FINETUNING WEATHER FOUNDATION MODEL TO DEVELOP CLIMATE MODEL PARAMETERIZATIONS

Sujit Roy[1,†], Aman Gupta[2,†], Johannes Schmude[3], Vishal Gaur[1], Weiji Long[4], Manil Maskey[5], Rahul Ramachandran[5], Aditi Sheshadri[2]

[1]Earth System Science Center, UAH, Huntsville, AL
[2]Department of Earth System Science, Stanford University, Stanford, USA
[3]IBM Research
[4]Development Seed, Washington, DC, USA
[5]NASA MSFC, Huntsville, AL
[†]Equal Contribution,
{sujit.roy}@nasa.gov

ABSTRACT

Climate prediction models parameterize a multitude of atmospheric-oceanic processes like clouds, turbulence, and gravity waves. These *physical parameterizations* are a leading source of uncertainty and strongly influence future projections of global temperature rise. We present a fresh approach to developing parameterizations for coarse-climate models by leveraging pre-trained AI foundation models (FMs) for weather and climate. A pre-trained encoder and decoder from a 2.3 billion parameter FM (Prithvi WxC) — which contains a latent probabilistic representation of atmospheric evolution — is fine-tuned to create a data-driven predictor of atmospheric gravity waves (GWs). GWs exert a profound influence on the atmospheric circulation but their momentum forcing in climate models is parameterized. We create an ML parameterization which learns GW fluxes from high-resolution "GW resolving" climate models to represent them in "GW-missing" coarse-climate models. The GW fluxes predicted by our fine-tuned model are comprehensively evaluated using three different tests. Comparison with a baseline (Attention U-Net) reveals the superior capability of the fine-tuned model to predict GW momentum fluxes throughout the atmosphere. Moreover, the fine-tuned model offers better predictions even in the upper stratosphere, which is a region excluded during the FM pre-training. This is quantified using the Hellinger distance which is 0.11 for the baseline and 0.06, i.e., roughly half, for the fine-tuned model. Foundation models are largely unexplored in climate science. Our findings emphasize the versatility and reusability of FMs (which require substantial compute) to accomplish a range of weather- and climate-related downstream tasks, especially in a low-data regime. These FMs can be further leveraged to create data-driven parameterizations for other earth-system processes.

1 INTRODUCTION

Accurate prediction of future climate is a trillion-dollar challenge that can critically impact the world economy, food security, global health, and urban planning. State-of-the-art future climate projections are highly uncertain. Obtaining reliable projections of future climate requires urgent improvements in current climate models, many of which are strongly influenced by parametric uncertainty, scenario uncertainty, and structural uncertainty (Morrison & Lawrence, 2020; Lee et al., 2023). This study aims at exploring the untapped potential of AI foundation models to improve climate models by addressing one of the leading sources of climate model uncertainty: physical parameterizations.

Foundation models (FMs) can be broadly defined as large self-supervised learning models which can be fine-tuned to perform a broad range of *downstream* tasks (Bommasani et al., 2022). FMs allow leveraging the pre-trained learnings of a big AI model to perform an array of sub-tasks. A good example is OpenAI's ChatGPT, which is first pre-trained on large language datasets and is

subsequently fine-tuned to perform many other language-related tasks. FMs are largely unexplored in climate science. Only a couple of weather-related FMs exist to date (AtmoRep (Lessig et al., 2023), ClimaX (Nguyen et al., 2023), and Prithvi (Jakubik et al., 2023). Otherwise, the use of large AI models in meteorology, like FourCastNet, PanguWeather, and GraphCast (Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023) is mostly restricted to weather prediction. Simply put, despite their huge training costs, these models have been limited to accomplishing just one task: synoptic-scale weather forecasting. In this study, we propose using a state-of-the-art FM, Prithvi WxC (CITE), developed for weather and climate applications, and use it for a downstream task of developing data-driven physical parameterizations for climate models.

Numerical climate models couple together multiple components of the earth system (atmosphere, ocean, land, ice, etc.) to predict climate evolution over years, decades, centuries, and beyond. Climate models often operate at a grid resolution of 100-300 km. Such resolution is direly insufficient to fully represent, or even resolve, the smaller-scale processes like clouds, precipitation, boundary layer turbulence, small-scale gravity waves, etc. These processes are crucial for the global energy balance. The traditional approach is to couple the large-scale fluid solver with a suite of *physical parameterizations* to approximately capture the effect of each unresolved processes (Alexander & Dunkerton, 1999; Lott & Miller, 1997; Bogenschutz et al., 2012; Iacono et al., 2000, to name a few).

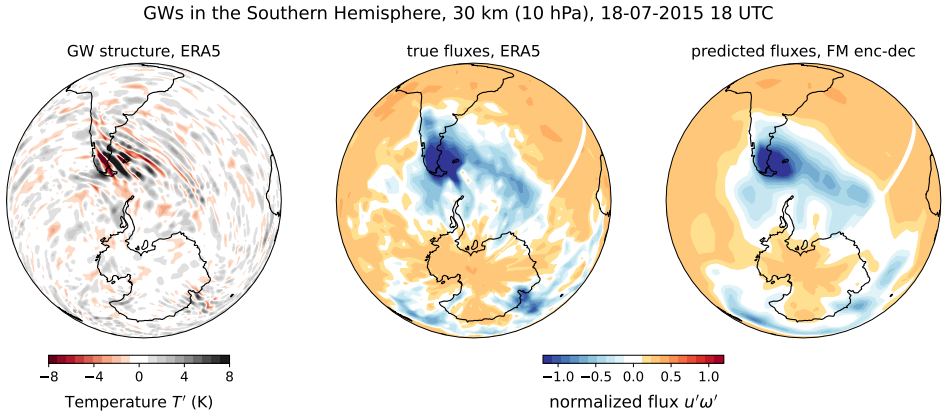

Figure 1: Comparing Baseline model and the fine-tuned model.

Most parameterizations are a rough, incomplete representation of the respective processes, in that, they are subject to a series of simplifications and idealizations regarding the physical evolution of the process Hourdin et al. (2017). This could be partly attributed to either a limited understanding of the processes or to climate model design considerations. Either way, errors stemming in parameterization design and parametric tuning of small-scale processes can often add up and lead to inaccurate dynamics, momentum imbalances, and uncertainties in future climate projections (Golaz et al., 2013; Mauritsen et al., 2012; Zhao et al., 2018).

Using AI to upgrade and improve climate model parameterizations is an area of active research (Mansfield et al., 2023; Eyring et al., 2024). Recent deep learning approaches include (a) learning process evolution from high-resolution models or parameterization data to represent it in coarse-resolution models (Espinosa et al., 2022; Chantry et al., 2021; Yuval & O'Gorman, 2023; Gupta et al., 2024b; Lu et al., 2024), (b) using equation discovery or similar techniques to learn analytical forms of sub-grid scale momentum closures (Zanna & Bolton, 2020; Jakhar et al., 2024), and (c) hybrid probabilistic combination of single-scenario high-fidelity data and multi-scenario low-fidelity data (Bhouri et al., 2023). Irrespective of the approach, the need for high-resolution high-quality training data and limited generalizability limits rapid progress in this domain.

In this study, we introduce a fresh approach to promote the development of AI-driven climate model parameterizations. We blend the trained encoders and decoders from the new Prithvi WxC FM and blend them with high-quality task-specific data, to create a lightweight, fine-tuned AI model capable of skillfully predicting subgrid-scale atmospheric variability in climate models. We demonstrate the effectiveness of our approach using atmopsheric gravity waves (GWs) as a test case.

GWs are intermittent, mesoscale and sub-mesoscale (spatial scale O(1)-O(1000) km) perturbations generated around thunderstorms, jet disturbances, flow over mountains, etc. (Fritts & Alexander, 2003). GWs couple the different layers of the atmosphere by carrying near-surface momentum and energy to stratospheric and mesospheric heights. GWs influence clear air turbulence, surface extremes, stratospheric circulation, and ocean heat transport. Thus, they are thus crucial to the earth's momentum budget yet are inadequately represented in climate models owing to limited grid resolution (Plougonven et al., 2020).

Scientific importance: a coarse O(100) km climate model practically misses all GW effects because it cannot resolve these small-scale waves. So, we develop an ML model that learns GW effects *from a high-resolution climate model* (which resolve a substantial portion of the GW effects/physics). This model can then be coupled to a coarse-resolution climate model to represent "missing" GW physics. The same principle can be generalized to develop more data-driven ML models to represent other missing processes (like clouds, precipitation) in coarser-climate models.

As shown in Figure 1, our fine-tuned model skillfully predicts the momentum fluxes associated with GWs for a given atmospheric background state. The structure of the excited GWs on 18 July 2015 is shown in Figure 1a. For the snapshot shown, the model predicts the intermittent intensification of the GW momentum fluxes around the Andes in South America and the Prince Charles Mountains in Antarctica. The region of flux intensification over the Andes extends over to the Drake Passage and parts of the Southern Ocean, indicating that the finetuned model can learn and represent the lateral propagation of the generated waves.

As discussed in more detail in the following sections, our approach takes advantage of the high-dimensional latent space of the NASA and IBM's Prithvi WxC FM trained on massive amounts of MERRA2 reanalysis data and clubs it with limited data on GW evolution to create a highly generalizable data-driven scheme for coarse-climate models:

- **Faster training, better performance:** Using the FM latent-space allows faster training of the fine-tuned model than a specialized Attention Unet baseline. The fine-tuned model outperforms the baseline in terms of global momentum flux distribution, regional flux distribution, and intermittent flux evolution.
- **Generalizable to new regions:** The fine-tuned model outperforms the specialized baseline model even in the middle-to-upper stratosphere region where Prithvi WxC was not trained.
- **Improved physics representation:** Since the model has been fine-tuned on GW resolving reanalysis data, our scheme represents key aspects of GW physics which traditional climate model parameterizations do not: transient evolution as opposed to steady-state evolution, and full three-dimensional evolution as opposed to pure vertical evolution (Plougonven et al., 2013).

Our results demonstrate a promising approach to leverage pre-trained large AI models to create smaller fine-tuned ones for a multitude of applications in weather and climate science. This approach allows blending data from multiple streams — high-resolution models, satellites, ground-based observations, etc. — for reliable process state prediction. For instance, Prithvi WxC has already been fine-tuned to perform a range of weather-related tasks incl. aviation turbulence prediction, hurricane intensity and track forecasts, identifying weather analogs, generate long-term precipitation forecasts, etc (CITE - see for more details).

2 MODELS AND DATA DESCRIPTION

2.1 PREPARING TRAINING DATA FOR GW FLUX PREDICTION

The fine-tuning data for GW flux prediction was extracted from ERA5 global reanalysis (Hersbach et al., 2020) at 25 km horizontal resolution, 137 vertical levels, and hourly-frequency. ERA5 resolves GWs with wavelengths longer than 150-200 km, but a typical climate model barely resolves any. Thus, we learn the fluxes from high-resolution ERA5 and plug them in a coarse-climate model to represent missing physics. ERA5 does not provide GW momentum fluxes as output. The GW fluxes were extracted by applying Helmholtz decomposition (HD) (as in Lindborg, 2015; Köhler et al., 2023) on the raw ERA5 output, to compute the directional GW flux covariances $u'\omega'$ and

$v'\omega'$. Essentialy, the horizontal winds (u and v) are decomposed into rotational and nonrotational components:

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

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

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

which are combined with the zonal mean removed vertical velocity (ω') to compute the GW momentum fluxes (g = –9.81 m/s^2 being the acceleration due to gravity):

$$\vec{F} = (F_x, F_y) = g^{-1}(u'_{div}\omega', v'_{div}\omega') \tag{3}$$

The procedure is applied to create the supervised training data. The top 15 out of the 137 vertical levels are discarded due to artificial model damping. All input-output pairs are coarse-grained from 25 km resolution to a 64 latitudes $\times$ 128 longitudes grid (roughly 280 km resolution) to obtain conservative wave averages.

The fluxes are computed at an hourly-resolution for four years: 2010, 2012, 2014, and 2015. For global training, this corresponds to roughly 35k training samples, which can arguably be classified as low-volume data.

Baseline: the input feature set consists of winds u, v, potential temperature θ, 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 ad 128 longitudes. Similarly, the output is fluxes $u'\omega'$ and $v'\omega'$, each on 122 vertical levels, 64 latitudes and 128 longitudes (Figure 2).

Finetuning: the input feature set consists of winds u, v, temperature T, and pressure (p), each on 122 vertical levels, 64 latitudes ad 128 longitudes. Similarly, the output is potential temperature θ (for validation), and fluxes $u'\omega'$ and $v'\omega'$, each on 122 vertical levels, 64 latitudes and 128 longitudes (Figure 4).

2.2 Baseline Model

An advanced baseline compared to standard MLP was created by training an Attention U-Net model (Oktay et al., 2018) on the training data from ERA5. The input is downsampled using four convolution blocks and then upsampled using four convolution blocks. The skip connection at each level comprises learnable attention layers. For every downsample (upsample), the number of channels increase (decrease) by a factor 2 but all spatial dimensions reduce (increase) by a factor 2. The total number of learnable parameters for the given baseline models was 35,103,456. HOW MANY LEARNABLE PARAMETERS? HOW MUCH TRAINING TIME?

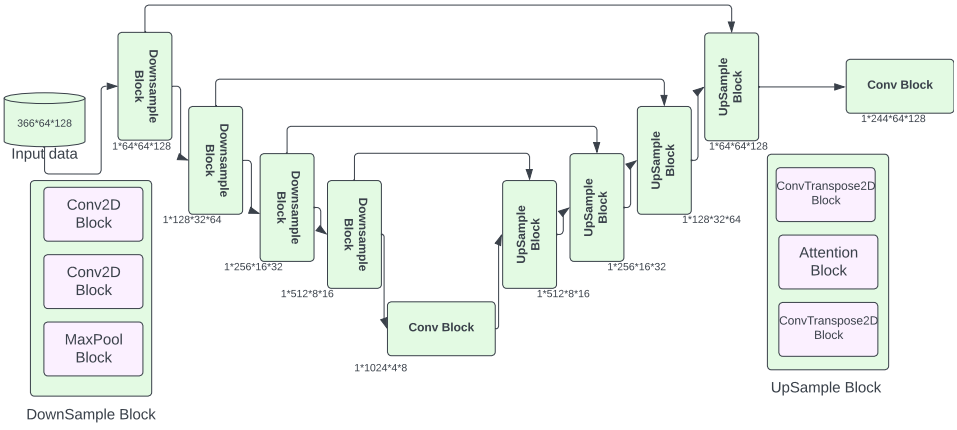

Figure 2: Architecture for the Attention U-Net baseline model.

2.3 THE UNDERLYING FOUNDATION MODEL: PRITHVI WXC

The Prithvi WxC FM was trained on (??) years of MERRA2 global reanalysis data (Gelaro et al., 2017) at roughly 50 km horizontal resolution and 3-hourly frequency. More details are provided in the Prithvi WxC FM paper (CITE). WHAT ALL LEVELS?

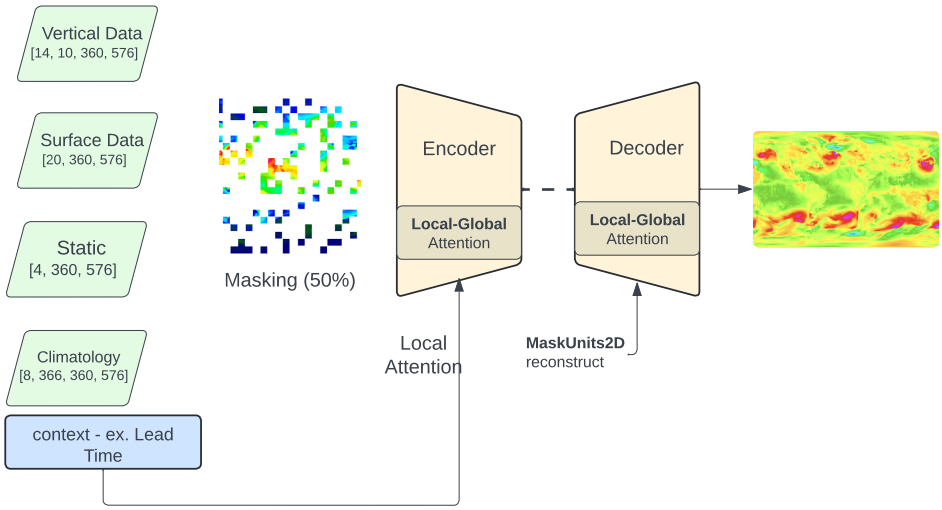

Figure 3: Pre-training Model Architecture

2.4 DESIGNING A FINETUNING MODEL

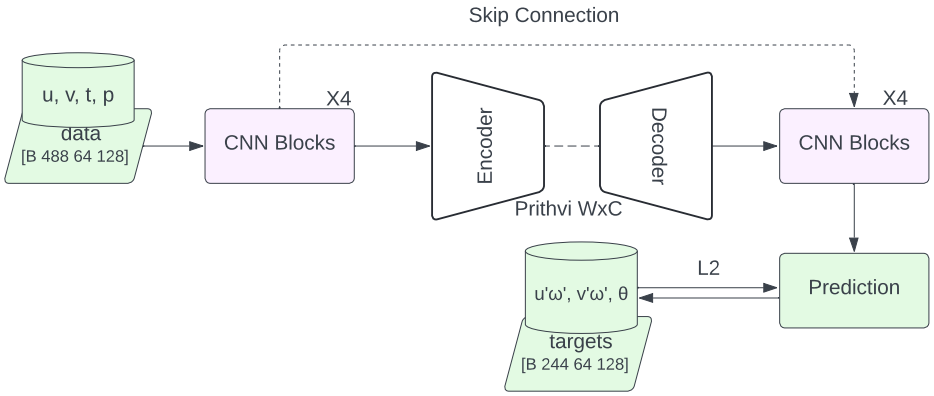

Figure 4: Finetuning architecture

During fine-tuning Prithvi WxC, we freeze the encoder and decoder part of the model. The frozen encoder is preceded by 4 new learnable convolution blocks each with an increasing number of hidden channels, i.e., C, $2C$, $4C$ and then $8C$, where $C = 160$. Likewise, the frozen decoder is succeded by 4 new learnable convolution blocks. HOW MANY LEARNABLE PARAMETERS? HOW MUCH TRAINING TIME?

Since gravity wave flux prediction is an instantaneous flux calculation task, we fix the lead time to be zero. The instantanous model input for fine-tuning has shape [1, 488, 64, 128] where the 488 channels comprise the four background variables u, v, t and p on 122 vertical levels, and on a 64×128 horizontal grid, as discussed above. The model was fine-tuned to produce an output with shape [366, 64, 128] comprising of the potential temperature, $u'\omega'$, and $v'\omega'$ on 122 vertical levels. The finetuned model leveraged a U-Net like architecture to allow the model to extract high-frequency

information from the given data source. We re-emphasize that Prithvi WxC was pre-trained on the MERRA2 dataset but the fine-tuning is achieved using the conservatively coarse-grained ERA5 dataset.

MSE Loss was used to train the model:

$$\mathcal{L}(\vec{x}, \vec{y}) = \sum_i (x_i - y_i)^2 \tag{4}$$

3 RESULTS

Climate model parameterizations are usually evaluated on various spatiotemporal scales to gauge their impact on planetary winds, temperature, and global momentum budget. Therefore, we test both the steady-state distribution of the generated flux and its time-evolving response. For GWs, the steady state analysis informs how well our models generate the tails of the momentum flux distribution, which are crucial to modeling atmospheric extreme responses. Likewise, the instantaneous time-evolving response informs how well the models learn the intermittent generation and evolution of GWs.

3.1 GLOBAL, STEADY STATE FLUX SPECTRUM

The observed and predicted global distribution of the GW momentum fluxes at different sampling frequencies is shown in Figure 5. The distributions represents the May 2015 monthly mean momentum flux over all the points in the troposphere and the stratosphere. Both the baseline and the fine-tuned models simulate the monthly mean distribution with remarkable accuracy both in the bulk of the distribution and its tails (Figure 5a). To quantify the difference between the two distributions, we use the Hellinger distance defined as follows.

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

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

By definition, $\mathcal{H} \in [0, 1]$. A Hellinger distance of 0 means the probability distributions are identical almost everywhere, while a Hellinger distance of 1 implies the distributions are disjoint, i.e., p is non-zero wherever q is zero, and vice versa. Based on our analysis, we consider a Hellinger distance of 0.05 or less to be pretty good.

Tail-Hellinger Distance. To quantify the accuracy around tails, we define a new metric, the Hellinger distance for distribution tails or the "tail-Hellinger" distance between p and q, $\mathcal{H}_{T,\epsilon}$, defined as:

$$\mathcal{H}_{T,\epsilon}(p, q) = \frac{1}{2} + \frac{1}{4\epsilon} \int_{x \in \mathcal{V}} p(x)dx - \frac{1}{2\epsilon} \int_{x \in \mathcal{V}} \sqrt{p(x)q(x)}dx. \tag{6}$$

Here $\mathcal{V} = (\infty, x_1] \cup [x_2, \infty)$ is a tail subset of X, and for cumulative distribution function F, $F(x_1) = \epsilon$ and $F(x_2) = 1 - \epsilon$. The tail Hellinger distance, unlike the regular Hellinger distance, can also be negative, and a negative value would imply a fatter tail of the predicted distribution than the true distribution. For $\epsilon = 0.5$, the tail-Hellinger distance yields the regular Hellinger distance. More details are provided in the Appendix.

The baseline and the fine-tuned model have a Hellinger distance of 0.005 and 0.003 from the true distribution suggesting that the two distributions are nearly identical to the underlying truth.

To consider the time-evolving fluxes which may have been averaged out in a monthly mean, we also considered the distribution of the daily sampled momentum fluxes (Figure 5b). The daily fluxes maintain similar accuracy (QUANTIFY) around the tails as the monthly mean, but the daily predicted fluxes from both models do not exhibit the minimum around 0 seen in the observed fluxes. Simply put, our models predict the bulk of the daily distribution quite well, but struggle a bit in learning/predicting small values. This is reminiscent of prevailing problems with even the state-of-the-art weather prediction models which predict large-scale features quite well but fail to project the same

level of accuracy in predicting the small-scales. As a result of this deviation, for daily sampling, the baseline and fine-tuned models have a degraded Hellinger distance of 0.116 and 0.062 respectively. This means that our fine-tuned model consistently outperforms the baseline model both on monthly mean and daily mean global statistics.

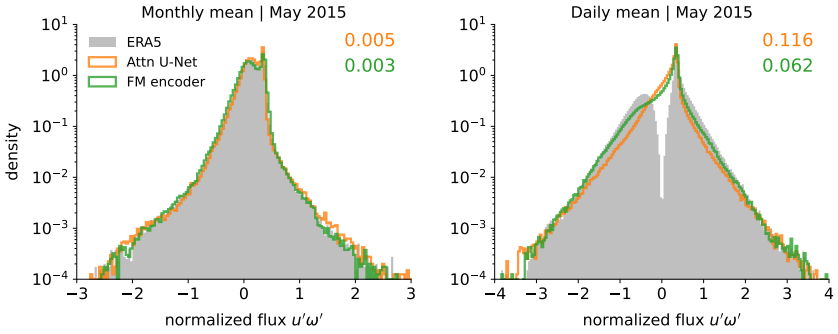

Figure 5: The distribution of the (left) May 2015 averaged and (right) daily averaged GW flux $u'\omega'$. Gray shading shows the true underlying distribution, orange the baseline prediction, and green the fine-tuning prediction. Numbers indicate the Hellinger distance for the respective predictions.

3.2 Region-wise, Steady State Flux Spectrum

The dynamics of the atmosphere and the evolution of GWs therein can notably vary with height, region (latitude and longitude), season, etc. The steady-state distributions conceal this. For a more stringent evaluation, we divide the global domain into 5 regions and 4 altitudes. The five regions comprise the two hemispheric poles, the two hemispheric midlatitudes, and the tropics. The four regions comprise the lower troposphere, the upper troposphere, the lower stratosphere, and the upper stratosphere.

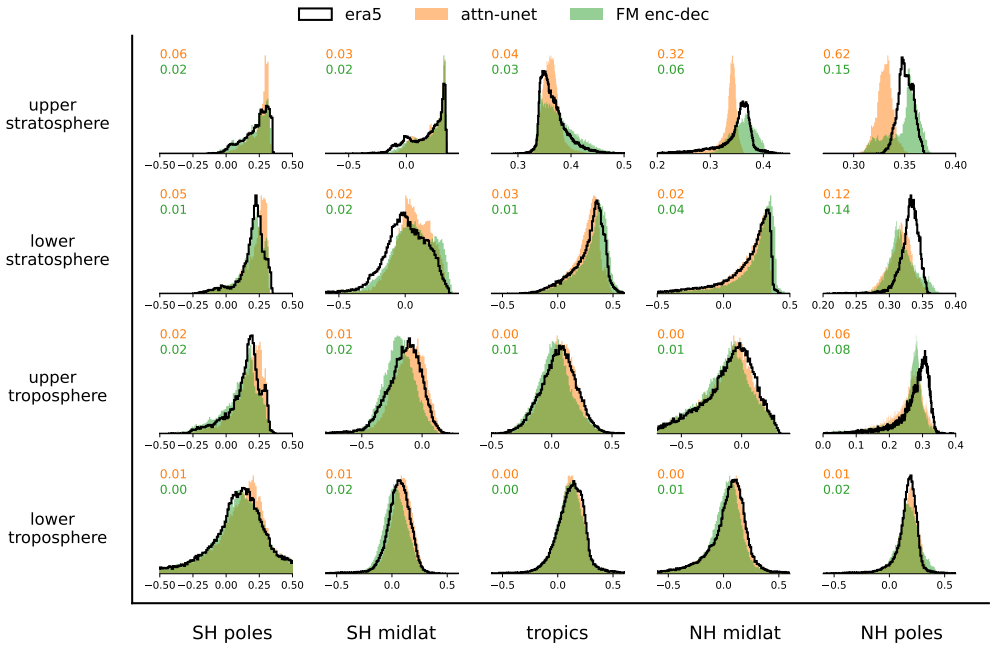

Figure 6: Global flux distributions segregated according latitude and height. The numbers indicate the respective Hellinger distances w.r.t the true distribution shown in black. For each latitude band, averaging is conducted over the whole latitude circle, i.e. over all longitudes.

The observed and predicted monthly mean flux distributions over the 20 slices are shown in Figure 6. The predicted and averaged fluxes agree quite well in all regions in the lower and upper troposphere. The Hellinger distances are less than 0.02 in all cases, except the northern hemispheric poles in the upper troposphere, where the ML models fail to predict the strong westerly fluxes. Within the troposphere, the baseline (in orange) has an ever so slightly better Hellinger distance from observations (in black) than the fine-tuned model (in green) but both models get the distributions fairly correct.

The Hellinger distances for the two models are higher in the stratosphere for all five regions. The tropics and midlatitudes in the lower stratosphere have distances within our 0.05 threshold, the polar regions have distances of up to 0.14. The deviations increase further in the upper stratosphere, but mostly for the baseline. In the northern hemisphere upper stratosphere, the baseline exhibits a Hellinger distance of up to 0.62, while the fine-tuned model is constrained within 0.15. Overall, we note that the fine-tuned model outperforms the baseline model in the stratosphere and the difference in their predictions is apparent both visually and quantitatively.

The baseline model consistently has a much lower variance than the fine-tuned model even as the fine-tuned model was not trained on upper stratospheric data at all. This is quite likely due to the higher amounts of data used to pre-train the model elsewhere as opposed to just four years of data used to train the baseline. Due to this, the encoder-decoders represent learning a much wider variability in the atmosphere than the baseline.

3.3 Instantaneous, Intermittent Evolution of Gravity Waves

Our final even more stringent test of the two models is the time-evolving response of the predictions and the ability of the models to predict the intermittent generation and evolution of GWs in the atmosphere.

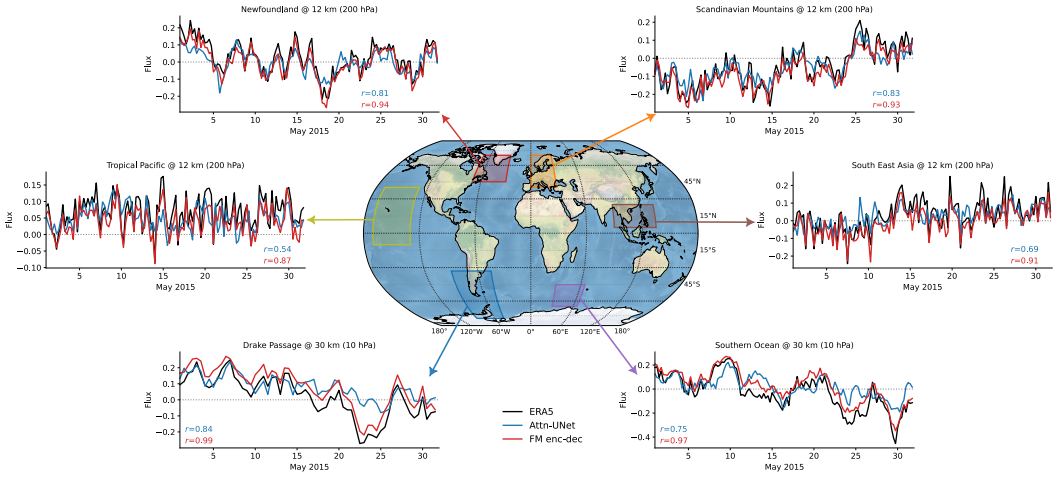

Figure 7: Instantaneous fluxes for May 2015 from ERA5 (black), and predictions from the baseline (blue) and the fine-tuned model (red) over six different hotspots. The numbers show the respective Pearson correlation coefficient w.r.t. ERA5 timeseries. The fluxes in the winter hemisphere are shown at 30 km, but the fluxes in the summer hemisphere are shown at 12 km, since GW activity in the summer at 30 km is substantially weaker.

Based on previously documented studies (Hindley et al., 2020; Wei et al., 2022), we select 6 known hotspots of GWs and analyze the time evolution of the box-averaged momentum fluxes for the month of May 2015 (Figure 7). Since we are interested in evaluating the nonlocal propagation of GWs, which is more prominent in the winter stratosphere (Sato et al., 2012; Gupta et al., 2024a), we focus on time evolution of fluxes in the upper winter stratosphere (10 hPa $\sim$ 30 km) as much as possible. For regions in the summer (northern) hemisphere, we analyze the fluxes in the upper troposphere (200 hPa $\sim$ 12 km).

The fine-tuned model generates significantly better prediction over all six hotspots for both the relatively smoother fluxes over Andes, Southern Ocean, Newfoundland and Scandinavian Mountains, and the rather noisier fluxes over the Pacific Ocean and Southeast Asia. Most notably for Andes (mountain waves) and the Southern Ocean (non-mountain waves), the predictions from the fine-tuned models bear a correlation coefficient (with the observed fluxes) of 0.99 and 0.97 respectively. In comparison, the respective correlation for the Attention U-Net baseline is 0.84 and 0.75 respectively. The correlation with observations is weakest over the Pacific when the fine-tuned and the baseline model ppredictions have a correlation of 0.87 and 0.54 respectively.

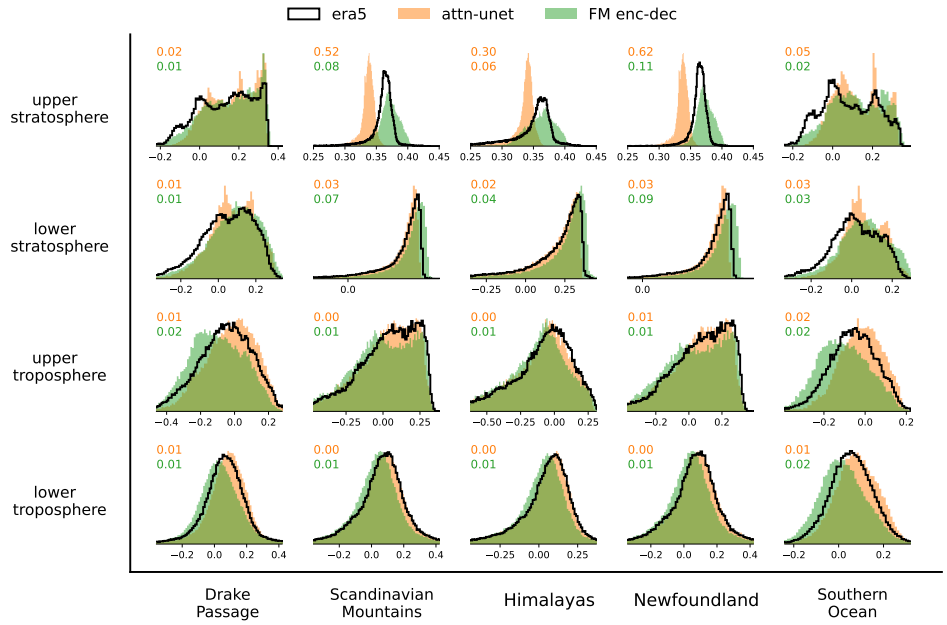

Figure 8: Global flux distributions similar to Fig 6 but segregated according to hotspots. Unlike Fig 5, the figure shows fluxes averaged over boxes outlined in Fig 7.

The remarkable skill of the fine-tuned model to predict both the short bursts of wave flux intensification in the tropics (due to GWs generated by tropical storms and convective systems) and the slower flux intensification over the midlatitudes (due to flow over mountains and due to storm tracks) points to the ability of the finetuned model to learn the intermittent and nonlocal evolution of medium-to-small scale atmospheric variability. This is further corroborated by the spatial structure of the predicted flux shown in Figure 1.

Revisiting the spatially segregated spectrum of Figure 6, this time exclusively for GW hotspots, we find something similar: the spectrum generated by the two ML models are not so different in the troposphere and lower stratosphere, but the spectrum generated by the fine-tuned model in the upper stratosphere is many degrees better than the spectrum generated by the baseline.

4 Discussion and Future Directions

The application of foundation models in climate science is largely unexplored. The series of three tests presented in this study clearly establish that the atmospheric evolution learned by large, transformer-based foundation models, can be leveraged to simplify, improve, and expedite the creation of physical parameterizations for climate models, ultimately improving climate prediction accuracy. In machine learning parlance, the latent encoder-decoder space of a weather and climate foundation model (here Prithvi WxC) contains a rich representation and learning of the atmospheric evolution due to training on large-amounts of data ranging from winds and temperature, to humidity and radiation, to soil moisture. Rather than creating task-specific artificial or convolutional neural networks ground up, these pre-trained encoders can be leveraged to create more accurate data-driven predictors of atmospheric processes.

Here we have demonstrated this point using one such atmospheric process, gravity waves, which dominate the mesoscale variability in the earth's atmosphere. Current climate model parameterizations of gravity waves have multiple deficiencies — missing horizontal propagation being the most prominent. Our fine-tuned model learns the nonlocal evolution of GWs and even generalizes remarkably to regions unseen during training. This provides a potentially improved pathway to representing gravity waves in climate models. Since the foundational models are typically trained on copious amounts of high-frequency data spanning a broad range of variables, the fined-tuned models for other atmospheric processes like boundary layer turbulence, precipitation, vertical mixing, can therefore be expected to outperform these respective baselines too. Since the fine-tuned models are less costly to train than the baselines — because only a fraction of parameters are retrained — our finding also has the potential to reduce the carbon emissions associated with the ground-up training of machine learning-based climate model parameterizations. (ADD NUMBERS)

4.1 LIMITATIONS:

Numerical climate models, climate reanalyses datasets, and data assimilation systems can have systematic biases. Training fine-tuned models on a given foundation model thus presents the danger of the inherent biases in training data to be carried over to the finetuned model. Such potential dangers can be alleviated by (a) using multiple data streams, for e.g., using data from multiple climate reanalyses datasets, (b) by combining high-resolution climate model data from model with a range of underlying numerics (spectral, finite-volume, spectral element, etc.), (c) climate model data from multiple climate-change scenarios, or (d) by using high-quality data during fine-tuning. As a straightforward experiment, the the encoder and decoder from large AI weather forecasting models (FourCastNet, PanguWeather, GraphCast etc.) can be used to develop a series of fine-tuned climate model parameterizations and their performance be compared (work in progress).

4.2 BROADER IMPACT AND FUTURE DIRECTIONS:

Climate prediction models participating in CMIP experiments are typically initialized in the 1850s and integrated to the end of the 21st century. Ideally we would want to run them at sub-kilometer resolutions, but due to computational constraints, these models are run at a coarse horizontal resolution (100-300 km), missing crucial information contained within sub-grid scales. Plugging the fine-tuned model presented here to a climate model and testing it in online climate simulations would be the final test for the model, as it would reveal nonlinear feedback the ML model has on the numerical climate model. This is work in progress. MORE MODELS - USE SATELLITE DATA ETC. ETC.

REFERENCES

M. J. Alexander and T. J. Dunkerton. A Spectral Parameterization of Mean-Flow Forcing due to Breaking Gravity Waves. *J. Atmos. Sci.*, 56(24):4167–4182, December 1999. ISSN 0022-4928. doi: 10.1175/1520-0469(1999)056⟨4167:ASPOMF⟩2.0.CO;2.

Mohamed Aziz Bhouri, Liran Peng, Michael S. Pritchard, and Pierre Gentine. Multi-fidelity climate model parameterization for better generalization and extrapolation, September 2023.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06185-3.

P. A. Bogenschutz, A. Gettelman, H. Morrison, V. E. Larson, D. P. Schanen, N. R. Meyer, and C. Craig. 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, November 2012. ISSN 1991-959X. doi: 10.5194/gmd-5-1407-2012.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022.

Matthew Chantry, Sam Hatfield, Peter Dueben, Inna Polichtchouk, and Tim Palmer. Machine Learning Emulation of Gravity Wave Drag in Numerical Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, 13(7):e2021MS002477, 2021. ISSN 1942-2466. doi: 10.1029/2021MS002477.

Zachary I. Espinosa, Aditi Sheshadri, Gerald R. Cain, Edwin P. Gerber, and Kevin J. DallaSanta. Machine Learning Gravity Wave Parameterization Generalizes to Capture the QBO and Response to Increased CO2. *Geophysical Research Letters*, 49(8):e2022GL098174, 2022. ISSN 1944-8007. doi: 10.1029/2022GL098174.

Veronika Eyring, William D. Collins, Pierre Gentine, Elizabeth A. Barnes, Marcelo Barreiro, Tom Beucler, Marc Bocquet, Christopher S. Bretherton, Hannah M. Christensen, Katherine Dagon, David John Gagne, David Hall, Dorit Hammerling, Stephan Hoyer, Fernando Iglesias-Suarez, Ignacio Lopez-Gomez, Marie C. McGraw, Gerald A. Meehl, Maria J. Molina, Claire Monteleoni, Juliane Mueller, Michael S. Pritchard, David Rolnick, Jakob Runge, Philip Stier, Oliver Watt-Meyer, Katja Weigel, Rose Yu, and Laure Zanna. Pushing the frontiers in climate modelling and analysis with machine learning. *Nat. Clim. Chang.*, pp. 1–13, August 2024. ISSN 1758-6798. doi: 10.1038/s41558-024-02095-y.

David C. Fritts and M. Joan Alexander. Gravity wave dynamics and effects in the middle atmosphere. *Reviews of Geophysics*, 41(1), 2003. ISSN 1944-9208. doi: 10.1029/2001RG000106.

Ronald Gelaro, Will McCarty, Max J. Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A. Randles, Anton Darmenov, Michael G. Bosilovich, Rolf Reichle, Krzysztof Wargan, Lawrence Coy, Richard Cullather, Clara Draper, Santha Akella, Virginie Buchard, Austin Conaty, Arlindo M. da Silva, Wei Gu, Gi-Kong Kim, Randal Koster, Robert Lucchesi, Dagmar Merkova, Jon Eric Nielsen, Gary Partyka, Steven Pawson, William Putman, Michele Rienecker, Siegfried D. Schubert, Meta Sienkiewicz, and Bin Zhao. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). July 2017. doi: 10.1175/JCLI-D-16-0758.1.

Jean-Christophe Golaz, Larry W. Horowitz, and Hiram Levy II. Cloud tuning in a coupled climate model: Impact on 20th century warming. *Geophysical Research Letters*, 40(10):2246–2251, 2013. ISSN 1944-8007. doi: 10.1002/grl.50232.

Aman Gupta, Aditi Sheshadri, M. Joan Alexander, and Thomas Birner. Insights on Lateral Gravity Wave Propagation in the Extratropical Stratosphere from 44 Years of ERA5 Data. *Geophysical Research Letters*, 2024a. ISSN 1944-8007. doi: 10.1029/2024GL108541.

Aman Gupta, Aditi Sheshadri, Sujit Roy, Vishal Gaur, Manil Maskey, and Rahul Ramachandran. Machine Learning Global Simulation of Nonlocal Gravity Wave Propagation, June 2024b.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. ISSN 1477-870X. doi: 10.1002/qj.3803.

N. P. Hindley, C. J. Wright, L. Hoffmann, T. Moffat-Griffin, and N. J. Mitchell. An 18-Year Climatology of Directional Stratospheric Gravity Wave Momentum Flux From 3-D Satellite Observations. *Geophysical Research Letters*, 47(22):e2020GL089557, 2020. ISSN 1944-8007. doi: 10.1029/2020GL089557.

Frédéric Hourdin, Thorsten Mauritsen, Andrew Gettelman, Jean-Christophe Golaz, Venkatramani Balaji, Qingyun Duan, Doris Folini, Duoying Ji, Daniel Klocke, Yun Qian, Florian Rauser, Catherine Rio, Lorenzo Tomassini, Masahiro Watanabe, and Daniel Williamson. The Art and Science of Climate Model Tuning. *Bulletin of the American Meteorological Society*, 98(3):589–602, March 2017. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-15-00135.1.

Michael J. Iacono, Eli J. Mlawer, Shepard A. Clough, and Jean-Jacques Morcrette. 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: Atmospheres*, 105(D11):14873–14890, 2000. ISSN 2156-2202. doi: 10.1029/2000JD900091.

Karan Jakhar, Yifei Guan, Rambod Mojgani, Ashesh Chattopadhyay, and Pedram Hassanzadeh. Learning Closed-form Equations for Subgrid-scale Closures from High-fidelity Data: Promises and Challenges. *J Adv Model Earth Syst*, 16(7):e2023MS003874, July 2024. ISSN 1942-2466, 1942-2466. doi: 10.1029/2023MS003874.

Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi, Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation Models for Generalist Geospatial Artificial Intelligence, November 2023.

Laura Köhler, Brian Green, and Claudia C. Stephan. Comparing Loon Superpressure Balloon Observations of Gravity Waves in the Tropics With Global Storm-Resolving Models. *Journal of Geophysical Research: Atmospheres*, 128(15):e2023JD038549, 2023. ISSN 2169-8996. doi: 10.1029/2023JD038549.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. GraphCast: Learning skillful medium-range global weather forecasting, August 2023.

Hoesung Lee, Katherine Calvin, Dipak Dasgupta, Gerhard Krinner, Aditi Mukherji, Peter Thorne, Christopher Trisos, José Romero, Paulina Aldunce, and Ko Barrett. *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, 2023.

Christian Lessig, Ilaria Luise, Bing Gong, Michael Langguth, Scarlet Stadtler, and Martin Schultz. AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning, September 2023.

Erik Lindborg. A Helmholtz decomposition of structure functions and spectra calculated from aircraft data. *Journal of Fluid Mechanics*, 762:R4, January 2015. ISSN 0022-1120, 1469-7645. doi: 10.1017/jfm.2014.685.

François Lott and Martin J. Miller. A new subgrid-scale orographic drag parametrization: Its formulation and testing. *Quarterly Journal of the Royal Meteorological Society*, 123(537):101–127, 1997. ISSN 1477-870X. doi: 10.1002/qj.49712353704.

Yixiong Lu, Xin Xu, Lin Wang, Yiming Liu, Tongwen Wu, Weihua Jie, and Jian Sun. 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, 2024. ISSN 1942-2466. doi: 10.1029/2023MS003611.

Laura A. Mansfield, Aman Gupta, Adam C. Burnett, Brian Green, Catherine Wilka, and Aditi Sheshadri. 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, 2023. ISSN 1942-2466. doi: 10.1029/2023MS003715.

Thorsten Mauritsen, Bjorn Stevens, Erich Roeckner, Traute Crueger, Monika Esch, Marco Giorgetta, Helmuth Haak, Johann Jungclaus, Daniel Klocke, Daniela Matei, Uwe Mikolajewicz, Dirk Notz, Robert Pincus, Hauke Schmidt, and Lorenzo Tomassini. Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, 4(3), 2012. ISSN 1942-2466. doi: 10.1029/2012MS000154.

Monica Ainhorn Morrison and Peter Lawrence. Understanding Model-Based Uncertainty in Climate Science. In Gianfranco Pellegrino and Marcello Di Paola (eds.), *Handbook of Philosophy of Climate Change*, pp. 1–21. Springer International Publishing, Cham, 2020. ISBN 978-3-030-16960-2. doi: 10.1007/978-3-030-16960-2_154-1.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate, December 2023.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning Where to Look for the Pancreas, May 2018.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, February 2022.

Riwal Plougonven, Albert Hertzog, and Lionel Guez. Gravity waves over Antarctica and the Southern Ocean: Consistent momentum fluxes in mesoscale simulations and stratospheric balloon observations. *Quarterly Journal of the Royal Meteorological Society*, 139(670):101–118, 2013. ISSN 1477-870X. doi: 10.1002/qj.1965.

Riwal Plougonven, Alvaro de la Cámara, Albert Hertzog, and François Lott. How does knowledge of atmospheric gravity waves guide their parameterizations? *Quarterly Journal of the Royal Meteorological Society*, 146(728):1529–1543, 2020. ISSN 1477-870X. doi: 10.1002/qj.3732.

Kaoru Sato, Satoshi Tateno, Shingo Watanabe, and Yoshio Kawatani. Gravity Wave Characteristics in the Southern Hemisphere Revealed by a High-Resolution Middle-Atmosphere General Circulation Model. *Journal of Atmospheric Sciences*, 69(4):1378–1396, April 2012. ISSN 0022-4928, 1520-0469. doi: 10.1175/JAS-D-11-0101.1.

Junhong Wei, Fuqing Zhang, Jadwiga H. Richter, M. Joan Alexander, and Y. Qiang Sun. 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, October 2022. ISSN 0022-4928, 1520-0469. doi: 10.1175/JAS-D-21-0173.1.

Janni Yuval and Paul A. O'Gorman. Neural-Network Parameterization of Subgrid Momentum Transport in the Atmosphere. *Journal of Advances in Modeling Earth Systems*, 15(4): e2023MS003606, 2023. ISSN 1942-2466. doi: 10.1029/2023MS003606.

Laure Zanna and Thomas Bolton. Data-Driven Equation Discovery of Ocean Mesoscale Closures. *Geophysical Research Letters*, 47(17):e2020GL088376, 2020. ISSN 1944-8007. doi: 10.1029/2020GL088376.

M. Zhao, J.-C. Golaz, I. M. Held, H. Guo, V. Balaji, R. Benson, J.-H. Chen, X. Chen, L. J. Donner, J. P. Dunne, K. Dunne, J. Durachta, S.-M. Fan, S. M. Freidenreich, S. T. Garner, P. Ginoux, L. M. Harris, L. W. Horowitz, J. P. Krasting, A. R. Langenhorst, Z. Liang, P. Lin, S.-J. Lin, S. L. Malyshev, E. Mason, P. C. D. Milly, Y. Ming, V. Naik, F. Paulot, D. Paynter, P. Phillipps, A. Radhakrishnan, V. Ramaswamy, T. Robinson, D. Schwarzkopf, C. J. Seman, E. Shevliakova, Z. Shen, H. Shin, L. G. Silvers, J. R. Wilson, M. Winton, A. T. Wittenberg, B. Wyman, and B. Xiang. 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, 2018. ISSN 1942-2466. doi: 10.1002/2017MS001208.

A Appendix

You may include other additional sections here.