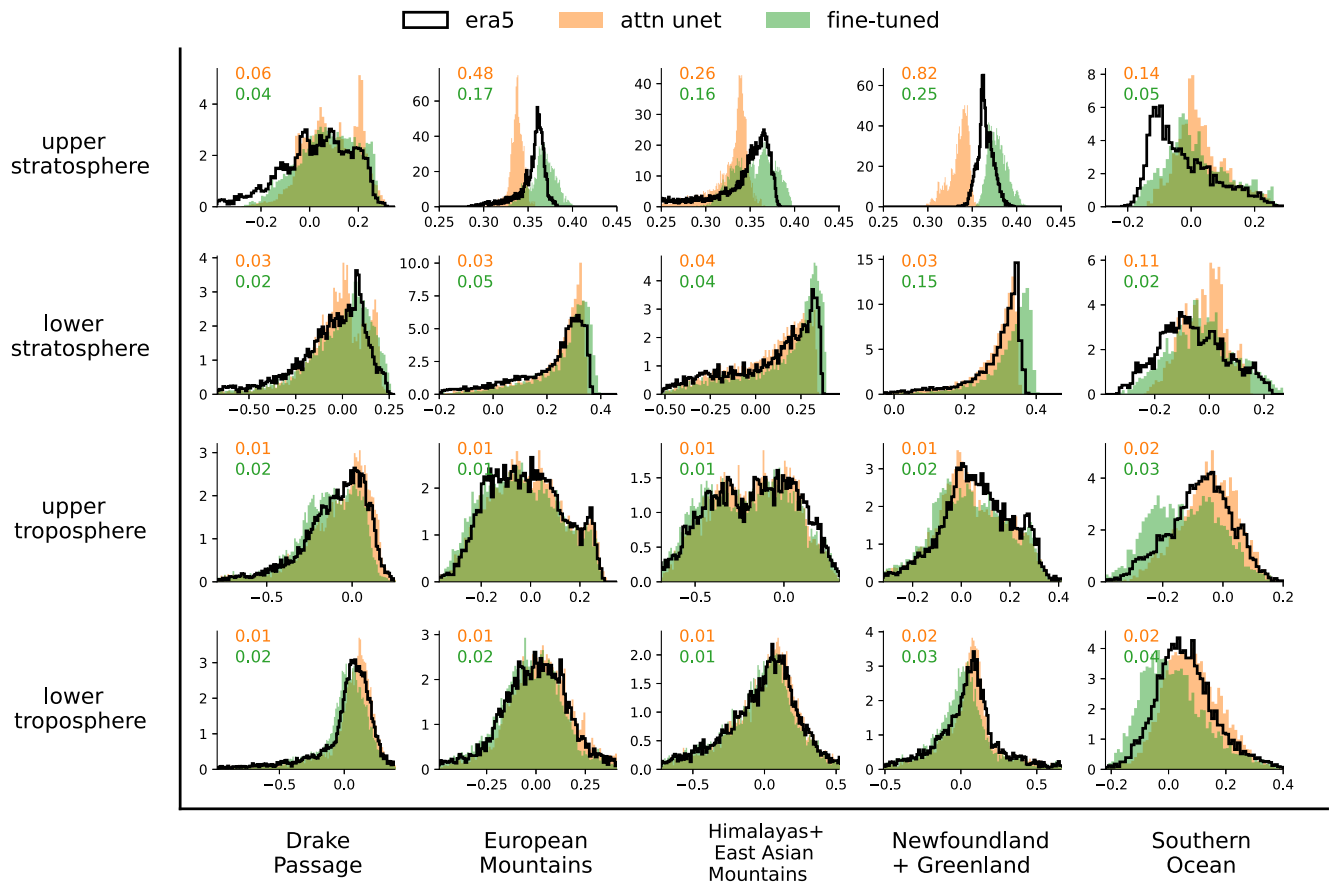


**Figure 4.** May 2015 averaged gravity wave momentum flux distributions divided according to hotspots and vertical regions in the atmosphere. The figure shows non-dimensional fluxes as predicted by the models for ease of comparison. The *y*-axis is the distribution density. The fluxes are averaged over the corresponding boxes outlined in A3. The numbers in orange and green indicate the Hellinger distances for the time-averaged flux distributions for the attn unet and the foundation model, respectively. Lower troposphere: 500 hPa to surface (0–10 km height), upper troposphere: 100–500 hPa (10–16 km height), lower stratosphere: 30–100 hPa (16–25 km height), upper stratosphere: 10–30 hPa (25–45 km height).

The fine-tuned FM generates substantially better predictions over all six hotspots. Most notably, for the Drake Passage (predominantly orographic waves) and the Southern Ocean (nonorographic waves), the Pearson correlation coefficients of the predictions from the fine-tuned model (vs. ERA5) are as high as 0.99 and 0.97, respectively. In comparison, the respective correlations for the attn unet are 0.84 and 0.76. The correlation with ERA5 is the weakest over the Tropical Pacific Ocean, but even then, the fine-tuned model has a higher correlation of 0.85, higher than attn unet's 0.76. The results in the lower stratosphere are mixed. Even though the fine-tuned model has a higher correlation over the tropical box, flux magnitudes from attn unet match better with ERA5. Noisier fluxes due to tropical convective GWs with a broad range of phase speeds appear to be more challenging to predict than extratropics GWs. Expanding the feature set to include diabatic heating or precipitation-related information could potentially lead to performance gains in the region.

The successful prediction of spontaneous bursts of flux intensification in both the tropics (from tropical storms and convective systems) and the midlatitudes (from mountains and storm tracks) shows that the fine-tuned model proficiently learns the intermittent excitation and horizontal evolution of medium-to-small-scale atmospheric variability directly from data. This is further corroborated by the spatial structure of the predicted flux in Figure 1, which shows that the model predicts both the fluxes over the Southern Andes and the laterally propagated fluxes in its vicinity. The wave packets preserve their coherence in time as they propagate away from their sources of excitation (see the animation provided as Supporting Information, https://doi.org/10.17605/OSF.IO/8W6AZ).

### 3.2. Regionwise Averaged Flux Distribution

The dynamical evolution of atmospheric GWs can vary substantially with height (troposphere vs. stratosphere), region (latitude and longitude), and season (summer vs. winter). Figure 4 shows the monthly-averaged predicted and "true" GW flux distributions partitioned by individual hotspots and varying atmospheric altitudes. To also focus on tropical orography, here we show the flux distribution over the Himalayas and the East Asian mountains, instead of the tropical Pacific and Southeast Asian hotspots in Figure 3. The fine-tuned model captures the entire range of flux magnitudes over the different GW hotspots (Figure 4). In the troposphere and the lower stratosphere, the models provide comparable performance. In fact, in some regions, such as the lower stratosphere over Newfoundland, and the troposphere over the Southern Ocean, the Hellinger distances are slightly better for the attn unet model. In the upper stratosphere, however, the fine-tuned model generates a substantially more consistent distribution than the baseline. Such distributions are challenging to replicate, as the waves excited near the surface are progressively filtered and dissipated as the waves propagate to stratospheric and mesospheric altitudes.

The Hellinger distances of the distributions for both the baseline and the fine-tuned model (w.r.t. ERA5) are shown for each hotspot and height. A Hellinger distance of 0 indicates that the predicted distribution is identical to the distribution from ERA5. In the stratosphere, the fine-tuned model outperforms the baseline, yielding a lower Hellinger distance in all regions except the lower stratosphere over Newfoundland and the European Mountains. The improvement is more evident in the upper stratosphere. Both models generate low Hellinger distances in the troposphere and most of the lower stratosphere, indicating a distribution similar to ERA5, at least in a cumulative sense. However, all regions in the upper stratosphere have higher Hellinger distances than down below, with Hellinger distances reaching up to 0.82 for the baseline over Newfoundland, revealing key biases in the summer hemisphere. The GWs in the summer upper atmosphere are likely much smaller due to filtering below by the easterly winds. Since the RMSE training loss used for finetuning would penalize the large scales more, one might expect these smaller waves to be less accurately captured by the neural nets.

Most interestingly, the baseline model has a lower variance (and hence poorer predictive skill) than the fine-tuned model in multiple stratospheric blocks, even though Prithvi was initially not trained on upper atmospheric data; Prithvi's vertical spacing is shown in Figure A1. This highlights another benefit of using an FM's encoder-decoder that allows the creation of a consistent mapping between the FM's learned embedding space and the fine-tuning data. The performance improvement, then, can be attributed to a combination of two factors. First, the substantially higher volume (40+ years) of data used for pre-training, as opposed to merely four years of ERA5 data used for fine-tuning and training the baseline. Second, the fine-tuning model efficiently leverage the latent space of the pre-trained Prithvi and unify the learning from both MERRA2 during pre-training and ERA5 during fine-tuning. As a result, the fine-tuning model substantially outperforms the attn unet baseline when trained on the same set of fine-scale data. Despite not being trained on upper atmospheric data during development, training on over four decades of atmospheric data on a masked reconstruction objective (as described in Section 2.1) likely allows more consistent mappings between Prithvi's embedding space and the fine-tuning input.

A similar partition of the ERA5 and predicted monthly mean distributions, but partitioned by different latitude bands, is shown in Figure A4.

### 3.3. Vertical Mean Profile and Variability

While both models generate mean vertical profiles that are very similar to ERA5 over the five hotspots, the fine-tuned model generates both richer and more accurate variability in the stratosphere than the baseline (Figure 5). The difference in variability is substantial in the stratosphere. Both models generate weaker stratospheric variability than ERA5 in the summer stratosphere (European mountains, Himalayas, and Newfoundland) owing to weak GW activity in the region. The shading in Figure 5 shows the range of the true and predicted fluxes. Yet, the wintertime stratospheric variability over the Drake Passage and Southern Ocean in the fine-tuned model is more consistent with ERA5. This is consistent with the lower variance noted for baseline predictions in the upper stratosphere. The observed and predicted mean vertical flux profiles and their variability over different hotspots for the meridional flux are shown in Figure A5.

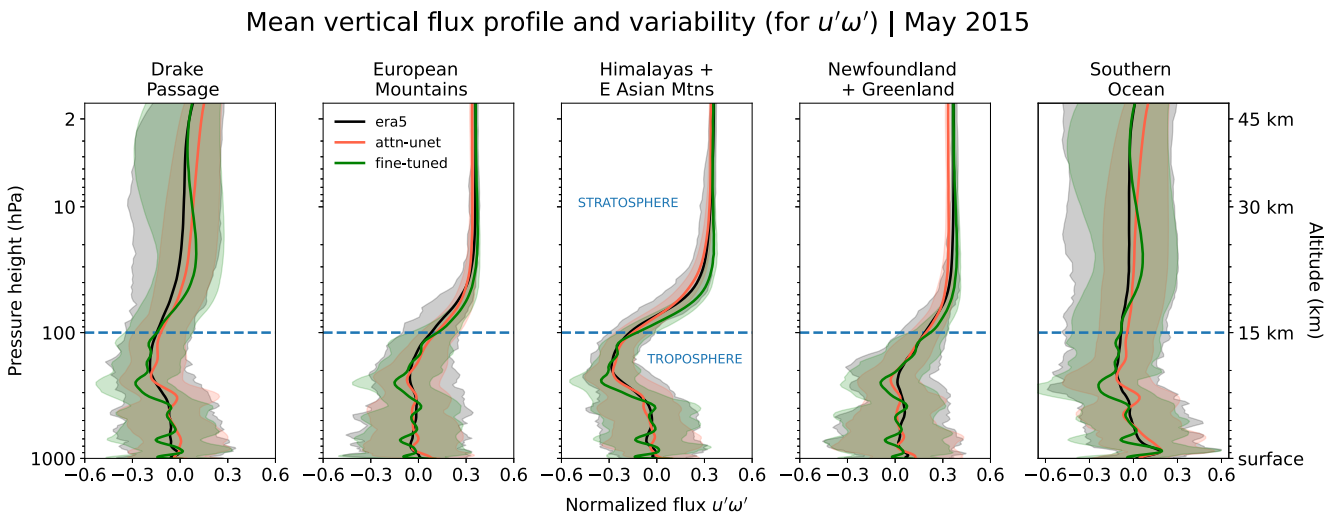


**Figure 5.** May 2015 mean ERA5 and predicted vertical profiles of the normalized (unitless) zonal flux, $F_x = u'\omega'$ over five hotspots. The exact boundaries of the hotspots are shown in Figure 3. The (normalized) ERA5 flux is shown in black, the prediction from attn unet is shown in orange, and the prediction from the fine-tuned model is shown in green. The gray, orange, and green shadings show the range of flux variability in the respective models. The regional extent for the hotspots is shown in Figure A3. The fluxes converge to near-zero in the stratosphere over all hotspots, due to successive wave filtering and dissipation, but appear to converge to non-zero values due to non-dimensionalization.

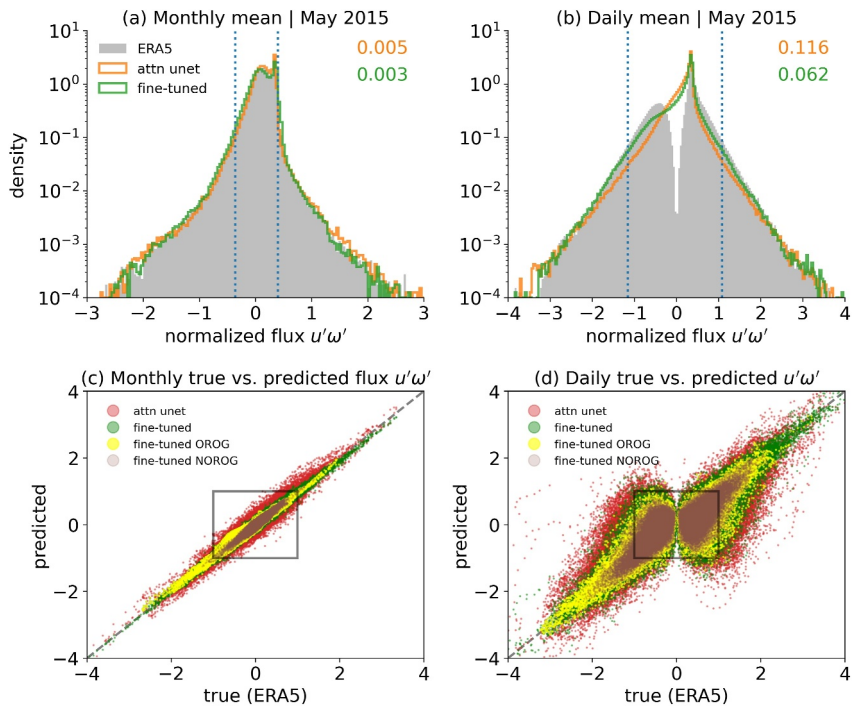


**Figure 6.** Histogram of the (a) May 2015 averaged and (b) daily averaged gravity wave flux $u'\omega'$. Gray shading shows the underlying ERA5 distribution, orange is the attn unet prediction, and green is the fine-tuning prediction. Numbers indicate the Hellinger distance for the corresponding predictive model. The dotted lines show the 2.5th and 97.5th percentile, respectively (note the log-scaled $y$-axis). Scatter plot of the ERA5 versus predicted flux *at each grid point* for (c) May 2015 monthly average and (d) daily average. Red and green markers show the scatter for the baseline and fine-tuned model. The scatter for four orographic (Drake Passage, Himalayas, Newfoundland, and European Mountains) and nonorographic hotspots (tropical Pacific, North Atlantic, Southeast Asia, and Southern Ocean) for the fine-tuned model is shown in yellow and brown, respectively. The regional extent of the hotspots is shown in Figure A3. A gray box is added for reference over the $[-1, 1] \times [-1, 1]$ interval.

### 3.4. Global Averaged Flux Distribution

Global flux distributions provide insight into how well our models generate the possible range of flux responses globally, which are crucial to modeling extreme GW events. The observed and predicted global distribution of the GW momentum fluxes at different sampling frequencies (monthly vs. daily averages) is shown in Figures 6 and A6. The histogram represents the distribution of the May 2015 monthly mean momentum flux globally, that is, over all points in the troposphere and stratosphere. Both the baseline and the fine-tuned models simulate the distribution on monthly time scales with remarkable accuracy, both in the bulk of the distribution and its tails (Figure 6a). The baseline and the fine-tuned model have a Hellinger distance of 0.005 and 0.003 from the underlying training (ERA5) distribution, suggesting that despite clear differences in predictive skill, the two distributions are nearly identical to the underlying ERA5 distribution. The fine-tuned model emulates the distribution tails slightly better than the baseline.

For monthly averages, the fine-tuned model provides excellent prediction of the mean flux field, as is gauged by the scatter plot in Figure 6c. The fine-tuned (green) model exhibits a reduced dispersion compared to attn unet (red). The fluxes over orographic hotspots (yellow) account for a larger scatter, and fluxes from nonorographic hotspots (brown) are clustered around smaller values.

Both models, however, struggle to accurately capture the daily-sampled histogram around small values. In the $[-0.5, 0.5]$ interval, the models fail to accurately learn the small values and instead predict close-to-zero values more frequently. This implies that the models learn the strong GW events more readily than the weak ones. Accordingly, the daily sampled flux distributions from both attn unet and the fine-tuned models produce higher Hellinger distances of 0.116 and 0.062, respectively. Our choice of loss function, that is, training to minimize the root-mean square error, could partly explain this magnitude-selective learning.

Even for daily samples, the attn unet model exhibits a visibly larger scatter against ERA5 compared to the fine-tuned model (Figure 6d, red markers vs. green markers). The points from the non-orographic regions (NOROG; brown markers) exhibit less scatter in panels (c) and (d), implying a tighter clustering around the diagonal compared to points from the orographic hotspots (OROG; yellow markers). Regardless, this indicates that the misrepresented fluxes around zero in panel (b) are due to the misrepresentation of both orographic and nonorographic GWs.

## 4. Conclusion and Discussion

Our analysis establishes that the atmospheric evolution learned by large transformer-based FMs (developed for weather research) can be leveraged to improve and expedite the creation of subgrid-scale parameterizations for climate models. The FM parameterization more accurately predicts lateral propagation effects than traditional parameterizations and outperforms the advanced attn unet benchmark with fewer learnable parameters, even in upper atmospheric regions where the foundation model was not pre-trained. This provides a fresh avenue to develop AI-driven representations of small-scale processes that have the potential to replace traditional parameterizations, which for over four decades have been envisioned as single-column plugins that often neglect key process physics. Coupling these parameterizations with existing climate models can promote and expedite the development of hybrid climate prediction models.

Since Prithvi is trained on large amounts of data, its latent encoder-decoder space contains a rich abstract representation of atmospheric evolution. The training data includes atmospheric variables like winds, humidity, radiation, and even leaf area index and soil moisture. The application of this approach transcends GWs and, with appropriate observational data for fine-tuning, could be used to create parameterizations for other atmosphere-ocean-land processes unseen during pre-training. As an added benefit, this approach allows using data from multiple streams for finetuning—high-resolution model data, satellite trains, terrestrial remote sensing data, ground observations, etc.

In the context of atmospheric GWs, the FM outperforms the attn unet benchmark in representing GW effects in the upper stratosphere. It learns the effects of three-dimensional propagation and dissipation of GWs in the atmosphere better than the attn unet. As a result, this model is capable of representing the missing GW effects in coarse-climate models. These effects are critical to getting a more realistic middle atmospheric circulation and seasonal wind transitions in climate models and in alleviating the "cold-pole bias" of the stratosphere that undermines the accuracy of these models.

Recent studies, for instance Bretherton et al. (2022), have discussed using deep learning models trained on high-resolution model simulations to "bias-correct" coarse-climate models. While effective, these techniques are less interpretable, as it is challenging to learn the true source of prevailing biases and to distinguish structural model errors from model errors due to inaccurate physics. Given the versatility of FMs, our approach enables the development of process-specific emulators for representing a suite of subgrid-scale processes, resulting in bias correction through a data-driven representation that is closer to observations.

To demonstrate the central idea, we have fine-tuned our models on limited years of ERA5 reanalysis data spanning a few variables. As noted above, the GWs in ERA5 are not assimilated but model-generated, and owing to a 25 km resolution, they miss a broad range of mesoscale GWs (with wavelengths shorter than 150–200 km). In addition, while the fluxes computed from ERA5 using HD better reflect fluxes carried by GWs, in the troposphere, it could include contributions from convection in regions with precipitation (Alexander et al., 2006; Wei et al., 2022). As a consequence, there is much scope for improvement in the parameterization. These improvements can be achieved by fine-tuning on longer periods of high-resolution (high-fidelity) data sets, through better estimates of tropospheric GW fluxes, and by including more convection-related variables (for instance, humidity, diabatic heating, and latent heat fluxes). This will be the focus of future work, where the fine-tuning will be accomplished using a kilometer-scale, high-resolution model output and an expanded feature set.

Admittedly, the nonlocal architecture adopted in this study presents a contrast with the widely adopted column-based discretization used by climate models. Yet, such a coupling is possible, and work is underway to couple the nonlocal fine-tuned scheme to an atmospheric model (National Center for Atmospheric Research (NCAR)'s CAM7) and evaluate its online performance on atmospheric variability and generalizability on warming scenarios. To this end, in collaboration with the Institute of Computing for Climate Science at the University of Cambridge and NCAR, we have identified inbuilt pooling and discretization functions that make this coupling possible using *ftorch*, without adding latency (Atkinson et al., 2025; Chapman & Berner, 2025).

Nevertheless, foundational models open avenues to using multisource observations to facilitate not just AI-powered weather research but also climate research. Due to constraints on computing power, we are still far away from being able to run climate models (such as those participating in CMIP) multiple decades and centuries into the future at kilometer or sub-kilometer resolutions. This means climate prediction will continue to miss crucial sub-grid physics and will continue to rely on physical parameterizations of unresolved processes.

We have demonstrated an appealing application of an existing foundation model in improving the sub-grid scale physics representation in a climate model. These emulators are not just intended to be coupled to a climate model, but can also serve as standalone plugins to improve small-scale variability in AI-based weather models and other fine-tuned models. In principle, FMs (like Prithvi WxC) could be strategically applied to address a range of climate applications, including, but not limited to, heat wave prediction, land-use trend detection, and cross-domain learning. This can be accomplished by fine-tuning the base FM on event-specific data sets, to enhance performance on rare events as, for instance, is discussed in Cui et al. (2025). First, the FMs can be pre-trained on global forecasts to optimize the global mean squared error. Subsequently, parts of the FM, when retrained on event-specific data using a specialized loss function, can be used to quickly re-train and develop specialized task-emulators. Used in conjunction with other FMs, such as Prithvi HLS (Jakubik et al., 2023), this approach can be leveraged to create lightweight, fine-tuned models for key weather and climate applications, including the prediction of wildfires, hurricane storm surge, and regional heatwave impacts, potentially improving climate change preparedness.

## Appendix A: Additional Model Information and Meridional Flux Predictions

See Figures A1–A6.

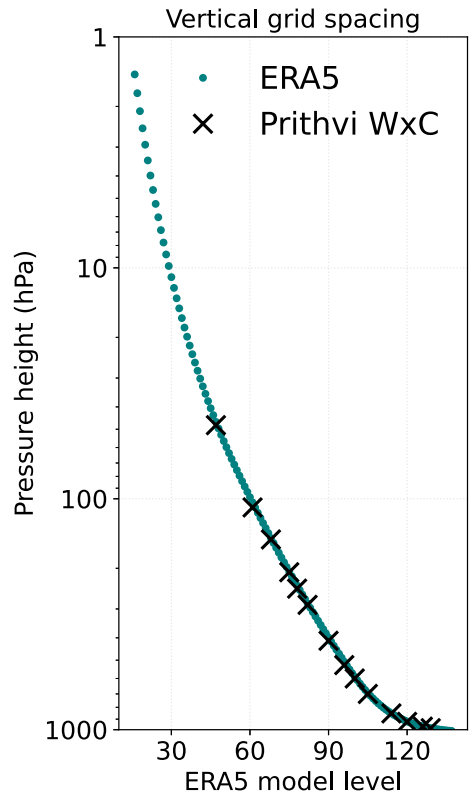


**Figure A1.** Prithvi was pre-trained on very sparse data in the vertical. The ERA5 fine-tuned data were computed on 137 model levels, and the top 15 model levels (i.e., levels above 1 hPa ~ 45 km) were discarded due to an artificial model sponge imposed at those levels. So, effectively 122 model levels between 1,000 hPa (surface) to 1 hPa (45 km) height were used. In contrast, Prithvi is trained on MERRA-2 data interpolated to 14 vertical levels: [985, 970, 925, 850, 700, 600, 525, 412, 288, 245, 208, 150, 109, 48] hPa. No training data were provided between 50 hPa and 1 hPa during pre-training. This means that the frozen encoder-decoder has no prior knowledge about the dynamic evolution of gravity waves at these heights. Still, as shown in Figures 4 and 5, the fine-tuned model outperforms the baseline in this region.

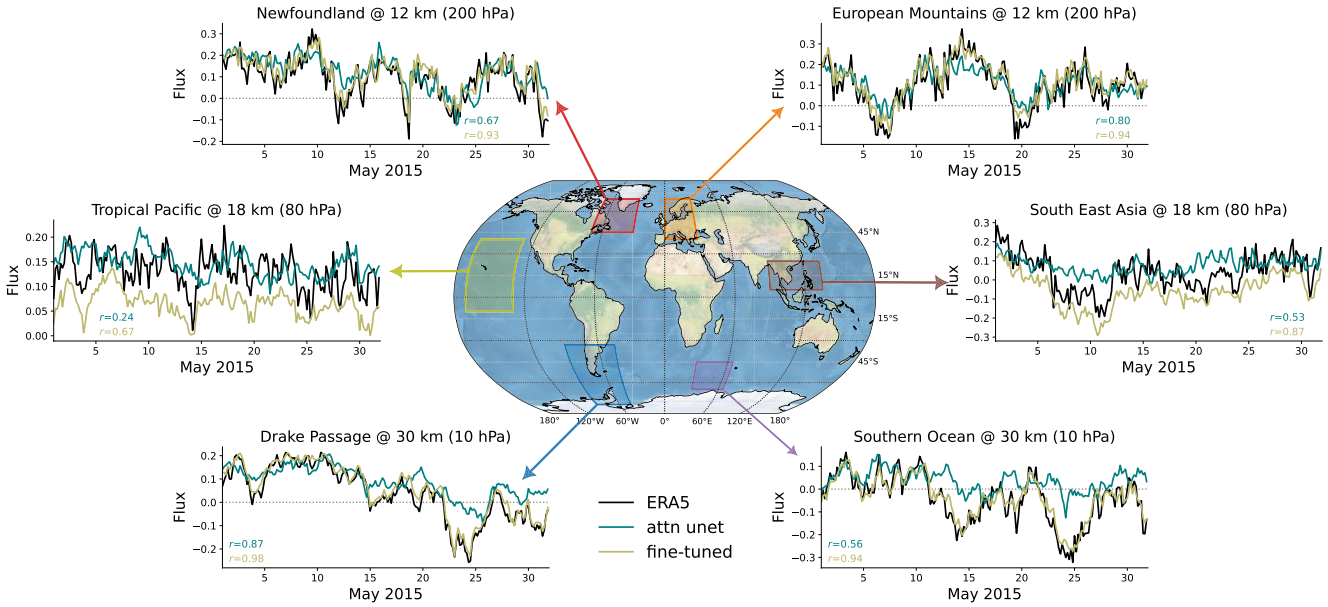


**Figure A2.** Same as Figure 3 but for $v'\omega'$ - instantaneous fluxes for May 2015 from ERA5 (black), and predictions from the baseline (teal) and the fine-tuned model (light green) over six different hotspots.

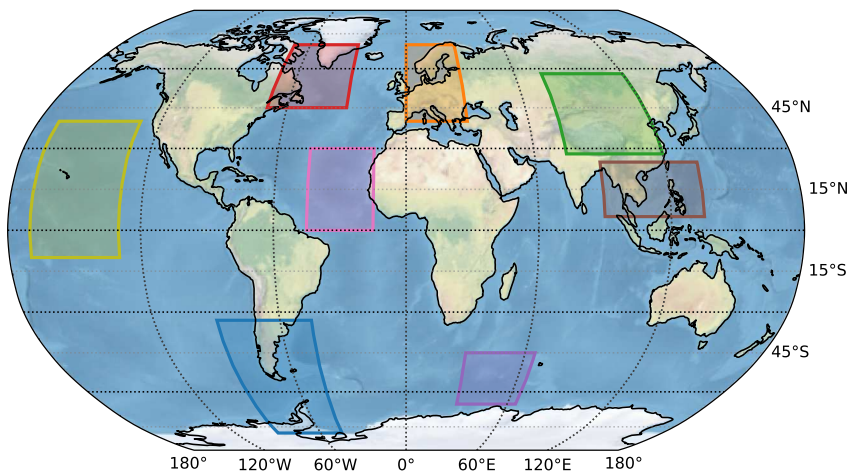


**Figure A3.** Figure showing all the gravity wave hotspots considered for regional analysis in this work. Yelllow: Tropical Pacific (170°W, 130°W) × (10°S, 40°N), Red: Newfoundland Mountains + Southern Greeland (70°W, 30°W) × (45°N, 70°N), Orange: European Mountains (0°, 30°E) × (40°N, 70°N), Green: Himalayas and East Asian Mountains (75°E, 120°E) × (28°N, 58°N), Light Pink: Northern Atlantic (45°W, 15°W) × (0°, 30°N), Brown: Southeast Asia (90°E, 135°E) × (5°N, 25°N), Blue: Drake Passage (90°W, 45°W) × (78°S, 33°S), Dark Pink: Southern Ocean (30°E, 65°E) × (65°S, 45°S).

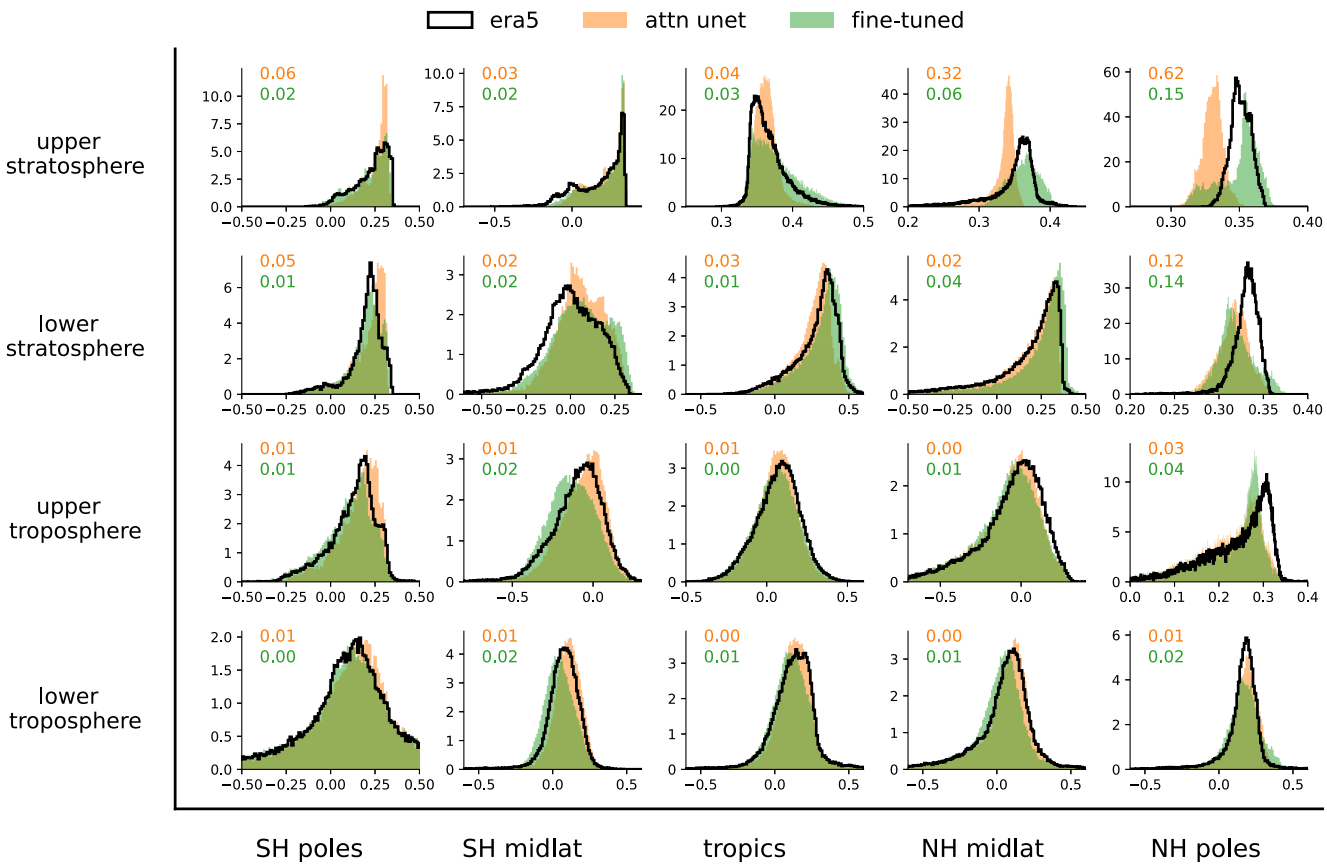


**Figure A4.** Gravity wave flux distributions similar to Figure 4, but divided according to latitude and height. The poles are defined as latitudes 60°–90°, the midlatitudes as 30°–60°, and the tropics as 15°S–15°N. The numbers indicate the respective Hellinger distances w.r.t. the distribution from ERA5 (black). For each latitude band, averaging is conducted over the whole latitude circle, that is, over all longitudes.

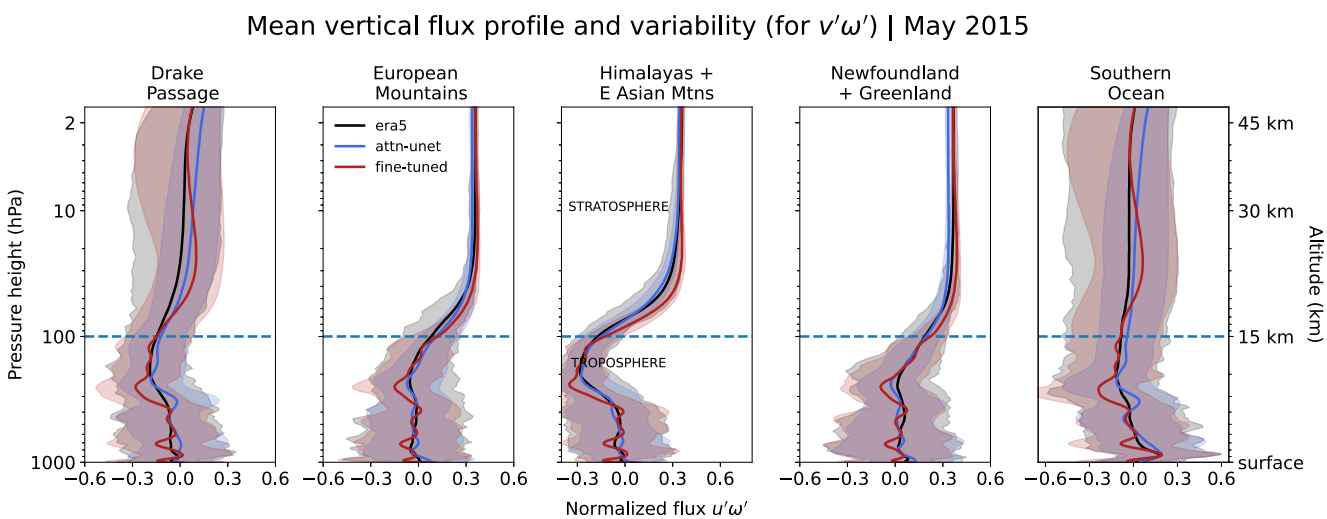


**Figure A5.** Same as Figure 5 but for the meridional flux $v'\omega'$.

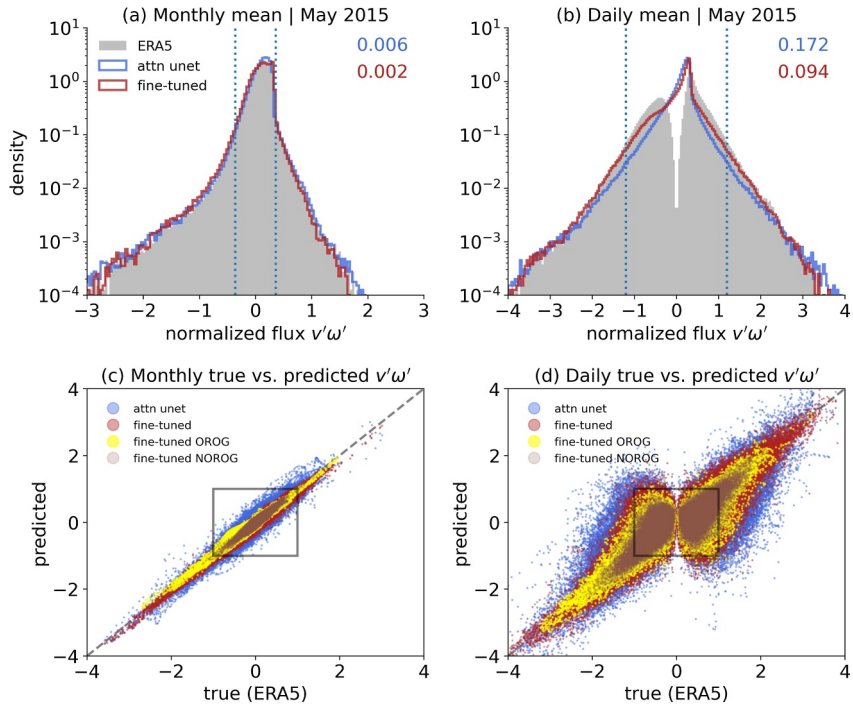


**Figure A6.** Same as Figure 6 but for $v'\omega'$. The distribution of the (a) May 2015 averaged and (b) daily averaged gravity wave flux $v'\omega'$. Gray shading shows ERA5's underlying distribution, orange is the baseline prediction, and green is the fine-tuning prediction. Numbers indicate the Hellinger distance for the corresponding predictive model. The dotted lines show the 2.5th and 97.5th percentile, respectively (note the log-scaled $y$-axis). The bottom row shows the respective scatter of the ERA5 flux and the predicted meridional flux $v'\omega'$.

## Conflict of Interest

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

## Data Availability Statement

- **ERA5 data:** ECMWF's ERA5 data (Hersbach et al., 2023) can be freely accessed from https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels.
- **Prithvi WxC FM and fine-tuning:** The code for the Prithvi WxC model is available at https://huggingface.co/ibm-nasa-geospatial/Prithvi-WxC-1.0-2300M. The fine-tuning code for climate model parameterization for GW flux is available at Roy and Gupta (2025): https://doi.org/10.5281/zenodo.16666812. Python scripts to compute GW momentum fluxes from the publicly available ERA5 reanalysis are available at Roy et al. (2025): https://doi.org/10.5281/zenodo.16666707.
- **Python packages:** 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 coarsegraining is publicly available at: https://xesmf.readthedocs.io/en/stable/.
- **Animation:** An animation showing the performance of the fine-tuned scheme over the Southern Ocean, and how it compares to resolved fluxes in ERA5 is available at: https://doi.org/10.17605/OSF.IO/8W6AZ.

**Acknowledgments**
Aditi Sheshadri and Aman Gupta are supported by Schmidt Sciences, LLC, 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 Earth Science Division's Earth Science Scientific Computing, Earth Science Data Systems Program, and the Earth Science Modeling and Analysis Program. NASA Grant (80MSFC22M004).

## 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

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

Alexander, M. J., Richter, J. H., & Sutherland, B. R. (2006). Generation and trapping of gravity waves from convection with comparison to parameterization. *Journal of the Atmospheric Sciences*, *63*(11), 2963–2977. https://doi.org/10.1175/JAS3792.1

Atkinson, J., Elafrou, A., Kasoar, E., Wallwork, J. G., Meltzer, T., Clifford, S., et al. (2025). FTorch: A library for coupling PyTorch models to Fortran. *Journal of Open Source Software*, *10*(107), 7602. https://doi.org/10.21105/joss.07602

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

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., et al. (2025). A foundation model for the Earth system. *Nature*, *641*(8065), 1180–1187. https://doi.org/10.1038/s41586-025-09005-y

Bogenschutz, P. A., Gettelman, A., Morrison, H., Larson, V. E., Schanen, D. P., Meyer, N. R., & Craig, C. (2012). Unified parameterization of the planetary boundary layer and shallow convection with a higher-order turbulence closure in the Community Atmosphere Model: Single-column experiments. *Geoscientific Model Development*, *5*(6), 1407–1423. https://doi.org/10.5194/gmd-5-1407-2012

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2022). On the opportunities and risks of foundation models (No. arXiv:2108.07258). *arXiv*. https://doi.org/10.48550/arXiv.2108.07258

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., et al. (2022). Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations. *Journal of Advances in Modeling Earth Systems*, *14*(2), e2021MS002794. https://doi.org/10.1029/2021MS002794

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

Chapman, W. E., & Berner, J. (2025). Improving climate bias and variability via CNN-Based state-dependent model-error corrections. *Geophysical Research Letters*, *52*(6), e2024GL114106. https://doi.org/10.1029/2024GL114106

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

Cui, Z., Meng, F., & Luo, J. (2025). Breaking through tropical cyclone intensity prediction: A foundation model Prithvi-TC. *Frontiers of Computer Science*, *19*(12), 1912369. https://doi.org/10.1007/s11704-025-41268-6

Dörnbrack, A., Leutbecher, M., Kivi, R., & Kyrö, E. (1999). Mountain-wave-induced record low stratospheric temperatures above northern Scandinavia. *Tellus A: Dynamic Meteorology and Oceanography*, *51*(5), 951–963. https://doi.org/10.3402/tellusa.v51i5.14504

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

Fritts, D. C., & Alexander, M. J. (2003). Gravity wave dynamics and effects in the middle atmosphere. *Reviews of Geophysics*, *41*(1), 1–64. https://doi.org/10.1029/2001RG000106

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., et al. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). https://doi.org/10.1175/JCLI-D-16-0758.1

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. https://doi.org/10.1029/2002GL014756

Golaz, J.-C., Horowitz, L. W., & Levy, H., II. (2013). Cloud tuning in a coupled climate model: Impact on 20th century warming. *Geophysical Research Letters*, *40*(10), 2246–2251. https://doi.org/10.1002/grl.50232

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*(3), 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., Roy, S., Gaur, V., Maskey, M., & Ramachandran, R. (2024). Machine learning global simulation of nonlocal gravity wave propagation. https://doi.org/10.48550/arXiv.2406.14775

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

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

Hirschfeld, A. (2024). Singapore Airlines death: Is climate change making air turbulence worse. https://www.aljazeera.com/economy/2024/5/23/singapore-airlines-death-is-climate-change-making-air-turbulence-worse

Hoffmann, L., Spang, R., Orr, A., Alexander, M. J., Holt, L. A., & Stein, O. (2017). A decadal satellite record of gravity wave activity in the lower stratosphere to study polar stratospheric cloud formation. *Atmospheric Chemistry and Physics*, *17*(4), 2901–2920. https://doi.org/10.5194/acp-17-2901-2017

Höpfner, M., Larsen, N., Spang, R., Luo, B. P., Ma, J., Svendsen, S. H., et al. (2006). MIPAS detects Antarctic stratospheric belt of NAT PSCs caused by mountain waves. *Atmospheric Chemistry and Physics*, *6*(5), 1221–1230. https://doi.org/10.5194/acp-6-1221-2006

Iacono, M. J., Mlawer, E. J., Clough, S. A., & Morcrette, J.-J. (2000). Impact of an improved longwave radiation model, RRTM, on the energy budget and thermodynamic properties of the NCAR community climate model, CCM3. *Journal of Geophysical Research*, *105*(D11), 14873–14890. https://doi.org/10.1029/2000JD900091

Jakubik, J., Roy, S., Phillips, C. E., Fraccaro, P., Godwin, D., Zadrozny, B., et al. (2023). Foundation models for generalist geospatial artificial intelligence (No. arXiv:2310.18660). *arXiv*. https://doi.org/10.48550/arXiv.2310.18660

Kim, Y.-H., & Chun, H.-Y. (2015). Momentum forcing of the quasi-biennial oscillation by equatorial waves in recent reanalyses. *Atmospheric Chemistry and Physics*, *15*(12), 6577–6587. https://doi.org/10.5194/acp-15-6577-2015

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

Kruse, C. G., Alexander, M. J., Hoffmann, L., van Niekerk, A., Polichtchouk, I., Bacmeister, J. T., et al. (2022). Observed and modeled Mountain waves from the surface to the mesosphere near the Drake Passage. *Journal of the Atmospheric Sciences*, *79*(4), 909–932. https://doi.org/10.1175/JAS-D-21-0252.1

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al. (2023). GraphCast: Learning skillful medium-range global weather forecasting (No. arXiv:2212.12794). *arXiv*. https://doi.org/10.48550/arXiv.2212.12794

Lear, E. J., Wright, C. J., Hindley, N. P., Polichtchouk, I., & Hoffmann, L. (2024). Comparing gravity waves in a kilometer-scale run of the IFS to AIRS satellite observations and ERA5. *Journal of Geophysical Research: Atmospheres*, *129*(11), e2023JD040097. https://doi.org/10.1029/2023JD040097

Lee, H., Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P., et al. (2023). *Climate change 2023: Synthesis report. Contribution of working groups I, II and III to the sixth assessment report of the intergovernmental panel on climate change*. The Australian National University.

Lessig, C., Luise, I., Gong, B., Langguth, M., Stadtler, S., & Schultz, M. (2023). AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning (No. arXiv:2308.13280). *arXiv*. https://doi.org/10.48550/arXiv.2308.13280

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., & 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

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., et al. (2012). Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, *4*(3). https://doi.org/10.1029/2012MS000154

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

Morrison, M. A., & Lawrence, P. (2020). Understanding model-based uncertainty in climate science. In G. Pellegrino & M. Di Paola (Eds.), *Handbook of philosophy of climate change* (pp. 1–21). Springer International Publishing. https://doi.org/10.1007/978-3-030-16960-2_154-1

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

Pahlavan, H. A., Wallace, J. M., & Fu, Q. (2023). Characteristics of tropical convective gravity waves resolved by ERA5 reanalysis. *Journal of the Atmospheric Sciences*, *80*(3), 777–795. https://doi.org/10.1175/JAS-D-22-0057.1

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

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

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., et al. (2025). Probabilistic weather forecasting with machine learning. *Nature*, *637*(8044), 84–90. https://doi.org/10.1038/s41586-024-08252-9

Procházková, Z., Kruse, C. G., Alexander, M. J., Hoffmann, L., Bacmeister, J. T., Holt, L., et al. (2023). Sensitivity of Mountain wave drag estimates on separation methods and proposed improvements. *Journal of the Atmospheric Sciences*, *80*(7), 1661–1680. https://doi.org/10.1175/JAS-D-22-0151.1

Roy, S., & Gupta, A. (2025). Amangupta2/gravity-wave-finetuning-james: Release v1.0.0—Gravity wave flux fine-tuning with Prithvi WxC [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.16666812

Roy, S., Kumar, A., Shinde, R., Gupta, A., & simonpf. (2025). Amangupta2/WxC-Bench: Release v1.0.0—WxC-Bench: A benchmark suite for AI foundation models in weather and climate [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.16666707

Sato, K., Tateno, S., Watanabe, S., & Kawatani, Y. (2012). Gravity wave characteristics in the Southern hemisphere revealed by a high-resolution middle-atmosphere general circulation model. *Journal of the Atmospheric Sciences*, *69*(4), 1378–1396. https://doi.org/10.1175/JAS-D-11-0101.1

Schmude, J., Roy, S., Trojak, W., Jakubik, J., Civitarese, D. S., Singh, S., et al. (2024). Prithvi WxC: Foundation model for weather and climate. https://doi.org/10.48550/arXiv.2409.13598

Stephan, C. C., Strube, C., Klocke, D., Ern, M., Hoffmann, L., Preusse, P., & Schmidt, H. (2019). Intercomparison of gravity waves in global convection-permitting models. *Journal of the Atmospheric Sciences*, *76*(9), 2739–2759. https://doi.org/10.1175/JAS-D-19-0040.1

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

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*.

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

Wu, Y., Miao, C., Fan, X., Gou, J., Zhang, Q., & Zheng, H. (2022). Quantifying the uncertainty sources of future climate projections and narrowing uncertainties with bias correction techniques. *Earth's Future*, *10*(11), e2022EF002963. https://doi.org/10.1029/2022EF002963

Yoshida, L., Tomikawa, Y., Ejiri, M. K., Tsutsumi, M., Kohma, M., & Sato, K. (2024). Large-Amplitude inertia gravity waves over Syowa station: Comparison of PANSY radar and ERA5 reanalysis data. *Journal of Geophysical Research: Atmospheres*, *129*(22), e2023JD040490. https://doi.org/10.1029/2023JD040490

Zhao, M., Golaz, J.-C., Held, I. M., Guo, H., Balaji, V., Benson, R., et al. (2018). The GFDL global atmosphere and land model AM4.0/LM4.0: 1. Simulation characteristics with prescribed SSTs. *Journal of Advances in Modeling Earth Systems*, *10*(3), 691–734. https://doi.org/10.1002/2017MS001208